

5-6-2016

## Implications of Computational Cognitive Models for Information Retrieval

Joshua Caleb Goodwin

*University of Texas Health Science Center at Houston*, [joshua.c.goodwin@uth.tmc.edu](mailto:joshua.c.goodwin@uth.tmc.edu)

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/uthshis\\_dissertations](https://digitalcommons.library.tmc.edu/uthshis_dissertations)



Part of the [Bioinformatics Commons](#), and the [Medicine and Health Sciences Commons](#)

---

### Recommended Citation

Goodwin, Joshua Caleb, "Implications of Computational Cognitive Models for Information Retrieval" (2016). *UT SBMI Dissertations (Open Access)*. 34.

[https://digitalcommons.library.tmc.edu/uthshis\\_dissertations/34](https://digitalcommons.library.tmc.edu/uthshis_dissertations/34)

This is brought to you for free and open access by the School of Biomedical Informatics at DigitalCommons@TMC. It has been accepted for inclusion in UT SBMI Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact [digitalcommons@library.tmc.edu](mailto:digitalcommons@library.tmc.edu).

Implications of computational cognitive models for information retrieval

A  
Dissertation

Presented to the Faculty of  
The University of Texas  
Health Science Center at Houston  
School of Biomedical Informatics  
in Partial Fulfilment of the Requirements for the Degree of  
  
Doctor of Philosophy

By  
  
J. Caleb Goodwin

University of Texas Health Science Center at Houston

2015

Dissertation Committee:

Elmer Bernstam, MD<sup>1</sup>, Advisor  
Todd R. Johnson, PhD<sup>1</sup>  
Trevor Cohen, PhD<sup>1</sup>  
Thomas Rindflesch, PhD<sup>2</sup>

<sup>1</sup>The School of Biomedical Informatics

<sup>2</sup>Lister Hill National Center for Biomedical Communications

Copyright by  
J. Caleb Goodwin  
2015

## Dedication

Katie and little Eukie.

## Acknowledgements

Dr. Bernstam has been extremely patient and has been instrumental in me completing this. He has the reputation for being tough. He will most likely be pleased to read I think it is true. But the secret is that he cares. Thank you. Dr. Cohen has been wonderful to work with and is a true scholar. I always enjoy hearing about his most recent work and learning from him as he forges deeper and deeper into dark chasms of the vector spaces. I had the pleasure of spending two summers at Lister Hill with Dr. Rindflesch. It was an absolute pleasure to work with you. Your leadership style has been a great example for me as I have grown my teams in industry. Dr. Johnson sent me down the path of what would turn out to be this dissertation. I vividly remember when you brought up the idea of using ACT-R for information retrieval with me in your office. I think we both felt at that moment that it would be turn into something exciting. Since then it has grown, shifted, and morphed, but I think that we were right. Thanks.

Quite literally this research could not have happened without Chris Young. The experiments in this dissertation require click stream data and without that..... Chris went above and beyond in helping me get the data I needed to do this research. I would also like to thank one of my early mentors Dr. Russomano. The lessons that you taught me early on have served me well. Finally, Elco! Thanks for the hours of laughter, still being awake at 4 AM when I was done working on this for the day, and Yucko.

## Table of Contents

Dedication .....	ii
Acknowledgements .....	iii
Table of Contents .....	iv
List of Tables .....	v
List of Figures .....	vii
Chapter 1: Introduction .....	1
Chapter 2: Background .....	266
Chapter 3: Related Work .....	62
Chapter 4: Predicting Document Clicks Using Desirability .....	95
Chapter 5: Predicting Document Clicks Using Information Scent .....	158
Chapter 6: Predicting Document Clicks Using Information Scent and Desirability .....	193
Chapter 7: Conclusion and Future Research .....	217
References .....	230

## List of Tables

Table 2.1. Properties of growth mechanisms.....	39
Table 2.2. Relationship between language model and TF-IDF ranking .....	46
Table 2.3. Instantiations of Prob <sub>1</sub> metric .....	49
Table 2.4. Instantiations of Prob <sub>2</sub> metric .....	49
Table 2.5. Instantiations of term frequency normalization .....	49
Table 2.6. Methods for extracting pairwise judgments.....	54
Table 2.7. Example results using pairwise judgments.....	57
Table 2.8. Results from sample data .....	61
Table 3.1. Bind, bundle, and release operators for different methods .....	79
Table 3.2. Overview of cognitive models using quantum probabilities .....	82
Table 3.3. Example text data from (Radiohead, 2011) .....	91
Table 3.4. Word by context vector.....	92
Table 4.1. Statistical properties of generated graphs .....	102
Table 4.2. Results of analysis for PubMed document access distribution.....	124
Table 4.3. Aggregate results for each day in PLOS data set.....	128
Table 4.4. Information content of each data source.....	149
Table 4.5. Correlation between coverage of data sets .....	150
Table 4.6. Desirability performance from different data sets .....	152
Table 4.7. Performance from JIF .....	152
Table 4.8. Performance from graph metrics computed on click graph.....	152
Table 4.9. Results for existing document ranking models.....	153
Table 4.10. Desirability results from combining multiple data sources .....	154

Table 5.1. Top 20 activated terms for “bank” and “money” in decreasing order of activation value .....	173
Table 5.2. Top 20 activated concepts for Bank $\oplus$ money in descending order of activation value .....	174
Table 5.3. Manual term classification.....	176
Table 5.4. Number of pairwise judgments for parameter tuning and evaluation .....	182
Table 5.5. First nine topics for LDA model.....	183
Table 5.6. Top ten most related terms from RI.....	184
Table 5.7. Results on test data for all pairwise judgments.....	190
Table 5.8. Results on test set for pairwise judgments extracted for document downloads .....	190
Table 6.1. Desirability results from combining multiple data sources .....	200
Table 6.2. Number of pairwise judgments for parameter tuning and evaluation .....	207
Table 6.3. Information scent and desirability results for all pairwise judgments .....	211
Table 6.4. Information scent and desirability results for predicting downloads.....	212
Table 6.5. Summary of performance increase of LM_TM + Desirability compared to existing IR models .....	213
Table 6.6. Summary of performance increase of LM_TM compared to existing IR models .....	215
Table 6.7. Summary of findings for different studies .....	216



## List of Figures

Figure 1.1 Lawyer-shark network.....	12
Figure 1.2. Information scent and the WWW .....	14
Figure 2.1. Example $MAP_{10}$ for two ranking functions.....	52
Figure 2.2. Example power law distribution.....	59
Figure 2.3. Example log-log plot .....	59
Figure 3.1. Semantic network structure .....	68
Figure 3.2. Spreading activation .....	68
Figure 3.3. Sample activation functions .....	69
Figure 3.4. Example connectionist network .....	74
Figure 3.5. Example of binding role and filler.....	77
Figure 3.6. Representation of computer and architecture .....	87
Figure 3.7. Representation of computer architecture.....	88
Figure 4.1. BA frequency effect .....	104
Figure 4.2. Log-log BA frequency effect.....	105
Figure 4.3. BA recency effect .....	106
Figure 4.4. Log-log BA recency effect .....	106
Figure 4.5. Average degree centrality for each recency bin .....	107
Figure 4.6. BA+triad frequency effect.....	106
Figure 4.7. Log-log BA+triad frequency effect.....	106
Figure 4.8. BA+triad recency effect .....	106
Figure 4.9. Log-log BA+triad recency effect .....	106
Figure 4.10. Average degree centrality for each recency bin .....	106
Figure 4.11. ER model frequency effect.....	107

Figure 4.12. Log-log ER model frequency effect .....	107
Figure 4.13. ER model recency effect .....	107
Figure 4.14. Log-log ER model recency effect.....	107
Figure 4.15. Average degree centrality for each recency bin .....	108
Figure 4.16. ER+triad frequency effect .....	109
Figure 4.17. ER+triad frequency effect .....	109
Figure 4.18. ER+triad recency effect.....	109
Figure 4.19. Log-log ER+triad recency effect .....	109
Figure 4.20. Average degree centrality for each recency bin .....	110
Figure 4.21. Preferential attachment for quotes from news cycle .....	114
Figure 4.22. Preferential attachment for predication graph .....	114
Figure 4.23. Preferential attachment for high energy physics network .....	114
Figure 4.24. Preferential attachment email communication network .....	114
Figure 4.25. Preferential attachment for Twitter hash tag network .....	114
Figure 4.26. Preferential attachment for Twitter communication.....	114
Figure 4.27. Frequency effect for quotes from news cycle.....	115
Figure 4.28. Frequency effect for predication graph .....	115
Figure 4.29. Frequency effect for high energy physics network .....	115
Figure 4.30. Frequency effect for email communication network.....	115
Figure 4.31. Frequency effect for hash tag network .....	116
Figure 4.32. Frequency effect for Twitter communication.....	116
Figure 4.33. Recency effect for quotes from news cycle.....	116
Figure 4.34. Recency effect for predication graph .....	116
Figure 4.35. Recency effect for high energy physics network .....	117
Figure 4.36. Recency effect for email communication network.....	117
Figure 4.37. Recency effect for hash tag network .....	117
Figure 4.38. Recency effect for Twitter communication .....	117
Figure 4.39. Distribution of document accesses from PubMed.....	123
Figure 4.40. Log-log distribution of document accesses from PubMed.....	123
Figure 4.41. Log-log plot of frequency versus odds of access .....	124

Figure 4.42. Recency with a 7 day training window .....	125
Figure 4.43. Recency with a 30 day training window .....	125
Figure 4.44. Recency with a 180 day training window .....	125
Figure 4.45. Recency with a 365 day training window .....	125
Figure 4.46. Preferential attachment for PubMed document accesses .....	126
Figure 4.47. PLOS log-log odds as a function of frequency .....	129
Figure 4.48. PLOS log-log odds as a function of recency .....	129
Figure 4.49. Sub-network of click graph extracted from HAM-TMC data set consisting of edge weights > 20.....	134
Figure 4.50. Configuration for desirability experiments .....	137
Figure 4.51. Configuration for experiments with existing IR models .....	138
Figure 4.52. Desirability experiments with sliding window.....	140
Figure 4.53. Experiments with existing IR models with sliding window.....	140
Figure 4.54. Distribution for documents viewed by year .....	148
Figure 4.55. Information content of data sources from 1970-2013 .....	150
Figure 4.56. Information content of data sources from 2000-2013 .....	151
Figure 5.1. Example result from PubMed.....	161
Figure 5.2. Mapping information scent to PubMed search.....	162
Figure 5.3. Sub-network created by the terms bank and money with independent activation calculation .....	173
Figure 5.4. Sub-network created by the terms bank and money with context dependent activation. ....	175
Figure 5.5. Activation values by category for bank and money .....	177
Figure 5.6. Activation values by category for bank, money, and bank $\oplus$ money .....	177
Figure 5.7. Experiment for computing information scent.....	179
Figure 5.8. Sliding window for calculating information scent .....	181
Figure 5.9. Smoothing level for language model with exact matching .....	185
Figure 5.10. Smoothing level for language model with partial matching using a topic model.....	186
Figure 5.11. Smoothing level for language model with partial matching using RI...	187

Figure 5.12. Match penalty for ACT-R using a topic model .....	188
Figure 5.13. Match penalty for ACT-R using RI.....	189
Figure 6.1. Information scent and the WWW .....	198
Figure 6.2. Overview of experiments for the combination of the information scent and desirability models .....	204
Figure 6.3. Overview of experiments for the existing IR models.....	205
Figure 6.4. Example of sliding window evaluation for the combination of information scent and desirability.....	206
Figure 6.5. Results for parameter tuning for BM25.....	209
Figure 6.6. Smoothing level for language model with partial matching using a topic model.....	210
Figure 7.1. Example result from PubMed.....	221
Figure 7.2. Personal information scent model .....	226
Figure 7.3. Potential levels for modeling desirability.....	229

## **Chapter 1: Introduction**

Thanks to science and technology, access to factual knowledge of all kinds is rising exponentially while dropping in unit cost... [we are] are drowning in information, while starving for wisdom.

(E. O. Wilson, 1992)

Clinicians and researchers can no longer keep up-to-date with literature manually, even in specialized domains. This problem of extracting knowledge from the rapidly created literature was declared as precluding the existence of experts in medical sub-disciplines in the appropriately titled article “On the Impossibility of Being an Expert” (Fraser, 2010). The authors argued that expertise could theoretically be obtained just as it was time to retire. One method to help cope with the increasing information overload is information retrieval (IR) systems that help users identify relevant information within large document collections. IR systems become increasingly important as the volume of scientific literature increases.

The National Library of Medicine’s (NLM) PubMed is the most widely used IR tool for accessing the MEDLINE database of biomedical literature (Falagas, Pitsouni, Malietzis, & Pappas, 2008). PubMed provides access to over 19 million articles and processes over 1.5 billion queries a year (Islamaj Dogan, Murray, Neveol, & Lu, 2009). By default, PubMed ranks the results by reverse chronological order<sup>1</sup>. Reverse chronological order ranking is

---

<sup>1</sup> At this time this research was done PubMed did not provide relevance ranking.

only sufficient if the user is seeking the most recent articles. Other information needs such as finding important articles are not well served by reverse chronological ranking. In addition, results sets returned from the PubMed IR system can be very large. For example, a query for “breast cancer” returns over 200,000 citations. Clearly this result set is too large for manual review. Ranking by importance or relevance could assist the user in finding articles that are relevant for their information need. In addition, users on average look at only the first ten results making ranking by relevance to the query a priority (Islamaj Dogan, et al., 2009).

Numerous approaches exist for ranking documents (Canfora & Cerulo, 2004) and similarly numerous approaches exist for characterizing the information seeking behavior of IR system users (Bates, 1989; Canfora & Cerulo, 2004; Ingwersen & Jarvelin, 2010; Marchionini, 1995). For the purposes of this dissertation, the following definition of information seeking behavior from (T. D. Wilson, 2000) is used.

*Information Seeking Behavior is the purposive seeking for information as a consequence of a need to satisfy some goal.*

Information seeking studies are largely concerned with characterizing how users utilize and interact with IR systems. This literature is rooted primarily in social and library sciences. In contrast, IR research is typically rooted in computer science and largely focuses on the development of algorithms that should, in theory, result in improved user experience by improving the search technology. A well-noted chasm exists between the information seeking literature and the IR literature (Bates, 1989; Belkin, 1993, 2008; Ingwersen, 1992; Ingwersen & Jarvelin, 2010; Saracevic, 1997; Sparck Jones, 1988). According to Ingwersen & Jarvelin, “the two camps do not communicate much with each

other and it is safe to say, that one camp generally views the other as too narrowly bound with technology whereas the other regards the former as an unusable academic exercise” (Ingwersen & Jarvelin, 2010). An alarming artifact of this chasm is that the performance gains of IR systems found in controlled laboratory experiments do not necessarily translate to real-world user satisfaction (Al-Maskari, Sanderson, Clough, & Airio, 2008; Allan, Carterette, & Lewis, 2005; W. Hersh et al., 2001; Jarvelin, 2009; Macdonald & Ounis, 2009; Sanderson, Paramita, Clough, & Kanoulas, 2010; Smith & Kantor, 2008; Smucker & Jethani, 2010; Su, 1992; Turpin & Scholer, 2001, 2006; Urbano, McFee, Downie, & Schedl, 2012).

The application of computational cognitive modeling in IR is an emerging area of research that seeks to bridge the information seeking and IR viewpoints. According to Pirolli, the role of computational cognitive modeling in IR has having the following general goals (P. Pirolli & Card, 1999b).

**Goal 1:** Explain and predict how people will best shape themselves for their information environments

**Goal 2:** Understand how information environments can best be shaped for people

This dissertation is primarily concerned with the second goal. In Chapter 4, I analyze the aggregate document accesses by a population of IR users and show that the statistical regularities of these aggregate accesses can be used to predict future accesses by individual users. In Chapter 5, I present the first application of computational cognitive modeling in the biomedical domain. Finally, in Chapter 6 I present an IR system based on the insights from the experiments in Chapters 4 and 5.

The remainder of this chapter presents the theoretical background and research questions that I pursued in this dissertation. Section 1.1 and Section 1.2 present an overview of the Adaptive Character of Thought-Rational (ACT-R) and the Information Foraging theory, which are necessary for understanding the contributions of this dissertation given that they are the theoretical foundation. Section 1.3 provides an overview of the research questions pursued in this dissertation and the experiments conducted to explore the questions. Section 1.4 presents the main contributions of this dissertation. Finally, Section 1.5 presents an overview of the structure of the rest of this dissertation.

### **1.1 ACT-R Theory of Human Associative Memory**

The term “cognitive architecture” was first introduced to cognitive science in 1971 (Bell & Newell, 1971). According to Anderson (J. Anderson, 2007), a cognitive architecture is a “specification of the structure of the brain at a level of abstraction that explains how it achieves the function of the mind”. ACT-R is a cognitive architecture based on a theory of how human cognition works (J. Anderson, 2007). ACT-R has been applied to gain insight into diverse areas of human cognition including perception (Brumby, Salvucci, & Howes, 2007) and problem solving (Danker & Anderson, 2007). The ACT-R theory asserts that the mind is comprised of structural modules and these modules correspond to brain regions. Example modules include the declarative memory module and visual perception module. The function of the mind (cognitive processes), according to the ACT-R theory, emerges through interaction of the modules. Thus, a cognitive model within the ACT-R cognitive architecture is a specification of the interaction of modules. For example, the Information Foraging Theory (discussed in Section 1.2) is a cognitive model created within ACT-R and



is used to understand a specific cognitive process, which is information seeking behavior in environments such as the World Wide Web (WWW).

This dissertation involves only the long-term memory theory of ACT-R, which Anderson developed the rational analysis approach that he invented (J. R. Anderson, 1989). The rational analysis approach emphasizes understanding the structure and dynamics of the environment, which leads to an understanding of how the cognitive system would perform tasks given these constraints. A summary of the rational analysis approach below is described in (J. R. Anderson, 1991).

1. Precisely specify the goals of the agent.
2. Develop a formal model of the environment to which the agent is adapted.
3. Make minimal assumptions about the computational costs.
4. Derive the optimal behavior of the agent considering items 1-3.
5. Test the optimality predictions against data.
6. Iterate.

(J. R. Anderson, 1991)

Interestingly, Anderson began with the insight that the human memory and IR systems are both attempting to solve the same computational challenge, that is, to retrieve the optimal set of items from an expansive set (J. R. Anderson & Milson, 1989). This parallel is summarized by (Steyvers & Griffiths, 2010) as follows:

For a search engine, the retrieval problem is finding the set of documents that are most relevant to a user query. In human memory, the retrieval problem can be construed in terms of assessing the relevance of items stored in the mind to a memory probe (either internally generated or based on environmental cues). The

common structure of these problems suggests a simple analogy between human memory and computer-based information retrieval: items stored in memory are analogous to documents available in a database of text (such as the world-wide Web) and the memory probe is analogous to a user query.

(Steyvers & Griffiths, 2010)

The rational analysis approach generated a branch of cognitive science with the aim of understanding many facets of cognition through this approach (Chater & Oaksford, 2008; Oaksford & Chater, 2007). In the case of human memory, Anderson & Schooler (1991) assert that human memory is an adaptation to the statistical properties of information in the environment. In addition, Anderson makes the (sometimes controversial) claim that the human memory system is optimal within the view of bounded rationality (for in-depth discussions see (Gigerenzer & Selten, 2002; Simon, 1956). Bounded rationality pertains to optimization under constraints (Stigler, 1961). That is, human beings must make decisions within the constraints of time, physical constraints of the cognitive system, and based on uncertain or incomplete information. For example, a Neanderthal with the goal of surviving in the savannah would be more likely to survive by falsely identifying an object as a lion and fleeing than taking additional seconds or minutes to insure that the classification is correct. Simon (Simon, 1956) provided a metaphor of a pair of scissors to describe bounded rationality where one blade is “cognitive limitations” and the other is “structure of the environment”. According to Simon, “a great deal can be learned about rational decision making. By taking into account the fact that the environments to which it must adapt possess properties that permit further simplification of its choice mechanisms” (Simon, 1956).

Anderson proposes a Bayesian solution based on rational analysis to model the human memory retrieval problem. Equation 1.1 shows the log odds form of Bayes' Theorem. In Equation 1.1,  $H$  corresponds to the hypothesis that a given memory item is needed and  $E$  corresponds to the evidence. The parameter  $\log \frac{P(H)}{P(\bar{H})}$  corresponds to the prior odds that a given memory item is needed. The parameter  $\sum_{j \in E} \log \frac{P(j|H)}{P(j|\bar{H})}$  corresponds to the log-likelihood that a given memory item is need. The log-likelihood is the context-sensitive component whereas the prior odds is independent of the context. Context in this chapter refers to an utterance or a sentence. For example, if the terms “money” and “bank” were viewed within the same context, it would be appropriate to retrieve memory items that pertain to the “financial institution” sense and not the “body of land near a river” sense. In summary, the ACT-R theory proposes that memory items have a prior probability distribution representing how likely a memory is to be needed in the future based on past use. Given a memory probe such as an utterance (analogous to a query in IR), the prior probabilities for the memory items are updated with the current evidence from the probe (likelihood based on each cue in the utterance) and the memory with the highest posterior probability is retrieved.

$$\log \frac{P(H|E)}{P(\bar{H}|E)} = \log \frac{P(H)}{P(\bar{H})} + \sum_{j \in E} \log \frac{P(j|H)}{P(j|\bar{H})} \quad (1.1)$$

In mapping back to bounded rationality, the ACT-R theory of long-term memory makes two major assertions regarding the structure of the environment to which the human

memory system has adapted. The first is that the prior probability of a given memory item being needed is learned from the statistical properties of information in the environment. The second is that the likelihood is sampled from the environment and the memory structure reflects the statistical co-occurrence of information in the environment. Sections 1.1.1 and Sections 1.1.2 present in detail the theoretical foundation of the prior probability and likelihood respectively.

### **1.1.1 The prior probability distribution for memory items reflects the statistical properties of information to which the memory system has adapted**

Quentin Burrell first defined the notion of desirability in the context of a library as “the average number of times an item is borrowed per unit time” (Burrell, 1980, 1985; Burrell & Cane, 1982; Burrell & Fenton, 1994). Burrell used a desirability function based on the frequency of past circulation to predict how likely a book was to be borrowed in the near future. The motivation of this model was to identify books that were not likely to be checked out such that they could be placed in storage.

Anderson & Schooler were interested in a similar proposition for human memory (J. R. Anderson & Schooler, 1991). That is, whether it possible to create a desirability model for human memory. Based on the rational analysis approach discussed previously, Anderson & Schooler hypothesized that analyzing the statistical properties of information in the environment would reveal a structure, and whatever this structure happened to be, would be reflected in the human memory system (J. Anderson, 2007). This reflection of the environment in memory would exist since human memory is an evolved system, and according to bounded rationality, would provide constraints that would influence the optimization of human memory in the environment (J. R. Anderson & Schooler, 1991).

Anderson & Schooler (J. R. Anderson & Schooler, 1991) investigated the statistical regularities of information in different environments. Specifically they looked at how past frequency (number of times an item appeared in the past) and recency (how recently a given item last appeared) influenced the probability that the item would appear in the future. This is known as the recency and frequency (recency-frequency) effect. Anderson & Schooler looked at the appearance of words in *New York Times* headlines, utterances spoken by children as a function of past utterances heard (MacWhinney & Snow, 1990), and email correspondences. In all of these situations, the relationship between probability of an item appearing in the future has a power law relationship with the past recency and frequency of appearance. Based on the results of the analysis, Anderson & Schooler developed a desirability model based on the recency-frequency effect that predicts the probability of a memory item being needed in the future. Anderson & Schooler showed that their model could accurately account for the long observed recency-frequency effect in human memory (Ebbinghaus, 1885).

### **1.1.2 The context sensitivity of human memory is learned based on past experience**

The context sensitivity of the ACT-R theory of human memory is very similar to the distributional hypothesis, which asserts that the meaning of a word can be defined based on the contexts in which it occurs (Harris, 1954) and Hebbian learning (Hebb, 1940, 1961) which asserts that “cells (in this case concepts) that fire together, wire together” (Doidge, 2007). For example, the utterance “my lawyer is a shark” is the context and the individual cues (ignoring stop terms) are “lawyer” and “shark”. Once this phrase is encountered, the memory system would strengthen the relationship between the concepts “lawyer” and

“shark”. According to the ACT-R theory of long-term memory, the conditional probabilities between concepts are learned through experience based upon the contexts in which the concepts appear. These conditional probabilities reflect the likelihood from Equation 1.1, which is known as association strength in the ACT-R theory. Schooler & Anderson (Schooler, 1993; Schooler & Anderson, 1997) provide the example of the associative effect for the terms “AIDS” and “virus” from the New York Times headlines. They found that the term “AIDS” was included in 1.8% of the headlines and the term “virus” was included in approximately 75% of the headlines. However, if the headline contains the term “virus”, the term “AIDS” was 41 times more likely to occur.

The context sensitivity of the ACT-R theory of human memory emerges through the spreading activation equation, which combines influence of the contexts provided by the cues. Numerous models of human memory have utilized a spreading activation component (J. R. Anderson, 1983; J. R. Anderson & Bower, 1973; Collins & Loftus, 1975). Pitkow offers the following intuitive explanation of spreading activation:

One way to conceptually understand spreading activation is to imagine a system of water reservoirs connected via a set of pipes, with the diameter of the pipes determining the rate of water flowing between reservoirs. When a large amount of water is injected into the system from a particular source reservoir or set of source reservoirs, after a period of time, the water levels in all the reservoirs will settle in a particular pattern. Based upon this final pattern, each reservoir can be inspected and the ones with the most water selected. If one views the flow rates between reservoirs as a measure of their connectedness (association), then the reservoirs

with the most water at the end are in a sense the ones more connected (related) to the source reservoir.

(Pitkow, 1997)

To illustrate the computational problem involved in selecting the correct context given the cues consider the problem highlighted in Figure 1.1 (adapted from (Glucksberg, 1998)). This figure maps the interpretation of the phrase “My lawyer is a shark” to the context sensitive solution proposed by the ACT-R theory of long-term memory. For simplicity, only the terms “lawyer” and “shark” are only considered. According to the ACT-R theory of long-term memory, the association strengths ( $S_{ji}$ ) between the concepts are accumulated based on the past contexts in which these terms have appeared. The association strengths between the cues “lawyer” and “shark” and the connected concepts (e.g. “law”) act as inhibitory and excitatory links and the connected concepts compete for activation. In this example, the concept “shark” would ideally inhibit the concepts “client”, “lawsuit”, and “law”, and the concept “lawyer” would ideally inhibit the concepts “gills”, “fins”, “fast-swimmer”, and “leathery skin”. The concepts “lawyer” and “shark” both have excitatory connections with “aggressive” and “vicious”, which results in both of these concepts being the candidates for retrieval. According to (W. Kintsch, 2000), the computational goal in this specific type of metaphorical reasoning is to select “those features of the (metaphoric) predicate (i.e., shark) that are appropriate for it (i.e., lawyer) and inhibit the features that do not apply or apply less aptly”. That is, the interpretation of the phrase “my lawyer is a shark” is an online process through which the meaning of “my lawyer is a shark” is constructed.

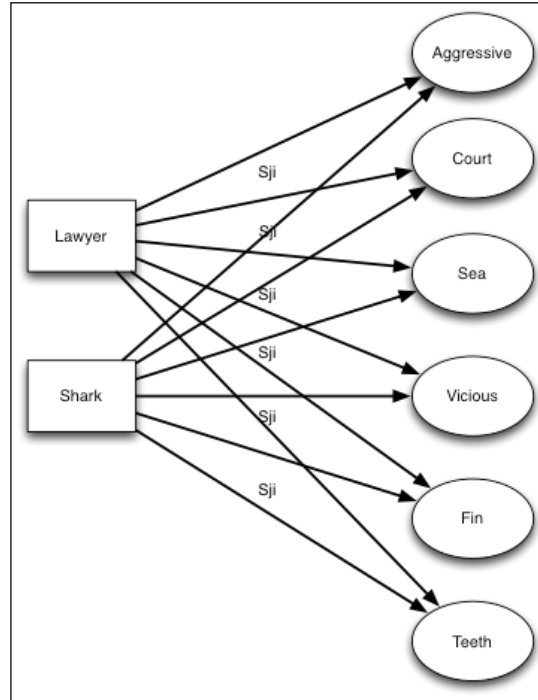


Figure 1.1. Lawyer-shark network. Adapted from (Glucksberg, 1998)

## 1.2 Information Foraging Theory

The Information Foraging Theory is based on the optimal foraging theory (Stephens & Krebs, 1986) and the ACT-R theory of long-term memory. Optimal foraging theory has been applied to describe the foraging behavior of numerous species of birds (Green, 1980), mammals (Kie, Evans, Loft, & Menke, 1991), reptiles (Huey, Bennett, John-Alder, & Nagy, 1984), and insects (Waddington & Holden, 1979). The basic concept behind the optimal foraging theory is that animals maximize the caloric intake per unit time while minimizing the energy expenditure to obtain the calories. Similarly, the information foraging theory models the information seeking behavior of humans with the assumption that they seek to maximize the intake of information while minimizing the effort taken to



obtain the information. The information foraging theory has several components. The “maximizing the intake of information per unit time” constraint from the optimal foraging theory predicts the amount of time a person will spend on a Web page before abandoning it (Huberman, Pirolli, Pitkow, & Lukose, 1998), which yields the Zipf-like distribution of the number of pages of user visits within a Web page (Islamaj Dogan, et al., 2009). Interestingly, recent studies have found that the search mechanism in human semantic memory shares some characteristics with optimal foraging, which strengthen the relation between the Information Foraging Theory and the optimal foraging theory (Hills, Todd, & Jones, 2009; Rhodes & Turvey, 2007).

The component of the Information Foraging Theory that is relevant for this work is information scent. Information scent is the utility of an information item, which can be thought of as a “rational analysis of categorization of cues according to their expected utility” (P. Pirolli & Card, 1999b). In the case of the Web, cues refer to “World Wide Web links or bibliographic citations, that provide users with concise information about content that is not immediately available” (P. Pirolli & Card, 1999b). According to the Information Foraging Theory, users attend to the cues with the highest expected utility given their information need. For example, consider the search results of a typical search engine shown in Figure 1.2. According to Information Foraging Theory, the user will select the hyperlink with the highest information scent based on proximal cues such as the Web Page title to maximize the probability of satisfying the information need with the distal information content (e.g., the Web page associated with a hyperlink).

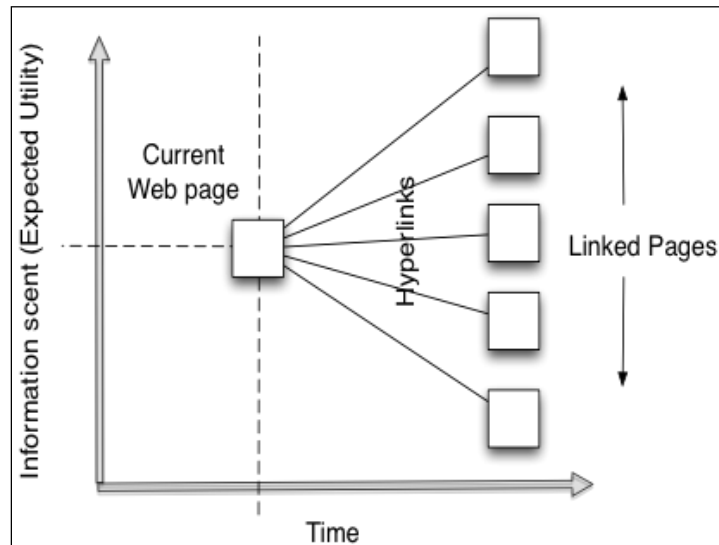


Figure 1.2. Information scent and the WWW. Adapted from (P. Pirolli, 2009)

The Information Foraging Theory can be viewed as a cognitive model within the ACT-R cognitive architecture. In fact, much of the Information Foraging Theory was developed and conducted in the SNIF-ACT cognitive model (P. Pirolli & W-T., 2006), which was implemented within the ACT-R framework. In this work and many previous applications of the Information Foraging Theory, information scent is calculated using the ACT-R's spreading activation model (J. Anderson, 2007). It is important to note that the task environment of the ACT-R human memory theory and the Information Foraging Theory differ, but the rational analyses of the computational tasks are similar and consequently the mathematical framework is identical. The ACT-R theory of human memory is based on the hypothesis that the human memory system actively predicts the memory items most likely to be needed based on the current context and past access of memory items (J. R. Anderson & Milson, 1989). In contrast, the Information Foraging Theory is based on the

idea that the forager is assessing the utilities (information scent) of distal information (e.g., Web pages) based on proximal cues (e.g., textual description of a hyperlink) and selects the proximal cue (i.e., hyperlink) that will most likely satisfy the user's information need. In both cases, the computational problem faced by the agent is calculating the utility of distal information given the proximal cues.

### **1.3 Research Questions**

This section is organized according to the research questions that I address in each chapter of this dissertation. Section 1.3.1 presents the research questions explored in Chapter 4, which pertain to the recency-frequency effect. Section 1.3.2 presents the research question explored in Chapter 5, which involves using information scent to predict document clicks for PubMed users. Finally, Section 1.3.3 presents the research question explored in Chapter 6, which involves combining the insights from Chapter 4 and Chapter 5 for predicting the document accesses of PubMed users.

#### **1.3.1 Leveraging the recency-frequency effect for IR.**

As discussed, an important goal of IR systems is to prioritize documents likely to be needed from an expansive corpus. In investigating this optimization problem, we looked at similar domains where models exist for estimating the probability of items being retrieved from a large set. The first domain is library science where the problem is predicting the book most likely to be checked out based on past use (Burrell, 1980). The second domain is cognitive science where the problem is modeling how human memory selects the memory with the highest probability of being needed based on past use (J. R. Anderson & Schooler, 1991). Burrell (1980) showed that library book circulation could be predicted based on past use. Anderson & Schooler linked the optimization problems faced in predicting library book

circulation and human memory by adapting Burrell's model to predict the probability of a memory item being accessed as a function of the recency and frequency of past use (J. R. Anderson & Schooler, 1991).

In Chapter 4, I explore the implications of the Anderson & Schooler model for predicting future documents accesses based on the past use. In essence, this work links a long line of information science research beginning with Burrell that predicted library book circulation, which was adapted by Anderson & Schooler to determine the prior probability of a given memory item being retrieved, and is explored here to predict document accesses in digital repositories. Chapter 4 seeks to answer the following questions regarding the use of desirability for document ranking.

**Research Question 1:** *Why does the recency-frequency effect exist?*

**Research Question 2:** *Do document accesses from digital repositories adhere to the recency-frequency effect?*

**Research Question 3:** *Can the recency-frequency effect be exploited to improve document ranking performance?*

The remainder of this section explores the experiments used to answer these questions.

### **1.3.1 Research Question 1: Why does the recency-frequency effect exist?**

Anderson & Schooler showed that the recency-frequency effect existed in a wide variety of different domains including email correspondence, language acquisition of children (CHILDES data set), and New York Times headlines (J. R. Anderson & Schooler, 1991). Based on these findings, Anderson & Schooler asserted that the recency-frequency effect was a natural property of information in the environment. They further hypothesized that, given that human memory evolved in this environment, human memory would also display

the recency-frequency effect. This hypothesis was experimentally validated when Anderson & Schooler showed that the probability of an item being accessed in memory adhered to the recency-frequency effect. However, Anderson & Schooler did not explain *why* information in the environment would have such a property.

Each of the domains that Anderson & Schooler investigated can be conceptualized as an evolving network where new nodes appear and new edges form between nodes in the network. For example, when looking at the language acquisition of children, new terms appear in the utterances of children, which add new nodes to the network and new edges are formed between existing nodes as they co-occur in different utterances. Recent research in network science has shown that many real-world networks including the WWW (Barabasi & Albert, 1999), metabolic networks (Jeong, Tombor, Albert, Oltval, & Barabasi, 2000; D. S. Lee et al., 2008), and social networks (Barabasi et al., 2002; Csanyi & Szendroi, 2004; Liljeros, Edling, Amaral, Stanley, & Aberg, 2001; Lusseau, 2003) can be modeled as scale-free networks. In scale-free networks, the distribution of the degree centrality of the nodes in the network is a power law. Degree centrality is a graph theory metric that reflects the importance of a node based on the number nodes to which it is connected (Opsahi, Agneessens, & Skvoretz, 2010). The mechanism that gives rise to the power law distribution was shown by Barabasi (Barabasi & Albert, 1999) to be a preferential attachment mechanism. The preferential attachment mechanism means that new nodes are more likely to connect with nodes that have larger degree centrality values. Recent work has shown that the types of data sets that Anderson & Schooler investigated can be characterized or at minimum have been theorized to be scale-free networks. Anderson & Schooler investigated the patterns of email correspondence of the first author

Anderson and found that the recency-frequency effect was present in the data. Recent studies have investigated much larger data sets and found that the correspondence between humans through mailed letters (Oliveira & Barabasi, 2005) and through email can be characterized as scale-free networks (Barabasi, 2005; Ebel, Mielsch, & Bornholdt, 2002). Anderson and Schooler analyzed the appearance of terms in the New York Times headlines and found the presence of the recency-frequency effect in this data. A study that was similar in nature to using the titles of the New York Times headlines was conducted by Pereira et al. (Pereira, Fadigas, Senna, & Moret, 2011). Pereira et al. conducted an analysis of a network extracted from the titles of scientific articles and found that it is a scale-free network. Anderson & Schooler investigated the utterances of children using the CHILDES corpus and found the presence of the recency-frequency effect. Numerous studies have found that a network extracted from the utterances of children in the CHILDES corpus is a scale-free network (Corominas-Murta, Valverde, & Sole, 2009; Sole, Murta, Valverde, & Steels, 2006). Finally, recent studies have provided evidence that the structure of human long-term memory is a scale-free network (Deyne & Storms, 2008; Griffiths, Steyvers, & Firl, 2007; Morais, Olsson, & Schooler, 2013; Steyvers & Griffiths, 2010; Steyvers & Tenenbaum, 2005).

Section 4.2 explores the idea that the recency-frequency effect is an artifact of scale-free network growth. The observation that the recency-frequency effect coexisted in data sets that numerous studies characterized as scale-free networks generated the initial hypothesis. However, the co-occurrence is not evidence of a causal relationship. To test this hypothesis, I generated a variety of networks using network growth rules that are known to yield certain properties. I performed experiments on the generated data from each network to determine

the presence of the recency-frequency effect. The experiment found that scale-free networks were the only type that exhibited the recency-frequency effect. This offers a potential mechanistic explanation for why Anderson & Schooler observed the recency-frequency effect in a wide variety of different domains. Furthermore, this finding has a possible implication for human memory. The recency-frequency effect is a well-known property of human memory dating back to the work of Ebbinghaus in 1885 (Ebbinghaus, 1885). This finding supports the hypothesis that the acquisition of concepts by human memory can be modeled by the growth of a scale-free network.

### **1.3.2 Research Question 2: Do document accesses from biomedical digital repositories adhere to the recency-frequency effect?**

There is some evidence that the recency-frequency effect exists for documents accessed online. Recker & Pitkow showed that the recency-frequency effect was present for documents accessed on the WWW (Recker & Pitkow, 1996). Dezso et al. investigated the access of news articles from a major news portal (Dezso et al., 2006). Dezso et al. did not specifically investigate the recency-frequency effect, but nonetheless found results that could indicate its presence. For example, Dezso et al. found that the visitation rates for documents decayed over time as a power law.

In Section 4.3, I investigate whether the recency-frequency effect exists for document accesses for two different populations. The first data set is comprised of documents accessed using the PubMed IR system from the users of the Houston Academy of Medicine Texas Medical Center (HAM-TMC) library. The HAM-TMC library provides access to information resources for over 50 institutions including numerous hospitals, medical schools, nursing schools, public health, and dentistry among others (Center, 2013). The

second data set is comprised of documents accessed through the Public Library of Science (PLOS) website. PLOS is an open-access publisher that currently has seven journals, which have published approximately 60,000 articles. The two data sets offer complimentary features for testing for the recency-frequency effect for document accesses in digital repositories. By default, PubMed ranks documents in reverse chronological order. In contrast, the PLOS search engine ranks documents by similarity to the query (PLOS, 2013). If the experiments found the recency-frequency effect in both data sets, this would provide evidence that the effect occurs regardless of the type of ranking function that used by the IR system. In these experiments, I found that the recency-frequency effect was present in both data sets.

### **1.3.3 Research Question 3: Can the recency-frequency effect be exploited to improve document ranking performance?**

In the previous experiments discussed in Sections 1.3.1 and 1.3.2, I found a possible explanation for why the recency-frequency effect exists and found that the recency-frequency effect was present for documents accessed through two different types of IR systems. In and of themselves, these studies are interesting, but they do not necessarily mean that the recency-frequency effect can be used to improve document ranking.

To address this question, I evaluated using desirability computed from document accesses from multiple crowd-sourced data sources to improve document ranking. The definition of desirability used in this dissertation is “probability of an item receiving attention” (Recker & Pitkow, 1996). The desirability function leverages the recency-frequency effect to calculate the prior probability of a given document being accessed. I compared the results of the desirability model with the Journal Impact Factor (JIF) metric. The JIF is a



bibliometric value that reflects the average number of citations for each article in a journal (Garfield, 2006). I found that the desirability model, which used information from multiple crowd-sourced data sources achieved an accuracy of 68.01% whereas JIF achieved an accuracy of 56.97%. A t-test found that the results were statistically significant ( $p < 0.05$ ). Additionally, desirability computed on the multiple crowd-sourced data sources outperformed all of the existing document ranking functions that were used as a benchmark. These experiments provided the first evidence that a desirability model that leverages the recency-frequency effect can improve document ranking.

#### **1.4 Predicting Document Clicks Using Information Scent**

The information scent model, which is based on the ACT-R spreading activation component, has been used to model the interaction of humans in information environments such as online browsing (Card et al., 2001; Chi, Pirolli, Chen, & Pitkow, 2001; Chi, Pirolli, & Pitkow, 2001), literature-based discovery (Chen et al., 2009; Goodwin, Cohen, & Rindflesch, 2012), debugging during programming (Lawrance, Bellamy, & Burnett, 2007a; Lawrance, Bellamy, Burnett, & Rector, 2008a, 2008b), and tag use and tagging behavior in on-line environments (Fu, 2008; S. Zhang, Farooq, & Carroll, 2009). Currently, no studies have investigated using information scent to model information seeking behavior in the biomedical domain. Additionally, the majority of past studies using information scent for click prediction were from the general user population and did not focus on modeling expert users. For example, only recently have researches explored using information scent to model expert behavior such as finding errors in programs (Lawrance, Bellamy, & Burnett, 2007b; Lawrance, Bellamy, Burnett, & Recker, 2008; Lawrance et al., 2013). The user population in this study, constrained to users in the Texas Medical

Center, has a high percentage of expert users since the user population is composed primarily of graduate students, clinicians, and researchers. Additionally, this chapter presents an updated mathematical framework for calculating information scent based recent insights from statistical IR models, which provides for an interpretation of information scent that more closely adheres to the Bayesian theory of the ACT-R theory and Information Foraging Theory. The following research question is the subject of Chapter 5.

**Research Question 4:** *Can information scent be used to predict biomedical document accesses?*

I conducted several experiments involving information scent on the documents accessed by HAM-TMC users through the PubMed IR system. The experiments were conducted separately for documents that received clicks (document clicks) and for documents that were downloaded (document downloads). The motivation here was to determine how well these models can predict accesses that resulted in downloads since downloads can be considered a stronger signal of relevance than document clicks alone. For example, a user can click a link for a document, view the abstract, and determine from that abstract text that they are not interested in reading the full text. A request for the full text is not necessarily a relevance judgment, but is an indication that the user wanted to read more of the document than just the abstract.

In all experiments, the best performing model was the new information scent model based on recent insights from research in statistical document ranking models. For the purpose of this section I will refer to the new information scent model as (IS-S) and the original model as IS. For document clicks, the IS-S model achieved the best performance with an accuracy

of 68.14%. However, the performance increase was not statistically significant as compared with the IS model, which achieved an accuracy of 65.16% ( $p > 0.05$ ). For document downloads, the best performing model was the IS-S model, which achieved an accuracy of 73.18%. In this instance, the model achieved statistically significant performance improvement over the IS model, which achieved an accuracy of 67.83% ( $p < 0.05$ ). In summary, these results support the hypothesis that information scent can be used for predicting document accesses in the biomedical domain.

### **1.5 Predicting Document Clicks Using information Scent and Desirability**

The research presented in Chapter 6 is the culmination of the research presented in Chapter 4 and Chapter 5. The specific goal of this chapter is to evaluate the combination of the desirability and information scent models for predicting document clicks. Thus, the research question explored in Chapter 6 is the following.

**Research Question 5:** *Will combining information scent and desirability improve click prediction accuracy?*

In the experiments presented in this Chapter 6, I found that the combination of information scent and desirability improved performance over the existing ranking functions. For document clicks, the combination of information scent and desirability improved performance over existing IR models by 9.81% and it improved performance by 6.9% for predicting document downloads. In both cases, the performance increase was found to be statistically significant ( $p < 0.05$ ).

### **1.6 Contributions**

The following are the main contributions of the work presented in this dissertation.

1. Proposed a theory to explain why the recency-frequency effect is present in data collected from a wide variety of domains
2. Demonstrated that the recency-frequency effect exists for documents accessed using different types of retrieval functions and different populations of users
3. Demonstrated that the recency-frequency effect can be leveraged to improve document ranking
4. Demonstrated that the combination of information scent and desirability improves ranking over existing state of the art ranking functions

## **1.7 Dissertation Outline**

The structure of the remaining chapters of this dissertation is outlined below. Each item in the list provides a brief summary of the main purpose of the chapter.

**Chapter 2: Background** – This chapter contains background information necessary to understand the experiments, evaluation techniques, and alternative information retrieval models used in this dissertation.

**Chapter 3: Related Works** – This chapter provides an overview of alternative approaches to modeling human information seeking behavior and IR applications that draw insight from cognitive science.

**Chapter 4: Predicting Document Clicks Using Desirability** – This chapter contains the experiments related to the recency-frequency effect. This chapter provides a mechanistic explanation for the cause of the recency-frequency effect (Research Question 1). In addition, this chapter shows that the recency-frequency effect is present for documents accessed through IR systems (Research Question 2) and can be leveraged for improving document ranking (Research Question 3).

**Chapter 5: Predicting Document Clicks Using Information Scent** – This chapter explores using information scent to predict document clicks (Research Question 4). In addition, this chapter presents a new interpretation of information scent based recent insights from research in probabilistic document ranking models and compares the performance to the existing information scent model.

**Chapter 6: Predicting Document Clicks Using Information Scent and Desirability** – This chapter explores using the combination of information scent and desirability (Research Question 5) for predicting document clicks. The performance is compared to a variety of existing ranking functions.

**Chapter 7: Conclusion** – This chapter summarizes the work in this dissertation and discusses limitations, contributions, and directions for future research.

## **Chapter 2: Background**

This chapter presents background information to contextualize the work in this dissertation. Section 2.1 presents a brief overview of the history of computational cognitive modeling; including the mathematical framework underlying the ACT-R theory of long-term memory and information scent. Section 2.2 provides an introduction to graph theory. Section 2.3 presents an overview of relevant document ranking methods. Section 2.4 provides and overview of the dimensionality reduction techniques used in this dissertation. Section 2.5 provides an overview of the evaluation methods used in this dissertation. In particular, this section provides an introduction to using query logs for the evaluation of document ranking approaches. Finally, Section 2.6 presents an overview of the method used in this dissertation for determining if empirical data obey a power law distribution.

### **2.1 Overview of Computational Cognitive Modeling**

#### **2.1.1 A brief history of computational cognitive modeling**

After the war, together with a small group of selected engineers and mathematicians, Johnny built, at the Institute for Advanced Study, an experimental electronic calculator, popularly known as Joniac, which eventually became the pilot model for similar machines all over the country. Some of the basic principles developed in the Joniac are used even today in the fastest and most modern calculators. To design the machine, Johnny and his co-workers tried to imitate some of the known operations of the live brain. This is the aspect which led him to study

neurology, to seek out men in the fields of neurology and psychiatry, to attend meetings on these subjects, and, eventually, to give lectures to groups on the possibilities of copying an extremely simplified model of the living brain for man-made machines.

(von Neumann, 1958).

Turing knew perfectly well the job he had to do, which was to manufacture or design a machine that would do the complicated sort of mathematics that had to be done in the Mathematical Division of [the National Physical Laboratory]. But he had all sorts of interesting things that he liked to do: for example, he was really quite obsessed with knowing how the human brain worked and the possible correspondence with what he was doing on computers .... Turing thought that the machine should be made quite simple, and at the same time should make everything possible that could be done. His particular purpose was to permit the writing of programs that modify programs, not in the simple way now common but rather in the way that people think.

(Newman, 1994)

As indicated by the above quotes, the idea of leveraging insights from cognitive science to inform the development of information systems as well as using information systems to understand the mind is not new. Aside from being examples of the earliest thinking in terms of the parallels between computers and the mind, the views of John von Neumann and Alan Turing represent the dominant research approaches in the field of computational cognitive modeling.

John von Neumann consulted what were at the time modern theories of the mechanisms of the brain when developing the central processing unit, which became utilized in almost all computers (von Neumann, 1958). John von Neumann was concerned primarily with simulating and developing hardware that mimicked the neural computation of the brain. In his posthumously published work, “The Computer and the Brain” (von Neumann, 1958), von Neumann discussed in detail how the crisp Boolean operators present in his von Neumann machines were inadequate as a model of neural processing and theorized as to how such devices could be developed or simulated. John von Neumann’s interests were very much in line with the connectionist view of cognition. In contrast, Alan Turing was interested in the algorithms of the mind and the general question of whether or not machines could think (Turing, 1950a, 1950b, 1956, 1999). The work of Turing became a foundation for the symbolic view of computational cognitive modeling. However, it should be noted that Turing did anticipate the role of learning and connectionist systems (Copeland & Proudfoot, 1996).

For many years, the connectionist and symbolic views of cognition existed as adversaries, which resulted in passionate debate (J. Fodor, 1997; J. A. Fodor & MCLAughlin, 1990; J. A. Fodor & Pylyshyn, 1988; Smolensky, 1987). The symbolic view is primarily concerned with the development of models using symbol manipulation (Newell & Simon, 1976b). The following presents the physical symbol system hypothesis proposed by Newell & Simon (1976).

A physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (symbol structure). Thus, a symbol structure is composed of a number



of instances (or tokens) of symbols related in some physical way (such as one token being next to another).

(Newell & Simon, 1976a)

The symbolic paradigm resulted in several early cognitive models in the 1960s and 1970s (Sternberg, 1966). Examples of early successes include the Logic Theorist (Newell & Simon, 1956) and the General Problem Solver (Newell, Shaw, & Simon, 1959). The major criticisms during the early days of symbolic cognitive models included brittleness when presented with new topics (frame problem (McCarthy & Hayes, 1969)), difficulty in ascribing meaning to symbols (symbol grounding problem (Harnad, 1990; Searle, 1980)), difficulty handling quantitative data (e.g. vision), difficulty with robust learning, and biological plausibility.

The connectionist<sup>2</sup> approach to cognitive modeling largely began in the 1980s with the publication of “Parallel distributed processing: Explorations in the microstructures of cognition” (McClelland, Rumelhart, & Group, 1986). The basic connectionist model is described by (Dawson & Shamanski, 1994) as follows.

PDP models are defined as networks of simple, interconnected processing units. A single processing unit is characterized by three components: a net input function which defines the total signal to the unit, an activation function which specifies the unit's current "numerical state", and an output function which defines the signal sent by the unit to others. Such signals are sent through connections between processing units, which serve as communication channels that transfer numeric

---

<sup>2</sup> Connectionist networks are commonly referred to as artificial neural networks (ANN) or parallel distributed processing (PDP)

signals from one unit to another. Each connection is associated with a numerical strength, which is used to scale transmitted signals. Connection strengths can be modified by applying a learning rule, which serves to teach a network how to perform some desired task.

(Dawson & Shamanski, 1994)

At the time, connectionist approaches seemed to be a viable alternative to symbolic modeling. By modeling cognition at the neuronal level, these models were able to excel at learning, which was a challenge for symbolic systems. Early on, the connectionist approach to cognitive modeling had numerous successes including modeling reading (Hinton & Shallice, 1991; Seidenberg & McClelland, 1989), sentence production (Dell, 1986), and verb learning (Rumelhart & McClelland, 1986). Connectionist models were eventually criticized (see (J. A. Fodor & Pylyshyn, 1988) for a particularly incendiary attack) because they could not incorporate background knowledge in learning, could only learn causal relationships (e.g. could not learn other semantic relations), and were not biologically plausible (e.g. is back-propagation a biologically plausible learning mechanism?).

Both connectionist and symbolic systems were able to overcome some of the initial criticisms through subsequent research. For example, tensor product variable binding connectionist systems overcame the inability of ANN to encode symbolic knowledge (Smolensky, 1990). Many of current approaches, including ACT-R, can be considered hybrid systems (i.e. symbolic-subsymbolic systems). Numerous researchers have argued the advantages of symbolic-subsymbolic systems (Kelly, 2003; Simen & Polk, 2010; Sun, 2001; A. Wilson & Hendler, 1993). The basic motivation for combining both approaches is straightforward. Ideally, one could leverage the learning capabilities and general

robustness of connectionist systems while taking advantage of the capabilities of symbolic systems (e.g. humans are able to follow symbolic reasoning, thus symbolic systems can “explain” how they arrived at a conclusion). However, this is a challenging task and is the focus of the majority of the research in developing symbolic-subsymbolic cognitive architectures.

### **2.1.2 ACT-R and Information Foraging Theory**

The computational models used in this dissertation can be classified as symbolic-subsymbolic models. The ACT-R theory, which is a major influence of the work in this dissertation, is a symbolic-subsymbolic model. From one view, ACT-R is a production system (rule-based system). The memory representation and the production rules in ACT-R are familiar to anyone who has worked with expert systems. The memory representation allows a developer to encode a network of concepts with typed relations between the concepts. Additionally, the developer can construct production rules to retrieve items from memory or to perform procedural tasks such as addition. The production rules in ACT-R provide access to information from different modules in the ACT-R architecture (e.g. visual module) as well as access to the long-term memory structure to satisfy a high-level goal. The role of symbols in the ACT-R architecture is summarized by Newell (Newell, 1990).

Symbols provide distal access to knowledge-bearing structures that are located physically elsewhere within the system. The requirement for distal access is a constraint on computing systems that arises from action always being physically local, coupled with only a finite amount of knowledge being encodable within a finite volume of space, coupled with the human mind’s containing vast amounts of

knowledge. Hence, encoded knowledge must be spread out in space, whence it must be continually transported from where it is stored to where processing requires it. Symbols are the means that accomplish the required distal access.

(Newell, 1990)

Anderson (2007) described symbols in the ACT-R architecture as being analogous to high-speed fiber cables in the brain allowing access to distal information.

Many components in the ACT-R architecture are “coated” (to use the term from Smolensky’s argument (Smolensky, 1987)) with subsymbolic functions, which enabled the architecture to learn and to emulate the general flexibility and adaptation seen in the human cognitive system. (J. Anderson, 2007) describes the role of symbolic and subsymbolic representations in the ACT-R theory of long-term memory as follows.

The symbolic level in ACT-R is an abstract characterization of how brain structures encode knowledge. The subsymbolic level is an abstract characterization of the role of neural computation in making that knowledge available.

(J. Anderson, 2007)

One role of subsymbolic computation in the ACT-R framework is to determine what distal information is accessed and how quickly the information is made available. In the case of long-term memory, symbolic structure encodes the relationships between items. Each item in memory has a prior probability function that describes how likely the item is to be needed based on the past access of the item (recency and frequency effect described in Chapter 1). This likelihood encodes the probability that a given item is needed given the current context (memory probe). It is this integration of the symbolic and subsymbolic that enables the ACT-R to replicate the results of numerous human memory experiments (J. R. Anderson,

Fincham, & Douglass, 1999; J. R. Anderson & Reder, 1999; J. R. Anderson, Reder, & Lebiere, 1996; Pavlik & Anderson, 2005; P. L. Pirolli & Anderson, 1985).

As discussed in Chapter 1, the information scent calculation used in the Information Foraging Theory is based on ACT-R's spreading activation function. For the purpose of this discussion, let the context (query terms within the context of information scent) be  $Q$  and let the proximal information cues be represented by  $D$ . Equation 2.1 presents the log form of Bayes' Theorem. The parameter  $\log \frac{P(D)}{P(\overline{D})}$  corresponds to the prior odds that a given document (in the context of Information Foraging Theory) or memory item (in the context of the ACT-R theory of long-term memory) would be accessed based on the past access patterns of that item. In the ACT-R framework, this is based on the recency-frequency effect discussed in Chapter 1. In the ACT-R terminology, this parameter is known as base-level activation. Generally, within the context of information scent, the prior odds are uniform to reflect the fact that people are not generally aware of the access patterns of documents. There are of course exceptions such as Google Scholar, where the search engine displays the number of citations for each article returned to the user (to the extent that accesses and citations can be seen as a reflection of one another). The parameter  $\sum_{j \in Q} \log \frac{P(j|D)}{P(j|\overline{D})}$  corresponds to the likelihood, which is a measure of how likely a given URL (in the context of Information Foraging Theory) or memory item (in the context of the ACT-R theory of long-term memory) is to be needed based on the relationship between the terms in the probe and the terms in the URL or memory item. In the ACT-R terminology, the likelihood is known as the association strength.

$$\log \frac{P(D|Q)}{P(\bar{D}|Q)} = \log \frac{P(D)}{P(\bar{D})} + \sum_{j \in Q} \log \frac{P(j|D)}{P(j|\bar{D})} \quad (2.1)$$

ACT-R and Information Foraging Theory make the simplifying assumption that the base rate probability of a given cue (query term in the context of IR)  $j$  occurring will not vary substantially from when the term appears and a given item  $D$  is not needed. This reduces the log odds calculation to  $\log \left( \frac{P(Q|D)}{P(Q)} \right)$ . After making the simplifying assumption, the following transformation in Equation 2.2 is applied to yield the approximation in Equation 2.3.

$$\frac{P(Q|D)}{P(Q)} = \frac{P(Q \cap D)}{P(D)} * \frac{1}{P(Q)} = \frac{P(D|Q)P(Q)}{P(D)} * \frac{1}{P(Q)} = \frac{P(D|Q)}{P(D)} \quad (2.2)$$

$$\log \frac{P(D|Q)}{P(\bar{D}|Q)} \approx \log \frac{P(D)}{P(\bar{D})} + \sum_{j \in Q} \log \frac{P(D|j)}{P(\bar{D}|j)} \quad (2.3)$$

Following from Equation 2.3, the activation function used by ACT-R is presented in Equation 2.4. The  $A_i$  parameter is the posterior odds of an item such as a document or term  $i$  being needed based on the context  $Q$ . The  $B_i$  equation reflects the context-independent prior odds from Equation 2.3 of an item  $i$  being needed independent of the current context. The parameter  $\sum_{j \in Q} W_j S_{ji}$  reflects the likelihood from equation 2.3. The  $W_j$  parameter is the attentional weight, which is used in the ACT-R framework to specify the validity of a piece of evidence in the context  $Q$ . For example, a given source of evidence may be noisy

(e.g. a face in a dimly lit room) and should be given less attentional weight than other cues in the context. For information scent, the attentional weight can be used to reflect the importance of the terms in the query. The  $S_{ji}$  parameter is the association strength and reflects the probability of an item being needed given the context.

$$\text{Activation equation} \quad A_i = B_i + \sum_{j \in Q} W_j S_{ji} \quad (2.4)$$

The base-level learning equation (prior odds) is shown in Equation 2.5. In this equation,  $d$  is a decay parameter and  $t_k$  is the time since the  $k$ th presentation of the item  $i$ . This function takes into account the recency and frequency effect discussed in detail in Chapter 1. This function requires that each access of the item be stored along with a time stamp for each access. In practice, Equation 2.6 is used, which requires storage of the creation date  $d$  of an item and the aggregate number of accesses  $n$  (Petrov, 2006).

$$\text{Base-level learning equation 1} \quad B_i = \log \left( \sum_{k=1}^n t_k^{-d} \right) \quad (2.5)$$

$$\text{Base-level learning equation 2} \quad B_i = \log \left( \frac{n}{1-d} t_n^{-d} \right) \quad (2.6)$$

The  $S_{ji}$  parameter is the context-dependent association strength (likelihood from Equation 2.3) and is estimated using Equation 2.7. Given the simplifying assumptions made by ACT-R and the information foraging theory, the likelihood estimation is equivalent to pointwise mutual information shown in Equation 2.8.

$$\text{Association strength equation} \quad S_{ji} \approx \log \left( \frac{P(i|j)}{P(i)} \right) \quad (2.7)$$

$$\text{Pointwise mutual information} \quad pmi(x; y) = \log \left( \frac{p(y|x)}{p(y)} \right) \quad (2.8)$$

## 2.2 Overview of Graph Theory

This section provides an overview of the graph growth mechanisms and graph metrics used in this dissertation. A graph is defined as a set of  $N$  vertices (also known as nodes) and set of  $K$  edges. Edges connect the vertices in the network and can be directed or undirected. Neighbors are defined as two vertices that are connected by an edge.

A frequently used metric for analyzing graphs is degree centrality. The degree centrality of vertices in the network have long been used as a measure of importance and numerous approaches exist for computing centrality over a graph including eigenvector centrality (Bonacich, 1972) and betweenness centrality (Freeman, 1977). In a directed network, the centrality of a vertex is known as in-degree and out-degree centrality and defines the



number of incoming edges and the number of outgoing edges respectively. If the graph is undirected, the in-degree and out-degree centrality measures are identical. The centrality metric used in this dissertation is shown in Equation 2.9. In Equation 2.9, the parameter  $deg(v_i)$  is the number of vertices connected to the vertex  $v_i$  and  $n$  is the number of vertices in the graph.

$$C(v_i) = \frac{deg(v_i)}{(n - 1)} \quad (2.9)$$

Another common metric used for characterizing graphs is the clustering coefficient. The clustering coefficient measures the extent to which the neighbors of a vertex tend to form cliques, which are regions in a graph where all of the vertices are connected. The clustering coefficient ( $\gamma_v$ ) shown in Equation 2.10 is a metric that characterizes the extent to which neighbor vertices of a vertex  $v$  are also neighbors of each other (Watts & Strogatz, 1998). In Equation 2.10,  $|E(\Gamma_v)|$  is the number of edges that are neighbors of  $v$  and  $\left(\frac{k_v}{2}\right)$  is the total number of possible edges in  $\Gamma_v$ .

$$\gamma_v = \frac{|E(\Gamma_v)|}{\left(\frac{k_v}{2}\right)} \quad (2.10)$$

The structure of a graph is an important feature, which can provide insight into the constraints that resulted in the formation of the graph. The networks that are pertinent for

this dissertation are random networks, small-world networks, and scale-free networks. Random graphs were pioneered by Erdős & Rényi (Bollobas, 1985; Erdos & Renyi, 1959, 1960, 1961). Random graphs are constructed by connecting the set of vertices in the graph to each other at random. Random networks are characterized as having a normal degree distribution<sup>3</sup>, small clustering coefficient, and short average path length. Small-world networks are characterized as having a short average path length (Watts & Strogatz, 1998). Finally, scale-free networks are characterized as having a majority of vertices that are loosely connected and a few rare vertices that are highly connected. The distribution of centrality of the concepts (nodes) in a scale-free network obey a power law distribution. In contrast to normal distributions, power-law distributions have a large number of small events and a few very large events. For example, if the heights of humans followed a power-law distribution, then the majority of people would be a foot tall and a few rare people would be hundreds or thousands of feet tall (Barabasi, 2003).

Several growth methods are used in this work to generate graphs with desired properties for experimentation. The motivation behind these types of studies is to investigate underlying growth rules that give rise to global properties of the graph such as having a power law degree distribution or having a high average clustering coefficient. Table 2.1 presents an overview of the graph growth methods used in this work and the emergent properties generated by the growth mechanisms.

Table 2.1

---

<sup>3</sup> If a random growth process is used in a temporal graph, the degree distribution will be Poisson.

### *Properties of growth mechanisms*

	Power law degree centrality distribution	High clustering coefficient	Short average path length
BA model	XXX		XXX
BA + Triad formation	XXX	XXX	XXX
ER model			XXX
ER model + Triad formation		XXX	XXX

The first, and most simple, graph growth mechanism is based on the Erdős & Rényi (Bollobas, 1985; Erdos & Renyi, 1959, 1960, 1961) random growth model (henceforth ER model). This growth process results in a degree distribution that obeys a Poisson distribution. The clustering coefficient generated by this growth process is small and the growth process generates a network with small average path length. The ER growth model works as follows.

1. Generate random network with  $M$  nodes where the connections are wired randomly with  $k$  connections per node.
2. For each new node added to the network, connect the new node with  $k$  existing nodes at random.

The Barabasi & Albert growth model (henceforth BA model) relies upon preferential attachment for connecting edges between new nodes and existing nodes (Barabasi & Albert, 1999). Preferential attachment has a relatively long history in the literature. The origins of the model are generally attributed to Herbert Simon (Simon, 1955), who in 1955 showed that the preferential attachment model could be used to account for Zipf's Law (Zipf, 1949). The preferential attachment model works by assuming that there are a set of objects that have some quantity attributed to them (e.g. amount of money or number of

connections in a graph). The future distribution of the quantity is a function of how much the existing quantity a given object has. That is, objects that have more of a given quantity will have a higher probability of receiving more quantity than objects that have less. This basic model is colloquially referred to as a “rich get richer” model. The contribution of Barabasi & Albert was that they showed that the preferential attachment model could account for the power law degree distribution observed in many real-world networks such as the WWW. The BA model works as follows.

1. Generate initial random network with  $M$  nodes where the connections are wired randomly with  $k$  connections per node.
2. For each new node added to the network, connect  $k$  edges from the new node to the existing nodes where the connection probability is calculated

as: 
$$\frac{\text{Number of nodes connected to } i}{\text{Number of nodes in network}}.$$

The result of a graph generated by the BA model yields the power law degree distribution seen in many real world networks and a short average path length among the nodes in the network. However, the clustering coefficient produced in a graph generated by the BA model is orders of magnitude smaller than what is seen in many real-world networks.

To overcome the shortcoming of the BA model to produce high clustering coefficients, several approaches have looked at augmenting the growth processes with triad formation (Holme & Kim, 2002; Sousa, 2005; Volz, 2004). The intuition behind triad formation is motivated by observations from real-world networks. For example, in a social network, if *person A* and *person B* are connected then there is a high probability that there is a third person that both *person A* and *person B* know. These formations are known as triads. Adding a triad formation step to the growth process can result in graphs with average

clustering coefficients that are magnitudes larger than those generated by either the ER model or the BA model. The basic triad formation process is presented below.

1. Generate initial random network with  $M$  nodes where the connections are wired randomly with  $k$  connections per node.
2. Attach new node to a node  $N$  from the existing network using either BA model or ER model.
3. Attach new node to  $K$  neighbors of node  $N$  at random.

## **2.3 Overview of Relevant IR models**

This section presents an overview of the IR models that were used in this dissertation. Section 2.3.1 presents an overview of language models, which are used extensively in this dissertation. The other models covered in the review section were used for performance comparison. Section 2.3.2 presents an overview of the TF-IDF ranking function. Section 2.3.3 presents an overview of the BM25 ranking function. Finally, Section 2.3.4 presents an overview of the divergence from randomness ranking function.

### **2.3.1 Language models**

Language models originated in machine translation research (Brown et al., 1990) and speech recognition (Jelinek, 1997) and were first applied to information retrieval by Ponte and Croft (Ponte & Croft, 1998). Language models have many desirable properties. For example, they provide theoretical justification for commonly used heuristics such as term frequency (TF), inverse document frequency (IDF) weighting, and document length normalization (Hiemstra, 2000a; Hiemstra & Kraaij, 1998; Singhal, Buckley, & Mitra, 1996). The term language model refers to a probabilistic model of text and underlies much

of the work in statistical natural language processing (Manning & Schtze, 1999) and probabilistic topic models (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004).

The basic language model proposed by Ponte and Croft is shown in Equation 2.11. Let  $Q$  be the query,  $D$  be the document, and  $\theta_D$  be a language model estimated on document  $D$ . Documents are returned to the user based on the likelihood of the document generating the query ( $score(Q|\theta_D)$ ). The focus of language model research is on estimating the language model for a document  $\theta_D$ . In general, the focus is on smoothing language models. For example, the maximum likelihood (ML) estimator is shown in Equation 2.12. In this equation,  $c(w_i, D)$  represents the counts of word  $w_i$  in document  $D$  and  $|D|$  represents the length of document  $D$ .

$$score(Q, D) = p(Q|\theta_D) \tag{2.11}$$

$$p(X_i = 1|D) = \frac{c(w_i, D)}{|D|} \tag{2.12}$$

The problem with the ML estimate is that a document with an unseen word will receive a likelihood score of zero, which has the result of reducing the retrieval to an exact match system. Thus, the majority of language model research focuses on smoothing the language model estimate such that zero probabilities are not assigned to documents with unseen words. Numerous statistical approaches have been proposed for smoothing language models. Examples of smoothing approaches are Dirichlet prior smoothing (C. Zhai & Lafferty, 2001b) and Kullback–Leibler (KL) divergence (Lafferty & Zhai, 2001; C. Zhai & Lafferty, 2001a). Equation 2.13 presents the multinomial distribution model proposed

by (Hiemstra & Kraaij, 1998; D. H. Miller, Leek, & Schwartz, 1999), which relies upon the ML estimator, but smoothes the estimate using a background language model. This model is known as the Jelinek-Mercer (JM) model. The  $p(w|C)$  is the probability of the word occurring in the entire document collection. The parameter  $\lambda$  is a smoothing parameter in the range  $[0,1]$ , which controls the influence of the ML estimate and the background language model in the linear integration.

$$p(w|D) = (1 - \lambda) \frac{c(w, D)}{|D|} + \lambda p(w|C) \quad (2.13)$$

An alternative method that has gained interest is the use of Dirichlet smoothing, which is shown in Equation 2.14. In Dirichlet smoothing the pseudocount parameter  $\mu$  is set to a large number (some report as high as 1500). Dirichlet smoothing is generally thought to outperform JM smoothing (Smucker & Allan, 2005; C. Zhai & Lafferty, 2001c). In Dirichlet smoothing the amount of smoothing is a function of the length of a document. For example, longer documents receive less smoothing whereas shorter documents receive more smoothing.

$$p(w|D) = \frac{p(w; d) + \mu p(w|C)}{|d| + \mu} \quad (2.14)$$

Recent work has investigated using term associative networks for smoothing. (Mei, Zhang, & Zhai, 2008) proposed a generalized framework for smoothing using networks which is shown in Equation 2.15. The parameter  $w(u)$  represents the importance of a vertex in a graph, which can be measured using a number of methods such as PageRank, degree centrality, or betweenness centrality. The parameter  $sim(u, v)$  is a measure of the similarity between the nodes  $u$  and  $v$ . The similarity  $sim(u, v)$  can be computed any number of methods such as TF-IDF or distributional semantics methods. The parameter  $f_v$  represents the smoothed value based on the network, which has the same role as the background language model in traditional smoothing. The parameter  $\bar{f}_u$  is the non-smoothed document language model that is typically computed using the ML estimate. Taking the first-order partial derivative of  $O(C)$  yields Equation 2.16. Finally, letting  $\frac{\partial O(C)}{\partial f_u} = 0$  yields Equation 2.17, which is the ranking function used for smoothing using an associative network. (Mei, et al., 2008) showed that the procedure for document ranking using associative networks in Equation 2.17 improved performance over both JM and Dirichlet smoothing.

$$O(C) = (1 - \lambda) \sum_{u \in V} w(u) (f_u - \bar{f}_u)^2 + \lambda \sum_{(u,v)} sim(u, v) (f_u - f_v)^2 \quad (2.15)$$

$$\frac{\partial O(C)}{\partial f_u} = 2(1 - \lambda) Deg(u) (f_u - \bar{f}_u) + 2\lambda \sum_{v \in V} sim(u, v) (f_u - f_v) \quad (2.16)$$

$$f_u = (1 - \lambda) \bar{f}_u + \lambda \sum_{v \in V} \frac{sim(u, v)}{Deg(u)} f_v \quad (2.17)$$



### 2.3.1.1 Relationship between language models and TF-IDF ranking

This section discusses the relationship between language models and TF-IDF ranking. A frequently provided motivation for using language models for document ranking is that they provide a probabilistic justification for many of the components of TF-IDF ranking including document length normalization, term frequency, and term importance. The purpose of this section is to demonstrate how the language model accounts for these metrics since it is not necessarily obvious from a cursory glance. The derivation of language models presented here follows from (Hiemstra, 2000b; Hiemstra & de Vries, 2000). Equation 2.18 shows the basic JM language model. Next I divide both sides of the equation by  $\prod_{i=1}^n (1 - \lambda)P(T_i)$  to yield Equation 2.19. Both  $\lambda$  and  $P(T_i)$  are constants and dividing by both quantities will not impact the ranking. Equation 2.20 is the term frequency rank for a term normalized by the total number of terms in the document. Equation 2.21 presents the background language model, which is the ratio between the frequency for a term in the collection divided by the total term frequency for all of the terms.

$$P(T_1, T_2 \dots T_N | D) = \prod_{i=1}^n ((1 - \lambda)P(T_i) + \lambda P(T_i | D)) \quad (2.18)$$

$$P(T_1, T_2 \dots T_N | D) \propto \prod_{i=1}^n 1 + \frac{\lambda P(T_i | D)}{(1 - \lambda)P(T_i)} \quad (2.19)$$

$$P(T_i = t_i | D = d) = \frac{tf(t_i, d)}{\sum_t tf(t, d)} \quad (2.20)$$

$$P(T_i = t_i) = \frac{df(t_i)}{\sum_t df(t)} \quad (2.21)$$

Equation 2.22 presents the log form of Equation 2.19 updated with Equations 2.20 and 2.21. From this form, Equation 2.22 can be broken down into different components that map closely to the TF-IDF weighting function as shown in Table 2.2.

$$P(T_i = t_i, T_2 = t_2 \dots T_n | D) \propto \sum_{i=1}^n \log \left( 1 + \frac{\lambda * tf(t_i, d) * \sum_t df(t)}{(1 - \lambda) * df(t_i) * \sum_t tf(t, d)} \right) \quad (2.22)$$

Table 2.2

*Relationship between language model and TF-IDF ranking*

$\frac{tf(t_i, d)}{df(t_i)}$	TF-IDF weight
$\frac{1}{\sum_t tf(t, d)}$	Document length normalization for document $d$
$\sum_t df(t)$	Constant for any document $d$ and term

### 2.3.2 TF-IDF overview

The TF-IDF ranking function is one of the oldest ranking functions (Sparack Jones, 1972), but remains competitive when compared to more recent ranking functions (W. R. Hersh et al., 2006). The TF-IDF ranking function has three major components: term frequency (TF), inverse document frequency (IDF), and document length normalization. The  $TF$  reflects the frequency of a given term in the document (i.e. how well does the term define the

document) and the *IDF* is a measure of term selectivity (i.e. how well does the term discriminate between documents), which takes into account how frequently the term occurs in the entire corpus. The motivation for normalization is that longer documents will naturally have higher *TF* values. The normalization essentially turns the term vector for the document into a unit vector. The TF-IDF ranking function used in this dissertation is shown in Equation 2.23. The IDF function used in this dissertation is shown in Equation 2.24.

$$score(q, d) = \sum_{t \in q} \frac{tf(t)}{length(d)} * idf(t) \quad (2.23)$$

$$idf(t) = 1 + \log \left( \frac{\text{number of documents}}{1 + \text{number of documents with } t} \right) \quad (2.24)$$

### 2.3.3 BM25 overview

The BM25 algorithm is a probabilistic retrieval function based on the probabilistic retrieval framework (K. S. Jones, Walker, & Robertson, 2000). The BM25 ranking function was implemented as part of the Okapi IR system and is often referred to as Okapi BM25 in the literature. Equation 2.25 presents the BM25 ranking function. The BM25 model is similar to the TF-IDF ranking function, but is based on probabilistic estimates of the parameters.

$$w_j(\bar{d}, C) = \frac{tf * (k_1 + 1)}{k_1 \left( (1 - b) + b \frac{dl}{avdl} \right) + tf} * \log \left( \frac{N - df(t) + 0.5}{df(t) + 0.5} \right) \quad (2.25)$$

$k_1$  Is a free parameter that controls the non-linear tf effect in the model.

$b$  Is a free parameter that controls the document length normalization.

$tf$  is the term frequency of a term in the document

$dl$  is the document length

$avdl$  is the average document length

$N$  is the number of documents

### 2.3.4 Divergence from randomness overview

The divergence from randomness model originated in (Amati & van Rijsbergen, 2002).

The divergence from randomness is a non-parametric model that measures the divergence of the term distribution from random. The divergence from randomness model shares similarity with language models in that it takes into account a document probability and a collection probability. Equation 2.26 presents the divergence from randomness model. The parameter  $Prob_1$  represents the information content of the term in a document. The parameter  $Prob_2$  corresponds to the information gain of a term. There are many different instantiations of the model. The instantiations of  $Prob_1$  and  $Prob_2$  used in this dissertation are shown in Table 2.3 and Table 2.4 respectively. The term frequency normalization techniques used in this dissertation are shown in Table 2.5.

$$Score(Q, D) = \sum_{t \in Q} weight(t, D) = \sum_{t \in Q} (1 - Prob_2) * (\log Prob_1) \quad (2.26)$$

Table 2.3

*Instantiations of  $Prob_1$  metric*

Laplacian normalization	$P_{risk}(t D) = \frac{1}{tf + 1}$	(2.27)
Bernoulli normalization	$P_{risk}(t D) = \frac{tf_{collection}}{dft * (tf + 1)}$	(2.28)

Table 2.4  
*Instantiations of Prob<sub>2</sub> metric*

Bose-Einstein	$P_m(t Collection) = \left( \frac{N}{tf_{collection} + N} \right) \left( \frac{tf_{collection}}{tf_{collection} + N} \right)^{tf}$	(2.29)
TF-IDF randomness	$P_m(t collection) = \left( \frac{tf + 0.5}{N + 1} \right)^{tf}$	(2.30)

Table 2.5  
*Instantiations of term frequency normalization*

$\hat{tf} = tf * \log \left( 1 + \frac{avdl}{dl} \right)$	(2.31)
$\hat{tf} = tf * \log \left( 1 + c * \frac{sl}{dl} \right)$	(2.32)

## 2.4 Evaluation Techniques

### 2.4.1 Evaluation using the Cranfield method

The vast majority of IR evaluations such as many of the ad-hoc document ranking competitions of TREC follow the protocol created by Cleverdon at Cranfield University during the 1960s, which are often referred to as the Cranfield experiments (Cleverdon,

1960, 1967; Cleverdon & Keen, 1966). The standard protocol established by the Cranfield experiments is to have judges rate the relevance of documents given particular queries. Once the relevance judgments are obtained, different retrieval functions can be evaluated on the same queries. The hypothesis of this protocol is that the performance gains found by algorithms on these test collections will translate to real-world performance gains. The remainder of this section will review the metrics used for Cranfield-based evaluation.

Precision is the fraction of retrieved documents that are judged as relevant. Recall is the fraction of relevant documents that are retrieved. In many of the evaluations in this dissertation, the precision and recall are evaluated at different cutoff points. The motivation behind evaluating IR systems at different cutoff points is that numerous studies have shown that the majority of users only look at the first 1-2 pages of search results (Islamaj Dogan, et al., 2009). Thus, it is important that the documents in the first 1-2 pages are relevant to the query. The mean average precision (MAP) (also known as “average precision at seen relevant documents”) is shown in Equation 2.33. MAP measures the precision at each point when a new relevant document is retrieved for each query and averages the scores for all queries. Specifically, in this work the MAP at  $N$  ( $MAP_N$ ) is used where a threshold is specified and the MAP for the top  $N$  documents are used for evaluation. For example,  $MAP_{10}$  would evaluate the performance of a ranking algorithm for the first 10 documents retrieved for all queries.

$$MAP = \frac{1}{N} \sum_{j=1}^N \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(doc_i) \quad (2.33)$$

$Q_j$  is the number of relevant documents for query  $j$   
 $N$  is the number of queries  
 $P(doc_i)$  is the precision at  $i$ th relevant document

One problem with looking only at MAP for evaluation is that it does not take into account recall. Figure 2.1 presents example results for two different ranking functions for the same query. In this example, assume that there are four relevant documents for the query. The  $MAP_{10}$  results for ranking function 1 is 0.6425 and ranking function 2 is 0.835. The  $MAP_{10}$  result for ranking function 2 is higher even though ranking function 1 returned more relevant documents. Thus looking only at  $MAP_N$  can be misleading. A solution to this problem is to take the harmonic mean of  $MAP_N$  and  $recall_N$ . The harmonic mean is calculated by  $2 * \frac{MAP_N * recall_N}{MAP_N + recall_N}$ . In the examples in Figure 2.1, the harmonic mean for ranking function 1 is 0.7823 and ranking function 2 is 0.6255, which reflects the general intuition regarding which ranking algorithm achieved the best performance.

Query 1 with ranking function 1		
Rank	Relevance	$P(doc_i)$
1	X	1.00
2		
3	X	0.67
4		
5		
6	X	0.5

Query 1 with ranking function 2		
Rank	Relevance	$P(doc_i)$
1	X	1.00
2		
3	X	0.67
4		
5		
6		

7		
8		
9		
10	X	0.4
Average		0.6425

7		
8		
9		
10		
Average		0.835

Figure 2.1. Example  $MAP_{10}$  for two ranking functions

### 2.4.2 Evaluation using query logs

One of the drawbacks of Cranfield inspired experiments is that studies have shown that the performance gains of IR systems using this protocol do not necessarily translate to real-world user satisfaction (Al-Maskari, et al., 2008; Allan, et al., 2005; W. Hersh, et al., 2001; Jarvelin, 2009; Macdonald & Ounis, 2009; Sanderson, et al., 2010; Smith & Kantor, 2008; Smucker & Jethani, 2010; Su, 1992; Turpin & Scholer, 2001, 2006; Urbano, et al., 2012). An often-proposed solution to this problem is to evaluate IR systems in the real world with many users. Toward this aim, Thorsten Joachims developed a methodology for evaluating search engines using query logs (Joachims, 2003; Joachims, Granka, Bing Pan, et al., 2007; Joachims, Granka, Pan, Hembrooke, & Gay, 2005; Joachims, Granka, Pan, et al., 2007; Radlinski & Joachims, 2006, 2007; Radlinski, Kurup, & Joachims, 2008). Commercial IR systems automatically collect query logs that contain information such as IP addresses, user emails, user queries, and the documents clicked in response to the query. Thus, query logs provide the potential to collect information automatically from thousands of users that can later be used to evaluate different ranking functions.

Joachims showed that document clicks could not be interpreted as absolute relevance judgments; however, document clicks can be interpreted as relative relevance judgments



within the context of the other documents in the result set. For example, consider a result set where *document A* is at rank 1 and *document B* is at rank 2. If a user clicked *document B* and not *document A*, Joachims showed that a pairwise preference of *document B* > *document A* can be extracted with high precision as compared to explicit judgment.

The remainder of this section is organized as follows. Section 2.4.2.1 presents an overview of the methods for extracting pairwise judgments developed by Joachims. Section 2.4.2.2 presents an overview of how these pairwise judgments can be used for evaluating the performance of ranking functions.

#### **2.4.2.1 Extracting pairwise judgments**

Table 2.5 presents the rules developed by Joachims for extracting pairwise judgments from query logs. The rules were evaluated within one page of query results. Joachims developed additional rules for extracting pairwise judgments from query chains (i.e. multiple queries pertaining to same information need) (Joachims, et al., 2005). Accurately segmenting query logs into query chains is an area of current research (for a review see (Risvik, Mikolajewski, & Boros, 2003)). The current methods have accuracy ranging from 75%-90% depending upon the test collection used. Thus, segmenting query logs into sessions can inject additional noise into the pairwise judgment extraction process and I did not utilize this approach in this dissertation. The column labeled “Accuracy by abstract judgment” presents the accuracy of the extracted pairwise judgments based on explicit human judgments made by looking at the abstracts. In these experiments, the Google search engine was used and the abstracts are title of the Web page, URL, and first 1-2 sentences from the Web page. In these experiments, the inter-judge agreement had a correlation of 82.5. The column titled “Accuracy by Web page judgment” presents the accuracy of the

extracted pairwise judgments and explicit human judgments made by looking at the actual Web page content. In these experiments, the inter-judge agreement had a correlation of 86.4.

Table 2.6

*Methods for extracting pairwise judgments*

Pairwise judgment extraction method	Accuracy by abstract judgment	Accuracy by Web page judgment
Click > Skip Above	88.0	78.2
Last Click > Skip Above	89.7	80.9
Click > Earlier Click	75.0	64.3
Click > Skip Previous	88.9	80.7
Click > No Click Next	75.6	67.4

Joachims advised that the rules *Click > earlier click* and *Click > no click next* not be used. Additional experiments looked at order effects imposed by Google's ranking by reversing the rankings. These studies showed that the accuracy of the *Click > earlier click* rule dropped to 28.6% and the accuracy of the *Click > No click next* rule dropped to 70%. The other rules were not significantly impacted by the reverse rankings. The rules *Click > Skip Previous* and *Last Click > Skip Above* both generate pairwise judgments that are a subset of the pairwise judgments extracted by the rule *Click > Skip Above*. Joachims found that the accuracy improvements of the *Click > Skip Previous* and *Last Click > Skip Above* rules were not statistically significant as compared to *Click > Skip Above* rule. Thus, only the rule *Click > Skip Above* is used for extracting pairwise judgments in this dissertation.

The following is an example of a query with seven documents in the result set. The clicked documents are denoted with a “\*”. The example is adapted from (Radlinski & Joachims, 2005).

$$Q1: I_1^* I_2 I_3^* I_4 I_5^* I_6 I_7$$

The *Click > Skip Above* rule states that clicked documents are more relevant than any documents that were skipped that preceded it. (Radlinski & Joachims, 2005) formally defined the *Click > Skip Above* rule as follows.

*For a ranking  $(I_1, I_2, I_3 \dots)$  and a set  $C$  containing the ranks of the clicked-on links, extract a preference example  $rel(I_i) > rel(I_j)$  for all pairs  $1 \leq j < i$ , with  $i \in C$  and  $j \notin C$ .*

The *Click > Skip Above* rule would extract the following pairwise judgments from  $Q1$ .

$$I_3 > I_2, I_5 > I_4, I_5 > I_2$$

#### 2.4.2.2 Evaluation using pairwise judgments

This section reviews the use of pairwise judgments for evaluation. Given a corpus of query logs, the query sessions are segmented by IP address or email address. For each query issued by a user, the rule *Click > Skip Above* is applied to extract pairwise judgments for the user.

Table 2.7 presents the results for two different data sources that are used for predicting document accesses of five users. In this example, assume that the two methods involve the use of citation counts and past document accesses for predicting user accesses. Two measures are of interest when using this type of information for ranking: click precision (Equation 2.34) and click coverage (Equation 2.35). As an example, consider the following pairwise judgment extracted for a user:  $doc1 > doc2$ . If the citation count for  $doc1$  was 7

and the citation count for *doc2* was 5, this would result in a correct ordering if the results were ranked by citation count. The click coverage metric measures the number of pairwise judgments where at least one item in the pairwise judgment has information in a given data set to enable ranking. For example, if a pairwise judgment is extracted where *doc1* > *doc2*, and the data set used for ranking contains information for *doc1*, the click coverage for the data set would be increased. The click coverage metric primarily pertains to using information such as citation counts or document downloads for ranking. Document ranking functions such as TF-IDF will produce a ranking for each document thus click precision will only matter for these experiments.

$$\text{Click Precision} = \frac{\text{Number of correct pairwise judgments}}{\text{Number of correct} + \text{number of incorrect}} \quad (2.34)$$

$$\begin{aligned} \text{Click Coverage} & \quad (2.35) \\ &= \frac{\text{Number of pairwise judgments where at least one item is ranked}}{\text{Total number of pairwise judgments}} \end{aligned}$$

Table 2.7 presents an example of using pairwise judgments to evaluate two data sources (historical document accesses versus citation counts) for five users. Since each user has a different number of extracted pairwise judgments, weighted averages are used to compute the results. The click precision is higher for the historical document access data, but the click coverage is much lower than using citation counts. The harmonic mean for historical document accesses is 0.7721 whereas the harmonic mean is 0.7895 for citation counts. For

these five users, it can be concluded that the citation count data are more effective for predicting document accesses than historical document access data.

Table 2.7

*Example results using pairwise judgments*

User ID	Number of pairwise judgments	Document accesses			Citation counts		
		Click Precision	Click Coverage	Harmonic Mean	Click Precision	Click Coverage	Harmonic Mean
User_1	500	75.0%	81.0%	0.7788	65.0%	99.0%	0.7848
User_2	101	68.0%	82.0%	0.7435	70.0%	98.0%	0.8167
User_3	260	77.0%	79.0%	0.7799	68.0%	96.0%	0.7961
User_4	50	76.0%	70.0%	0.7288	75.0%	99.0%	0.8534
User_5	300	72.0%	83.0%	0.7711	65.0%	95%	0.7719
	<b>Weighted Average</b>	0.7414	0.8070	0.7721	0.6647	0.9728	0.7895

## 2.5 Finding Power Law Distributions in Empirical Data

The original method for determining if a data set obeys a power law distribution was established by (Pareto, 1964). The first step is to create a log-log plot of the histogram. A linear regression is used to fit the data and if the  $R^2$  is above some threshold, then it is asserted that the data follow a power law distribution. Additionally, the slope of the fitted regression serves as the estimate of the scaling parameter.

There are several problems with this method. First, real-world data sets rarely follow a power law for each point in the data set. The nature of a power law distribution implies that there are very large rare events that can inject noise and severely impact a linear regression fit. The data set used as an example in this section is the PLOS document accesses for one day. Consider the log-log plot of a power law distribution in Figure 2.2. The linear regression fit has an  $R^2$  of 0.7993 and is greatly impacted by a few outliers at the tail of the

distribution. Many researchers remove the points at the tail and claim that the data obey a power law distribution within the truncated range. For example, consider Figure 2.3 which shows a truncated log-log plot where documents that have accesses higher than 500 are removed. The truncated data set resulted in a linear regression fit of 0.9708. One problem with this approach is that the cutoff threshold is determined on an ad-hoc basis. An unbiased method for automatically determining the cutoff point is desirable. Additionally, alternative distributions such as exponential or lognormal can produce nearly straight lines on a log-log plot. This method does not compare the fit of the data to alternative distribution, which could describe the data as well as the power law distribution.

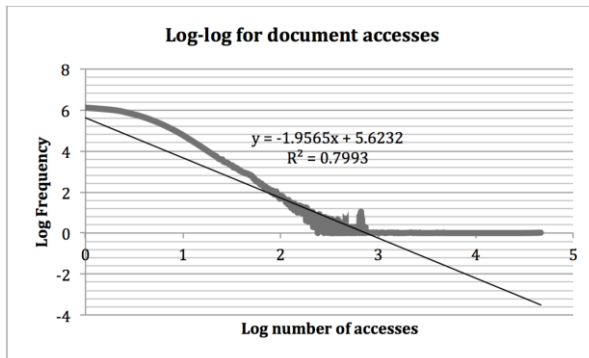


Figure 2.2. Example power law distribution

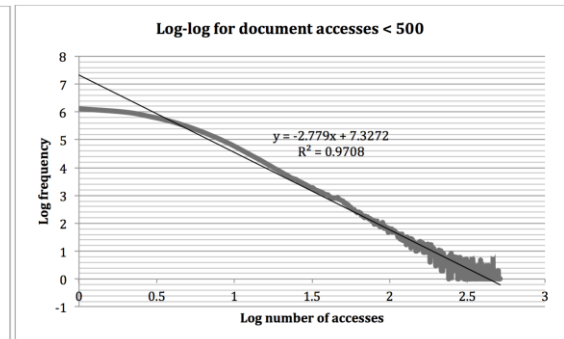


Figure 2.3. Example log-log plot

The method developed by (Clauset & Shalizi, 2009) seeks to address the shortcomings of traditional methods discussed previously. Clauset & Shalizi (2009) demonstrated that their method provided more accurate estimates of the scaling parameter  $\alpha$  than the traditional method based on Pareto's work. For example, they generated a data set with known  $\alpha =$

2.5. They attained an estimate of  $\alpha = 2.50$  using their method. Estimates attained using regression fits to the log-log plot of the histogram varied widely in the range  $1.39 \leq \alpha \leq 2.50$ . Additionally, Clauset & Shalizi (2009) analyzed twenty-four data sets where previous literature had found power law distributions. They found that the presence of a power law distribution was inconclusive for approximately 30% of these studies using the more precise method. The summary of Clauset & Shalizi's method is provided below.

1. Estimate the parameters  $X_{min}$  and  $\alpha$  of the power-law model. This method involves testing each point in the empirical data set to find the point where the Kolmogorov-Smirnoff (KS) statistic is minimized. Maximum Likelihood Estimate (MLE) is used to obtain an estimate for  $\alpha$ .
2. Calculate the goodness-of-fit between the truncated empirical data and the power law. The first step is to generate a large number of synthetic data sets using the estimated parameters from Step 1. (Clauset & Shalizi, 2009) recommend 2,500 synthetic data sets. The KS statistic is used to compare the synthetic data set and the empirical data set. In this case, the null hypothesis is that the two distributions come from the same distribution. If the p value is above 0.1, then we fail to reject the null. For 2,500 experiments, it is expected that more than 90% should fail to reject the null hypothesis in order for the power law to be a good fit.
3. Compare the power law with alternative hypotheses. This step involves comparing the empirical data with alternative distributions. Step 1 and Step 2 are repeated for each alternative distribution under consideration.

The alternative distributions used for comparison in this dissertation are the exponential (Equation 2.36) and log normal (Equation 2.37) distributions. In the exponential

distribution, the parameter  $\lambda$  is called the rate parameter and is estimated from the empirical data. In the log normal distribution, the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) are both estimated from the empirical data.

$$y = f(x; \lambda) = \lambda e^{-\lambda x} \quad (2.36)$$

$$y = f(x|\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \quad (2.37)$$

I briefly demonstrate how the Clauset & Shalizi method is applied using the example data in Figure 2.2. The estimated parameters are shown for each model of comparison in Table 2.8. The “xMin” parameter is the cutoff used to truncate the data set. The “xMin results” are the results of the tests where the truncated data are used with the estimated xMin parameter. The “All points results” column are the results of the experiments where all of the data points are used. For each case, 2500 experiments were performed following the advice from (Clauset & Shalizi, 2009). The power law distribution passed 98.76% of the experiments with 181 data points in the truncated distribution. The exponential distribution passed 96.28% of the experiments, but only seven data points in the truncated data set obeyed the distribution. The log normal distribution can be completely ruled out since neither of the cases passed any of the statistical significance tests. Based on the results of this analysis, it can be concluded that the truncated data set obeys a power law. Additionally, the power law distribution fits many more data points than the exponential distribution. Thus, the power law is the best model of those tested to describe the empirical data.



Table 2.8

*Results from sample data*

Power law			Exponential				Log normal				
$\alpha$	xMin Results		$\lambda$	xMin Results		All points result	$\mu$	$\sigma^2$	xMin Results		All points result
	xMin	Result		xMin	Result				xMin	Result	
1.33	181	2469	2.33e-06	7	2407	0	11.73	1.87	15	0	0

## **Chapter 3: Related Work**

This chapter presents an overview of research that has leveraged either ACT-R (Adaptive Control of Thought—Rational) or Information Foraging Theory in the design of information systems (Section 3.1). Additionally, this chapter presents an overview of alternative computational theories of cognition that have had impact on IR systems (Section 3.2). The review in Section 3.2 is purposefully limited to computational models. There are numerous qualitative models that describe information seeking behavior. For example, the berry picking model is a qualitative model of information seeking behavior with the idea that a user's information need is satisfied through successive queries with evolving information needs (Bates, 1989). While these types of models are useful for understanding information seeking behavior they are omitted from this review unless they provide insight into how one can model this phenomenon from a computational viewpoint.

### **3.1 Impact of ACT-R and Information Foraging Theory on the Design of Information Systems**

This section focuses on applications of the ACT-R and Information Foraging Theory for the development of information systems. Section 3.1.1 presents studies that have leveraged aspects of these theories to develop personal document management systems. Section 3.1.2 presents studies that have leveraged aspects of these theories to develop recommendation systems. Section 3.1.3 presents an overview of algorithms that leveraged the recency-frequency effect described in Chapters 1 and 2. Finally, Section 3.1.4 presents an overview

of technologies developed based on these theories to improve the ability of users to more effectively browse.

### **3.1.1 Personal document management**

The earliest work, which applied aspects of the ACT-R theory for IR was the Memory Extender (W. P. Jones, 1986a, 1986b). The Memory Extender was a personal document management system that utilized the ACT-R spreading activation function for retrieving and ranking documents from a personal database of documents. The spreading activation function ranked documents based on the terms in the documents and “context terms” that the user could assign to documents based upon the context of use. For example, I could assign a context term such as “dissertation literature review” to the documents (W. P. Jones, 1986a, 1986b).

The Memory Extender used the decay mechanism of the ACT-R theory. The association between terms and documents slowly decayed over time until they hit zero. At this point, Memory Extender notified the user that the document was a candidate for deletion and the user would have the option of deleting the document or strengthening the relationships in the network to prevent deletion. There were no formal evaluations or user studies conducted on the Memory Extender.

### **3.1.2 Recommendation systems**

Several studies have investigated the use of the ACT-R theory of long-term memory for collaborative filtering (Van Maanen & Marewski, 2009; Van Maanen et al., 2009). One system is the Personal Publication Assistant, which accepts a set of talks to be given at a conference and recommends talks in which a user may be interested based on a user profile (Van Maanen & Marewski, 2009; Van Maanen, et al., 2009). The user profile is constructed

based on the user's previous publications. The terms extracted from more recent articles are weighted higher than older publications to reflect the fact that research interests evolve over time. The user's profile serves as a query and the ACT-R spreading activation function ranks the conference abstracts based on the weighted terms in the profile.

The Personal Publication Assistant was compared with human judgments and a strong correlation was found (Van Maanen & Marewski, 2009; Van Maanen, et al., 2009). In a follow-up study, the ACT-R inspired recommender was compared to six other models originating in decision theory (Van Maanen & Marewski, 2009). The ACT-R inspired method was outperformed by a method known as the take-the-best heuristic (Czerlinski, Gigerenzer, & Goldstein, 1999), which is a simple heuristic for selecting the best of two possible options.

(Woodruff, Gossweiler, Pitkow, Chi, & Card, 2000) developed a personalized book recommender based on the spreading activation mechanism of ACT-R. Woodruff et al. (2000) proposed to imbed the recommender within an electronic book (eBook) to provide users with suggestions of related content that may be of interest. The user profile is comprised of books that a given reader has read. The evaluation of the method showed that the spreading activation recommendation engine had a high correlation (0.8) with human judgments.

### **3.1.3 Document prior probability estimation**

There has been some interest in using the prior probability function of the ACT-R long-term memory theory to predict document accesses. Pitkow & Recker (1994) analyzed the access patterns of Web pages on the WWW and found that these access patterns had the recency and frequency effect (Recker & Pitkow, 1996). Based on this finding, Recker &

Pitkow (1994) developed an algorithm for caching documents based on the probability of future accesses (Pitkow & Recker, 1994). The authors did not compare the new method to alternative caching approaches so it is not possible to determine if the method resulted in improved performance. However, the method did result in a patent (P. L. Pirolli & Pitkow, 2000).

#### **3.1.4 Tools to support browsing**

This section focuses on applications of the Information Foraging Theory and ACT-R long-term memory theory to develop tools and algorithms to improve information access. Substantial research has focused on applying insights from the Information Foraging Theory to improve electronic books (Chi, Gumbrecht, & Hong, 2007; Chi, Hong, Gumbrecht, & Card, 2005; Chi, Hong, Heiser, & Card, 2004; Chi, Hong, Heiser, Card, & Gumbrecht, 2007; Woodruff, et al., 2000). (Chi, Gumbrecht, et al., 2007; Chi, et al., 2005; Chi, Hong, et al., 2007) performed several studies of a method known as ScentHighlights, which supports the skimming of text by highlighting conceptually related terms and sentences in response to a user query. Skimming is a type of reading where the individual quickly scans text in order to extract specific information. The goal of ScentHighlights is to improve the speed of skimming and to decrease over-looked information. ScentHighlights works by accepting a user query and then highlights related concepts and sentences using the spreading activation mechanism of ACT-R. The utility of ScentHighlights was demonstrated by its ability improved fact-finding and comprehension (Chi, Hong, et al., 2007).

(Chi, et al., 2004; Chi, Hong, Heiser, & Card, 2006) developed an eBook utility known as ScentIndex. The goal of ScentIndex is to automatically generate an index for an eBook

based on the user query. The ScentIndex generates the custom index using the spreading activation component of ACT-R. The user studies found that ScentIndex improved fact-finding, comparison, and comprehension tasks (Chi, et al., 2006).

(Olston & Chi, 2003) developed ScentTrails, which sought to help users find information on the WWW. The overall goal was to augment the process of browsing with additional cues to assist the user in making navigational choices. For example, in browsing a user navigates from Web page to Web page using the hyperlinks within the pages. ScentTrails allows the user to enter a query and in response, the spreading activation mechanism calculates the information scent of the linked pages. ScentTrails then uses the information scent values to highlight the links with high information scent to assist the user in finding relevant linked pages. When using ScentTrails the amount of highlighting is proportional to the information scent to which the hyperlink connects. (Olston & Chi, 2003) found that ScentTrails improved browsing performance by allowing users to find information more quickly.

### **3.2 Overview of Computational Cognitive Models and Applications to Information Systems**

This section presents an overview of computational cognitive models that have had an impact on the development of information systems. The remainder of this section is organized as follows. Section 3.2.1 presents an overview of document ranking approaches that have leveraged insights from the spreading activation theory of human memory. Section 3.2.2 presents an overview of document ranking methods based on connectionist theory. Section 3.2.3 introduces Vector Symbolic Architectures (VSA), which are a recent development in cognitive modeling with numerous IR applications. Section 3.2.4

introduces an emerging research area that leverages insights from the mathematical framework of quantum probability theory in modeling cognition. Finally, Section 3.2.5 provides an overview of latent semantic analysis (LSA), which is an unsupervised method that has been proposed as a model of human semantic memory.

### **3.2.1 Overview of spreading activation theory of human memory**

Spreading activation was first proposed as a mechanism involved in memory retrieval (Collins & Loftus, 1975) and has served as a component in numerous computational cognitive theories including the ACT-R theory of long-term memory. Previous applications of the spreading activation theory of human memory are particularly relevant for the work contained in this dissertation. A significant portion of the ACT-R theory of long-term memory involves the idea that spreading activation is a mechanism involved in retrieving memories. Additionally, the Information Foraging Theory heavily uses information scent that is based on a spreading activation mechanism. Several variations exist, but all follow the same basic principles as detailed here and differ in underlying assumptions and computational cost.

The spreading activation algorithm operates on a network data structure as shown in Figure 3.1. The network data structure can be an associative network where the connections represent the co-occurrence of terms in text or a semantic network where the relations are typed. The relationships between the concepts are denoted with weighted links ( $W_{ij}$ ). Figure 3.2 provides an overview of the spreading activation processing technique, which is comprised of four phases. The first phase is the preadjustment phase. In this phase, the algorithm divides the activation among the available connections to a given node. The next step is spreading the activation to concepts connected by associative links. For

example, in Figure 3.1, if  $i$  were the current activated concept then  $j$  would receive activation. After the activation spreads to the associated concepts, the algorithm calculates the activation levels using Equation 3.1. In Equation 3.1,  $I_j$  is the input of node  $j$ ,  $O_i$  is the output of node  $i$  connected to node  $j$ , and  $w_{ij}$  is the weighted link connecting node  $i$  to node  $j$ .

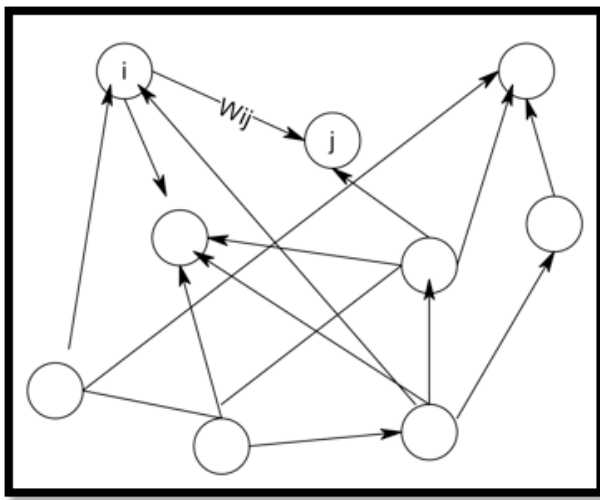


Figure 3.1. Semantic network structure

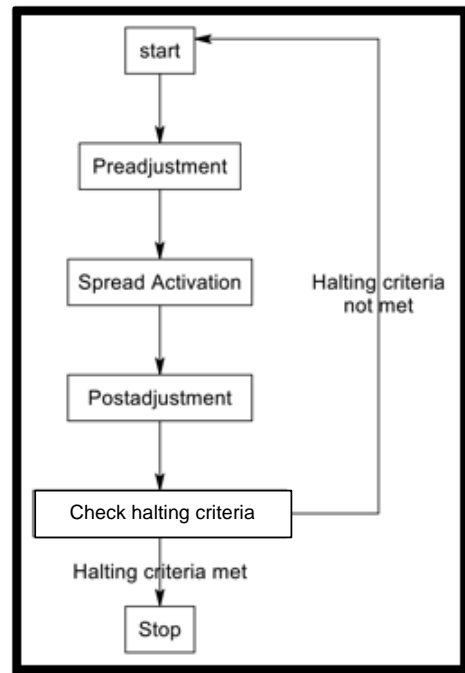


Figure 3.2. Spreading activation overview

Activation input equation

$$I_j = \sum_i O_i w_{ij} \quad (3.1)$$

Output activation equation

$$O_j = f(I_j) \quad (3.2)$$



After the input  $I_j$  is calculated, the output  $O_j$  is determined. Figure 3.3 shows common functions for determining the concept output. After the output is calculated, the algorithm spreads activation to all connected nodes in the network. The algorithm continues spreading activation until satisfying halting criterion such as a convergence criteria or a set number of nodes are processed.

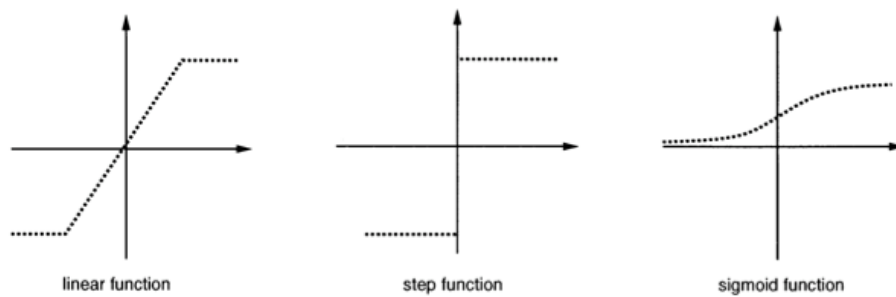


Figure 3.3. Sample activation functions (Crestani, 1997)

IR applications rarely use unconstrained spreading activation. (Berthold et al., 2009) showed that unconstrained spreading activation can converge at query independent solutions. A common way of enhancing the process is to use constrained spreading activation (CSA), which defines heuristics and rules on how to spread activation. This method allows for processing the network according to the semantics of the relations. For example, when the algorithm encounters a part-whole relation it may be appropriate to restrict the spread of activation to the concept designated as *whole* if the *part* receives activation. Below are a list of common constrains used for CSA (Crestani, 1997):

1. Distance constraint: Cease activation when the activation spreads a certain number of links away from the initial source of activation. This heuristic prevents activation of the entire network and prevents query-independent results.
2. Fan-out constraint: Cease activation when algorithm encounters a highly connected concept. This heuristic prevents the spread of activation to overly general concepts. The algorithm can also utilize information theory metrics to halt the spread of activation at concepts that have low information content.
3. Path constraint: Spread activation using preferred paths using inference rules and the semantics of the links. The algorithm can spread activation to meaningful links and restrict activation of links that are less informative.

#### **3.2.1.1 Overview of applications of CSA in IR**

A very thorough review of CSA can be found in (Crestani, 1997). The earliest works of applying CSA for IR were done in parallel by (Preece, 1981) and (Shoval, 1981). (Preece, 1981) showed that approaches such as the vector space model could be implemented using spreading activation and used relevance feedback in the search process. (Shoval, 1981) presented a CSA algorithm that utilized a thesaurus to expand the query terms. The approach utilized feedback allowing the user to indicate irrelevant expanded terms or spread activation to preferred terms. The work by (Preece, 1981) and (Shoval, 1981) can be considered seminal, but neither performed robust evaluations by comparing their algorithms to the state of the art.

The remainder of this section provides an overview of research that has explored using CSA for document ranking. This review is limited to the work that included evaluations of the results. The remainder of this section is organized as follows. Section 3.2.1.1.1 presents

an overview of CSA for document retrieval and ranking. Section 3.2.1.1.2 presents an overview of applications of CSA for ranking Web documents.

#### **3.2.1.1.1 Applications of CSA for bibliographic document ranking and retrieval**

(Salton & Buckley, 1988) performed the first robust evaluation of CSA for document ranking. The CSA model was constrained to traverse at most two links from the original source nodes contained in the query and was much simpler than most subsequent models. It did not include a notion of term importance such as inverse document frequency (IDF) or document length normalization. Both of these factors are crucial for achieving good IR performance. As a result, the vector space model significantly outperformed the CSA model in each experiment.

(Kimoto & Iwadera, 1989) proposed the use of what they described as a dynamic thesaurus. This work was an early attempt to personalize ranking using CSA starting with a static preexisting terminology. The dynamic network learns from the documents that the user marked as relevant. The learning procedure involves extracting term information from the relevant documents, which is used to strengthen existing link weights, strengthen node weights (analogous to a prior probability), and create new links between nodes. The nodes' weights are increased based on the occurrence of the terms in the relevant documents. The learning procedure creates a relation if the relevant documents contain relations between items that are not present in the static taxonomy. Similarly, the learning procedure strengthens the relations between items in the taxonomy if they co-occur within one of the relevant documents. In this work, the CSA algorithm operates on the dynamic thesaurus to

find terms related to those that are in the query. The evaluation found that the CSA method using the dynamic thesaurus improved performance over the static thesaurus.

(Ngo & Cao, 2011) developed a model for query expansion that worked in conjunction with the vector space model. The work used three knowledge sources: KIM (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004), WordNet (G. A. Miller, 1995), and YAGO (Suchanek, Kasneci, & Weikum, 2007, 2008). An entity extraction engine developed for mapping free text to the KIM knowledge source indexed the queries and terms. The mapping relations within the KIM ontology allow for connection to WordNet and YAGO. A CSA algorithm expands the query with knowledge from the three ontologies. They showed that the CSA query expansion algorithm improved the vector space model in terms of mean average precision (MAP) from 0.5099 to 0.5652.

#### **3.2.1.1.2 Applications of CSA for WWW document ranking and retrieval**

(Crestani, 1999; Crestani & Lee, 2000) apply spreading activation for retrieval of information on the WWW. The Web Search by Constrained Spreading Activation (WebSCSA) prototype treated hypertext links as associations among pages. The work also included a CSA mechanism to implement ostensive retrieval (also known as query by example) (Campbell & van Rijsbergen, 1996). This implementation allowed the user to give an example of the information need by selecting a document or documents and then utilize spreading activation to retrieve similar items. The study found a 30% improvement over the baseline and shows promise in applying both ostensive retrieval and CSA on the Web.

#### **3.2.2 Overview of connectionist models as IR systems**

Section 2.1 provided an overview of connectionist models of cognition. This section will focus on the use of connectionist models for document ranking. For an in-depth review see (Cunningham, Holmes, Littin, Beale, & Witten, 1997). This review is limited to methods that include formal evaluations.

The majority of the approaches are based on the network architecture shown in Figure 3.4 (Belew, 1989; Crouch, Crouch, & Nareddy, 1994; Jennings & Higuchi, 1992; Kowk, 1989; Pannu & Sycara, 1996; Wilkinson & Hingston, 1991; Wong, Cai, & Yao, 1993). A family of IR models known as inference networks utilize essentially the same approach, but provide a probabilistic interpretation (Turtle & Croft, 1990, 1991). In general, the so-called connectionist approach to IR bears little resemblance to artificial neural networks or ANNs used in modeling cognition. These approaches are more similar to the network topology and methods used in the CSA retrieval models presented in Section 3.2.1.

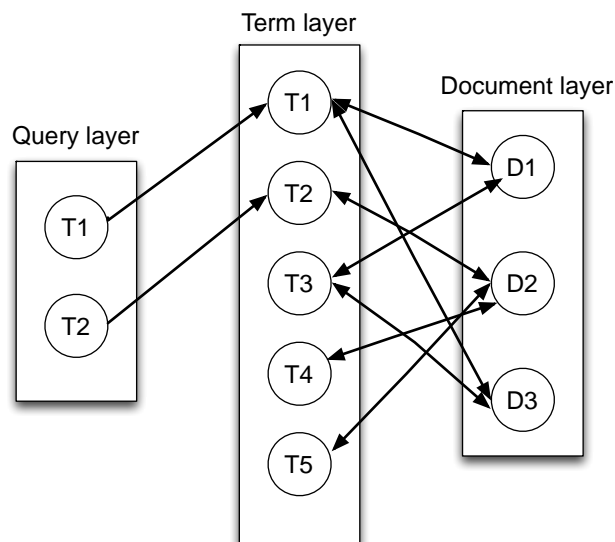


Figure 3.4. Example connectionist network

The network in Figure 3.4 is composed of three layers: one for the query terms, one for the document terms, and a third for the documents. Given a query to the system, the activation spreads from the query terms to the term layer and then from the term layer to the document layer. Equation 3.3 computes the activation level for each document that contains at least one query term. Equation 3.3 is the ranking function used in the classic vector space model. After computing activation values for the documents, activation flows from the top  $N$  documents to the terms in the network. This is essentially a theoretical justification for pseudo-relevance feedback algorithm (Cao, Nie, Gao, & Robertson, 2008). The basic idea is to select the content bearing terms (e.g., terms that constitute a significant percentage of the document's content) from the top  $N$  documents and spread activation (weighted by the document activation value and a measure of term importance) back to the term layer. After this step, activation is spread from the newly adjusted term layer, back to the document layer. The intuition behind this is that highly ranked documents that contain alternative terms such as synonyms that can improve ranking.

$$\sum_{i=1}^t \bar{w}_{i,q} \bar{w}_{i,j} = \frac{\sum_{i=1}^t w_{i,q} w_{i,j}}{\sqrt{\sum_{i=1}^t w_{i,q}^2} * \sqrt{\sum_{i=1}^t w_{i,j}^2}} \quad (3.3)$$

Aside from providing a theoretical justification for pseudo-relevance feedback, a promising feature of connectionist models is the ability to learn from feedback or from user interaction with the IR system. The majority of the studies have looked at modeling explicit feedback where the user is asked to judge the relevance of a document, and relevance

feedback is then used by a computational process to re-rank the documents (Belew, 1989; Bordogna & Pasi, 1996; Crouch, et al., 1994; Kowk, 1991; Kwok, 1989). Belew (1998) and Crouch, et al. (1994) exemplify the basic approach. These algorithms increase the activation level of the second phase of spreading activation and decrease the activation level for documents rated as not relevant.

### **3.2.3 Overview of Vector Symbolic Architectures**

Vector Symbolic Architectures (VSAs) are a recent family of symbolic-subsymbolic models that seek to implement characteristics that are typically associated with symbolic systems within connectionist systems (R. Gayler, 2003). Among other things, VSAs seek to encode semantic information using typed relations within a connectionist framework. In general, VSAs involve the use of high-dimensional vectors and mathematical operators to perform operations such as finding the nearest neighbors of a concept. VSAs began with Smolensky's tensor product variable binding networks (Smolensky, 1990). Smolensky described the motivation for this approach (which describes eloquently the motivation for VSAs in general) as follows.

A one-sentence summary of the implications of this view for AI is this: connectionist models may well offer an opportunity to escape the brittleness of symbolic AI systems, a chance to develop more human-like intelligent systems—but only if we can find ways of naturally instantiating the sources of power of symbolic computation within fully connectionist systems. If we ignore the connectionist approach, we may miss an excellent opportunity for formally capturing the subtlety, robustness, and flexibility of human cognition, and for elucidating the neural underpinnings of intelligence. If we ignore the symbolic

approach, we throw out tremendous insights into the nature of the problems that must be solved in creating intelligent systems, and of techniques for solving these problems; we probably doom the connectionist approach to forever grappling with simple cognitive tasks that fall far short of the true capacity of human intelligence. If we use connectionist systems merely to implement symbolic systems, we might get AI systems that are faster and more tolerant of hardware faults, but they will be just as brittle.

(Smolensky, 1990)

The following example of tensor production variable binding follows from (Blank, 1997; Smolensky, 1990). There are two basic operations common to all VSAs: binding and release. The binding function enables encoding relationships between variables. The release operator is the reverse of binding and decodes a relationship between two variables. In this section, the bind operation is represented as  $\oplus$  and the release operator is represented as  $\ominus$ . Smolensky required that items be broken down into roles and fillers. A role is a named position and the filler is the specific instance that fills the role. For example, the relationship “Barack Obama is president” would be represented by making *president* the role and *Barack Obama* the filler. Smolensky created a mathematical framework for binding a role with its filler. Smolensky represents each role and filler as a vector, which contains “activation” values. For example, the vector of activations for *Barack Obama* could be the vector  $\langle 0.1 \ 0.5 \ 0.9 \ 0.1 \rangle$  and the vector for the role *President* could be  $\langle 0.9 \ 0.1 \ 0.5 \ 0.1 \rangle$ . The binding between the filler and the role is produced by taking the outer product of the two vectors. The resulting matrix, which is shown in Figure 3.5, is



produced by the outer product and represents *Barack Obama* in the role *President* ( $\text{Barack Obama} \oplus \text{President}$ ).

	0.9	0.1	0.5	0.8	0.1
0.1	0.0	0.1	0.1	0.0	
0.5	0.1	0.3	0.5	0.1	
0.1	0.0	0.1	0.1	0.0	
		0.1	0.5	0.9	0.1
		Filler (Barack Obama)			

Figure 3.5. Example of binding role and filler

This basic approach can be used to encode sophisticated structures such as semantic networks with typed relations or sentences. A noted challenge for connectionist systems is the ability to perform analogical reasoning (Gentner & Markman, 1992). Gentner & Markman (1992) stated that the ability of a connectionist system to perform analogical reasoning would constitute a watershed moment. Numerous papers exist demonstrating the ability of VSAs to solve simple analogies such as the following where the goal is to retrieve “Peso” (Eliasmith & Thagard, 2001; R. W. Gayler & Levy, 2009; R. W. Gayler & Sandin, 2013; Halford, Wiles, Humphreys, & Wilson, 1993; Kanerva, 2010; Plate, 1994, 2000; W. H. Wilson, Street, & Halford, 1995).

*United States : Mexico :: Dollar : ?*

One problem with the tensor product variable binding networks is that the resultant vector after the binding is larger than the vectors involved in binding. This is a very undesirable property if one plans to extend such a framework to very large networks (R. Gayler, 2003). Later works such as Pentti Kanerva's binary spatter code (BSP) (Kanerva, 1994) and Tony Plate's Holographic Reduced Representation (HRR) (Plate, 1995) were successful in encoding semantic knowledge within a connectionist framework with fixed vector size. The motivation behind the basic operators for bind and release remain the same, but the mathematics behind the operations differ. Table 3.1 presents an overview of the bind and release operators for the different models.

Table 3.1

*Bind, bundle, and release operators for different methods*

	Bind $\oplus$	Release $\ominus$
Tensor product variable binding	Tensor product	$\cos \theta_{ji} \frac{\ r_j\ }{\ r_i\ }$
BSP	Exclusive OR	Exclusive OR
HRR	Circular convolution	Circular correlation

The remainder of this section focuses on applications of VSAs for document ranking. The review is limited to works that presented formal evaluations. Carrillo has conducted research on using HRR to encode syntax to improve document ranking (Carrillo et al., 2009; Carrillo, Eliasmith, & Lopez-Lopez, 2009; Carrillo et al., 2010; Symonds, 2013). The

motivation behind this work is that the majority of ranking algorithms represent text as bag of words, which ignores the relationships between the terms. Numerous researchers have proposed that ranking can be improved through a more granular representation that includes relationships between the terms in a given text, but this approach increases the modeling complexity and has not led to consistent improvements. The approach taken by Carrillo is to use HRR to bind the terms with their roles within a given text. (Carrillo, et al., 2009) provides the example of how the relationship between the terms *information* and *retrieval* would be bound if they were encountered within a sentence. In Carrillo's work, the term vectors are trained using Random Indexing (Sahlgren, 2005), which produces a reduced dimensional space similar to that of Latent Semantic Analysis (LSA). In this case, the vectors are *information* ( $\vec{r_1}$ ) and *retrieval* ( $\vec{r_2}$ ). There are two roles involved, which are *right noun* ( $\vec{role_1}$ ) and *left noun* ( $\vec{role_2}$ ). Equation 3.4 shows the generation of the vector for *information retrieval* from its constituents. (Carillo, et al., 2009) found that this representation approach resulted in a statistically significant performance improvement of approximately 7%.

$$\vec{R} = (\vec{role_1} \oplus \vec{r_1} + \vec{role_2} \oplus \vec{r_2}) \quad (3.4)$$

(Fishbein & Eliasmith, 2008) explored using an HRR to encode syntax to improve text classification. The actual method for encoding syntax is nearly identical to the work of (Carillo, et al., 2009; Carrillo, et al., 2009; Carrillo, et al., 2010; Symonds, 2013). They

found that by encoding syntax using HRR, they could improve performance over bag-of-words representation.

### **3.2.4 Quantum probability theory and models of cognition**

The use of quantum probability theory to develop cognitive models is an emerging research field that is rapidly gaining attention (Buchanan, 2011). This research is relevant to the research contained in this dissertation for several reasons. First, the mathematical framework of quantum probability theory is an alternative to Bayesian probability theory, which is the mathematical framework of the work contained in this dissertation. Like Bayes' theorem, quantum probability theory allows for probabilistic updating of evidence and the capability to integrate evidence from multiple sources. Second, the use of quantum probability theory is rapidly becoming a valid framework for modeling cognition. An excellent review of the motivation behind modeling cognitive processes using the framework of quantum probability theory is (Busemeyer & Bruza, 2012; Pothos & Busemeyer, 2013). In general, the motivation behind using the mathematical framework of quantum probability theory to model cognition stems from numerous studies that have shown that people do not make decisions according to classical probability theory (Kahneman & Tversky, 1979; Shafir & Tversky, 1992; Tversky, 1977; Tversky & Kahneman, 1974; Tversky & Shafir, 1992).

Quantum probability theory is more general than classical probability theory and provides a mathematical language for modeling ambiguity and uncertainty. Quantum probabilities are based on a geometric model where events are modeled as regions in a vector space (known as a Hilbert space within the quantum probability framework). One difference between quantum probability and classical probability is commutativity in conjunction

(Pothos et al., 2011). In classical probability,  $Probability(A\&B) = Probability(B\&A)$ , but this commutativity property does not necessarily hold in quantum probability as it can be impacted by order effects or context effects. An additional difference is the law of total probability in classical probability theory, which holds that  $Probability(A) = Probability(A\&X) + Probability(A\&\bar{X})$ . In quantum probability theory, the law of total probability will not necessarily hold since interference effects may be present.

Currently, quantum probability theory in modeling cognition has been able to account for several experiments where humans were shown to not behave according to classical probability theory (Aerts, Aerts, & Gabora, 2009; Bruza et al., 2012; Conte et al., 2009; Khrennikov & Haven, 2009; Pothos & Busemeyer, 2009b) (Table 3.2). Each of the models in Table 3.2 are very intricate and in-depth explanation of each is beyond the scope of this review. Instead, one classical probability violation (conjunction fallacy) and the quantum probability model to explain the phenomena will be explored in-depth to provide an example of how quantum probability theory can be used to develop cognitive models.

Table 3.2

*Overview of cognitive models using quantum probabilities*

Name of violation	Description
Failures of commutative in decision making	A well-established property of classical probability theory the commutative property, which states that $Probability(A\&B) = Probability(B\&A)$ . However, numerous psychological experiments have shown that the order in which questions are posed can greatly impact probability judgments (Feldman & Lynch, 1988; Moore, 2002; Schuman & Presser, 1981; Tourangeau, Rips, & Rasinski, 2000). Wang, Solloway, & Busemeyer developed a quantum model to account for the

	commutativity violation (Wang, Solloway, & Busemeyer, 2013).
Violations of the sure-thing principle	The sure-thing principle asserts the following: if you prefer action A over B under state of the world X, and you also prefer A over B under the complementary state $\bar{X}$ , then you should prefer A over B when the state is unknown (Savage, 1954). Tversky & Shafir showed in several experiments that humans violate the sure-thing principle (Shafir, 1994; Shafir & Tversky, 1992; Tversky & Shafir, 1992). Pothos & Busemeyer developed a quantum model to account for violations of the sure-thing principle (Pothos & Busemeyer, 2009a).
Asymmetry in human similarity judgments	Intuitively, one would think that the similarity between object A and object B would be the same as the similarity between object B and object A. Tversky (1977) showed that the symmetry assumption was frequently violated for human similarity judgments (Tversky, 1977). Pothos & Busemeyer (2011) developed a quantum model to account for symmetry violations in human similarity judgments (Pothos & Busemeyer, 2011).

A basic tenet of classical probability is that the probability of a conjunction such as  $P(A \& B)$  cannot exceed the probability of the constituents  $P(A)$  and  $P(B)$ . Tversky and Kahneman (1983) showed that humans violated this basic property of classic probability in reasoning (Kahneman & Tversky, 1979). Tversky & Kahneman presented the subjects with the following description of a woman named Linda.

*Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with the issue of discrimination and social justice, and also participated in antinuclear demonstrations.*

After presenting the subjects with the above statement, they were asked which of the following is more probable.

*Option 1. Linda is a bank teller.*

*Option 2. Linda is a bank teller and is active in the feminist movement.*

Tversky & Kahneman (1974) found that a vast majority of the subjects chose the conjunction (Option 2). From a classical probability theory perspective, a conjunction can never be more likely than one of its constituents (Option 1). Similar results have been obtained with different stories and different situations (Gavanski & Roskos-Ewoldsen, 1991; Sides, Oshershon, Bonini, & Viale, 2002; Stolarz-Fantino, Fantion, Zizzo, & Wen, 2003; Tentori & Crupi, 2012; Wedell & Moro, 2008).

Bussemeyer et al. (2011) developed a cognitive model of this task using quantum probability theory. The model, after reading the description of Linda, was constructed as follows to reflect the initial prior state before reading Option 1 and Option 2. The initial state vector ( $|\psi\rangle$ ) is very near the vector for *feminist* ( $|feminist\rangle$ ). The vector for *bank teller* ( $|bank\ teller\rangle$ ) is oriented such that it lies at a non-orthogonal distance from the vector  $|feminist\rangle$  reflecting that it is possible for a feminist to have such a job, but it is not necessarily highly likely. The next step of the model is to simulate the results after reading Option 1 and Option 2. The event vector is projected onto the feminist vector  $|feminist\rangle$ , and is then projected onto the bank teller vector  $|bank\ teller\rangle$ . The result of these operations is that Option 2 is the most likely instead of Option 1 as predicted by classical probability theory. These results are explained by Bussemeyer et al. (2011) as follows.

Psychologically, the QP model explains the conjunction fallacy in terms of the context dependence of probability assessment. Given the information participants receive about Linda, it is extremely unlikely that she is a bank teller. However, once participants think of Linda in more general terms as a feminist, they are more able to appreciate that feminists can have all sorts of professions, including being bank tellers. The projection acts as a kind of abstraction process, so that the projection

on to the feminist subspace loses some of the details about Linda, which previously made it impossible to think of her as a bank teller. From the more abstract feminist point of view, it becomes a bit more likely that Linda could be a bank teller, so that while the probability of the conjunction remains low, it is still more likely than the probability for just the bank teller property. Of course, from a QP theory perspective, the conjunctive fallacy is no longer a fallacy, it arises naturally from basic QP axioms.

(Busemeyer, Pothos, Franco, & Trueblood, 2011)

#### **3.2.4.1 Implication of quantum probability theory for IR**

The application of quantum probability to IR largely began with (van Rijsbergen, 2004). A review of the motivation for using quantum probabilities for IR is provided by (Piwowarski, Frommholz, Lalmas, & Rijsbergen, 2010; Yaoyong & Cunningham, 2008). Many mathematical frameworks based on quantum probability theory have been proposed which include modeling polyrepresentation in documents (Frommholz et al., 2010; Piwowarski, Frommholz, Lalmas, & van Rijsbergen, 2010; Zellhofer, Frommholz, Schmitt, Lalmas, & van Rijsbergen, 2011), modeling user interaction (Buccio, Melucci, & Song, 2011; Piwowarski & Lalmas, 2009), and modeling context (Melucci, 2007; Melucci & White, 2007a, 2007b). The following summary of quantum probability theory to IR is provided by (Piwowarski & Lalmas, 2009).

Our working hypothesis is that a pure, in the sense that we know exactly what the user is looking for, user interaction can be represented as a system in quantum physics, i.e. as a unit vector in a Hilbert space, and that this state evolves while the user is interacting with the system. According to the quantum probability



formalism, this interaction vector generates a probability distribution over the different subspaces of the Hilbert space. We make the hypothesis that among other possible uses, such subspaces can be related to the relevance of documents, therefore enabling the computation of a relevance score for a document, and to user interactions (like typing a query or clicking on a document), making it possible to exploit them.

The remainder of this section presents an overview of the applications of quantum probability for IR. This review is limited to studies that conducted formal evaluations. Section 3.2.4.1.1 presents the use of quantum probability theory for query expansion. Section 3.2.4.1.2 presents the use of quantum probability theory for representing additional information such as syntax for documents. Finally, 3.2.4.1.3 presents an overview of quantum negation for document ranking.

#### **3.2.4.1.1 Quantum probability theory and query expansion**

Zhang, Song, Zhao, & Hou (2011) present an approach to query expansion based on the analogy of photon polarization (P. Zhang, Song, Zhao, & Hou, 2011). The full description of the photon polarization experiment can be found in (Rieffel & Polak, 2000). In the polarization experiment, the experimenter inserts polarization filters between the light source (source of photons) and a screen. Quantum probability theory and not classical probability theory accurately describes the amount of light on the screen. In the work by Zhao et al. (2011), the documents are modeled as photons and the original and expanded queries are modeled as the polarization filter.

A noted problem in query expansion is query drift (Zighele & Kurland, 2008). The problem occurs when terms are automatically appended to the query, which can cause the

query to shift from the original intended meaning. In this work, the document is modeled as passing through two polarization filters (the original query and the expanded query). The motivation is to fuse the evidence from the original query and the expanded query. For query expansion, the top 50 documents are selected and the top 100 terms with the highest probability<sup>4</sup> are appended to the query. The quantum probability model was compared to several alternative models for integrating the results from the original and expanded query. In the majority of the cases, the quantum probability model resulted in superior performance.

#### 3.2.4.1.2 Quantum probability theory and document representation

(Sordoni, Nie, & Bengio, 2013) explored using the quantum probability theory to encode terms and relationships between terms in the same space which they call a quantum language model. Consider the example from (Sordoni, et al., 2013). In this example,  $n = 3$  and the vocabulary in the corpus is composed of  $\{computer, architecture, games\}$ . Assume a document  $W_d$  where  $W_d = \{\epsilon_{computer}, \epsilon_{architecture}\}$ . If the terms are modeled separately, this results in the disjoint set of projectors shown below.

$$\epsilon_{computer} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \epsilon_{architecture} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Figure 3.6. Representation of computer and architecture

---

<sup>4</sup> The top  $N$  documents are treated as a context. The probability is computed based on the probability of the term appearing in this context versus the probability of the term occurring within the corpus as a whole.

If a term dependency is detected in a document, the relationship is modeled using Equation 3.5. The parameter  $\sigma_i$  must be chosen such that  $\sum_i \sigma_i^2 = 1$ . The parameter can be used to reflect corpus statistics to emphasize important terms. In this example,  $\varepsilon_{computer,architecture}$  is calculated by  $\sqrt{2/3} |e_{computer} + \sqrt{1/3} |e_{architecture}$  resulting in the projector shown in Figure 3.7. The important difference here is that the individual terms as well as the term relationship are represented within the same space. The results were compared to a language model using Dirichlet smoothing. They found that the quantum language model was able to improve performance over the language model using Dirichlet smoothing in some cases.

$$\sum_{i=1}^K \sigma_i |e_{w_i} \tag{3.5}$$

$$\varepsilon_{computer,architecture} = \begin{pmatrix} \frac{2}{3} & \frac{\sqrt{2}}{3} & 0 \\ \frac{\sqrt{2}}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Figure 3.7. Representation of computer architecture

(Zuccon & Azzopardi, 2010; Zuccon, Azzopardi, & van Rijsbergen, 2009) proposed a framework for modeling the dependency among documents for subtopic retrieval. The goal

of subtopic retrieval is to cover all possible subtopics and present the user with unique and relevant information quickly in the search results. For example, assume that document  $A$  is the first document in a search result. Another document  $B$  is relevant, but contains significant overlapping information with document  $A$ . On the other hand, document  $C$  is relevant and contains no overlapping information with document  $A$ . In this case, the goal is to identify that documents  $A$  and  $B$  contain significant duplicate information and rank document  $C$  higher than document  $B$ . The ranking function is shown in Equation 3.6. The parameter  $P(d_i)$  can be estimated using any probabilistic ranking function. The parameter  $\sum_{d_x \in RA} I_{d_x, d_i}$  is a measure of the inteference (i.e. overlapping information) of a document  $d_i$  and any document that is ranked above (RA in Equation 3.6) it. That is, documents are penalized if they contain duplicate information and boosted if the information is previously unseen in the ranked list. (Zuccon & Azzopardi, 2010; Zuccon, et al., 2009) showed that modeling the document dependencies resulted in performance improvement for subtopic document retrieval in the majority of the experiments.

$$d = P(d_i) + \sum_{d_x \in RA} I_{d_x, d_i} \quad (3.6)$$

$$I_{A,B} = P_A + P_B + 2\sqrt{P_A}\sqrt{P_B} \cos \theta_{AB} \quad (3.7)$$

(Wittek, Koopman, Zuccon, & Daranyi, 2013; Zuccon, Piwowarski, & Azzopardi, 2011) propose the use of complex numbers within the quantum probability framework to encode different semantic representations. The previous applications of quantum probability

theory to IR assume real valued vector spaces. They mapped text in the query and documents to SNOMED-CT using MetaMAP. The distributional information for the terms was encoded using Random Indexing, which is an approach similar to LSA. In the complex representation space, the real component encoded distributional semantics using Random Indexing whereas the imaginary component is based on the concept space from SNOMED-CT. They found that combining different semantic representations of text within a complex Hilbert space improved performance over either knowledge source alone. However, it should be noted that combining representations is known to improve retrieval (Croft, 2002). The model was not compared to simpler approaches such as linear integration for combining the evidence from multiple representations.

#### **3.2.4.1.3 Negation and document ranking**

(Widdows & Peters, 2003) presented a novel form of negation within a vector space using insights from quantum probability theory. The negation is implemented by finding the orthogonal subspace using Equation 3.8. In traditional Boolean systems, negation works by removing the negated term. For example, the query *a NOT b* would remove only the term *b*. According to (Widdows & Peters, 2003), quantum negation is best understood as finding “those features of *a* to which *b* is irrelevant”. For example, a query such as *rock NOT band* would ideally return terms related to the *stone* sense of the term *rock* and remove the terms related to the *music* sense. (Widdows, 2003) evaluated the use of quantum negation in document retrieval and found that quantum negation was the best model of those evaluated for removing unwanted documents.

$$a \text{ NOT } b \equiv a - \frac{a \cdot b}{|b|^2} b \quad (3.8)$$

### 3.2.5 Latent Semantic Analysis

LSA originated in the computer science literature (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). LSA relies upon the distributional hypothesis which asserts that the meaning of a word can be defined based on the contexts in which it occurs (Harris, 1954). The first step in LSA is to represent the text as a matrix where each row stands for a unique word and each column represents the count of the terms in a given context such as a passage, sentence, or paragraph.

Consider the sample text in Table 3.3. In this example, each sentence is treated as context. Table 3.4 presents the term-context vector generated from the text. The rows in Table 3.4 represent the unique terms from the example text in Table 3.3. The columns represent each context (sentence in this case) and the term frequency count of the terms in each context. The similarity between two terms can be computed by taking the cosine between two term vectors. Similarly, the similarity between the contexts is computed by the cosine between the two column vectors.

Table 3.3

*Example text data from (Radiohead, 2011)*

C1: Slowly we unfurl
C2: As lotus flowers
C3. Cause all I want is the moon upon a stick
C4. Just to see what if
C5. Just to see what is

Table 3.4

*Word by context vector*

	<b>C1</b>	<b>C2</b>	<b>C3</b>	<b>C4</b>	<b>C5</b>
<b>a</b>	0	0	1	0	0
<b>all</b>	0	0	1	0	0
<b>as</b>	0	1	0	0	0
<b>cause</b>	0	0	1	0	0
<b>flowers</b>	0	1	0	0	0
<b>i</b>	0	0	1	0	0
<b>if</b>	0	0	0	1	0
<b>is</b>	0	0	1	0	1
<b>just</b>	0	0	0	1	1
<b>lotus</b>	0	1	0	0	0
<b>moon</b>	0	0	1	0	0
<b>see</b>	0	0	0	1	1
<b>Slowly</b>	1	0	0	0	0
<b>stick</b>	0	0	1	0	0
<b>the</b>	0	0	1	0	0
<b>to</b>	0	0	0	1	1
<b>unfurl</b>	1	0	0	0	0
<b>upon</b>	0	0	1	0	0
<b>want</b>	0	0	1	0	0
<b>we</b>	1	0	0	0	0

<b>what</b>	0	0	0	1	1
-------------	---	---	---	---	---

In practice, the term-context matrix can be very large. For example, the 2013 MEDLINE corpus contains 2,864,711 unique terms and nearly 20 million documents (NLM, 2013). This would result in a matrix with 2,864,711 rows and 20 million columns. Additionally, the matrix is very sparse. For example, a MEDLINE abstract will contain only a small percentage of the possible 2,864,711 terms. LSA involves the application of a linear algebra technique known as Singular Value Decomposition (SVD) to the sparse term-context matrix (Golub & Reinsch, 1970). SVD is a dimensionality reduction technique that creates a reduced dimensional approximation of the full term-context matrix. This lower dimensional space is often referred to as the “latent” space and improves similarity in part by removing noise that is present in the sparse high dimensional space. Deerwester et al. (1990) describes the strengths of LSA as being able to handle synonymy, polysemy, and term dependence.

LSA has impacted nearly all areas of IR and NLP and has been used in countless applications including query expansion and document ranking. A full review of all of the applications of LSA for IR is beyond the scope of this review (e.g. a Google Scholar query for "latent semantic analysis" *AND* "information retrieval" retrieved over 13,000 citations). Instead, the remainder of this review will focus on the role of LSA in cognitive science.

After the initial introduction to the computer science community, LSA began to be proposed as a general theory learning and meaning (Landauer & Dumais, 1997). One particular problem that LSA has been proposed to solve is Plato’s problem, which refers to



the general problem of “how we can know what in fact we do know” (Chomsky, 1980). This problem is a hypothesis, which asserts that language learners are not exposed to sufficient input to have the knowledge that they possess. The view of LSA as a theory of semantic memory is summarized by (Landauer & Dumais, 1997).

The other, more radical, interpretation of this result takes the mechanism of the model seriously as a possible theory about all human knowledge acquisition, as a homologue of an important underlying mechanism of human cognition in general. In particular, the model employs a means of induction – dimension optimization – that greatly amplifies its learning ability, allowing it to correctly infer indirect similarity relations only implicit in the temporal correlations of experience.

(Landauer & Dumais, 1997)

One of the motivations of LSA as a cognitive theory stems from its high correlation with human similarity judgments (Foltz, Kintsch, & Landauer, 1998; Landauer & Dumais, 1997; Till, Mross, & Kintsch, 1988; P. D. Turney, 2001a). Landauer & Dumais (1997) compared the performance of humans and LSA on the synonym portion of the *Test of English as a Foreign Language* (TOEFL) examination. LSA got 51.5% correct whereas human subjects got 51.6 correct. Other studies have shown that LSA mirrors the learning rates of humans (Landauer & Dumais, 1997; W. Nagy & Anderson, 1984; W. E. Nagy & Herman, 1987), has strong correlation with human graders (Islam & Hoque, 2010; Landauer, Laham, & Foltz, 2000; T. Miller, 2003; Villacorta & Jammalamadaka, 2009), and can be used to model text comprehension (W Kintsch, 1998).

## **Chapter 4: Predicting Document Clicks Using Desirability**

This chapter describes a study of the recency-frequency effect with a particular emphasis on predicting document accesses. Anderson & Schooler (1991) presented the original investigation of the recency-frequency effect and showed a strong parallel between human memory optimization (i.e., predicting the memory item most likely to be needed) and the statistical properties of information in a wide variety of domains. According to (J. R. Anderson & Milson, 1989), these results provide evidence of a universal law which governs the ebb and flow of information. (J. R. Anderson & Milson, 1989) summarize this hypothesis as follows.

Should we really believe that information retrieval by humans has the same form as library borrowings and file accesses? The fact that two very different systems display the same statistics suggests that there are “universals” of information retrieval that transcend device (library, file system, or human memory) and that these systems all obey the same form but differ only in parameterization.

(J. R. Anderson & Milson, 1989)

An unanswered question regarding the recency-frequency effect is what is the underlying mechanism that makes it such a widespread phenomenon? Understanding this mechanism is particularly important because it could provide insight into the workings of human memory and lead to new theories. Additionally, as demonstrated by (J. R. Anderson & Schooler, 1991), the recency-frequency effect is a widespread phenomena. Thus, gaining

insight into the cause of the recency-frequency effect has the potential to touch many disciplines. Section 4.1 presents an investigation of the mechanism that gives rise to the “universals” of information retrieval proposed by Anderson & Milson. In Section 4.1, I show that the preferential attachment mechanism is a sufficient condition for the observation of the recency-frequency effect. Additionally, I analyzed six real-world data sets and show that the recency-frequency effect co-occurs with the presence of preferential attachment. Together, these experiments provide strong evidence that the preferential attachment mechanism causes the recency-frequency effect.

The remainder of this chapter focuses on the use of the recency-frequency effect to predict biomedical document accesses. There are several motivations for this work. One motivation is enhancing Bayesian IR models, which are a particular type of probabilistic IR model based on Bayes’ theorem. Bayesian IR models require calculation of the prior probability of a document being relevant. The most common assumption is asserting that documents have an equal probability of access (uniform prior) (Turtle & Croft, 1990, 1991). In Section 4.2, I show that that the uniform prior assumption is sub-optimal. Additionally, I show that document access from two different IR systems and two different user populations display the recency-frequency effect. The recency-frequency effect provides a theoretically-motivated method for estimating the prior probability of a document being relevant. Additionally, in Section 4.3, I show that the non-uniform prior based on the recency-frequency effect improves prediction of biomedical document accesses.

A second motivation is that the most commonly used prior probability estimates in IR are based on utilizing the structure of a document network such as the hyperlink structure on

the WWW. Notably, these metrics are often not explicit probability estimates, but can be viewed abstractly as estimating the prior probability. An example is the PageRank algorithm (Page, Brin, Motwani, & Winograd, 1998), which has had considerable success in document ranking. Many domains cannot be modeled explicitly as a graph structure, which is required for these methods. The recency-frequency effect can be applied based on document accesses alone and does not require an explicit network structure. Additionally, the work in Section 4.1 provides evidence that the preferential attachment network growth mechanism generates the recency-frequency effect. This implies that the recency-frequency effect can be viewed as reflecting the degree centrality of the implicit and generally unobservable dynamic graph that is generating the document accesses.

In summary, this chapter presents the following contributions. Section 4.1 presents the hypothesis along with experimental evidence that the preferential attachment network growth mechanism generates the recency-frequency effect. This section provides a mechanistic explanation for the recency-frequency effect and a general explanation of why it is present in a wide variety of domains. Section 4.2 presents an analysis of documents accesses for two different populations of users. This study showed that the recency-frequency effect was present for both user populations. Section 4.3 presents an evaluation of using the recency-frequency effect for predicting document accesses in a large real-world data set. These results show that the recency-frequency effect can be used for predicting the future accesses.

#### **4.1 Relationship Between Recency-Frequency Effect and Preferential Attachment**

The recency-frequency effect has been documented for human memory as well as other areas such as email correspondence patterns and word learning in children (J. R. Anderson & Schooler, 1991). The recency-frequency effect is as follows:

1. The relationship between the odds of an item appearing in the future and the frequency of past occurrence is a power law.
2. The relationship between the odds of an item appearing in the future and the recency (i.e., how recently was the item last encountered) is a power law.

In human memory, this effect predicts retention based on the historical encounters with a given item. An analogy is caching in a computer system where items that are predicted to be needed in the future are stored in faster memory.

Anderson & Schooler hypothesized that human memory adapted to the statistical properties of the appearance of information in the environment. Anderson & Schooler looked at the statistical properties of information in the following environments (the data set descriptions follow from (J. R. Anderson & Schooler, 1991)).

1. New York Times headlines. Anderson & Schooler analyzed 730 days of New York Times headlines from January 1, 1986 to December 31, 1987.
2. Child early word learning. Anderson & Schooler looked at a subset of the CHILDES database (MacWhinney & Snow, 1990). The CHILDES database is a large corpus of recorded data from many studies that have looked at the development of language in children. According to Anderson & Schooler, every time someone says a word to a child, this is a demand on the child to retrieve the word's meaning.

3. Email correspondence. Anderson & Schooler looked at the electronic mail messages that the first author (J.A.) received from March 1985 to December 1989.

The study analyzed the communication of J.A.

In each of these scenarios, Anderson & Schooler demonstrated the presence of the recency-frequency effect. Subsequently, Anderson & Schooler showed that the recency-frequency effect held for human memory retrieval. Based on this finding, Anderson & Schooler proposed the hypothesis that human memory adapted to the statistical properties of information in the environment. That is, the human memory system attempts to make available the memory that is most likely to be needed. In doing so, the human memory system has taken advantage of the recency-frequency effect. In addition, Anderson & Schooler hypothesized that the ebb and flow of information obeyed a yet unknown universal law given that they found the recency-frequency effect in a variety of disparate domains (J. R. Anderson & Milson, 1989).

A yet unanswered question is what is the mechanistic cause of the recency-frequency effect? The work in this section attempts to provide insight into underlying mechanisms that gives rise to the observation of the recency-frequency effect in a wide variety of domains. Specifically, I propose the hypothesis that the recency-frequency effect is a byproduct of a preferential attachment growth mechanism. The preferential attachment growth mechanism asserts that the probability of a vertex in a graph receiving a new connection is proportional to its current degree centrality (Barabasi & Albert, 1999). Numerous studies have shown that preferential attachment can account for the emergence of scale-free networks found in areas including protein interaction networks (Eisenberg & Levanon, 2003), metabolic networks (Light, Kraulis, & Elofsson, 2005), numerous social

networks (Capocci et al., 2006; de Blasio, Svensson, & Liljeros, 2006), and the growth of the WWW (Barabasi & Albert, 1999). I propose that the preferential attachment growth mechanism can account for the observed recency-frequency effect as follows.

1. The preferential attachment mechanism implies that the future appearance (e.g., receiving a connection with a new node) of a vertex is a function of its degree centrality. This property will generate the frequency effect. That is, vertices with higher degree centralities have appeared with more frequency than vertices with lower degree centralities in the past. I hypothesize that if one looks at the frequency of appearance of vertices in a network generated with a preferential attachment mechanism, the odds of a vertex appearing in the future will have a power law relationship with the frequency of past appearances.
2. I hypothesize that the preferential attachment mechanism accounts for the recency effect as follows. The preferential attachment mechanism implies that recently accessed vertices tend to have higher degree centralities than those accessed long ago since the probability of a new connection is a function of the degree centrality. For example, vertices that have not been accessed (e.g., received new connections) within a 100-day window will tend to have lower degree centrality measures than those that were accessed within a week. Thus, if one bins the vertices by the most recent access (e.g., most recent new connection), each recency bin will correspond to an average increase in degree centrality for the vertices in the bins. If one calculates the odds of the vertices appearing in the future based on the recency bins, a power law relationship will exist between the recency and the odds of appearing in the future.

The remainder of this section is organized as follows. In Section 4.1.1, I analyzed the appearance of information generated using different types of network growth mechanisms (methods discussed in detail in Chapter 2.2) that are known to generate different types of graph statistics. In this analysis, I found that the preferential attachment growth mechanism is a sufficient condition for observing the recency-frequency effect. Additionally, in Section 4.1.2, I analyze six real-world data sets and confirmed that, at a minimum, the recency-frequency effect and preferential attachment co-occur in empirical data. Finally, in Section 4.1.3, I present the summary and discussion of the work presented in this section.

#### **4.1.1 Relationship between recency-frequency effect and network growth models**

##### **4.1.1.1 Methods**

I generated the networks using the Barabasi & Albert (BA) model, Erdős & Rényi (ER) model, BA+triad model, and ER+triad model. The background for each of these network growth models is presented in Chapter 2.2. Each network started with a seed network composed of 500 nodes with five edges randomly connected to the other nodes in the network. During each step, one hundred new nodes were added and five edges were created between the new nodes and the existing nodes using one of the network growth models. If a triad formation step was included, the new node was randomly connected to five neighbors of the existing node. In each case, the networks were generated with 5,000 steps. Table 4.1 presents the properties of the generated networks. The average shortest path length and the clustering coefficient was computed using the Gephi API (Bastian, Heymann, & Jacomy, 2009). The degree centrality distribution was calculated using the method described in Chapter 2.5 (Clauset & Shalizi, 2009).



Table 4.1

*Statistical properties of generated graphs*

	<b>Number of nodes</b>	<b>Number of edges</b>	<b>Average number of edges per node</b>	<b>Degree distribution</b>	<b>Clustering coefficient</b>	<b>Average shortest path length</b>
BA	500,100	2,500,000	10.00	Power law	1.838E-4	2.00
BA + Triad	500,100	5,500,000	21.99	Power law	0.107	1.84
ER	500,100	2,500,000	10.00	Log normal	3.054E-5	2.42
ER + Triad	500,100	5,500,000	21.99	None	0.107	1.95

These selected models provided a cross sectional view of the known degree distributions (power law versus non power law), average path length, and clustering coefficient (high versus low). I evaluated the recency-frequency effect in each of the networks using the methods developed by Anderson & Schooler (J. R. Anderson & Schooler, 1991; Recker & Pitkow, 1996). The evaluation of each network began at step 1000 and ended at step 5000. I started the evaluation at step 1000 as that it gave the network time to stabilize from the initial randomly created seed network. In these experiments, I used 100 steps of data to predict the next step. The 100-step window was defined as the training window and the subsequent step was defined as the testing window.

For the frequency effect, the nodes in the training window were binned based on frequency of appearance in the training set and the odds were calculated based on the number of vertices in a given bin that appeared in the testing window. For example, four nodes appear three times in the training window. Of the nodes that appear three times in the training

window, two were present in the testing window yielding a probability of 50% and odds of 1.0.

For the recency effect testing, the nodes in the training window were binned based on the last appearance of the vertex and the odds were calculated based on the number of vertices in a given bin that appeared in the testing window. For example, two nodes last appeared on step 60. Of the nodes that appeared on step 60, one was present in the testing window yielding a probability of 50% and odds of 1.0. For both the recency and frequency effect experiments, a sliding window was used. For example, the frequency effect was computed on steps 1000-1100 and tested on step 1101. In the next iteration, the frequency effect was computed on steps on 1001-1101 and tested on step 1102. This process was repeated until reaching step 5000.

#### **4.1.1.2 Results**

Figures 4.1 to 4.4 present the results for the recency-frequency effect for the BA model. The results from the frequency experiment in Figure 4.2 show a very strong correlation ( $R^2 = 0.9771$ ) with a power law function. Similarly, the results from the recency analysis show a very strong correlation ( $R^2 = 0.9952$ ) with a power law function. From this analysis, it can be inferred that the recency-frequency effect holds for data that is generated from the BA model. Figure 4.5 presents an analysis of the average degree centrality for the nodes that are grouped based on recency. This result confirms the hypothesis that nodes that are accessed longer ago will tend to have lower degree centrality than those that are more recently accessed. Since the new vertices connect to the older vertices based on their degree centrality, this observation explains the recency effect.

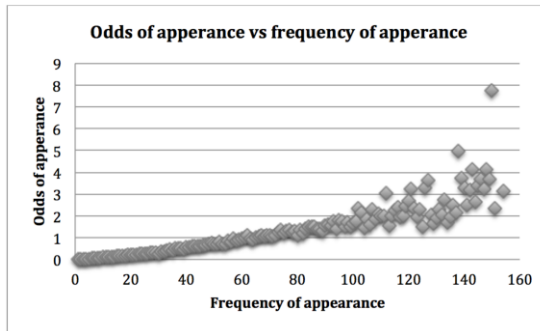


Figure 4.1. BA frequency effect

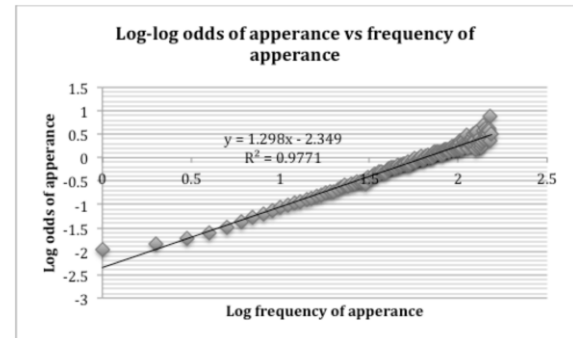


Figure 4.2. Log-log BA frequency effect

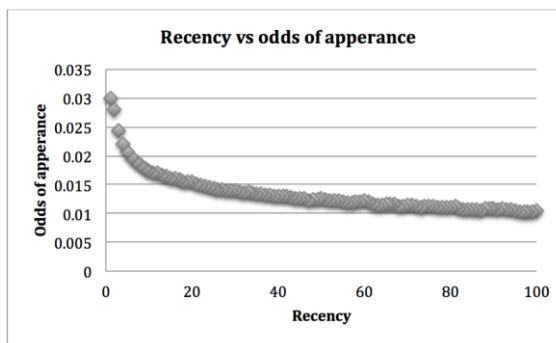


Figure 4.3. BA recency effect

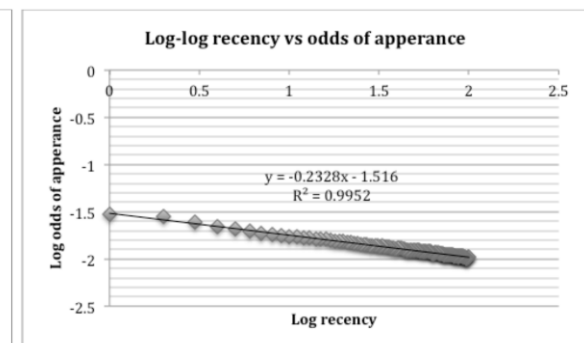


Figure 4.4. Log-log BA recency effect

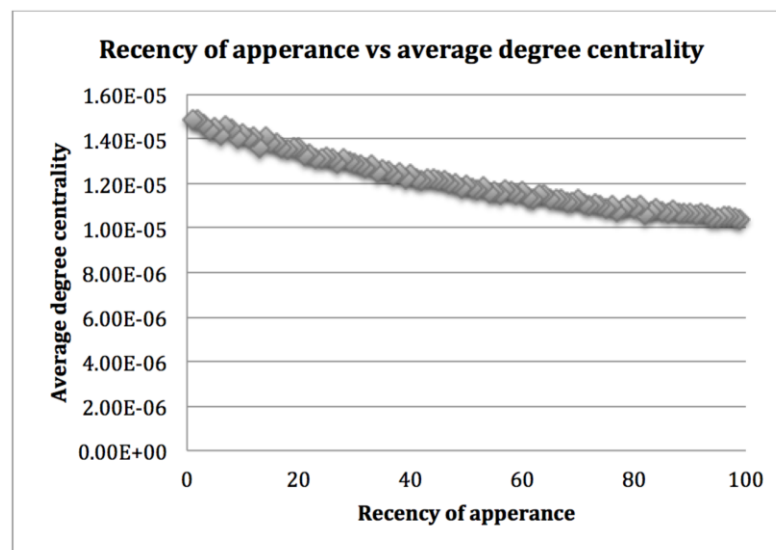


Figure 4.5. Average degree centrality for each recency bin

Figures 4.6 to 4.9 present the results for the recency-frequency effect for the BA+triad model. The results from the frequency experiment show a very strong correlation ( $R^2 = 0.9522$ ) with a power law function. Similarly, the results from the recency experiment show a very strong correlation ( $R^2 = 0.9937$ ) with a power law function. From this analysis, it can be inferred that the recency-frequency effect holds for data generated from the BA+triad model. Figure 4.10 presents an analysis of the average degree centrality for the nodes that are grouped based on recency. This result is consistent with the hypothesis that nodes that are accessed longer ago will tend to have lower degree centrality than those that are more recently accessed. Since the new vertices connect to the older vertices based on their degree centrality, this observation explains the recency effect.

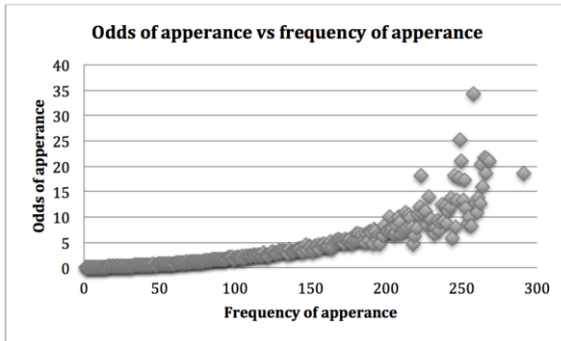


Figure 4.6. BA+triad frequency effect

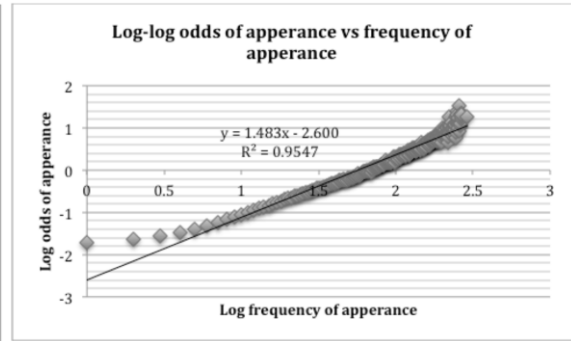


Figure 4.7. Log-log BA+triad frequency effect

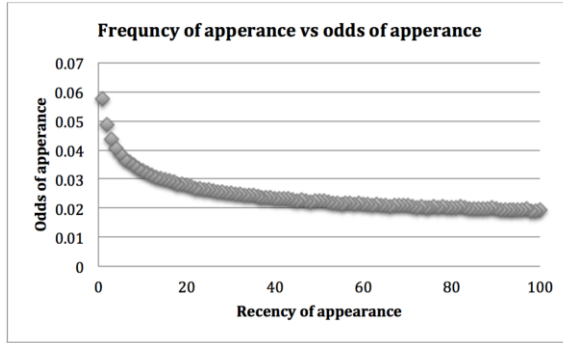


Figure 4.8. BA+triad recency effect

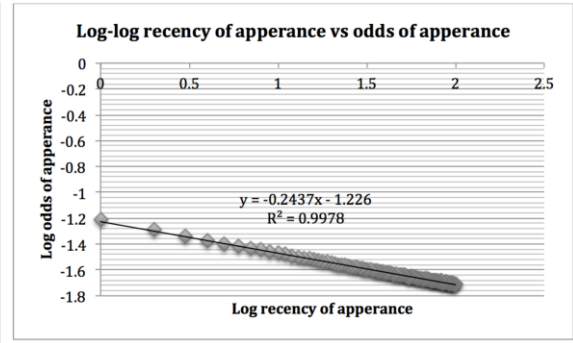


Figure 4.9. Log-log BA+triad recency effect

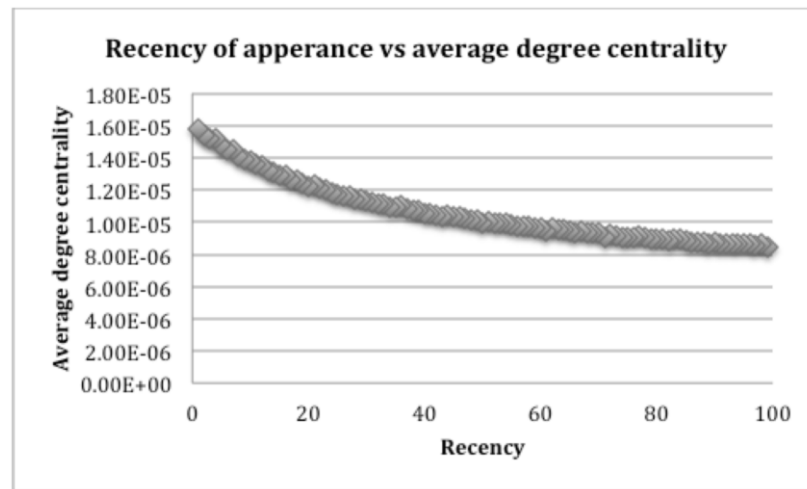


Figure 4.10. Average degree centrality for each recency bin

Figures 4.11 to 4.14 present the results for the recency-frequency effect for the graph generated by the ER growth process. Both the frequency and the recency experiments yield a fairly uniform yet noisy relationship. This is expected from the ER model since the connections are selected at random. Figure 4.15, presents that average degree centrality for the vertices grouped by recency. There is no clear pattern visible from this analysis.

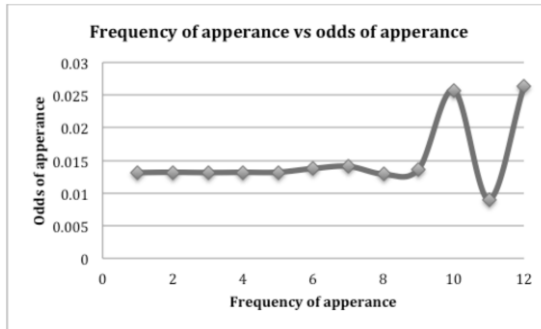


Figure 4.11. ER model frequency effect

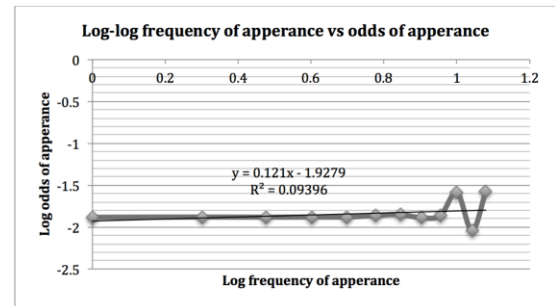


Figure 4.12. Log-log ER model frequency effect

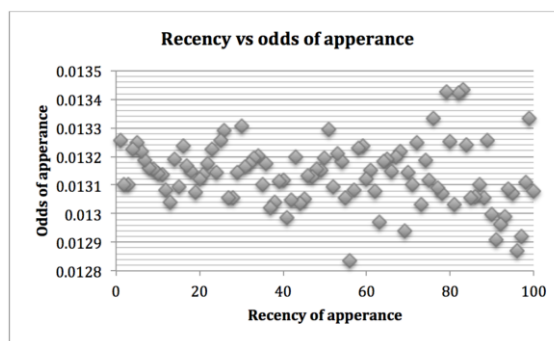


Figure 4.13. ER model recency effect

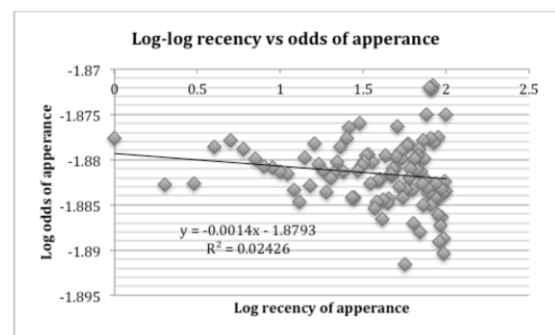


Figure 4.14. Log-log ER model recency effect

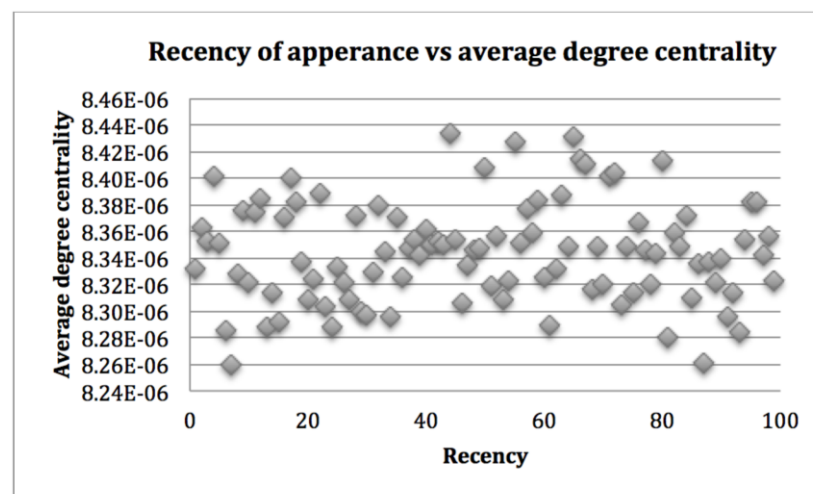


Figure 4.15. Average degree centrality for each recency bin

Figures 4.16 to 4.20 present the results for the recency-frequency analysis for the ER+triad model. In the case of the frequency experiment, a trend is visible whereby a more frequently accessed vertex is likely to receive connections in the future. This is explained by the ER+triad growth model. Vertices that have more connections have a higher probability of receiving new connections at random. That is, the vertices are selected at random, but there is a second step where the neighbors of the selected vertex are chosen. Thus if a vertex has many neighbors it has an increased probability of being selected at random. However, despite the increase in odds based on the frequency of past accesses, the results in Figure 4.17 show a curved plot with a relatively low correlation with a power law.

The recency experiment (shown in Figure 4.18) presents a linear relationship between the recency and odds of future appearance. This linear relationship is supported by Figure 4.20, which presents the average degree centrality where the vertices are binned by recency. The log-log plot in Figure 4.19 presents a curved line that is not a good fit ( $r^2 = 0.8444$ ) for a power law relationship. In conclusion, the ER+triad model can be seen as producing a very weak preferential attachment mechanism. That is, there is some increase in the odds of acquiring a new link based on the past recency and frequency of access, but the effect is not strong enough to produce a power law relationship.

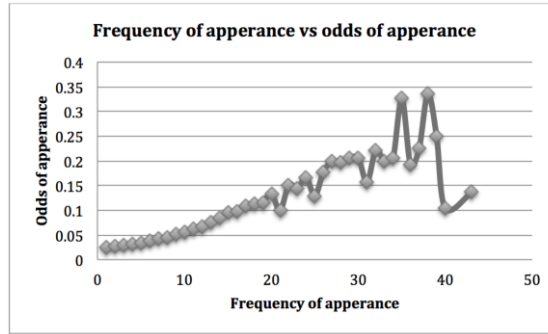


Figure 4.16. ER+triad frequency effect

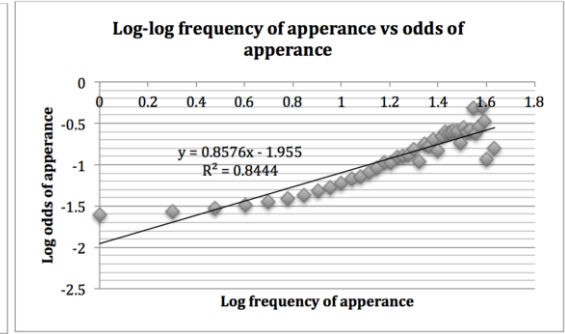


Figure 4.17. ER+triad frequency effect

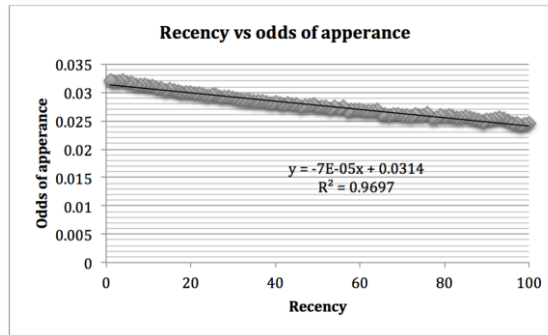


Figure 4.18. ER+triad recency effect

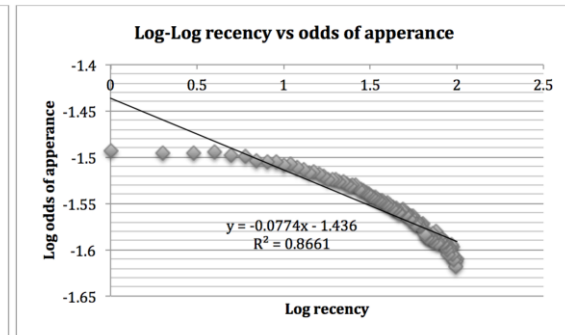


Figure 4.19. Log-log ER+triad recency effect

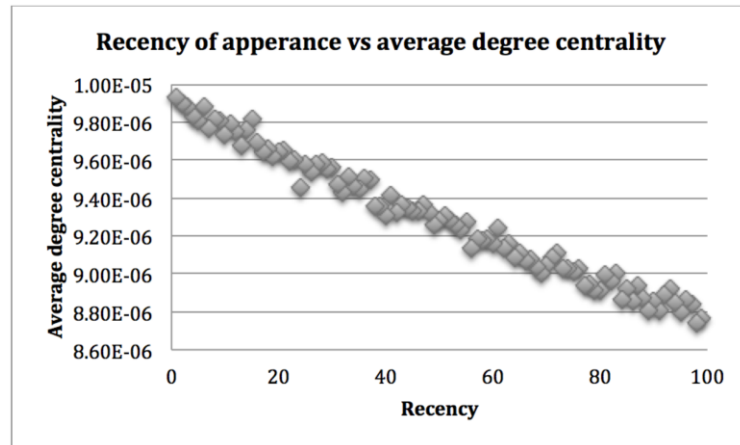


Figure 4.20. Average degree centrality for each recency bin



### **4.1.2 Evidence for relationship between preferential attachment and recency-frequency effect in empirical data**

In Section 4.1.1, I showed using simulation that preferential attachment is a sufficient condition for the recency-frequency effect. In this section, I analyze six real-world network data sets and demonstrate that the recency-frequency effect co-occurs when a preferential attachment growth mechanism is present. Of course, the co-occurrence in and of itself does not imply causality. These results should be interpreted within the context of the experiments in Section 4.1.1 that provided evidence for a causal relationship between preferential attachment and the recency-frequency effect.

#### **4.1.2.1 Methods**

Different training windows and testing windows were used for each data set. The reasoning behind this is that different data sets adapt at different speeds. For example, analyzing quotations from news articles and blogs will likely change on an hourly level, whereas studying the citation network for a corpus of scientific articles will adapt over a period of months or years due to the amount of time it takes to conduct research and create publications that cite the existing literature. For each data set, I look for the presence of preferential attachment and the recency-frequency effect. I conducted the preferential attachment experiments by constructing a network within the training window and then testing for new connections in the test window. A sliding window is used and the aggregate results are reported.

The relationship between degree centrality and the odds of the item appearing in the future does not necessarily have to follow a power law. However, if the network is scale free, a power law relationship between the degree centrality and the odds of receiving a new

connection will be observed. The same procedure described in Section 4.1.1.1 is used for testing the recency-frequency effect. With the exception of the email communication network, I discarded the data points if they were not present in at least 100 experiments within the sliding window. For the email communication network, data points were discarded if they were not present in at least 20 experiments. The motivation behind this is that rare data points do not have enough data and will inject noise.

**1. Quotations extracted from news data** – The data used in these experiments were extracted by MemeTracker (Leskovec, Backstrom, & Kleinberg, 2009). The MemeTracker extracted quotations from news articles and blogs from August 2008 to April 2009. During this period, MemeTracker analyzed over 17 million unique phrases from more than 900,000 news stories. In my experiments, the extracted quotations were analyzed on an hourly basis. For the preferential attachment experiments, I constructed the network by adding edges between the phrases and the articles that used a given phrase. For all of the experiments using this data set, I used a training window of 6 hours and a test window of 1 hour.

**2. Predication graph** - The predication graph data set is generated by the SemRep NLP tool, which extracts subject-predicate-object triples (predications) from the medical literature (Rindflesch & Fiszman, 2003). This data set contains the predications extracted from articles included in MEDLINE from January 1, 2006 to December 2010. I constructed the network by connecting the concepts with an undirected edge if a predication contains two concepts. The predication graph extracted during this period contains 211,566 concepts and 10,518,291 edges. The experiments used a training window of 36 months and a test window of one month.

**3. High energy physics citation network** – This data set covers 34,546 papers published in high-energy physics. The data set was originally released as part of the 2003 KDD Cup (Gehrke, Ginsparg, & Kleinberg, 2003). The data set contains citations only to documents within the corpus. In the preferential attachment experiments, I constructed the network by adding an edge between an article and the articles that it cites. The corpus contains 421,578 edges. In these experiments, I used a training window of 700 days and a test window of 7 days.

**4. Email communication network** - This data set was originally created in (Ebel, et al., 2002). The researchers constructed the data set from the e-mail server at Kiel University over a period of 122 days. The data set includes 5,165 student accounts, which had communication with 54,647 individuals. In the preferential attachment experiments, I constructed the network by adding edges between people if they communicated with each other through email. With self-emails removed, the data set is composed of 392,280 edges. In this work, I used a training window of 90 days and a test window of 1 day.

**5. Twitter hash-tag network and communication network** – This data set was originally collected for the work contained in (Li, Wang, & Chang, 2012; Li, Wang, Deng, Wang, & Chang, 2012). The data set contains 2,237,351 users and 18,407,690 communications among these users. The data set contains communications over a 1000-day period. I derived two different data sets from this data source. The first data set analyzes the communication patterns of the Twitter users. In the second data set, the use of hash tags was studied.

In the Twitter communication data set, the creator of the tweet and the other Twitter users mentioned in the tweet are used. For the preferential attachment study, I constructed the network by adding edges between two users that communicated. I used a training window

of 100 days and a test window of 1 day.

The second data set is constructed based on the hash tags that co-occur in the same tweet.

The data set contains 522,718 unique hash tags, which were used 41,079,412 times. In the preferential attachment experiments, I constructed the network by adding edges between hash tags that co-occurred within the same tweet. I used a training window of 100 days and a test window of 1 day.

#### 4.1.2.2 Results

Figures 4.21-4.26 present the results of the preferential attachment experiments using the data sets and parameters discussed in detail in Section 4.1.2.1. Each of the data sets clearly display an increase in the odds of a vertex receiving a new connection based on the degree centrality. Additionally, each of the data sets showed a high correlation (minimum  $R^2 = 0.9099$ ) with a straight line on the log-log plot indicating a power law relationship between degree centrality and the odds of an existing vertex receiving a new connection.

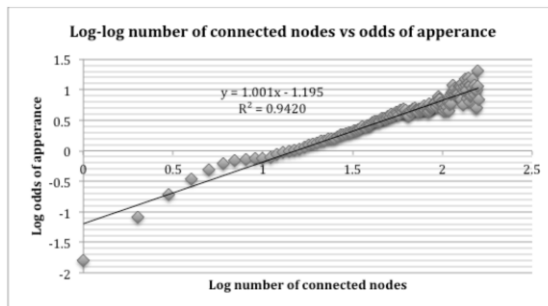


Figure 4.21. Preferential attachment for quotes from news cycle

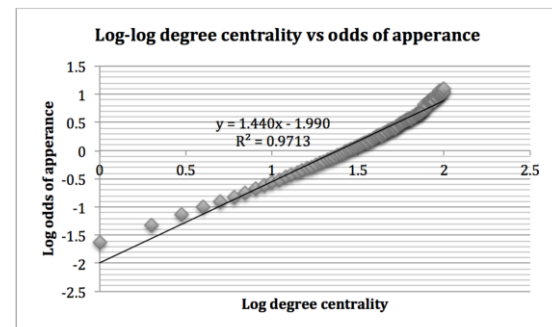


Figure 4.22. Preferential attachment for predication graph

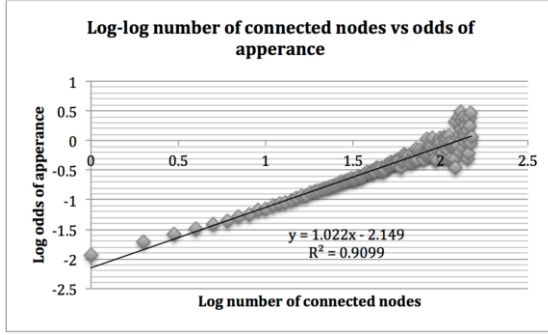


Figure 4.23. Preferential attachment for high energy physics network

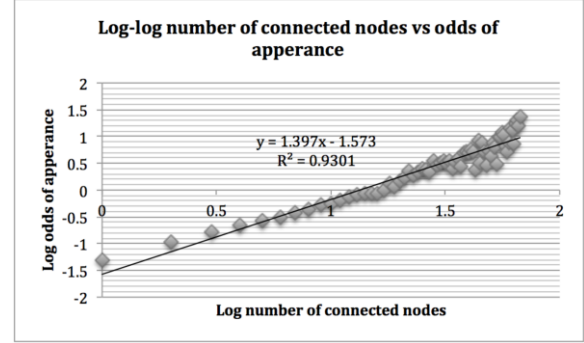


Figure 4.24. Preferential attachment email communication network

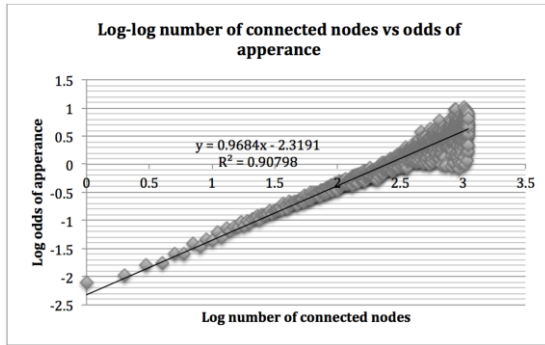


Figure 4.25. Preferential attachment for Twitter hash tag network

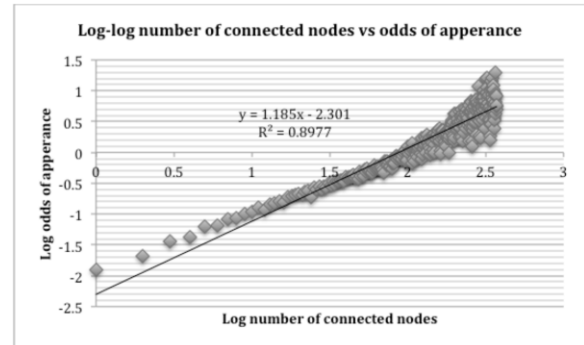


Figure 4.26. Preferential attachment for Twitter communication

Figures 4.27-4.32 present the results of the frequency experiments using the data sets and parameters discussed in detail in Section 4.1.2.1. In each of the cases, a high correlation (minimum  $R^2 = 0.9080$ ) was found with a linear regression fit between the frequency of appearance of an item and the odds of appearance of the item in the future. In all cases, the high  $R^2$  indicates that a power law relationship exists between the frequency of appearance of an item and the odds of appearance of the item in the future.

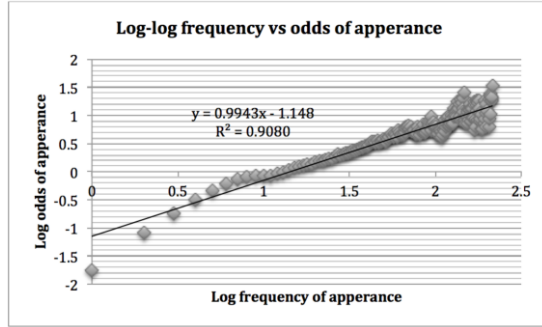


Figure 4.27. Frequency effect for quotes from news cycle

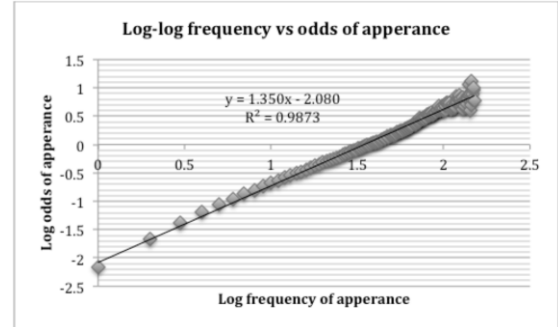


Figure 4.28. Frequency effect for predication graph

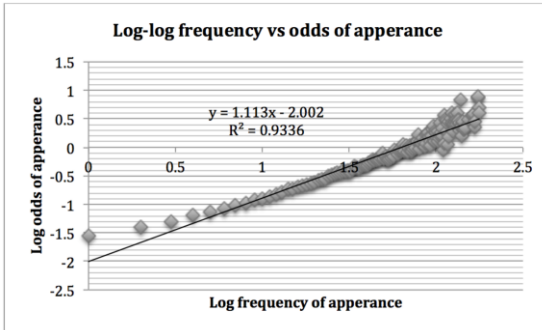


Figure 4.29. Frequency effect for high energy physics network

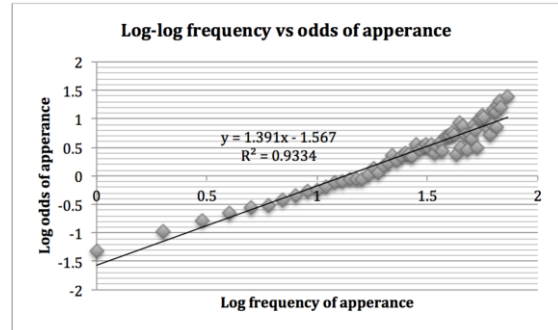


Figure 4.30. Frequency effect for email communication network

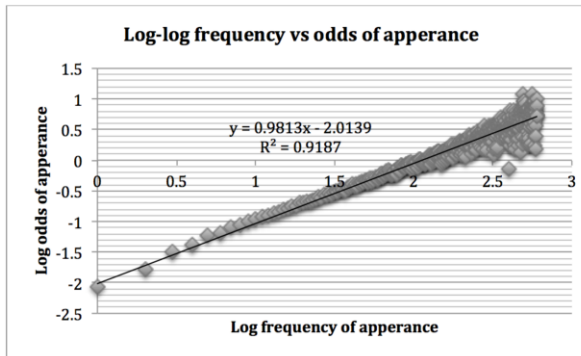


Figure 4.31. Frequency effect for hash tag network

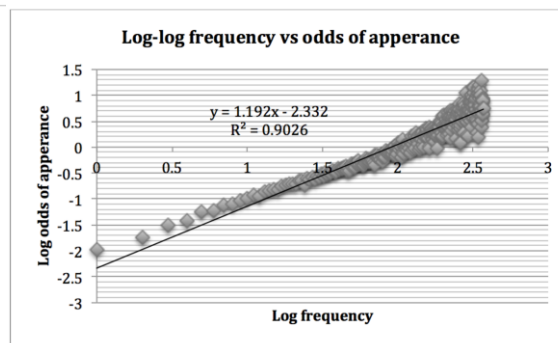


Figure 4.32. Frequency effect for Twitter communication

Figures 4.33-4.38 present the results of the recency experiments using the data sets and parameters discussed in detail in Section 4.1.2.1. In each experiment, the recency tests

revealed a strong power law relationship (minimum  $R^2 = 0.940$ ) between the recency of appearance and the odds of the item appearing in the future.

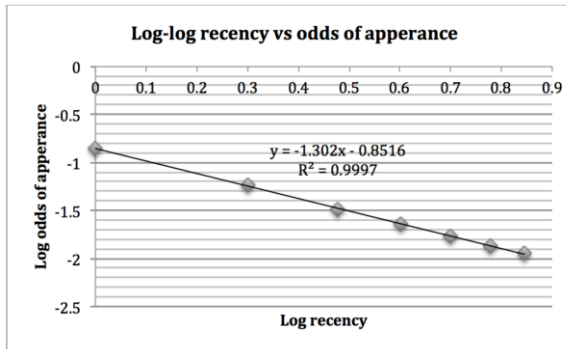


Figure 4.33. Recency effect for quotes from news cycle

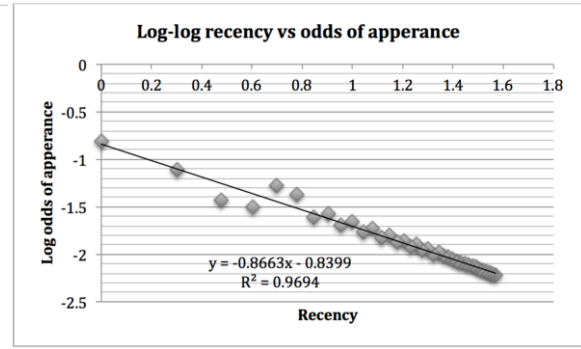


Figure 4.34. Recency effect for predication graph

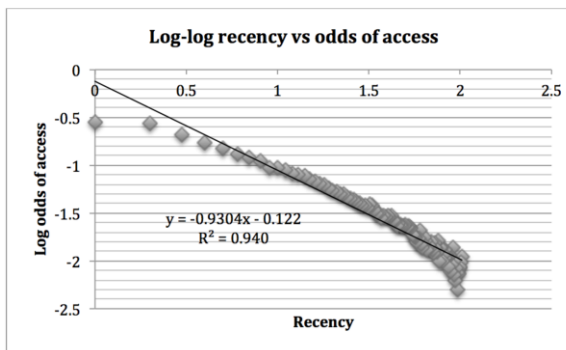


Figure 4.35. Recency effect for high energy physics network

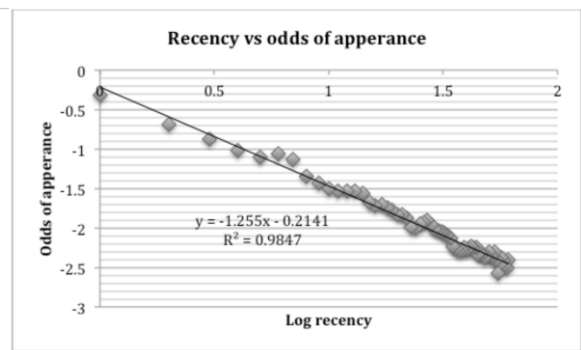


Figure 4.36. Recency effect for email communication network

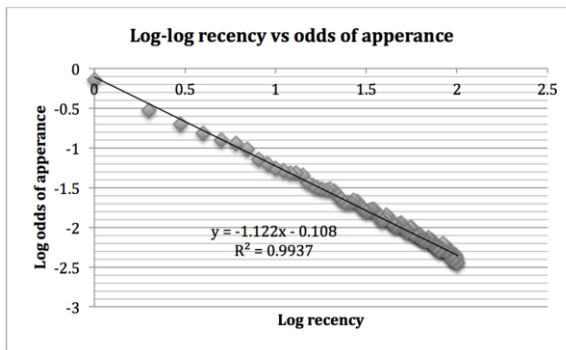


Figure 4.37. Recency effect for hash tag network

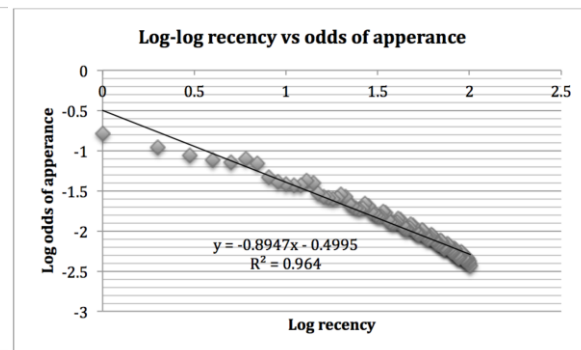


Figure 4.38. Recency effect for Twitter communication

### 4.1.3 Discussion

I performed several experiments to investigate the hypothesis that there is a relationship between preferential attachment and the recency-frequency effect. In the first set of experiments, I generated several graphs using different graph growth mechanisms. The results of these experiments were consistent with the hypothesis that the preferential attachment growth mechanism was a sufficient condition for the presence of the recency-frequency effect. In the next set of studies, I analyzed six real-world data sets and found that the recency-frequency effect and preferential attachment co-occurred in the data sets. The first set of experiments allowed me to isolate the parameters used to generate the networks and provides the strongest evidence of a causal link between preferential attachment and the recency-frequency effect. The second set of experiments provides empirical evidence that the two phenomena co-occur in real-world data sets.

The weakness of this study is that other, yet unknown, mechanisms could generate the recency-frequency effect. The empirical evidence from the real-world data sets shows a correlation between the recency-frequency effect and preferential attachment, which of course does not imply causality. In real world data sets, many complexities can be involved in the growth of the network. However, the preferential attachment mechanism can almost be viewed as a first principle. That is, if a dynamic network is scale free over a given time interval, the preferential attachment mechanism is, without any known exceptions, responsible for the emergent scale free properties of the network. Thus, in the simulated network experiments, I showed that the recency-frequency effect was present when the preferential attachment mechanism is present. This observation implies that the recency-



frequency effect would be present in any data set where preferential attachment has been found.

This study has several important implications. First, the work provides insight into the mechanisms that give rise to the recency-frequency effect. The recency-frequency effect has been validated through numerous experiments, but no mechanistic explanation for *why* information would have such properties have been proposed. The recency-frequency effect was first observed in human memory by Ebbinghaus in 1885 (Ebbinghaus, 1885). The model derived by Anderson & Schooler was the first computational model capable of yielding the experimental results of Ebbinghaus. Anderson & Schooler observed that the recency-frequency effect was present in a wide variety of domains and hypothesized that there was a universal principle that was giving rise to this effect. In this study, I have proposed and provided evidence (both empirical and experimental) that preferential attachment is the universal that is responsible for the observation of recency-frequency effect.

Another implication of this work is that it provides an update to Anderson & Schooler's rational theory of long-term memory. Anderson & Schooler hypothesized that human memory adapted to the statistical properties of information in the environment. Anderson & Schooler provided evidence for the validity of this hypothesis by looking at the statistical properties of information in the environment, which revealed that the presence of the recency-frequency effect in a wide variety of domains. Since the recency-frequency effect, according to Anderson & Schooler, is a universal property of information, they hypothesized that the human memory system adapted based on this environmental constraint. In other words, if the recency-frequency effect is often observed in the

environment and the goal of the human memory system is to make available the memory items most likely to be needed, it would serve the goals of the memory system to use the recency-frequency effect to predict the memory items most likely to be useful. In this work, I showed that a likely mechanistic explanation for the recency-frequency effect is the preferential attachment mechanism. Thus, in updating Anderson & Schooler's hypothesis, the preferential attachment mechanism is the universal and it is this universal to which the human memory system has adapted.

This work raises several questions for future research. First, what characteristics of human memory retrieval can be understood by viewing human memory as a dynamic network? It is widely theorized that human memory is a scale-free and small world network, which would make the presence of a preferential attachment mechanism very likely (Steyvers & Tenenbaum, 2005). In addition, previous work has shown that degree centrality plays a role in human memory retrieval (Griffiths, et al., 2007; P. Pirollo, 2005; Steyvers & Griffiths, 2010). The results in this study are previously undocumented, but could have been uncovered much sooner had the graph structure and evolution of the memory network been taken into account. The open question that could lead to additional insights into human memory is how much can be explained by modeling long-term memory as an evolving scale-free network?

## **Chapter 4.2: Statistical Properties of Document Accesses**

The goal of this section is to determine if the recency-frequency effect is observed in documents accessed using IR systems. The practical motivation of this work, which is explored in Chapter 4.3, is that if documents accessed from IR systems have the recency-frequency effect, this information can be exploited to improve document ranking.

In this section, I determine whether the recency-frequency effect is present for document accesses from two different IR systems. The differences in the data sets provide some evidence that recency-frequency effect is generalizable across different populations of users and different types of IR systems. The first data set is from the Houston Academy of Medicine-Texas Medical Center library (HAM-TMC). In this data set, the document accesses come from the users of PubMed. The HAM-TMC library is located in the Houston Texas Medical center, which is the largest medical center in the world. The HAM-TMC library provides access to published journals for numerous hospitals and universities in the Houston Texas Medical Center. PubMed is a Boolean IR system, which ranks the documents in reverse chronological order, and had no relevance ranking capabilities at the time the study was conducted. The second data set is documents accessed using the PLOS search engine. The PLOS search engine is built on top of the Lucene IR system (Hatcher & Gospondnetic, 2004). Thus, this data set is composed of document accesses made using a system that uses relevance ranking.

The remainder of this section is organized as follows. The first study, presented in Section 4.2.1, investigates whether documents accessed through PubMed have the recency-frequency effect. Additionally, in Section 4.2.1, I determine if a preferential attachment mechanism is present for the PubMed document accesses. In the second study, which is presented in Section 4.2.2, I determine whether documents accessed through the PLOS IR system have the recency-frequency effect. The users in the PLOS data set are not uniquely identified so it is not possible to determine if a preferential attachment mechanism is present for this data set. Finally, Section 4.2.3 presents a discussion and summary of the results of this section.

## **Chapter 4.2.1 Analysis of document accesses on PubMed**

In this section, I investigated whether documents accessed from PubMed have the recency-frequency effect. Additionally, I tested for a preferential attachment mechanism in the documents accessed by users of PubMed. This section is organized as follows. Section 4.2.1.1 presents the methods used in these experiments. Section 4.2.1.2 presents an analysis of the distribution for the documents accessed through PubMed. Section 4.2.1.3 presents the experiment to determine if the recency-frequency effect is present. Finally, Section 4.2.1.4 presents an experiment to determine the presence of a preferential attachment mechanism for the document accesses.

### **4.2.1.1 Methods**

The data set used in this analysis came from the HAM-TMC library, which is located in the largest medical center in the world and provides access to resources for numerous institutions. In this analysis, I used server logs which recorded PubMed use for 1,112 days (September 30, 2009 to October 17, 2012). The server logs recorded the query and the documents accessed in response to the query. The data set was comprised of 4,513,463 accesses over 2,107,806 unique documents.

The methods used to conduct the recency-frequency experiments were similar to the methods discussed in detail in Section 4.1.1.1. A sliding window was used for both the recency and frequency experiments. For the frequency experiments, the number of accesses for each document was counted and the documents were binned according to the number of accesses. The odds were calculated by computing the number of documents in each bin that were present in each test window. Similarly, for the recency testing, the documents were binned based on their most recent day of access. The odds in this case were computed

based on the number of documents in each bin that were present in each test window. For the frequency experiment, I used a window of 365 days and a 1-day window as the test. For the recency experiment, I used a 1-day window as a test. I used a variety of window sizes for the training window, which were 7 days, 30 days, 180 days, and 365 days. For the preferential attachment experiments, a bipartite network was constructed based on a 365-day training window. The edges in the network connected the users to the documents that they accessed. In the one-day test window, the new connections for the documents were extracted and the odds of a document receiving a new connection based on the degree centrality in the training window were calculated.

#### **4.2.1.2 PubMed document access distribution**

Figures 4.39 and 4.40 present the analysis of PubMed document accesses. In both cases, the plots are truncated based on the results of the analysis in Table 4.2. The results in Table 4.2 were generated using the method of Clauset & Shalizi, which is presented in detail in Chapter 2.5. From Table 4.2, the power law was the best fit for the most data points. The power law distribution fit 64 of the data points and passed 2,464 (98.56%) statistical significance experiments. The exponential function fit five of the data points well, but could only account for a small amount of the data. The log normal distribution can be completely ruled out. Based on these results, the power law distribution is the best fit for the distribution.

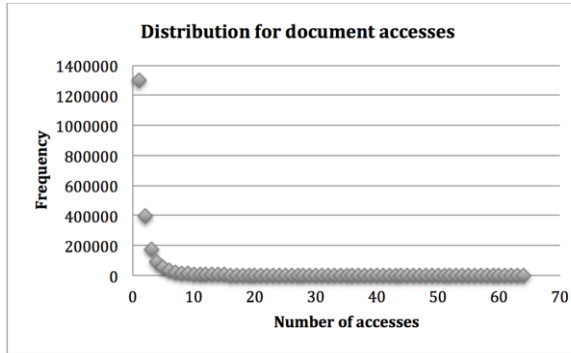


Figure 4.39. Distribution of document accesses from PubMed

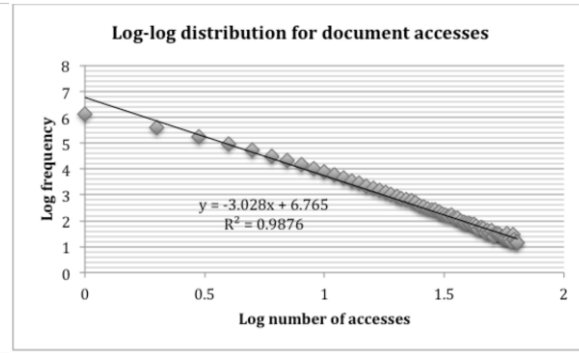


Figure 4.40. Log-log distribution of document accesses from PubMed

Table 4.2

*Results of analysis for PubMed document access distribution*

Power law			Exponential				Log normal				
$\alpha$	xMin Results		$\lambda$	xMin Results		All points result	$\mu$	$\sigma^2$	xMin Results		All points result
	xMin	Result		xMin	Result				xMin	Result	
1.309	14 (64)	2464	2.853e-06	51859 (5)	2484	0	3.698	9.872	1 (135)	330	0

#### 4.2.1.3 PubMed recency-frequency experiment

Figure 4.41 presents the result of the frequency experiment. The results indicate a strong power law relationship ( $R^2 = 0.9705$ ) between the frequency of past document access and the odds of a document being accessed in the future.

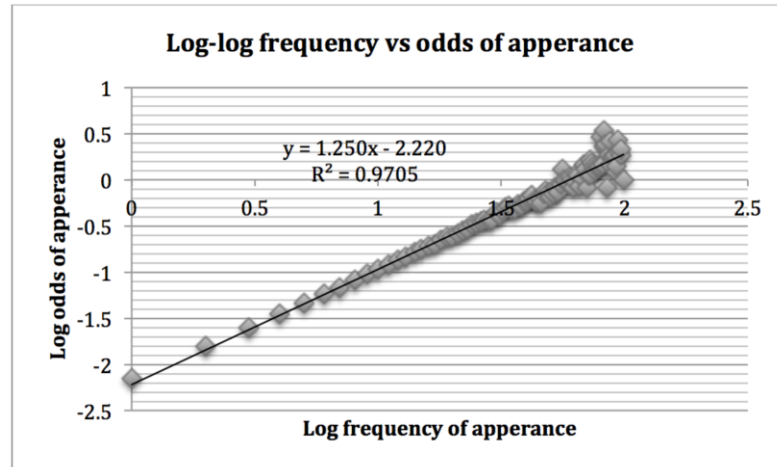


Figure 4.41. Log-log plot of frequency versus odds of access

Figures 4.42-4.43 present the result of the recency effect with varying training window sizes. The motivation behind using different training window sizes is that PubMed reverse-chronological ranking could potentially cause the recency effect. The vast majority of users look at only the first 1-2 result pages (Islamaj Dogan, et al., 2009). Thus if a wide training window were used, older documents would be buried within the search results and new articles will receive more clicks since they are ranked more highly. The varying window sizes allow for some control over the potential impact of the reverse chronological order ranking. For all training window sizes, a very strong power law relationship (minimum  $R^2 = 0.9926$ ) was found between the recency of document access and the odds of the document being accessed in the future.

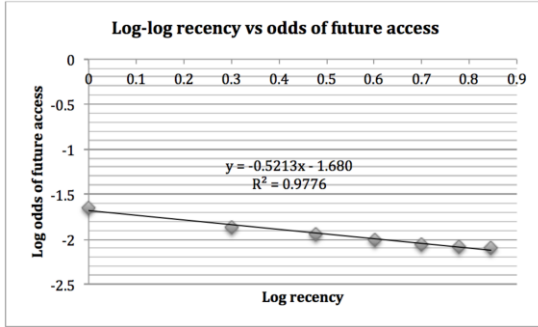


Figure 4.42. Recency with a 7 day training window

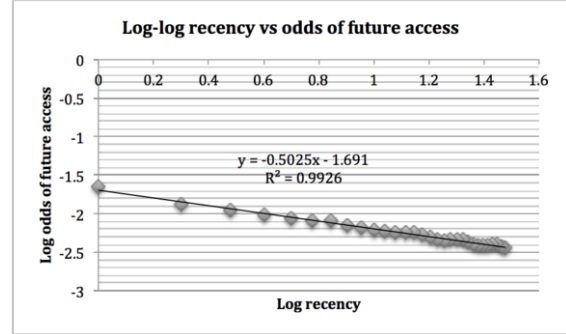


Figure 4.43. Recency with a 30 day training window

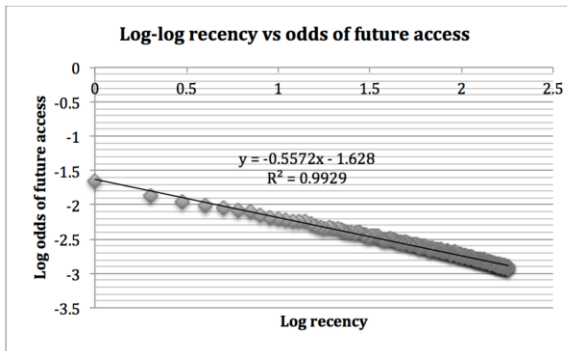


Figure 4.44. Recency with a 180 day training window

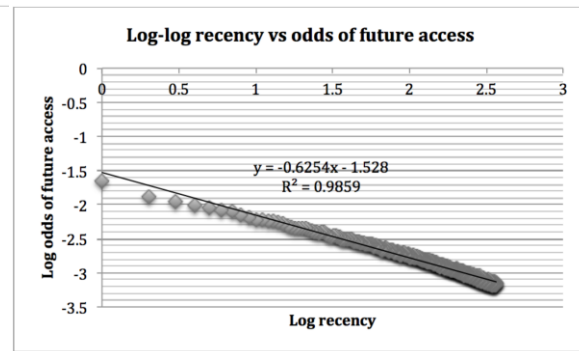


Figure 4.45. Recency with a 365 day training window

#### 4.2.1.4 PubMed preferential attachment experiment

Figure 4.46 presents the result for the preferential attachment experiment for PubMed document accesses. The results show a clear preferential attachment mechanism where the odds of a given vertex receiving a new edge increases with degree centrality. Additionally, a log-log plot reveals that the relationship between the degree centrality of a given vertex and the odds of receiving a new link is a power law.



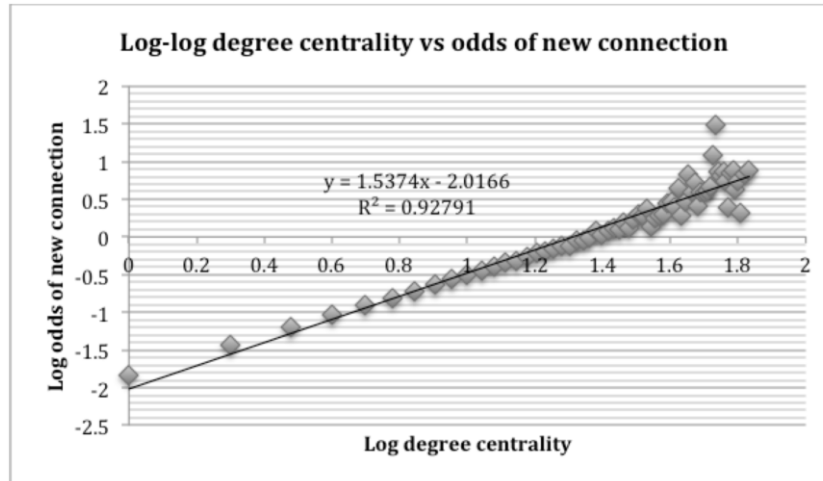


Figure 4.46. Preferential attachment for PubMed document accesses

## 4.2.2 Analysis of document accesses on PLOS

This section investigates whether documents accessed from PLOS have the recency-frequency effect. This section is organized as follows. Section 4.2.2.1 presents the methods used in these experiments. Section 4.2.2.2 presents an analysis of the distribution for the documents accessed through PLOS. Section 4.2.2.3 presents the experiment to determine if the recency-frequency effect is present.

### 4.2.2.1 Methods

The data set used in this analysis came from PLOS document accesses. PLOS is a nonprofit publisher of open-access journals. At the time of this investigation, PLOS published seven journals. PLOS makes available article-level metrics for all of their articles which includes usage information from the PLOS website, citations, social networking applications, and media coverage (Yan & Gerstein, 2011). In this study, I used only document accesses from

the PLOS website. The accesses included document downloads and document clicks. The data set contained 57,666 documents with 18,576,503 accesses. The article-article level metrics data covered June 30, 2012 to November 29, 2012.

The experiments followed the protocol for testing of the recency-frequency effect described in detail Chapter 4.1.1.1. A sliding window is used where a training window and test window are slid across the data set for the entire experimental test set. In the training window, the documents were binned based on the recency or frequency of access. The odds of a document appearing based on recency or frequency were calculated by computing how many of the documents in a given bin from the training window appeared in the test window. The training window in these experiments was 60 days and the test window was 1 day. The results were limited to document bins that were present in at least 50 experiments in the sliding window.

#### **4.2.2.2 PLOS document access distribution**

In this step, I looked at the distribution of document accesses for each individual day in the 152-day data set. Table 4.3 presents the aggregate results for each day in the PLOS data set using the Clauset & Shalizi method, which is presented in detail in Chapter 2.5. The “average number of data points” column presents the number of data points for each day that was a good fit for a given distribution. The column “percentage of experiments passed” is the number of experiments where a given distribution was determined to be a good fit. Based on the aggregate results, the power law distribution is the best fit for the data set as it described the most number of data points (average 43.48) and passed all of the statistical significance tests for each day. The exponential with cutoff is a good fit for a small number

of points (average 7.15) in the majority of the days. Finally, the lognormal distribution can be ruled out completely.

Table 4.3

*Aggregate results for each day in PLOS data set*

	Average number of data points	Percentage of experiments passed
Power law with cutoff	43.48	100.0%
Exponential with cutoff	7.15	98.46%
Exponential	99.15	0.0%
Lognormal with cutoff	13.05	0.0%
Lognormal	99.15	0.0%

#### 4.2.2.3 PLOS recency-frequency experiment

Figure 4.47 presents the results of the frequency experiment for the PLOS data set. The result indicates a strong power law relationship ( $R^2 = 0.9466$ ) between the frequency of past document access and the odds of a document being accessed in the future.

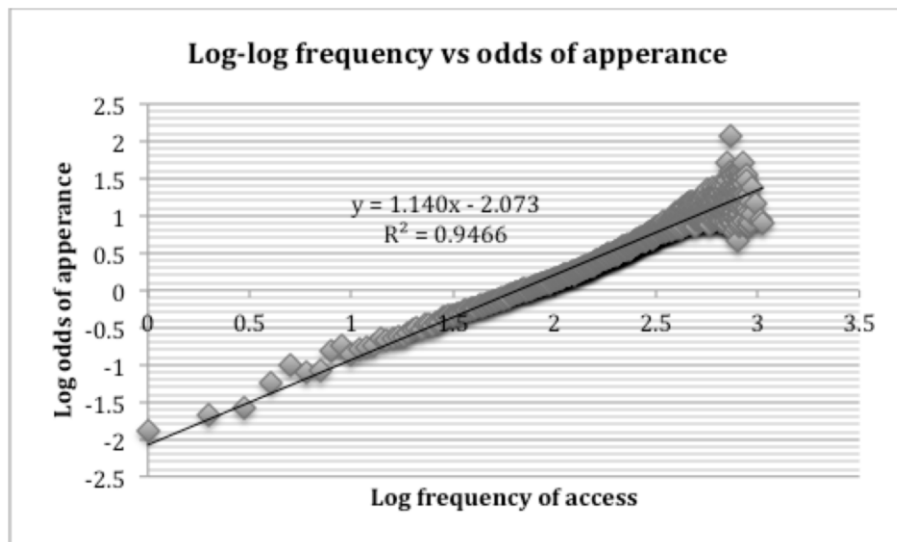


Figure 4.47. PLOS log-log odds as a function of frequency

Figure 4.48 presents the results of the recency experiment for the PLOS data set. The result indicates a strong power law relationship ( $R^2 = 0.9309$ ) between the recency of past document access and the odds of a document being accessed in the future.

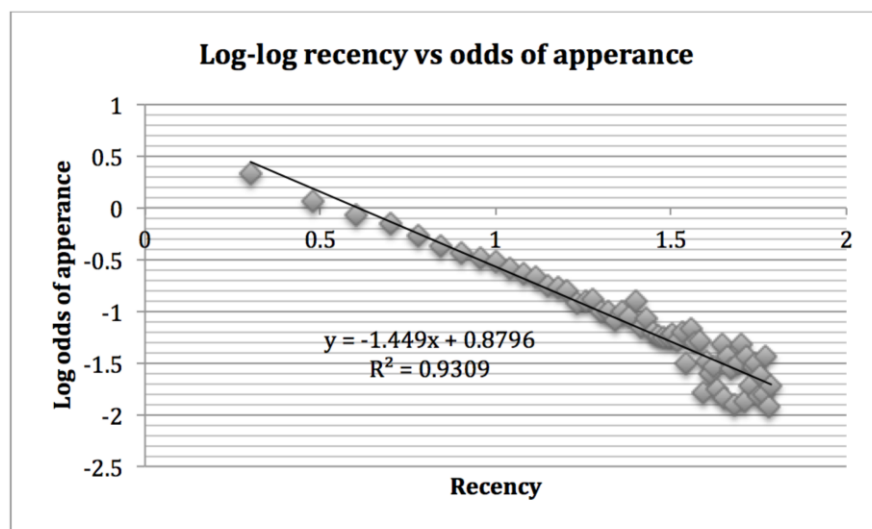


Figure 4.48. PLOS log-log odds as a function of recency

### 4.2.3 Discussion

In this study, I found that the recency-frequency effect holds for documents accessed through two different IR systems and two different populations of users. This provides evidence that the recency-frequency effect is generalizable and will hold across different user populations and different types of ranking functions. Additionally, I showed that a preferential attachment mechanism was present for documents accessed through PubMed.

The finding provides further empirical evidence to support the hypothesis presented in Section 4.1 that there is a causal relationship between preferential attachment and the recency-frequency effect.

The HAM-TMC data set is restricted to primarily the faculty and staff of the medical facilities and the universities in the Texas Medical Center. According to some estimates, approximately 30% of the PubMed users are estimated to be from the general public. In contrast, the HAM-TMC query logs record the document accesses of graduate students, scientists, and clinicians and is a population with a lower percentage of general public users. The constraints of the HAM-TMC data set eliminate much of the general public information seeking behavior. However, the PLOS data set does not have this limitation and is unconstrained in regards to who can access the documents. Finally, the underlying ranking functions used by the two IR systems differ, but the recency-frequency effect held despite the differences in the underlying ranking function used to access the documents.

This study is the first to show that the recency-frequency effect holds for documents accessed through IR systems in the biomedical domain. The most similar study is Recker & Pitkow (Recker & Pitkow, 1996) which showed that the assumptions of the Anderson & Schooler desirability model were valid for Web retrieval. The results of Recker & Pitkow (Recker & Pitkow, 1996) and the results presented here are mutually reinforcing and show that it is possible to model the desirability of documents in a variety of domains.

The work presented here has several practical applications. In particular, this work is important for probabilistic information retrieval approaches, which often make the assumption that all documents are equally likely to be accessed. This work shows that this assumption is false and I provide a method for efficiently and accurately predicting the

prior probability of a document being accessed. Another important contribution of this work is that I present methods for extracting meaningful information from server logs, which many search systems automatically collect. If the recency-frequency effect can improve document ranking, the methods presented here can be applied to automatically extract the information from query logs and improve the ranking of documents. This is the focus of the next section.

### **4.3 Evaluation of Using the Recency-Frequency Effect for Predicting Document Accesses**

In Section 4.1, I proposed the hypothesis and provided evidence to support the hypothesis that the recency-frequency effect arises from the growth of networks via a preferential attachment mechanism. In Section 4.2, I showed that the recency-frequency effect holds for documents accessed in bibliographic databases using two different IR systems and two different populations of users. This section seeks to address the currently unexplored question of whether the recency-frequency effect has utility for document ranking. Additionally, this section investigates whether document usage data from disparate data sources can be aggregated and used for predicting document accesses. To evaluate the performance of desirability computed on different data sets I extract pairwise judgments from query logs, which capture user interactions with the PubMed IR system. Using the pair-wise evaluation method discussed in detail in Chapter 2.4.2, I can determine which data sets and metrics best agree with the preferences of the IR system users from the query logs.

The remainder of this section is organized as follows. Section 4.3.1 presents the methods used in this study. Section 4.3.2 provides a descriptive analysis of the data sets used in this

study. Section 4.3.3 presents an evaluation of using recency-frequency effect from multiple data sources for predicting the documents accessed in response to user queries. Finally, Section 4.3.4 presents a discussion of the work presented in this section.

### **4.3.1 Methods**

#### **4.3.1.1 Description of data sets used for predicting document access**

This study utilized several different data sets, which were used for computing the probability that a document is accessed based on past use. The first data source is HAM-TMC document accesses, which included abstract views and document downloads. The HAM-TMC data set included 4,513,463 accesses over 2,107,806 documents from September 30, 2009 to October 17, 2012.

The second data source was the number of CiteULike users who had a given document saved in their reading list (CiteULike). CiteULike is a social networking application that allows scientists to manage reference libraries, discuss articles, and rate articles. This data source was obtained from the CiteULike website. The third data source was Mendeley, which is similar to CiteULike and allows scientists to manage their reference library, rate articles, and discuss articles (Curran, 2011; Henning & Reichelt, 2008; Zaugg, West, Tateishi, & Randall, 2011). The Mendeley data set contained the number of users that had a given document in their personal library. I obtained this data source using the Mendeley API, which allowed the download of article metrics such as the number of people that have a given document in their library (Mendeley-API, 2013). I obtained the final data source from Scopus. Scopus is a bibliographic database that contains citations for scientific articles from over 19,000 journals (Archambault, Campbell, Gingras, & Lariviere, 2009;

Burnham, 2006). The Scopus data source contained the number of citations for a given document.

I used additional data sources for comparison. One data source was the journal impact factor (JIF) from the 2012 Science Citation Index (most recent publically available Science Citation Index when the experiments were conducted) (Reuters, 2013). The JIF is a bibliometric value that reflects the average number of citations to each article in a journal (Garfield, 2006). The intuition behind using this metric is that people may click on articles from journals with high JIF (high-impact journals) over articles from journals with lower JIF. This provided a base-line measure for comparing the results in this section. An additional motivation is that researchers have explored using JIF in document ranking (Sidiropoulos & Manolopoulos, 2005; Vesely, Rajman, & Meur, 2008).

For further comparison, I extracted a second data source from the HAM-TMC data set by creating a network of documents that were clicked in response to the same query, which is referred to as the click graph. That is, if two documents were clicked in response to the same query an edge is created between the two documents. Figure 4.49 presents a sub-network of the click graph. Similar methods have been used where a document-document network is constructed by creating edges between documents that are highly similar (Kurland & Lee, 2005). Once the network was constructed, network metrics could be computed such as degree centrality or PageRank for the documents. In this particular case, the edge weights between the documents can be seen as human curated similarity judgments. That is, two documents connected with high edge weights indicate that many users clicked the two documents in response to the same query.



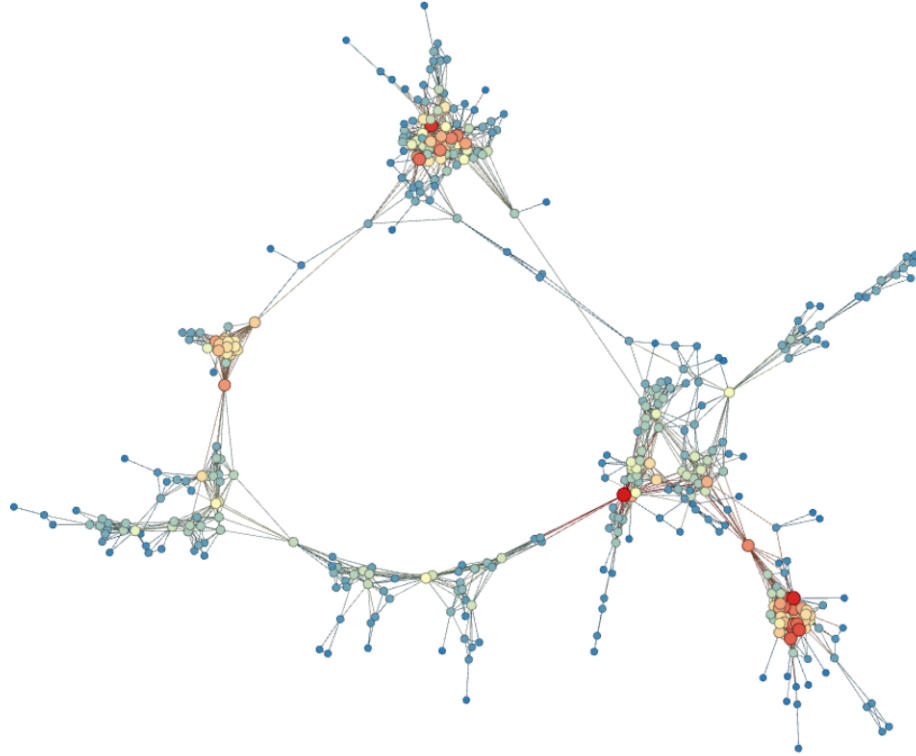


Figure 4.49. Sub-network of click graph extracted from HAM-TMC data set consisting of edge weights  $> 20$

#### **4.3.1.2 Description of experiments and evaluation method**

The purpose of this section is to discuss the preprocessing required for constructing the query log data set for evaluation and to discuss the design of the experiments. Section 4.3.1.2.1 presents the preprocessing required to extract the pairwise judgments from the HAM-TMC query logs. Section 4.3.1.2.2 presents the design of the experiments conducted using the pairwise judgments to evaluate different algorithms. Finally, 4.3.1.2.3 presents a review of the metrics used for evaluation.

##### **4.3.1.2.1 Preprocessing of query logs**

The data set used for evaluation was collected from HAM-TMC users that accessed PubMed from October 18, 2012 to November 4, 2012 (19 days). I distinguished informational from navigational queries. Informational queries are queries where the underlying information need is to gain information about a topic (Broder, 2002). For example, “link between fish oil and blood pressure” is an informational query. Navigational queries in contrast, are queries where the user is looking for a specific item, such as a specific paper or papers published by a particular author. For example, a query for the document title “Predicting biomedical document accesses” is a navigational query where the user is requesting a specific article. I used the following criteria for identifying navigational queries. Queries were considered to be navigational if:

1. The query contained only a PubMed document identifier.
2. The query contained a title of an abstract. Queries containing titles were identified by submitting the query to PubMed using eUtils (Sayers & Wheeler, 2004).
3. The query was composed only of the following: Author, journal name, year, or volume. This information was obtained by submitting the query to PubMed using eUtils (Sayers & Wheeler, 2004).

By using the above constraints, 4,665 queries were classified as navigational and removed from the data set. This left 11,880 queries which were either informational or mixed (informational with some component of navigational such as author). From these queries, the method for extracting pairwise judgments (described in detail in Chapter 2.4.2) was used to extract pairwise judgments for each query. The final data set consisted of 156,623 pairwise judgments for 2,960 users.

#### **4.3.1.2.2 Design of experiments**

Figure 4.50 presents an example of an experimental configuration for computing desirability. Specifically, Figure 4.50 presents the experimental configuration for computing desirability on the combination of the Scopus, CiteULike, HAM-TMC, and Mendeley data sets. The only change for each experiment was removing one of the data sets to attain the desired configuration. For each query, the documents viewed in response to the query were extracted from the query logs. The desirability score was computed for each document that was viewed based on the data sets being used in a given experimental configuration. Based on these desirability scores for the documents, the pairwise judgments extracted from the query logs were used to evaluate how well the documents' desirability scores reflect the preference of the user.

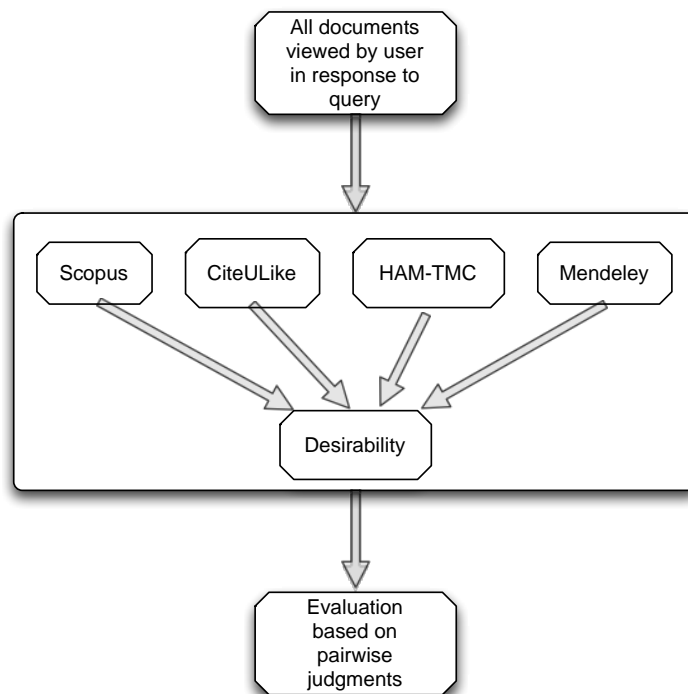


Figure 4.50. Configuration for desirability experiments

I used several existing query dependent ranking functions to benchmark the performance of desirability. Specifically, I compared the performance of desirability with TF-IDF, divergence from randomness Bose-Einstein (DFR\_BE), and divergence from randomness TF-IDF (DFR\_IDF). These functions were discussed in detail previously in Section 2.3 of this dissertation. For the IR models, the training window was used to compute the corpus statistics. The corpus statistics for all of the models was computed using the titles and abstracts from the MEDLINE corpus. I utilized the stop word list generated by Salton and Buckley for the SMART IR system for calculating the corpus statistics (Salton, 1971).

Figure 4.51 presents the experimental design for the existing IR models. Each of the existing IR models required different types of corpus statistics (e.g. number of documents containing a given term) in order to calculate the relevance score. For each query, the documents viewed in response to the query were extracted from the query logs. The score for each of the viewed documents was computed based on the similarity between the document and query using one of the existing IR models. I used the pairwise judgments extracted from the query logs to evaluate how well the documents' relevance scores reflect the preferences of the user.

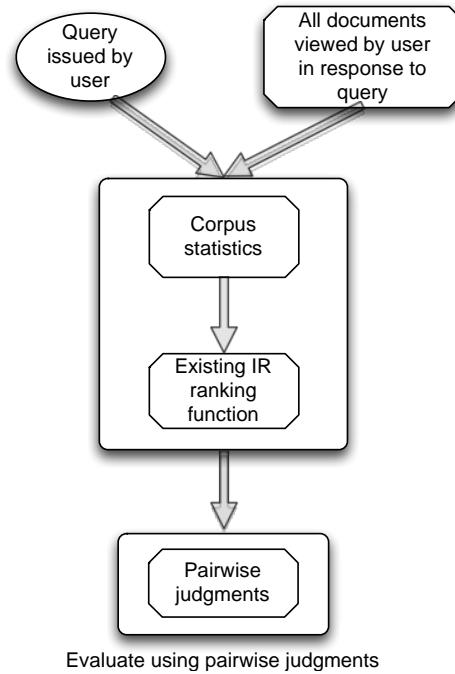


Figure 4.51. Configuration for experiments with existing IR models

A sliding window was used to perform the experiments. Figure 4.52 and Figure 4.53 present a pictorial representation of how the sliding window was utilized to conduct the experiments for desirability and existing IR models respectively. On each test day  $n$ , the pairwise judgments were extracted. For desirability, the training data was comprised of all document accesses from the data sources described in Section 4.3.1.2 for days occurring on or before  $n - 1$ . For document accesses, the training data is composed of days occurring on or before  $n - 1$  to reflect the real-world environment in which such a system would be deployed where past document accesses are used to rank the documents returned by current user queries. For the existing IR models, the corpus statistics were computed for documents published on days  $n$  or earlier. For the existing IR models, the data used during the training window was composed of documents published on days  $n$  or earlier because a document

has to be in the database in order to be returned in response to the query therefore it is reasonable to assume that the corpus statistics should reflect all of the documents currently in the database. For both desirability and the existing IR models, the information in the training window were used to rank the documents, which were subsequently evaluated using the pairwise judgments.

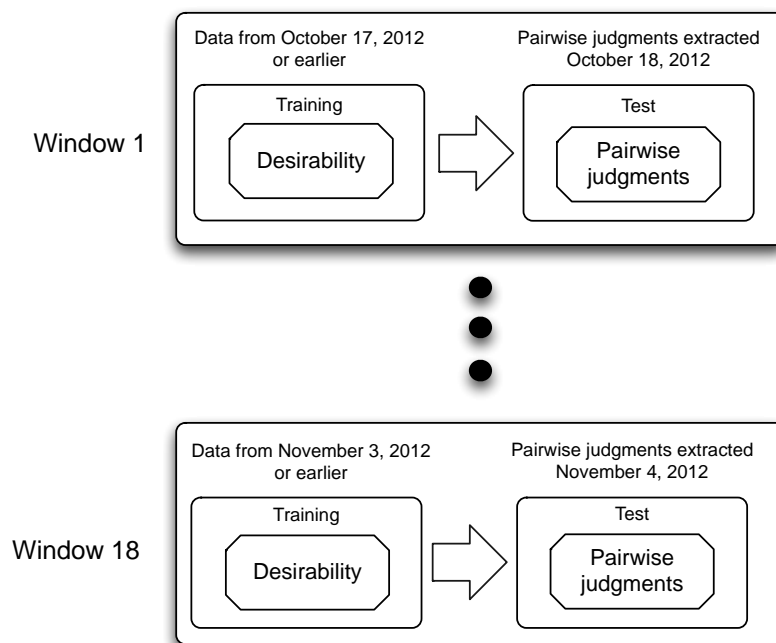


Figure 4.52. Desirability experiments with sliding window

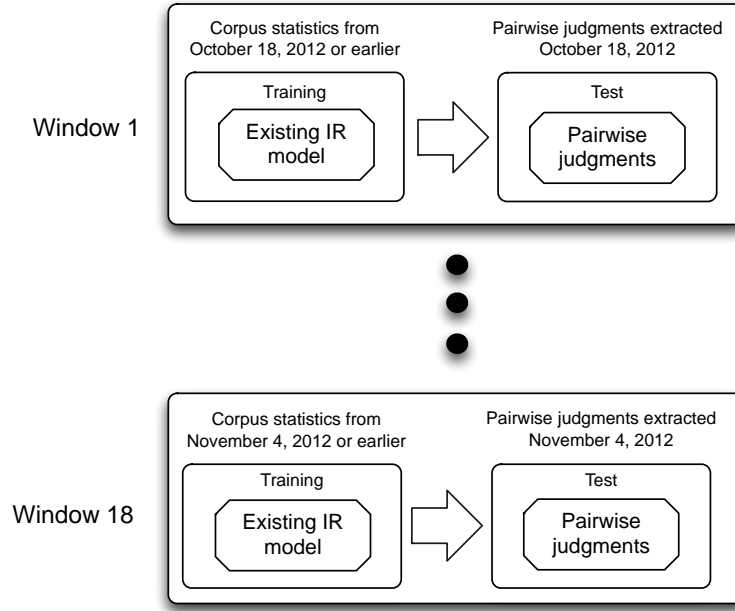


Figure 4.53. Experiments with existing IR models with sliding window

#### 4.3.1.2.3 Evaluation metrics

Evaluation using pairwise judgments extracted from query logs was previously discussed in detail in Chapter 2.4.2. An overview of the metrics used to evaluate the performance of the existing IR models and desirability is repeated here for reader convenience. Equation 4.1 presents the accuracy metric. The accuracy metric is conceptually similar to the notion of precision, which indicates how many of the extracted pairwise judgments were correctly ordered by the different algorithms. For example, consider *document A* which has 10 user clicks and *document B* which has 5 user clicks. Further, assume that the following pairwise judgment was extracted: *document A preferredOver document B*. If the documents were ordered based on the number of document clicks alone then this would result in an ordering that was consistent with the extracted pairwise judgment.

$$\frac{\text{Number of correct pairwise judgments}}{\text{Number of correct} + \text{incorrect}} \quad (4.1)$$

Equation 4.2 presents the coverage metric. The notion of the coverage metric is conceptually similar to recall. In this case, the goal was to determine how many of the extracted pairwise judgments could be ordered by the data available regardless of whether or not those orders were correct. This metric was used to characterize how much information from one given data source can be used for ranking documents on the HAM-TMC data set. For example, consider that there are two documents where *document A* was accessed 10 times in the past and *document B* was not accessed in the past. Since *document A* has some past access information this enables ranking thus the coverage would be increased. Now suppose that *document A* was not accessed in the past and *document B* was not accessed in the past. In this case there is no information in terms of document accesses that can be used to enable ranking, thus the coverage would be decreased.

The engaged reader will recall that the desirability equation includes a decay function, which includes a parameter that encodes the amount of time since publication of the document. Thus, in theory, if two documents contain no document accesses then it would be possible to rank them based on the decay parameter and the age of the document alone. For example, if two documents have no accesses, but one document is newer than the other then the newer document would be ranked higher. However, this would inject an unfair bias into the experiments given the nature of the HAM-TMC corpus. Recall that PubMed



ranks by reverse chronological order. To illustrate the problem consider that a document (*document A*) is clicked and a document higher (*document B*) in the result set is it not clicked. Due to the PubMed ranking *document A* will always be published after *document B*. If the desirability equation is applied to these two documents this will result in predicted pairwise ordering where *document B* > *document A* , which will be incorrect. In summary, an incorrect ordering will be predicted every time based solely upon the characteristics of the corpus. This bias would contaminate the results thus the desirability metric is not applied when no document access information is available.

$$\frac{\text{Number of pairwise judgments where at least one item can be ranked}}{\text{Total number of pairwise judgments}} \quad (4.2)$$

Equation 4.3 presents the harmonic mean, which was used to summarize the results for a given data set in terms of accuracy and coverage. The metric in 4.3 is exactly the metric used to compute the f-measure, which is a widely used metric for evaluating IR algorithms. In this work, the metric was referred to as harmonic mean to distinguish that the metric was computed using accuracy and coverage and not precision and recall as in traditional IR applications.

$$2 * \frac{\text{accuracy} * \text{coverage}}{\text{accuracy} + \text{coverage}} \quad (4.3)$$

A final metric that was used for evaluation is accuracy with ties broken at random. In this scenario, when two documents contain no document access information, they were ordered at random.

#### 4.3.1.3 Equations

The equations used for calculating the odds that a given document is accessed in the future based on the past accesses is known as the desirability function. Desirability is defined as the probability of an item being accessed (Pitkow & Recker, 1994). In this work, desirability was computed based on the recency-frequency effect. The two desirability equations used in this work are shown in Equations 4.4 and 4.5 (Petrov, 2006).

$$B_i = \log \left( \frac{n}{1-d} t_n^{-d} \right) \quad (4.4)$$

$$B_i = \log \left( \sum_{i=1}^k t_i^{-d} + \frac{(n-k)(t_n^{1-d} - t_k^{1-d})}{(1-d)(t_n - t_k)} \right) \quad (4.5)$$

In Equation 4.4, the parameter  $n$  represents the total number of accesses for a document. The parameter  $d$  is a decay parameter, which controls the overall shape of the power law function. For equations 4.4 and 4.5 in this work, the decay parameter was set to 0.1, which was experimentally determined in (Goodwin, Johnson, Cohen, Herskovic, & Bernstam, 2011). Finally, the parameter  $t_n$  represents the amount of time since the publication of a document. The desirability function in Equation 4.4 is an approximation that assumes that accesses are distributed evenly over time. This equation was used for the CiteULike,

Mendeley, and Scopus data sets since only the publication date and the number of accesses was known, but time of each individual access was not recorded.

The desirability function in Equation 4.5 does not assume a uniform rate for the accesses. In this work, a fixed sized window was used which stores the last  $k$  accesses for the documents. In this work,  $k = 1$  so only the last access day was stored. The influence of the recency effect is computed by the parameter  $\sum_{i=1}^k t_i^{-d}$ . The rest of the equation makes the same assumption as that of 4.4, which is that the document accesses are evenly distributed over the period of time  $t_n$ . The motivation for this function is that storing a time stamp for each document poses a computational challenge. For example, the PubMed search system processes millions of queries per day, which will produce many document clicks. In this case, each document click would require the storage of the time stamp for each access. This equation requires the storage of a fixed window of accesses, which eliminates (or at minimum allows for a priori knowledge of the storage and computational requirements for computing desirability) much of the computational burden and can be used in large systems.

The click graph was used to compute degree centrality and PageRank for each document in the network. Chapter 2.2 provides an in-depth discussion of relevant graph theory concepts. The degree centrality metrics used in this study is shown in Equation 4.6. The parameter  $deg(v_i)$  is the number of unique connections for the document  $v_i$ . The parameter  $n$  is the number of unique documents in the click graph. The parameter  $k$  represents the time since the publication of a given document and is used for normalization. In some domains, the  $k$  parameter may not be necessary. In this study, the pairwise judgments were extracted from a retrieval system that ranks in reverse chronological order.

Recall that the rule to extract the pairwise judgments in this work was *click > skip above*. In this particular data set, the clicked document (with very few exceptions) will be older than the documents skipped above it. Thus, metrics that are correlated with the age of the document are likely to result in improved performance in this particular data set. For example, older documents will have many more chances to receive clicks than newer documents simply due to them being around longer. Subsequently, these older documents will have many more chances to be clicked along with other documents in response to the same query, which results in a bias where older documents will tend to have higher degree centralities. The desirability functions in Equations 4.4 and 4.5 normalize the score by taking into account the number of accesses and the age of the documents. Thus, for fair comparison and to remove bias from the age of the document, Equation 4.6 was used to calculate the degree centrality of the document while taking into account the age ( $k$ ) of the document.

$$C(v_i) = \frac{\left(\frac{deg(v_i)}{(n-1)}\right)}{k} \quad (4.6)$$

In addition to degree centrality, PageRank was computed for each document in the click graph (Page, et al., 1998). Equation 4.7 presents the PageRank function used in this work. The parameter  $v \in V(v_i)$  is the set of documents that are connected to the document  $v_i$ . The parameter  $PR(v)$  is the PageRank for a connected document  $v$ . The parameter  $L(v)$  is the number of documents connected to  $v$ . The parameter  $k$  represents the time since the

publication of a given document and was used for normalization as discussed previously. The PageRank algorithm is an iterative algorithm given that the PageRank score for a given document is based on the PageRank score of the connected documents. I used the Gephi API to calculate the PageRank values for the documents in the click graph (Bastian, et al., 2009). The instantiation of Gephi's PageRank algorithm has two tunable parameters: restart probability and epsilon. The restart probability is the probability of jumping to a random vertex in the graph (i.e. the random surfer model) and was set to 0.85. Epsilon is the convergence criteria and halts the PageRank computation for the click graph. This parameter was set to 0.001.

$$PR(v_i) = \frac{\sum_{v \in V(v_i)} \frac{PR(v)}{L(v)}}{k} \quad (4.7)$$

### 4.3.2 Descriptive analysis of document viewed during testing window

This section analyzes the document accesses during the testing period from October 18, 2012 to November 4, 2012. Additionally, this section analyzes the differences between the data sets used for predicting document accesses discussed in Section 4.3.1.1. Shown in Figure 4.54 is a histogram for the documents that were viewed (i.e. clicked or not clicked) by HAM-TMC users from October 18, 2012 to November 4, 2012 binned by publication year. In total 116,450 documents were viewed within this time frame. A small number of publications have a publication date of 2013 and were available ahead of the official print date. From the histogram, we see that the vast majority of viewed documents tended to be

newer documents. This is an unsurprising result since PubMed ranks by reverse chronological order and it is known that the vast majority of users only look at only 1-2 pages of the search results (Islamaj Dogan, et al., 2009).

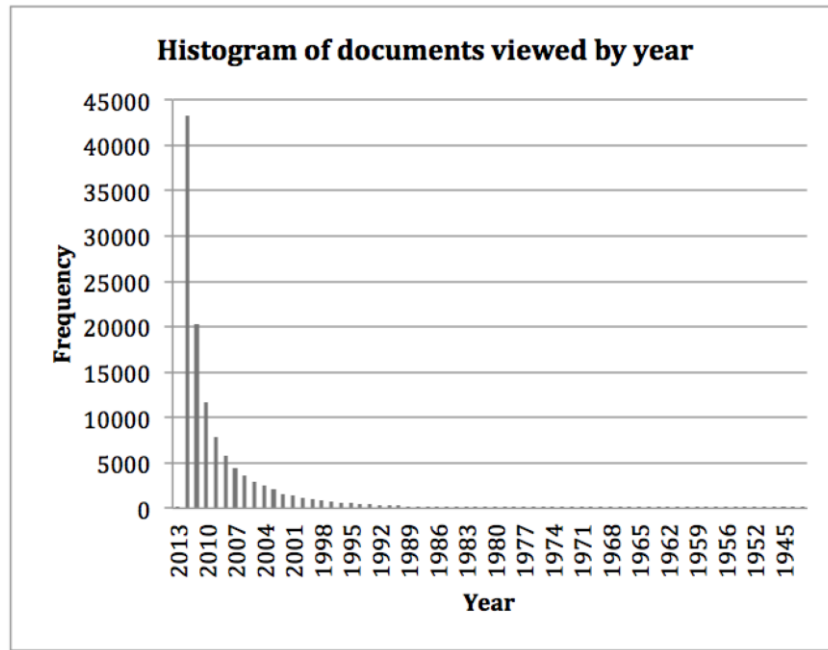


Figure 4.54. Distribution for documents viewed by year

Table 4.4 presents analysis of the information contained in each data set that is used for predicting accesses. The data set that contained the most information about the viewed documents was Mendeley, which contained information on 60.07% of the documents. The CiteULike data set contained very little information about the documents viewed. Overall, the four data sets contained information about 78.83% of the documents that were viewed in the test interval.

Table 4.4

*Information content of each data source*

<b>Data set</b>	<b>Percentage of documents viewed that were previously accessed</b>
Scopus	51.79%
Mendeley	60.07%
HAM-TMC	48.41%
CiteULike	7.57%
All data sets	78.83%

Looking at the aggregate information in Table 4.4 can be deceptive. For example, Figure 4.54 shows that a disproportionate number of views were for newer documents. If a data set contains a lot of information about recently published documents, then this data set would be particularly useful, as many pairwise judgments will involve newer documents. Figure 4.55 presents the number of documents viewed based on the year and the percentage of documents where a given data set had information about the viewed documents. For example, if 2,000 documents were viewed by PubMed users from October 18, 2012 to November 4, 2012 that had a publication date of 1995 and the Scopus data set had information about 1,000 documents, this would yield a percentage of 50%. The data points earlier than 1970 were omitted from the chart since these were relatively rare. Table 4.5 provides the correlation between the information contained in each of the data sets. There is considerable difference in the information contained in the data sets. For example, the

Mendeley and Scopus coverage is highly correlated (0.8577). The HAM-TMC document coverage differs greatly from both Mendeley and Scopus. This difference is highlighted by Figure 4.56, which shows the viewed documents and the information content of the data sources from 2000-2013. Notably, the documents accessed from 2000-2013 contain 93.31% of the views of the HAM-TMC users. In particular, the documents accessed from 2011-2013 contain a disproportionate number of views. It is this period of accesses where the HAM-TMC data set had better coverage than both Scopus and Mendeley.

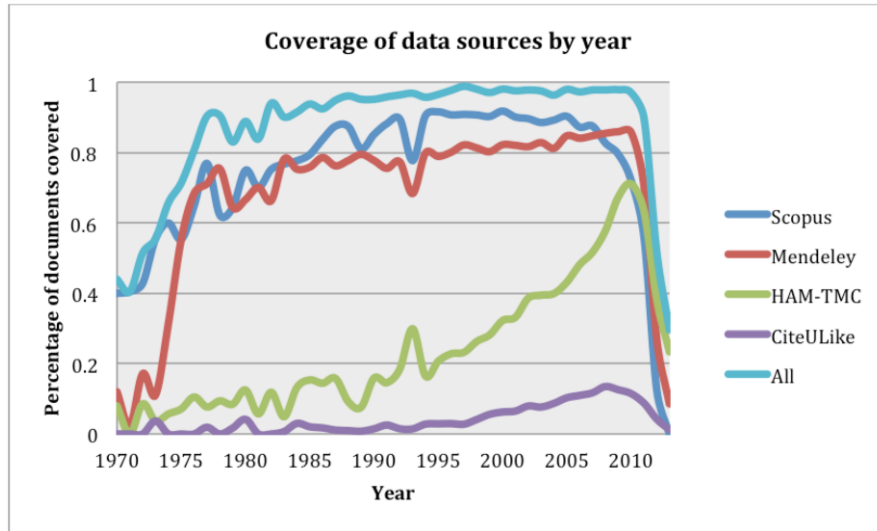


Figure 4.55. Information content of data sources from 1970-2013

Table 4.5

*Correlation between coverage of data sets*

	<b>Scopus</b>	<b>Mendeley</b>	<b>HAM-TMC</b>	<b>CiteULike</b>
<b>Scopus</b>	NA	0.8577	0.1849	0.3148
<b>Mendeley</b>	0.8577	NA	0.4153	0.4393
<b>HAM-TMC</b>	0.1849	0.4153	NA	0.9188



<b>CiteULike</b>	0.3148	0.4393	0.9188	NA
------------------	--------	--------	--------	----

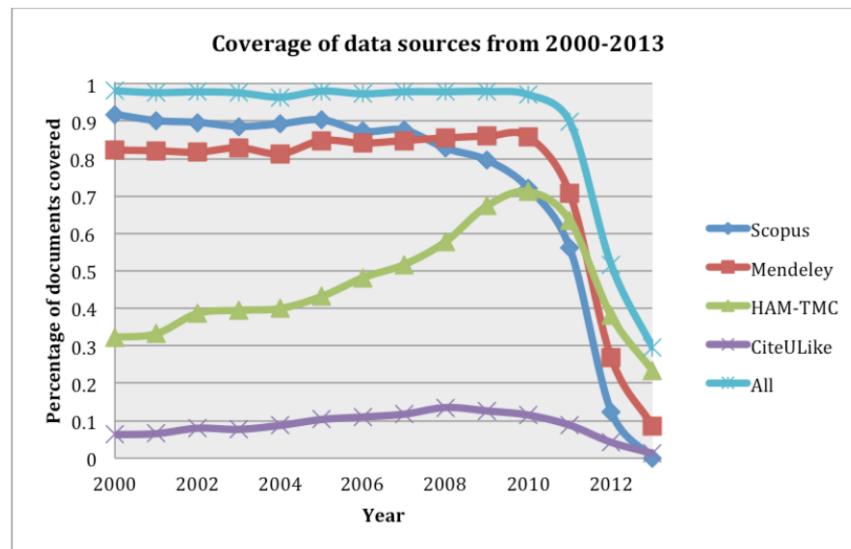


Figure 4.56. Information content of data sources from 2000-2013

### 4.3.3 Results

Table 4.6 presents the results for desirability computed on the Mendeley, Scopus, HAM-TMC, and CiteULike data sets individually. Table 4.7 presents the results of using JIF for predicting document accesses. Table 4.8 presents the results of using degree centrality and PageRank on the click graph for predicting accesses. Finally, Table 4.9 presents the results for existing IR models.

Table 4.6

*Desirability performance from different data sets*

Data sources	Accuracy and coverage			Accuracy with ties broken
	Accuracy	Coverage	Harmonic mean	
Mendeley	60.52%	71.27%	0.6546	57.54%
Scopus	<b>70.49%</b>	70.10%	<b>0.7029</b>	<b>64.36%</b>
HAM-TMC	63.02%	<b>73.31%</b>	0.6778	59.54%
CiteULike	69.22%	4.703%	0.0881	50.90%

Table 4.7

*Performance from JIF*

Data sources	Accuracy and coverage			Accuracy with ties broken
	Accuracy	Coverage	Harmonic mean	
JIF	57.28%	95.78%	0.7178	56.97%

Table 4.8

*Performance from graph metrics computed on click graph*

Data sources	Accuracy and coverage			Accuracy with ties broken
	Accuracy	Coverage	Harmonic mean	
Degree centrality (normalized by date)	57.68%	<b>68.48%</b>	0.6262	55.26%
PageRank (normalized by date)	<b>58.15%</b>	<b>68.48%</b>	<b>0.6289</b>	<b>55.58%</b>

Table 4.9

*Results for existing document ranking models*

<b>Model</b>	<b>Accuracy with ties broken</b>
TF-IDF	<b>64.20%</b>
DFR_BE	63.90%
DFR_IDF	63.90%

In terms of accuracy, the click graph metrics and desirability metrics using the HAM-TMC, Mendeley, and Scopus data sets outperformed JIF. However, JIF had a large coverage of 95.78%, which gave it a higher harmonic mean. The Scopus data set had the highest harmonic mean of the document access data sources. An unpaired t-test found that the results of the Scopus data set and the JIF in terms of the harmonic mean were not statistically significant ( $p$  value  $> 0.05$ ).

When looking at the accuracy with ties broken at random, desirability computed on the Scopus data set performed the highest of any of the methods in Tables 4.6-4.9 with an accuracy of 64.36%. A t-test found that the accuracy with ties broken at random for desirability computed using the Scopus data set was statistically significant compared to the performance of the click graph metrics and the JIF metric in terms.

All of the existing IR models outperformed the JIF metric and metrics computed from the click graph. In both cases, a t-test found the performance to be statistically significant ( $p < 0.05$ ). The existing IR models outperformed desirability computed on the HAM-TMC, Mendeley, and CiteULike data sets. The performance for each of the existing IR models

was statistically significant as compared to the performance of desirability computed on the HAM-TMC, Mendeley, and CiteULike data sets individually ( $p < 0.05$ ). Desirability computed on the Scopus data set outperformed all of the existing IR models. However, the increase in performance was small (e.g. 0.16% over TF-IDF) and was not statistically significant ( $p > 0.05$ ).

In the next set of experiments, I explore combining the data sources to improve coverage and accuracy for desirability. Table 4.9 presents the results of the experiments. Overall, the combination experiments improved both accuracy and coverage. The best performance, in terms of the harmonic mean was found using the combination of the CiteULike, HAM-TMC, Mendeley, and Scopus data sets, which attained a harmonic mean of 0.7877. A t-test found that the results were statistically significant as compared to the JIF metric, desirability computed on individual data sets, and the metrics computed on the click graph ( $p < 0.05$ ).

Table 4.10

*Desirability results from combining multiple data sources*

Data sources	Accuracy and coverage			Accuracy with ties broken
	Accuracy	Coverage	Harmonic mean	
CiteULike, HAM-TMC, Mendeley, Scopus	69.41%	<b>91.19%</b>	<b>0.7882</b>	67.70%
CiteULike, HAM-TMC, Mendeley	63.86%	88.38%	0.7415	62.25%
CiteULike, HAM-TMC, Scopus	70.34%	88.55%	0.7840	<b>68.01%</b>
CiteULike, HAM-TMC	62.86%	75.44%	0.6858	59.70%
CiteULike, Scopus	<b>71.32%</b>	69.62%	0.7046	64.84%
CiteULike, Mendeley	61.77%	72.83%	0.6685	58.57%
HAM-TMC, Scopus	69.77%	88.46%	0.7801	67.49%
HAM-TMC, Mendeley	61.77%	72.83%	0.6685	58.57%
HAM-TMC, Scopus, Mendeley	68.53%	91.29%	0.7829	66.92%

Mendeley, Scopus	67.96%	82.06%	0.7435	64.74%
------------------	--------	--------	--------	--------

In terms of accuracy with the ties broken at random, desirability computed on the CiteULike, HAM-TMC, and Scopus data set was the best with an accuracy of 68.01%. This combination outperformed the CiteULike, HAM-TMC, Mendeley, and Scopus data, which was the best in terms of harmonic mean, by 0.31%, which was not statistically significant. The combination outperformed desirability computed on the individual data sets from Table 4.6 for Mendeley (+10.47%), HAM-TMC (+8.56%), and CiteULike (+17.92%) ( $p < 0.05$ ). The combination outperformed desirability computed on the Scopus data set by 3.65%, but this increase was not statistically significant ( $p > 0.05$ ). Similarly, the combination improved performance over TF-IDF by 3.81%, but this increase was not statistically significant ( $p > 0.05$ ). Desirability computed on the CiteULike, HAM-TMC, and Scopus combination outperformed the DFR\_BE (+4.11%) and DFR\_IDF (4.11%) document ranking functions from Table 4.9. This performance increase was found to be statistically significant.

#### 4.3.4 Discussion

In this section, I performed an evaluation of using desirability computed on multiple data sources for predicting document accesses. For summarization purposes, I will only discuss the performance for accuracy with ties broken at random. The performance of desirability was compared with JIF, graph metrics computed on the click graph, and existing IR models. I found that desirability computed on the Scopus data set outperformed JIF, the existing IR models, and desirability computed on the other individual data sets. A t-test found that the performance increase of desirability computed on the Scopus data set was

statistically significant for JIF and desirability computed on the individual data sets ( $p < 0.05$ ). However, the performance increase over the existing IR models was not statistically significant ( $p > 0.05$ ). In the next set of experiments, I looked at the possible combinations of the individual data sets for computing desirability. The best performing combination was the CiteULike, HAM-TMC, and Scopus data sets. The combination of CiteULike, HAM-TMC, and Scopus outperformed desirability computed on the Scopus data set by 3.65%, but a t-test found that this improvement was not statistically significant ( $p > 0.05$ ). Similarly, the combination outperformed TF-IDF by 3.81%, but again was found to not be statistically significant ( $p > 0.05$ ). The combination outperformed DFR\_BE and DFR\_IDF by 4.11%, which was found to be statistically significant ( $p < 0.05$ ).

The primary weakness of this study is the availability of data itself. The goal of this study was to show that desirability had utility for document ranking. In that aim, I was successful. For example, I showed that desirability computed on the CiteULike, HAM-TMC, and Scopus data sets outperformed the existing document ranking algorithms (though the increase over TF-IDF was not statistically significant). This is a very interesting finding as TF-IDF (though dated) and BM25 are competitive ranking algorithms. Additionally, desirability is independent of the query so it relies on information not present in the query or in the document text, which means that it provides information that can be utilized in conjunction with traditional query dependent metrics for ranking. However, the study was not able to answer precisely how well desirability can perform if adequate data are available. With all of the available data sets combined, a coverage of 91.19% was attained. However, when looking at individual data sets, the document accesses from HAM-TMC alone attained a coverage of only 73.31%. This means that while a large number of accesses

were available from the historical query logs, there were still 26.69% of the data for which there was no access information. Referring back to Figure 4.54, I showed that a disproportionate amount of the viewed documents were from 2013. This is unsurprising as PubMed ranks by reverse chronological order. The coverage of the Scopus data set decreased dramatically for 2013 and had less than 10% coverage for that year as shown in Figure 4.56. This is unsurprising as it takes time for authors to generate new papers that cite newly published works. However, this is exactly where click data such as HAM-TMC can be very valuable. Once a paper is published, information can be quickly gathered about their usage if a service has the number of users as PubMed. To fully understand the utility of desirability experiments and especially the utility of leveraging document access information for ranking studies must be conducted over much larger data sets. To summarize, the data sparseness problem precludes drawing a definitive conclusion on how much performance can be squeezed from click data alone. However, there is an optimistic interpretation available. Even though the available data in this experiment were sparse, I was able to show that desirability was competitive with existing and well-established document ranking functions.

In this study, I showed that desirability was competitive with existing document ranking functions. However, desirability is a query independent (i.e. prior probability) score and is intended for use in conjunction with a query dependent score such as that produced by BM25 or TF-IDF. Specifically, this study did left unanswered the question of how much improvement can be gained by using a non-uniform prior in conjunction with a query dependent ranking function. I explore this question in Chapter 6.

## **Chapter 5: Predicting Document Clicks Using Information Scent**

According to the ACT-R theory of human memory, memory items are retrieved based on the prior probability and the current context. Chapter 4 presented an in-depth investigation of estimating the prior probability of a document being accessed and using this prior probability function for predicting document accesses. This chapter seeks to predict document accesses using context.

Before proceeding, I discuss the use of the term “context” in the literature as it pertains to IR. There are many different types of context (for in-depth discussion see (Ingwersen & Jarvelin, 2010)). The most straightforward example of context is the terms that comprise the user query. Based on the context provided by the user query, many IR systems rank the documents based on a measure of similarity between the documents in the corpora and the user query (D. L. LEE, Chuang, & Seamons, 1997). Another example of context that can be used in an IR system is information about the user who issued the query (Pohl, Radlinski, & Joachims, 2007). This work uses the user query to define the context.

Recall from previous discussions that Pirolli defines information scent as a rational analysis of the categorization of cues (P. Pirolli, 2009). The information scent calculation is a prediction of how likely a given user is to click a document based on the context (cues from the text of the document and the user’s information need). The click predictions are based on the interaction (through a spreading activation mechanism) of the textual inputs (such



as title of the documents) and the information need of the user (which is represented as the user query).

The Information Foraging Theory extends the ACT-R theory to predict the browsing behavior of users. The Information Foraging Theory's information scent calculation uses ACT-R's spreading activation theory, which ACT-R uses for predicting the memory items that will most likely be needed given the current context. The Information Foraging Theory uses the spreading activation mechanism to compute the information scent for a given document or URL given the user's information need (query used as proxy) and the information present to the user on the screen.

This work makes several contributions. From an applied viewpoint, this work is the first attempt to apply insights from computational cognitive modeling to model users as they interact with document retrieval systems in the biomedical domain. The previous applications of the Information Foraging Theory were applied entirely outside of the biomedical domain (Budiu, Pirolli, & Hong, 2009; Card, et al., 2001; Chi, Pirolli, Chen, et al., 2001; Chi, Pirolli, & Pitkow, 2001; Hong, Chi, Budiu, Pirolli, & Nelson, 2008; Huberman, et al., 1998; P. Pirolli, 2005, 2009; P. Pirolli & Card, 1995, 1999b; P. Pirolli & W-T., 2006; P. L. Pirolli & Anderson, 1985; P. L. Pirolli & Pitkow, 2000). Additionally, the majority of past studies using information scent for click prediction modeled the general user population rather than expert users. For example, only recently have researchers explored using information scent to model expert behavior such as finding errors in programs (Lawrance, et al., 2007b; Lawrance, Bellamy, Burnett, & Recker, 2008; Lawrance, et al., 2013). The user population in this study, constrained to users in the Texas Medical Center, has a high percentage of expert users since the user population is

composed primarily of graduate students, clinicians, and researchers. Additionally, this chapter presents an updated mathematical framework for calculating information scent based on the mathematical framework of language models, which provides an interpretation of information scent that more closely adheres to the underlying Bayesian mathematical framework of the ACT-R theory and Information Foraging Theory.

The remainder of this chapter is organized as follows. Section 5.1 presents an overview of using information scent to model user interactions with the PubMed retrieval system. Section 5.2 presents an overview of the mathematical frameworks for computing information scent. Section 5.3 discusses how spreading activation handles context sensitivity. Section 5.4 presents the methods which discuss the creation of the corpus statistics used in this work as well as the evaluation method used in this work. Section 5.5 presents the results of the experiments. Finally, Section 5.6 presents the discussion of the work in this chapter.

### **5.1 Overview of Modeling Biomedical Document Accesses Using Information Scent**

Figure 5.1 presents the information displayed for a typical document on PubMed. Examples of information that could influence whether or not a given document is clicked includes the title, date, authors, and journal name. The evidence used for computing information scent in this work is the title of the document. Future work will explore using additional information visible to the user through the PubMed interface, such as the authors and the journal in which the document was published.

1. [Predicting biomedical document access as a function of past use.](#)  
Goodwin JC, Johnson TR, Cohen T, Herskovic JR, Bernstam EV.  
J Am Med Inform Assoc. 2012 May-Jun;19(3):473-8. doi: 10.1136/amiajnl-2011-000325.  
Epub 2011 Sep 13.  
PMID: 21917645 [PubMed - indexed for MEDLINE] [Free PMC Article](#)  
[Related citations](#)

Figure 5.1. Example result from PubMed

Figure 5.2 presents an example of applying information scent to predict the document accesses of users browsing PubMed. In this example, the user has the high-level information need of finding documents that discuss predicting document accesses. The user compiles the information need into the query “predicting document accesses” which is submitted to the PubMed IR system. PubMed subsequently returns a list of six documents. According to the assumptions of the Information Foraging Theory, which is supported by studies of human browsing behavior (Granka, Joachims, & Gay, 2004), users browse the set of returned documents in descending order<sup>5</sup>. In this example, the user looks at the first item that is returned, which is the document “Predicting biomedical document access as a function of past use” (Goodwin, et al., 2011). The network in Figure 5.2 represents how the information scent is calculated for the first document.

---

<sup>5</sup> The Information Foraging Theory contains additional models that predict the amount of time a user will spend within a page of results before abandoning the result set or issuing another query (Huberman, et al., 1998). I did not explore this aspect of the Information Foraging Theory in this dissertation, but this is an area for future exploration.

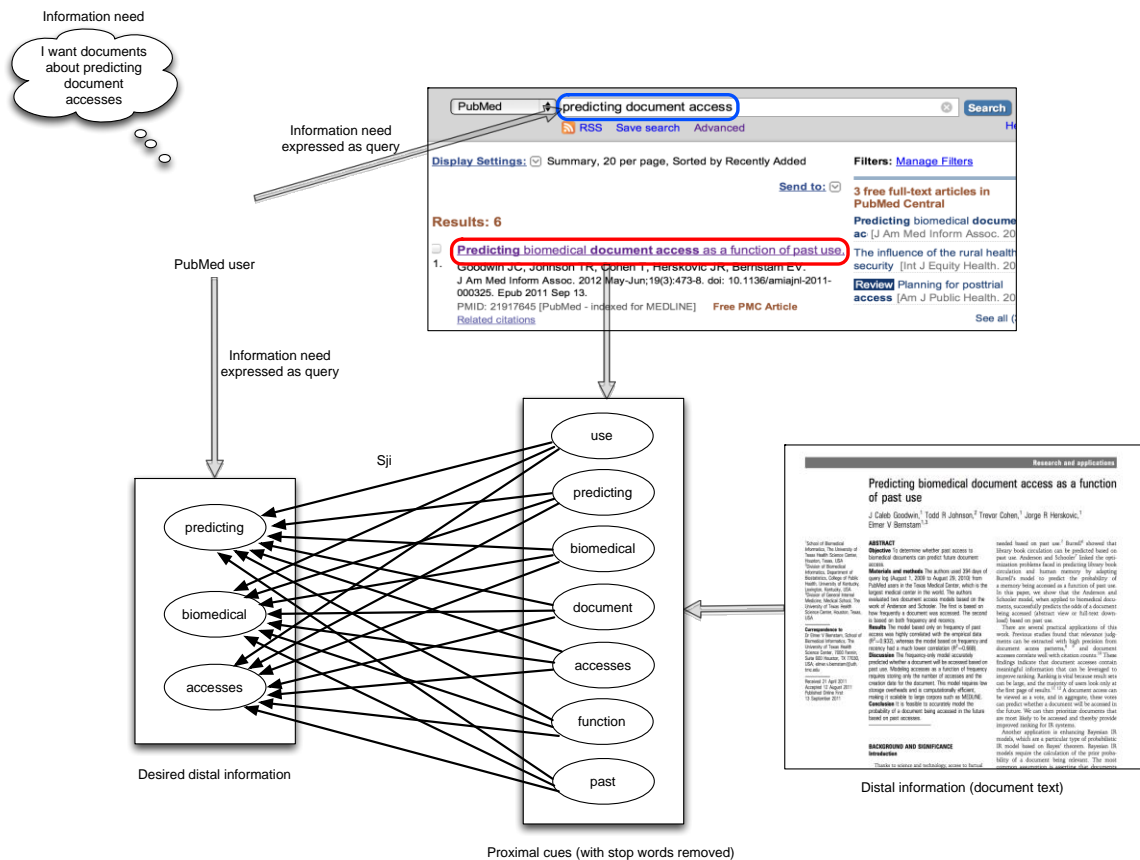


Figure 5.2. Mapping information scent to PubMed search

The network in Figure 5.2 is nearly identical to the network used to model spreading activation in human memory by the ACT-R theory of human memory. The only noted difference is the direction of the inference. The ACT-R model of human memory models the memory probe as originating from the external environment. For example, a person hears the phrase “my lawyer is a shark” and the activation is modeled as propagating from the nodes “lawyer” and “shark”. In the case of information scent, the information need (query terms used as proxy) is modeled as a goal state (e.g. finding documents about the goal (information need)) in the user’s mind. In this case, the information scent calculation models the activation as flowing from the proximal cues (e.g. title of an article) to the

information need. In terms of cognitive modeling, the goal is satisfied as the user finds documents that contain information about the information need. That is, the activation level of the goal represents the utility of the given link (e.g. how likely is a given link to satisfy the information need) and the user clicks the links with high utility values.

## **5.2 Overview of Information Scent Calculation**

This section presents an overview of information scent, discusses previous implementations, and presents the new interpretation based on recent research in probabilistic IR systems. Section 5.2.1 presents the mathematical framework that underlies information scent. Section 5.2.2 presents an overview of the previous implementations of information scent. Finally, Section 5.2.3 presents the new interpretation of information scent.

### **5.2.1 Mathematical framework for information scent**

The information scent component of the Information Foraging Theory is derived from the log-odds form of Bayes theorem provided in Equation 5.1. The symbol  $D$  represents the terms in a proximal cue. In this work the proximal cue is the title of the document, but it can include other information such as the authors or journal in which the document was published. The symbol  $Q$  represents the user query, which is a representation of the user's information need. In Equation 5.1, the component  $\log \left( \frac{P(Q)}{P(\bar{Q})} \right)$  is the prior odds of a given information need. The information scent calculation assumes that the prior odds are uniform which reduces the calculation to Equation 5.2. The component  $\log \left( \frac{P(D|Q)}{P(D|\bar{Q})} \right)$  represents the log likelihood ratio.

$$\log\left(\frac{P(Q|D)}{P(\bar{Q}|D)}\right) = \log\left(\frac{P(Q)}{P(\bar{Q})}\right) + \log\left(\frac{P(D|Q)}{P(D|\bar{Q})}\right) \quad (5.1)$$

$$\log\left(\frac{P(Q|D)}{P(\bar{Q}|D)}\right) \approx \log\left(\frac{P(D|Q)}{P(D|\bar{Q})}\right) \quad (5.2)$$

The information scent calculation in Information Foraging Theory makes the simplifying assumption that the base rate probability of an information need  $Q$  occurring will not vary substantially from when the information need is present and a given item  $D$  is not needed. In other they assert the following:  $P(D|\bar{Q}) \approx P(Q)$ . This reduces the log odd calculation to  $\log\left(\frac{P(D|Q)}{P(Q)}\right)$ . After making the simplifying assumption, the following transformation is applied.

$$\frac{P(D|Q)}{P(D)} = \frac{P(Q \cap D)}{P(D)} * \frac{1}{P(Q)} = \frac{P(Q|D)}{P(Q)}$$

The simplifying assumptions and the transformation yield the association strength ( $S_{ji}$ ) equation in Equation 5.3. Equation 5.4 presents the final activation ( $A_i$ ) equation used to calculate information scent.

$$S_{ji} \approx \log\left(\frac{P(i|j)}{P(i)}\right) \quad (5.3)$$

$$A_i = \sum_{j \in Q} \log \left( \frac{P(i|j)}{P(i)} \right) \quad (5.4)$$

Given the simplifying assumptions made by the Information Foraging Theory, the likelihood estimation is equivalent to pointwise mutual information (PMI) (Equation 5.5). PMI has shown promising correlations with human similarity judgments even when compared with more sophisticated methods such as latent semantic analysis (LSA) (Recchia & Jones, 2009; P. D. Turney, 2001b). Numerous studies utilized PMI as a component in IR systems with its primary role being finding synonyms and related terms (Aminul & Inkpen, 2006; Terra & Clarke, 2004; P. Turney, 2001). In addition, although PMI is a component in some ranking functions, it lacks document length normalization and a term importance measure which are commonly incorporated into ranking functions. For the purposes of document ranking, an interpretation of the likelihood component of Bayes Theorem that avoids a simplification to PMI while incorporating term importance and document length normalization would be ideal. I present this interpretation in Section 5.2.2.

$$pmi(x; y) = \log \left( \frac{p(y|x)}{p(y)} \right) \quad (5.5)$$

### 5.2.2 Previous information scent implementations

The actual implementations of the information scent calculation made additional assumptions than the reduction to PMI discussed previously in Section 5.2.1. This goal of this section is to discuss the actual implementation of the mathematical framework presented in the previous section.

Many implementations of Information Foraging Theory have been developed within the ACT-R environment (P. Pirolli, 2005; P. Pirolli & Card, 1998, 1999a; P. Pirolli, Chi, & Farahat, 2005; P. Pirolli & W-T., 2006). In these instances, the information scent calculation relied upon ACT-R's mathematical framework. In other instances, Pirolli implemented information scent using TF-IDF (Chi, Pirolli, Chen, et al., 2001). Additionally, Pirolli has explored using alternative models such as LSA for calculating information scent (Budiu, Royer, & Pirolli, 2007).

The implementation of Information Foraging Theory is not alone in its loose interpretation of the underlying Bayesian theory of ACT-R. First, ACT-R makes assumptions that reduce the strength of association calculation to PMI (Equation 5.5). ACT-R implementations further simplify the strength of association calculation by setting the  $S_{ij}$  value to 1 for query terms that are not present in the document. If the term is present, then the strength depends on the ratio of the number of documents that contain a term and the total number of documents in the corpus. In other words, ACT-R approximates the  $S_{ij}$  using the standard IDF equation shown in Equation 5.6. Thus, the actual implementation of the ACT-R activation function is described in Equation 5.7. This interpretation is extremely close to the standard TF-IDF interpretation shown in Equation 5.8, but lacks a length normalization and TF component.



$$IDF = \log \left( \frac{\text{number of documents}}{\text{number of documents with term}} \right) \quad (5.6)$$

$$A_i = \sum_{j \in Q \cap D} IDF_j \quad (5.7)$$

$$TF - IDF = \sum_{t \in q \cap D} \frac{tf}{|D|} * idf \quad (5.8)$$

An additional extension to the ACT-R theory called partial matching enables synonym matching through partial matching. In the ACT-R model, the partial matching mechanism enabled modeling tasks such as memory retrieval errors (Lebiere, Anderson, & Reder, 1994). Historically, partial matching is not generally used in the Information Foraging Theory implementations. Equation 5.9 presents the ACT-R function with partial matching<sup>6</sup>. The partial matching works by assigning scores to semantically related terms. Numerous methods such as LSA and PMI can compute a semantic relatedness score. The  $M_{ki}$  parameter reflects the semantic relatedness between an element in the goal (in this case a query term) and a given memory element (a term in a document in this case). In the ACT-R framework, these values take on the range  $[-1, 0]$ . The value -1 is used when there is no similarity between items. The  $P_k$  parameter is known as the mismatch penalty, which weights the amount of evidence given to partial matches. Equation 5.10 shows the equation implemented in this work.

---

<sup>6</sup> In this chapter, the prior probability ( $B_i$ ) is assumed equal for each document so this parameter can be ignored. The next chapter explores using a non-uniform  $B_i$ .

$$A_i = B_i + \sum_j W_j S_{ji} + \sum_k P_k M_{ki} \quad (5.9)$$

$$A_i = \sum_{j \in Q} IDF_j + \sum_k P_k M_{ki} \quad (5.10)$$

### 5.2.3 Interpretation of information scent based on language models

Two methods are explored in the work for estimating  $P(Q|D)$  based on the mathematical framework of language models. Previously, I presented an introduction to language models in Chapter 2.3.1. Equation 5.11 presents the first method which is based on Dirichlet smoothing (MacKay & Peto, 1995; C. Zhai & Lafferty, 2002). The parameter  $w$  represents an element of the query  $Q$ . For the purpose of information scent,  $D$  represents the proximal cue. The proximal cue that is used in this work is the document title since this information is visible to the user and influences whether or not a document is clicked. Additional information is available to the user such as the journal in which the article is published and the authors of the paper. These additional cues are not investigated in this dissertation and will be the focus of future research. Equation 5.12 and Equation 5.13 present the maximum likelihood estimate for the document language model and the background language model respectively. The maximum likelihood estimate in Equation 5.13 calculates the probability  $p(w|D)$  based on the number times  $w$  occurs in the proximal cue  $D$ . The background language model shown in Equation 5.13 is based on the frequency of occurrence of  $w$  and the frequency of all terms in the collection  $C$ . The parameter  $\mu$  is the pseudo count parameter, which controls the amount of smoothing.

$$P(w|D) = \frac{P(w|D) + \mu P(w|C)}{|D| + \mu} \quad (5.11)$$

$$P(w|D) \approx P_{ML}(w|D) = \frac{\text{numberOfRelations}(w, D)}{\text{numberOfRelations}(D)} \quad (5.12)$$

$$P(w|C) \approx P_{ML}(w|C) = \frac{\text{numberOfRelations}(w)}{\text{numberOfRelations}(C)} \quad (5.13)$$

The second version of the spreading activation model is based on the generalized framework for language model smoothing using graphs, which I previously discussed in Section 2.3.1. This is the analogue to ACT-R's partial matching discussed in Section 5.2.2. In Equation 5.14, the language model in Equation 5.11 is updated with evidence from semantic relatedness scores, which enables partial matching. Conceptually, one can view this as combining a score, which reflects the likelihood of an element of the query  $w$  given the proximal cue (e.g. document title) with the likelihood of the neighbors of  $w$  given the proximal cue. The semantic relatedness measure can be computed using several different methods such as LSA or topic modeling. Equation 5.15 presents the degree centrality metric used in this work, which is equivalent to the generalized measure for computing degree centrality in weighted networks (Barrat, Barthelemy, Pastor-Satorras, & Vespignani, 2004). The  $P(v|D)$  for the connected term  $v$  is calculated using Equation 5.11.

$$P(w|D) = (1 - \lambda)P(w|D) + \lambda \sum_{v \in V} \frac{w(w, v)}{\text{Deg}(v)} P(v|D) \quad (5.14)$$

$$Deg(v) = \sum_{u \in V} w(u, v) \quad (5.15)$$

### 5.3 Understanding Context Sensitivity and Spreading Activation

The ACT-R theory of human memory has primarily focused on the role of spreading activation in modeling memory retrieval and, with some exceptions (Budiu, 2001; Budiu & Anderson, 2000, 2002, 2004, 2006; Guhe, Smaill, & Pease, 2010), has not focused on the role of spreading activation in modeling text comprehension. In contrast, Kintsch has deeply explored spreading activation models for developing a computational theory of text comprehension as part of his construction-integration model (W Kintsch, 1998). The purpose of this section is to look at spreading activation through the lens of Kintch's theory to gain additional insight into how spreading activation models handle context using an illustrative example.

In this work and in Kintsch's construction-integration theory, the background knowledge of a user is constructed of a network of propositions (Kintsch refers to this as the "knowledge network"), which are theorized to be the fundamental unit of knowledge in theories of comprehension (W Kintsch, 1998). In this work, as well as in much of Kitsch's work, the propositions are unlabeled and encode the number of times that the terms occurred in the same context (the context in this work is terms co-occurring within the same abstract).

According to Kintsch, the meaning of a node in the network is based on its position in the network. That is, the meaning of the node is based on the strength of the connections of the given node with its direct neighbors in the network and indirectly through nodes that are

connected to the direct neighbors. Spreading activation is the theorized mechanism by which activation values are computed on the network given a probe. Given a probe (e.g. query term(s)), the spreading activation mechanism computes the activation of the connected nodes by aggregating the association strengths between the probe and the nodes connected to the probe. According to Kintsch, the spreading activation mechanism plays an integral role in text comprehension. This role is described by Kintsch by the following (W Kintsch, 1998).

Knowledge nets thus imply a commitment to a radical constructionist position in the controversy about the mental representation of word meanings. In a mental lexicon, one looks up the meaning of a word. In a knowledge net, there is nothing to look up. Meaning has to be constructed by activating nodes in the neighborhood of a word. This activation process is probabilistic, with activation probabilities being proportional to the strengths of connections among the nodes, and it may continue for a variable amount of time, spreading outward into the knowledge net from the source node. The meaning of the source word is then, the set of activated nodes in the knowledge net.

(W Kintsch, 1998)

To summarize, Kintsch views the meaning construction process and thus the comprehension of concepts to be highly contextual. In this framework, the meaning of concepts is not static and simply retrieved, but is constructed and the meaning, which is an activated sub-network, will vary based upon context.

The significance of Kintsch's work is that it provides a more general view of spreading activation than that of the ACT-R theory of human memory. In the ACT-R theory of human

memory, the spreading activation is theorized to be context-sensitive (J. Anderson, 2007; J. R. Anderson & Bower, 1973), but plays the role of calculating the activation values of the nodes connected to the memory probe to determine the single item that will be retrieved from long-term memory. Within Kintsch’s framework, the spreading activation model creates a sub-network that represents the meaning of the memory probe which facilitates text comprehension.

In the remainder of this section, I provide a demonstration of the context sensitivity of spreading activation using the Wikipedia abstracts (first two lines of a Wikipedia article). In this example, I use the spreading activation model based on language models presented previously in Equation 5.14. In this example, I explore the term space of the terms “bank” and “money”. Table 5.1 presents the top twenty terms associated with the terms “bank” and “money”. The term “bank” is composed of several different meanings which includes “financial institution”, “body of land near a river”, and “location known as West Bank”. The related terms for the concept “money” are primarily synonyms and activities done with money such as lending. Figure 5.3 presents a graphical representation of the sub-network formed by the terms “bank” and “money”. The activation values in this particular case are not context sensitive (i.e. the activation values are computed based on the terms “bank” and “money” in isolation of each other).

Table 5.1

*Top 20 activated terms for “bank” and “money” in decreasing order of activation value*

<b>Terms for bank</b>	<b>Terms for money</b>
banking, nablus, theban, ramallah, banco, savings, banque, szczecin, jenin, krka,	laundering, purses, pga, raise, banknotes, monetary, totalling, currency, majors,

regulator, kolpa, sava, hebron, tulkarm, luxor, bireh, hsbc, drava, banks	raises, lending, payment, cash, sums, interbank, fraud, borrow, borrowing, scam, debt
---	---

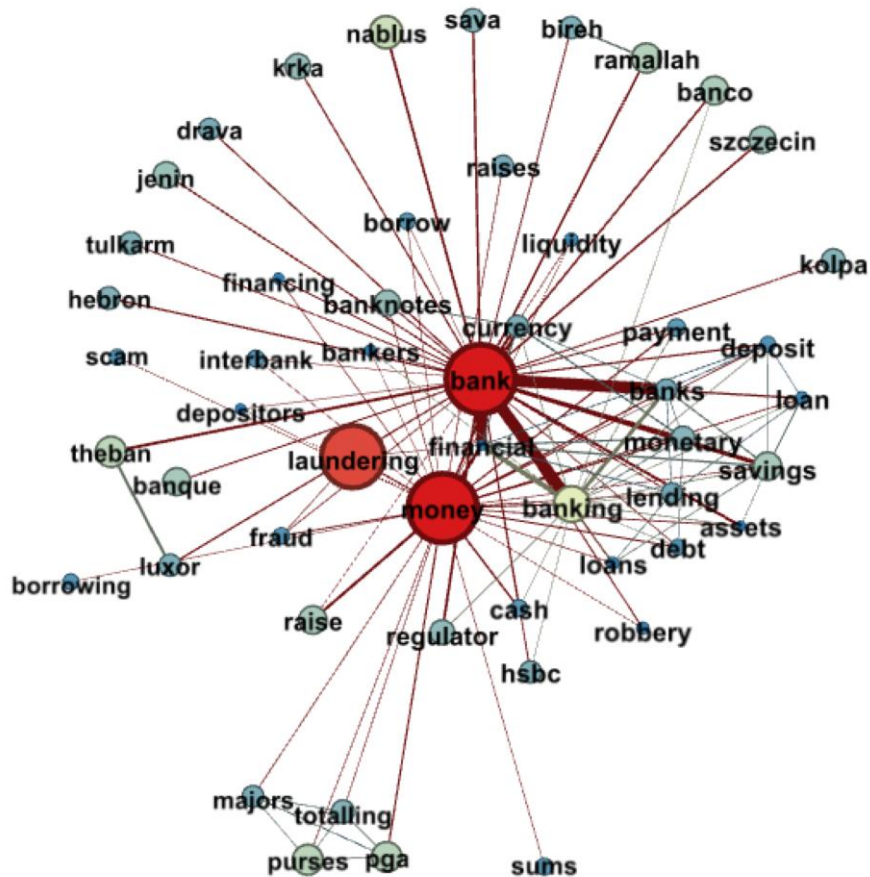


Figure 5.3. Sub-network created by the terms bank and money with independent activation calculation. The node size represents the activation value. The color of the node is on the scale high (red) to low (blue) activation. The line width represents the number of times the two terms co-occurred.

The symbol  $\oplus$  is used here to denote the combination of two terms. Table 5.2 presents the top twenty terms for the combination bank  $\oplus$  money. In this case, the spreading activation algorithm suppressed the unwanted meanings of the term “bank” such as the “location

known as West Bank”. The top terms deal primarily with the desired “banking as a financial institution” meaning. Figure 5.4 presents the same sub-graph as Figure 5.3, but in this case, the node size and colors reflect the activation values from the combination bank  $\oplus$  money. With a cursory glance, it is apparent that the combination bank  $\oplus$  money has suppressed many of the unwanted terms. In mapping this example back to Kintch’s view, the comprehension of the combination bank  $\oplus$  money is the subnetwork in Figure 5.4 with the activation values of the nodes in the network representing the level of contribution of each term to the combination’s meaning.

Table 5.2

*Top 20 activated concepts for Bank  $\oplus$  money in descending order of activation value*

banknotes, banking, monetary, savings, laundering, currency, deposit, lending, banks, loans, loan, interbank, liquidity, payment, robbery, financing, assets, financial, depositors, bankers
--



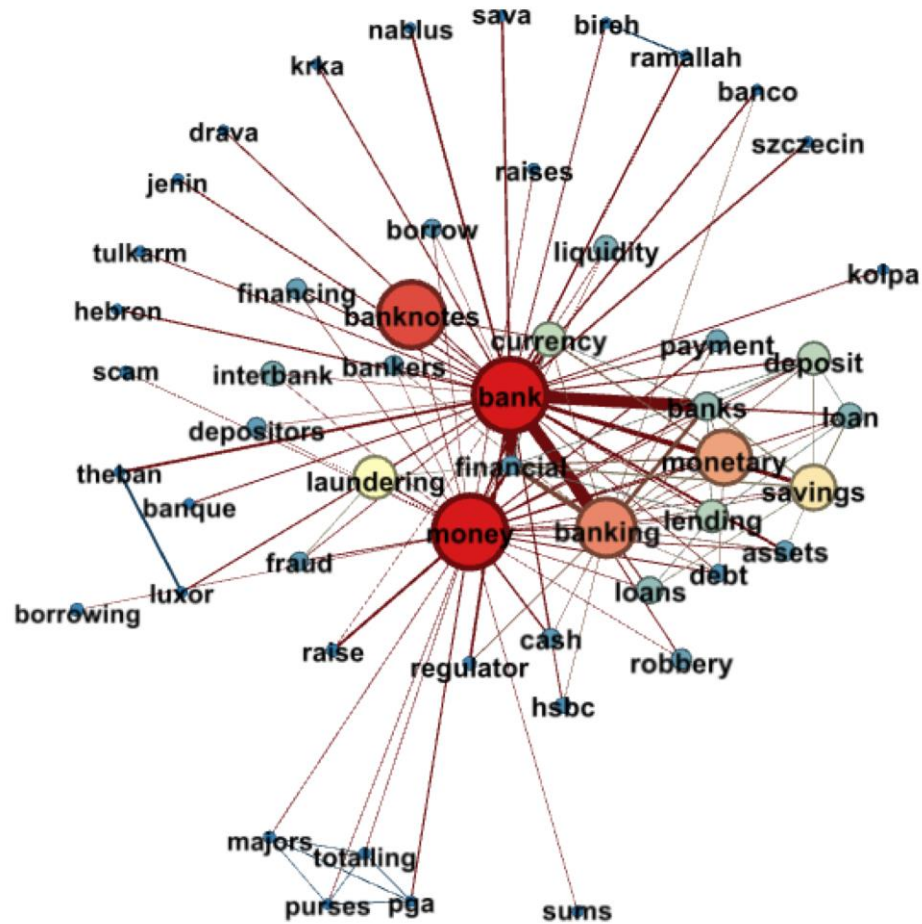


Figure 5.4. Sub-network created by the terms bank and money with context dependent activation. The node sizes reflect the activation value. The color of the node is on the scale high (red) to low (blue) activation. The line width represents the number of times the two terms co-occurred.

To further illustrate the effect of context sensitivity, I manually categorized the top 20 terms for the terms “bank” and “money” into topics (presented in Table 5.3). If a term (e.g. “purses”) would require the creation of a category where it would be the only instance in that category, I placed it into the “Other concepts” category for the purpose of this illustration.

Table 5.3

*Manual term classification*

Financial institution	Banking, banco, savings, banque, regulator, hsbc, laundering, banknotes, lending, interbank, borrow, borrowing, deposit, loans, robbery, financing, bankers, depositors, assets
West bank	Nablus, Ramallah, jenin, hebron, tulkarm, bireh
Thebes	Theban, luxor
River	Krka, kolpa, sava, drava
Name or type of money	Monetary, currency, payment, cash, debt, liquidity, financial
Tasks done with money	Raise, totaling, raises, sums, fraud, scam
Other concepts	Szczecin, majors, pga, purses

Figure 5.5 presents the activation values for each category from Table 5.3 for the probes “bank” and “money” in isolation of each other. For the term “bank”, the “financial institution” meaning comprises approximately 2% of the total possible activation. From Figure 5.5, it is noticeable that the term “bank”, while being predominately composed of the “financial institution” meaning, is spread across the other meanings of the term. Figure 5.6 compares the activation values for the combination bank  $\oplus$  money and the terms “bank” and “money” in isolation. The combination bank  $\oplus$  money suppressed the majority of the unwanted meanings such as “location known as West Bank” and “body of land near a river”. Most notable is the high concentration of the activation values for the category “financial institution”. In this case, the concentration of activation for the “financial institution” category was increased to nearly 20% from the maximum of around 2% for the “bank” term alone.

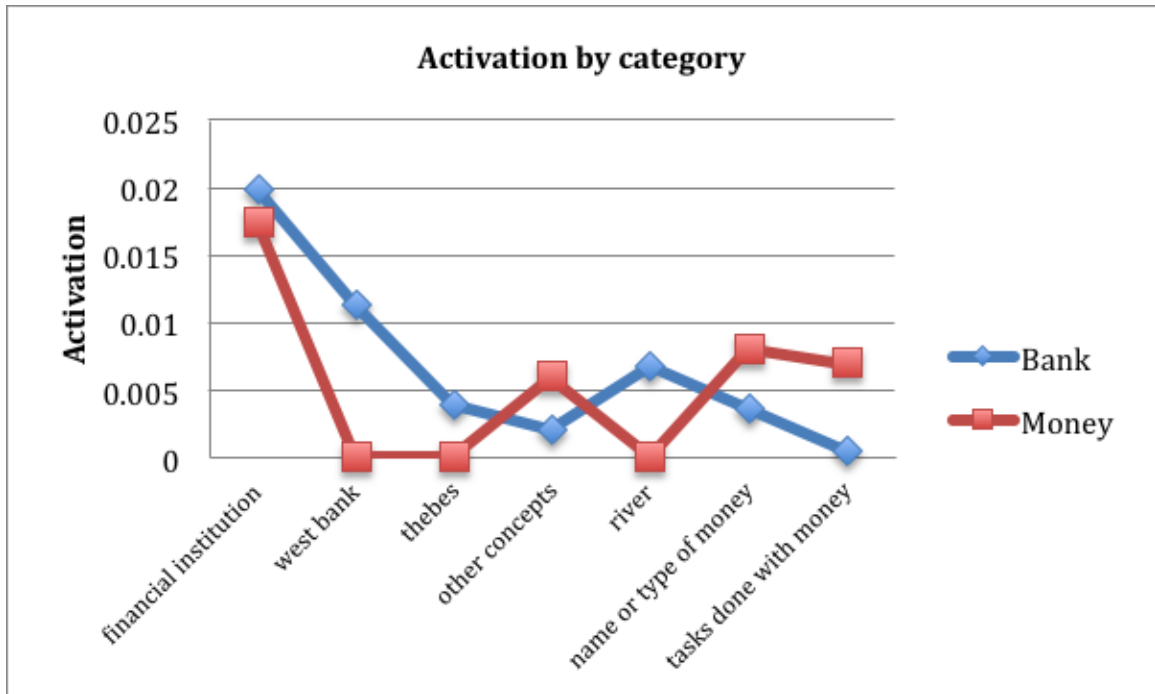


Figure 5.5. Activation values by category for bank and money

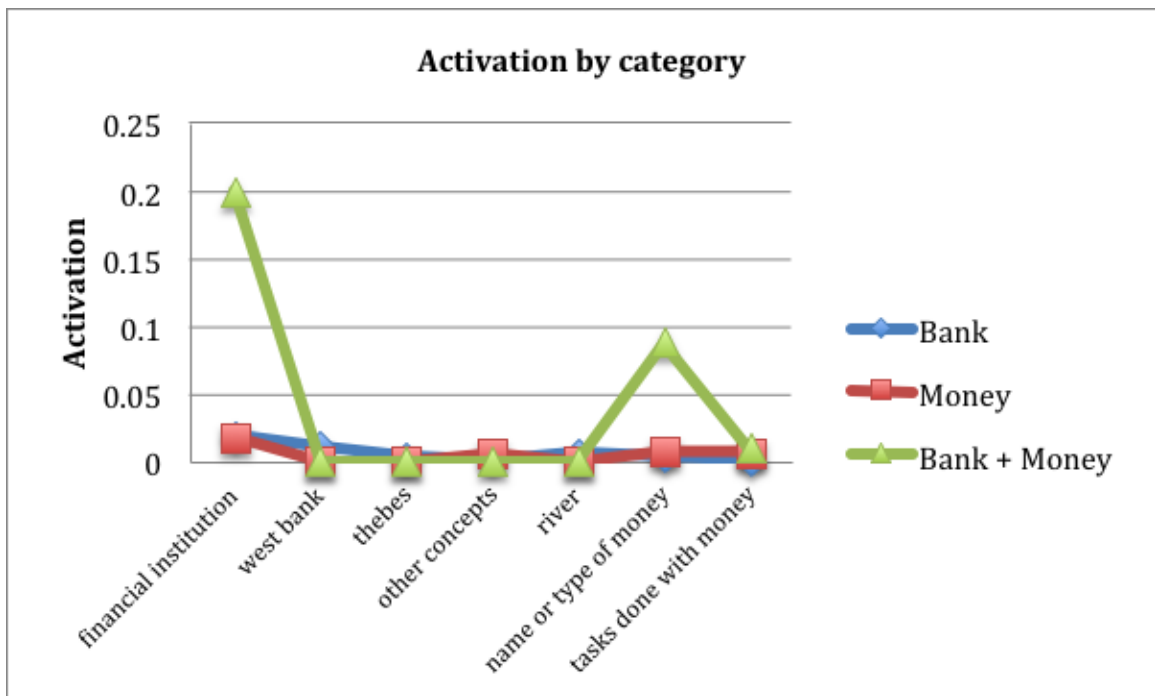


Figure 5.6. Activation values by category for bank, money, and bank  $\oplus$  money

## 5.4 Methods

### 5.4.1 Design of experiments

Figure 5.7 presents an overview of an experiment conducted for one query from the HAM-TMC pairwise judgment data set using information scent. In the experiments, the “information scent calculation” module in Figure 5.7 was replaced with a variation of the information scent calculation in this work. The experiment in Figure 5.7 relied upon the HAM-TMC pairwise judgment data set, which contained the query that was issued, the documents clicked by HAM-TMC PubMed users, and the pairwise judgments extracted for the documents that were clicked. I previously discussed the method for extracting pairwise judgments and using them to evaluate IR algorithms in Section 2.4.2. As a review, I extracted two sets of pairwise judgments for evaluating the models in this work. I extracted the first set of pairwise judgments between documents that were clicked and documents that were not clicked. The goal of this set of pairwise judgments was to evaluate how well a given model could predict the documents that receive clicks and is referred to as the document click pairwise judgments in the remainder of this chapter. The second set of pairwise judgments was extracted between documents that were downloaded and documents that were not downloaded. The goal of this set of pairwise judgments was to evaluate how well a given model could predict document clicks that resulted in a download and is referred to as the document download pairwise judgments in the remainder of this chapter. The motivation for the document download pairwise judgment data set was to determine how well the information scent models could predict accesses that resulted in downloads since downloads could be considered a stronger signal of relevance than

document clicks alone. For example, a user can click a link for a document, view the abstract, and determine from that abstract text that they are not interested in reading the full text. A request for the full text is not necessarily a relevance judgment, but is an indication that the user wanted to read more of the document than just the abstract. For all of the approaches, the documents that the HAM-TMC users viewed were assigned a ranking score by one of the models in the experiments. The ranking scores were used to determine how many of the extracted pairwise judgments from the document download pairwise judgment data set or document click pairwise judgment data set were correctly ordered (e.g., documentA is preferred over documentB) based on the ranking scores.

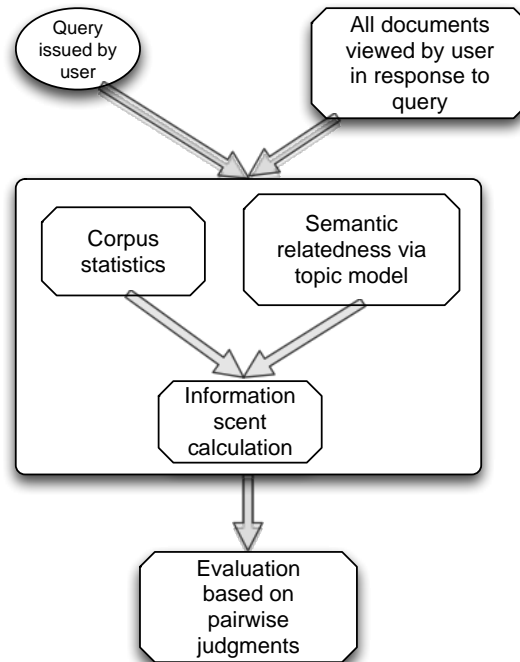


Figure 5.7. Experiment for computing information scent

I used a sliding window for evaluating all of the models in this work. In each window, I divided the data into training and test sets. The algorithms used in this work rely on corpus statistics such as IDF. The corpus statistics for all of the models were computed using the titles and abstracts from the MEDLINE corpus. I utilized the stop word list generated by Salton and Buckley for the SMART information retrieval system for calculating the corpus statistics (Salton, 1971). Some of the information scent models required the use of a semantic relatedness measure. The semantic relatedness measures are discussed in Section 5.4.2. In the training period, the information sent model used the corpus statistics to compute the information scent score for the documents, which were subsequently evaluated using the pairwise judgments extracted from the test sets. Each test set was comprised of one of the days from the HAM-TMC pairwise judgment data set. If the test window was day  $N$  then the training set was comprised of corpus statistics on day  $N$  or earlier. The data used during the training window was composed of documents published on days  $n$  or earlier because a document has to be in the database in order to be returned in response to the query therefore it is reasonable to assume that the corpus statistics should reflect all of the documents currently in the database. Figure 5.8 presents an example of how the sliding window could be used to evaluate the information scent model over each day in the HAM-TMC pairwise judgment data set.

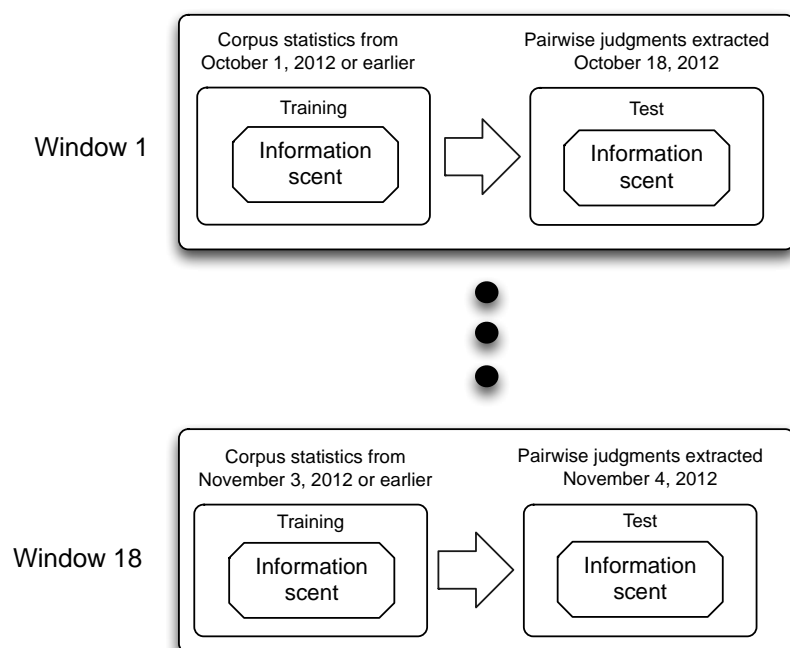


Figure 5.8. Sliding window for calculating information scent

The pairwise judgments were extracted from the query logs of HAM-TMC users, which captured their interactions with the PubMed IR system from October 18, 2012 to November 4, 2012 (19 days). I will refer to this data set as the HAM-TMC pairwise judgment data set. Some of the information scent models had free parameters that required tuning for a particular data set. To tune the parameters, I divided the HAM-TMC pairwise judgment data set into 50% for parameter tuning and 50% for evaluation. This resulted in nine windows for parameter tuning (October 18, 2012-October 26, 2012) experiments and nine windows for evaluation (October 27, 2012-November 4, 2012). Table 5.4 presents an overview of the number of pairwise judgments for the parameter tuning experiments and evaluation.

Table 5.4

*Number of pairwise judgments for parameter tuning and evaluation*

<b>Data set</b>	<b>Number of pairwise judgments</b>
Parameter tuning experiments – document click pairwise judgments	85,435
Evaluation – document click pairwise judgments	68,716
Evaluation – document download pairwise judgments	8,435

Section 2.4.2 presented in detail the methods used in this experiment to extract pairwise judgments and use them for evaluation. For reader convenience, Equation 6.8 presents the accuracy metric used to evaluate the performance of the models in these experiments. If a given algorithm resulted in a tie for the two documents in the pairwise judgment, the tie was broken at random.

$$Accuracy = \frac{\text{Number of correct pairwise judgments}}{\text{Number of incorrect pairwise judgments}} \quad (6.8)$$

### 5.4.2 Semantic relatedness models

I used two models for calculating semantic relatedness in this work. In Section 5.4.2.1 and 5.4.2.2, I discuss in brief the creation of the semantic relatedness models. Both models were generated on the same 10 million document randomly sampled subset of the MEDLINE corpus. The motivation behind creating the subset is that topic modeling is very computationally expensive and generating the topic model on the full corpus for each period in the sliding window could require months of computation. However, the models



developed from the subset (in this case approximately 50% of the available documents) could be used to perform inference to generate a topic distribution for text that was not in the training corpus. The random indexing model is more scalable than the topic model, but I used the same corpus as the topic model for comparison.

#### 5.4.2.1 Topic Modeling

I used the MALLET package to generate the LDA topic model (McCallum, 2002). I selected 500 topics for this model. The MALLET package inferred the topic distribution for new pieces of text using Gibb's sampling. Table 5.5 presents the first nine topics generated using LDA along with the top ten terms associated with each topic.

Table 5.5

*First nine topics for LDA model*

<b>Topic 1</b> liver, hepatic, hepatocytes, cirrhosis, hcc, hepatocellular, hepatitis, livers, fibrosis, portal	<b>Topic 2</b> plant, plants, arabidopsis, thaliana, transgenic, pollen, development, tobacco, ft, leaves	<b>Topic 3</b> acid, acids, ascorbic, uric, nucleic, acetic, lactic, arachidonic, cla, citric
<b>Topic 4</b> driving, military, traffic, vehicle, accidents, accident, medical, drivers, car, safety	<b>Topic 5</b> uptake, pet, imaging, emission, tomography, scintigraphy, spect, positron, performed, scan	<b>Topic 6</b> eyes, eye, corneal, lens, ocular, cataract, intraocular, visual, glaucoma, anterior
<b>Topic 7</b> pregnancy, women, pregnant, abortion, postpartum, pregnancies, fertility, birth, reproductive, maternal	<b>Topic 8</b> methylation, histone, chromatin, dna, epigenetic, cpg, gene, promoter, acetylation, h3	<b>Topic 9</b> tumor, tumors, malignant, carcinoma, metastatic, metastasis, metastases, carcinomas, primary, cancer

#### 5.4.2.2 Random Indexing

I used the Semantic Vectors package to generate the term-document RI model (Semantic Vectors Package, 2012; Widdows & Cohen, 2010; Widdows & Ferraro, 2008). I set the dimensionality for each vector to 500, the number of seeds to 20, and I utilized the stop word list generated by Salton and Buckley for the SMART IR system. Table 5.6 presents

several terms (loosely corresponding to the topics in Table 5.5) and the most related terms for each.

Table 5.6

*Top ten most related terms from RI*

<b>Liver</b> hepatic, livers, cirrhosis, tartarcontrol, fauci, aquaregia, hepatocyte, nonutilitarian, highld, virusmoloney	<b>Plant</b> plants, lymphotic, variable, Arabidopsis, clcells, doxcontaining, fibrillates, demethylate, aiken, mlkgday	<b>Acid</b> amino, acids, fatty, arachidonic, ascorbic, nonane, nonsimilar, sivinfected, noetia. solvolyses
<b>Driving</b> drivers, city, coutilized, homethanol, shouldered, nikolaus, preligand, amulv, 43oxosteroid, vkdependent	<b>PET</b> positron, fdg, disagreeableness, emission, ffdg, alpha latd, mfms, antilesion, dbcampstimulated, tillaux	<b>Eyes</b> intraocular, acuity, pacer, eye, macular, vitrectomy, fellow, detachment, choroidal, 18mers
<b>Pregnancy</b> pregnant, pregnancies, maternal, women, ironedoped, trimester, glucuronidates, opc21268, endosulfani, xeliri	<b>Methylation</b> mythylated, cpg, methylationspecific, epigenetic, hypermethylation, demethylating, 5aza2, icnp, ddstata, pupexposed	<b>Tumor</b> tumors, cancer, transection, carcinoma, metastasis, cells, leukaphereses, epitheliumbruch, prognostic, fosaprepitant

## 5.5 Information scent experiments

This section includes the results for the parameter fitting experiments and the results for the evaluation of information scent using the extracted pairwise judgments. Section 5.5.1 presents the results for the parameter fitting experiments. Section 5.5.2 presents the results for evaluation of information scent.

### 5.5.1 Parameter fitting experiments

This section presents the parameter fitting experiments on the training data. Figure 5.9 presents the performance results for the language model (henceforth denoted as LM) with

different settings for the Dirichlet smoothing in the range [10,5000]. I attained the best performance with a Dirichlet smoothing value of 10 with 65.11% accuracy.

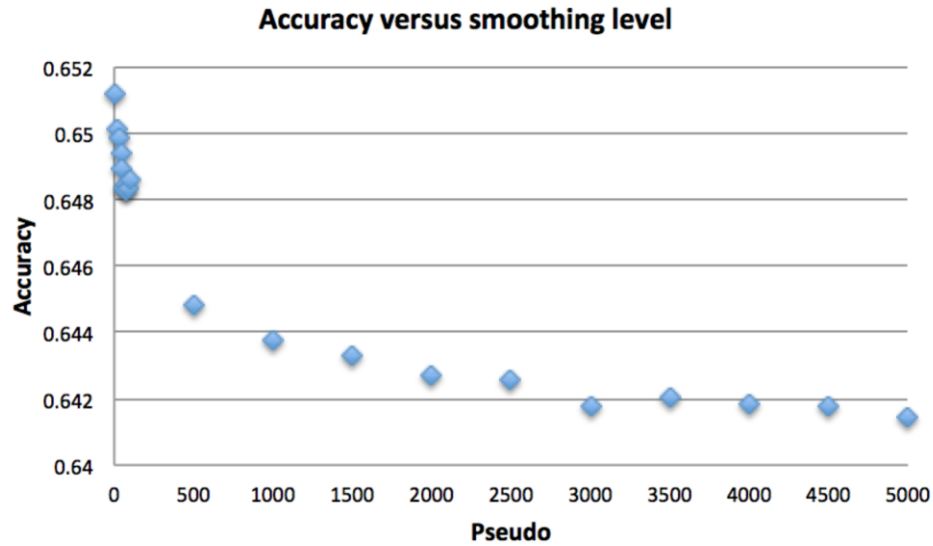


Figure 5.9. Smoothing level for language model with exact matching

Figure 5.10 presents the parameter fitting results for the language model with the topic model used for partial matching (henceforth denoted as LM\_TM). I tested the model in the range [10,5000]. I attained the best performance with a smoothing value of 10 with an accuracy of 66.49%.

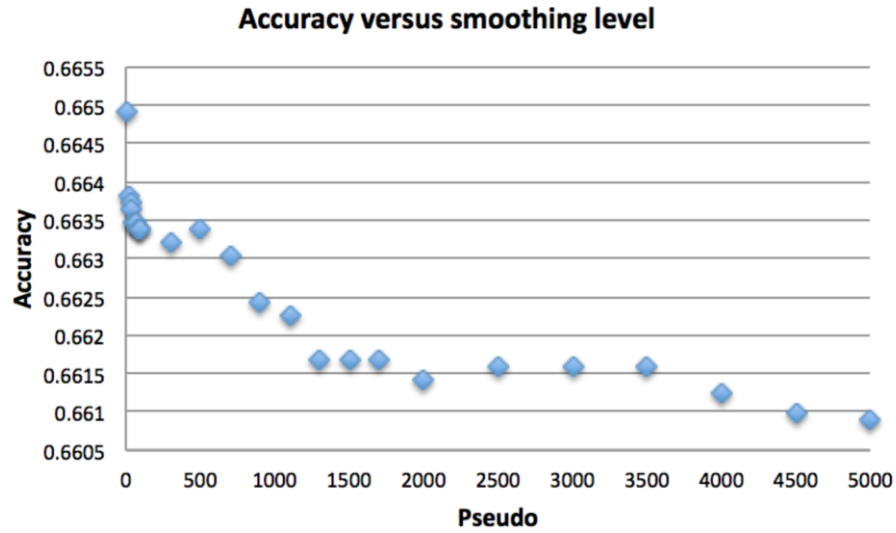


Figure 5.10. Smoothing level for language model with partial matching using a topic model

Figure 5.11 presents the result of the parameter fitting experiments for the language model with the RI model used for partial matching (henceforth denoted as LM\_RI). I tested the model in the range [10,5000]. I attained the best performance with a smoothing value of 10 with accuracy 65.87%.

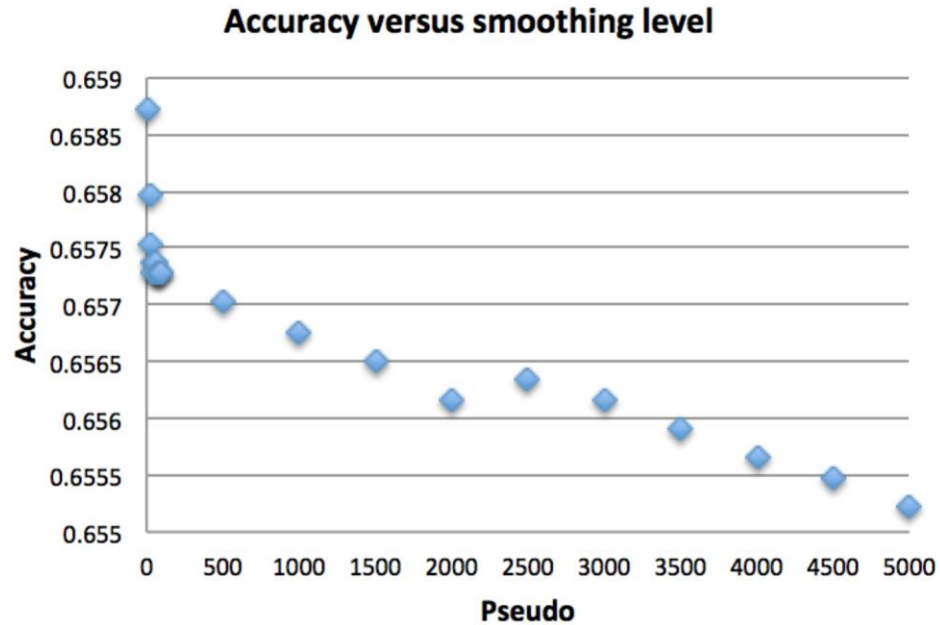


Figure 5.11. Smoothing level for language model with partial matching using RI

Figure 5.12 presents the results for the ACT-R model with the topic model used for partial matching (henceforth known as ACT-R\_TM). I tested the match penalty in the range [0.0,1.0]. I attained the best performance when the best performance when the match penalty had a value of 0.7 with an accuracy of 66.14%.

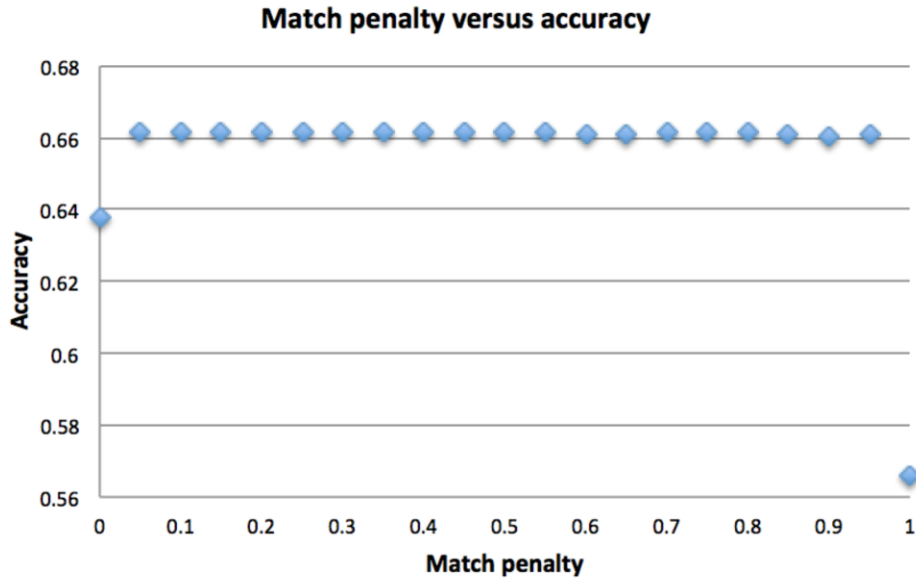


Figure 5.12. Match penalty for ACT-R using a topic model

Figure 5.13 presents the results for the ACT-R model with the RI used for partial matching (henceforth known as ACT-R\_RI). I tested the match penalty in the range [0.01,0.5]. I achieved the best performance when the match penalty had a value of 0.55 with an accuracy of 64.07%.

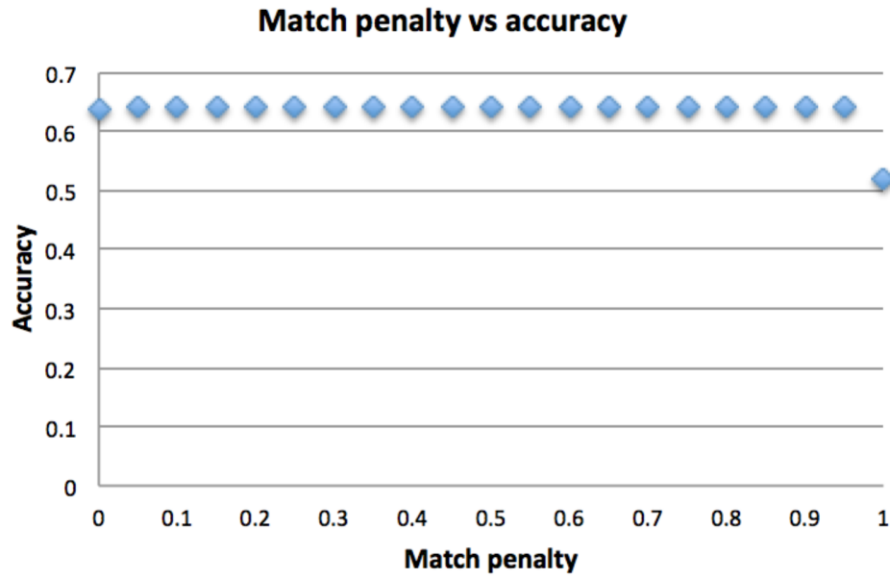


Figure 5.13. Match penalty for ACT-R using RI

## 5.5.2 Results

The experiments are divided according to the pairwise judgments used for evaluation. Section 5.5.2.1 presents the evaluation of all of the models using the document click pairwise judgments. Section 5.5.2.2 presents the evaluation of all of the models using the document download pairwise judgments.

### 5.5.2.1 Results for document click pairwise judgments

Table 5.7 presents the results for the information scent models evaluated using the document click pairwise judgments. The best performing model was LM\_TM. However, this model achieved a marginal improvement over the LM model. The LM model achieved a performance improvement of 2.89% over the ACT-R model. However, this performance increase was not statistically significant ( $p > 0.05$ ).

Table 5.7

*Results on test data for all pairwise judgments*

<b>Model</b>	<b>Accuracy</b>
ACT-R	65.16%
ACT-R_RI	67.65%
ACT-R_TM	67.66%
LM	68.05%
LM_RI	68.14%
LM_TM	<b>68.61%</b>

### 5.5.2.2 Results for document download pairwise judgments

Table 5.8 presents the results for all of the models for the document download pairwise judgments. The LM\_TM model achieved the best performance. The LM\_TM model achieved a 2.26% performance increase over the LM baseline, which was not statistically significant. However, the LM\_TM model achieved a 5.35% performance increase over the ACT-R baseline. A t-test found that this performance improvement for both the LM\_RI and LM\_TM models was statistically significant over the ACT-R model ( $p < 0.05$ ).

Table 5.8

*Results on test set for pairwise judgments extracted for document downloads*

<b>Model</b>	<b>Accuracy</b>
ACT-R	67.83%
ACT-R_RI	69.46%
ACT-R_TM	69.50%
LM	70.92%
LM_RI	72.92%
LM_TM	<b>73.18%</b>

## 5.6 Discussion



This chapter presented the first exploration of the Information Foraging Theory for predicting biomedical document accesses. This chapter had several goals. The first goal was to demonstrate that information scent could be used to predict document accesses in the biomedical domain. The top performance for document click pairwise judgments was the LM\_TM model with an accuracy of 68.14%. However, the performance increase was not statistically significant as compared with the performance of the LM and ACT-R models, which achieved an accuracy of 68.05% and 65.16% respectively. I performed the second experiment on the document download pairwise judgment data set. In this experiment, the best performing model was the LM\_TM model, which achieved an accuracy of 73.18%. In this instance, the model achieved statistically significant performance improvement over the ACT-R model, which achieved an accuracy of 67.83%. In summary, these results support the hypothesis that information scent can be used for predicting document accesses in the biomedical domain.

The second goal of this work was to propose a model that is more closely aligned with the Bayesian theory upon which the Information Foraging Theory relies. The ACT-R model, which has been leveraged in several implementations of the Information Foraging Theory, made a simplifying assumption that reduced the Bayesian mathematical theory to a PMI equivalence. Furthermore, in implementation, the model was further reduced to the product of IDF scores. IDF is generally regarded as a heuristic measure and a probabilistic interpretation of IDF is often debated, but is uncertain at best (Aizawa, 2003; Hiemstra, 2000c; Papieni, 2001; S. Robertson, 2004; Siegler & Witbrock, 1999). Additionally, the model did not include a *tf* component, which has a long history of use in IR (K. S. Jones, 1972). It is important to note that in this particular instance, where the text is a very short

title, the lack of a *tf* component will not be as significant since titles are not likely to contain many duplicate terms. In this corpus, the average length of the titles is 8.87 terms (ignoring stop terms) and only 15.0% of the documents had duplicate terms in the title. However, over longer texts, the lack of a *tf* component will probably become more evident.

The language model framework presented in this chapter enables computing the likelihood scores directly without having to rely on simplifying heuristics. At worst, the updated model was comparable to the ACT-R performance. In fact, the model generally outperformed the ACT-R model, but the performance improvement was not statistically significant in many of the cases. When looking at document download pairwise judgment data, the language model had statistically significant performance improvement over the ACT-R model ( $p < 0.05$ ).

An obvious weakness of this current chapter is that I did not compare the information scent models to traditional IR models. That is, this chapter showed that the language model interpretation of information scent had comparable performance to the ACT-R model (in some cases improved performance), but in and of itself these results do not clearly indicate that information scent is superior to traditional IR models. In the next chapter, I explore this question. In addition, I explore the full model, which takes into account the prior probability model discussed in detail in Chapter 4.

## **Chapter 6: Predicting Document Clicks Using Information Scent and Desirability**

The research presented in this chapter is the culmination of the research presented previously in Chapter 4 and Chapter 5. The specific goal of this chapter is to evaluate the combination of the desirability and information scent models. The function used for predicting document accesses in this chapter is presented in Equation 6.1<sup>7</sup>. This Bayesian function has been proposed by previous researchers for use in IR systems (Hiemstra & Kraaij, 1998; D. H. Miller, et al., 1999), but it is the instantiation of the parameters of this function that makes this work unique. The prior probability  $P(D)$  corresponds to the work presented in Chapter 4. Chapter 4 was motivated by research in cognitive science that showed the prior probability of a memory being retrieved can be calculated based on the recency and frequency of past accesses. The cognitive science literature refers to this property as the recency-frequency effect. For the purpose of document ranking this metric is known as desirability. In Chapter 4, I verified that the recency-frequency effect was present for documents accessed through the PubMed and PLOS search engines. Finally, I showed that desirability could be used to predict documents on which a user will click. The likelihood  $P(Q|D)$  corresponds to the work presented in Chapter 5. In Chapter 5, I introduced a novel instantiation of the information scent calculation based on recent insights from language models.

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)} \propto P(Q|D)P(D) \quad (6.1)$$

---

<sup>7</sup>  $P(Q)$  is assumed to be uniform.

The remainder of this chapter is organized as follows. Section 6.1 presents an overview of the work presented in Chapter 4 and Chapter 5. The purpose of this review is to refresh the reader on the results of the previous chapters. Section 6.2 presents a discussion of the motivation for the experiments in this chapter. Section 6.3 presents the methods used in this chapter. Section 6.4 presents the results of the experiments. Finally, section 6.5 presents the discussion of the results.

## **6.1 Review of Desirability and Information Scent**

This section contains a brief overview of the desirability and information scent studies conducted in Chapter 4 and Chapter 5. The goal of this section is to present the key results to contextualize the work in this chapter. Section 6.1.1 presents an overview of desirability. Section 6.1.2 presents an overview of information scent.

### **6.1.1 Review of desirability**

Quentin Burrell first introduced the notion of desirability and defined it for his particular use case as “the average number of times an item is borrowed per unit time” (Burrell, 1980, 1985; Burrell & Cane, 1982; Burrell & Fenton, 1994). Burrell used a desirability function based on the frequency of past circulation to predict how likely a book was to be borrowed in the near future.

Anderson & Schooler were interested in a similar proposition for human memory (J. R. Anderson & Schooler, 1991). That is, is it possible to create a desirability model for human memory? Anderson & Schooler (J. R. Anderson & Schooler, 1991) investigated the statistical regularities of information in different environments. Specifically they looked at how past frequency (number of times an item appeared in the past) and recency (how recently a given item last appeared) influenced the probability that the item would appear

in the future. This is known as the recency-frequency effect. In all of the situations that they investigated, the probability of an item appearing in the future has a power law relationship with the past recency and frequency of appearance. Based on the results of the analysis, Anderson & Schooler developed a desirability model based on the recency-frequency effect that predicts the probability of a memory item being needed in the future. Chapter 4 presented an in-depth investigation of desirability. Section 4.1 investigated the question of why the recency-frequency effect exists in such a wide variety of different environments. Section 4.1 explored the idea that the recency-frequency effect is an artifact of scale-free network growth. I generated the initial hypothesis from the observation that the recency-frequency effect coexisted in data sets that numerous studies characterized as scale-free networks. To test the hypothesis, I generated numerous networks using network growth models that are known to yield networks with certain statistical properties. I performed experiments on the generated data from each network to determine the presence of the recency-frequency effect. I found that the preferential attachment growth rule was the only one of the tested growth rules tested that exhibited the recency-frequency effect. This offers a potential mechanistic explanation for why Anderson & Schooler observed the recency-frequency effect in a wide variety of different domains.

In Section 4.2, I investigated whether the recency-frequency effect exists for document accesses for two different populations. The first data set was comprised of documents accessed using the PubMed IR system from the users of the Houston Academy of Medicine Texas Medical Center (HAM-TMC) library. The second data set was comprised of documents accessed through the Public Library of Science (PLOS) website. In these experiments, I found that the recency-frequency effect was present in both data sets.

The research presented in Section 4.3 is the most relevant for this chapter. In this section, I evaluated using desirability computed from document accesses from multiple crowd-sourced data sources for predicting document accesses. An in-depth description of all of the data sets is presented in Section 4.3.1.1. The data sets used for calculating desirability in Section 4.3.1.1 were CiteULike, HAM-TMC, Mendeley, and Scopus. The HAM-TMC data set contained the number of abstract views and document downloads. The CiteULike data set contained the number of users who had a given document saved in their reading list. The Mendeley data set contained the number of users that had a given document in their personal library. The Scopus data set contained the number of citations for a given document.

In Section 4.3.3, I presented an in-depth investigation of these data sets and evaluated them in numerous combinations. The desirability function in Equation 6.1 leverages the recency-frequency effect to calculate the prior probability of a given document being accessed. I used Equation 6.1 to calculate the desirability for the documents accessed in the HAM-TMC collection since the date of each access was known. The desirability function in Equation 6.2 assumes that the accesses were uniformly distributed. I used Equation 6.2 to calculate the desirability for the CiteULike, Mendeley, and Scopus data sets since only the frequency was known. The best performance resulted from combining evidence from CiteULike, HAM-TMC, Mendeley, and Scopus data sets.

$$B_i = \log \left( \sum_{i=1}^k t_i^{-d} + \frac{(n-k)(t_n^{1-d} - t_k^{1-d})}{(1-d)(t_n - t_k)} \right) \quad (6.1)$$

$$B_i = \log\left(\frac{n}{1-d} t_n^{-d}\right) \quad (6.2)$$

### 6.1.2 Review of information scent

Information scent is the utility of an information item, which can be thought of as a “rational analysis of categorization of cues according to their expected utility” (P. Pirolli & Card, 1999b). In the case of the Web, cues refer to “World Wide Web links or bibliographic citations, that provide users with concise information about content that is not immediately available” (P. Pirolli & Card, 1999b). According to the Information Foraging Theory, users attend to the cues with the highest expected utility given their information need. For example, consider the search results of a typical search engine shown in Figure 6.1. According to Information Foraging Theory, the user will select the hyperlink with the highest information scent based on proximal cues such as the Web Page title to maximize the probability of satisfying the information need with the distal information content (e.g., the Web page associated with a hyperlink).

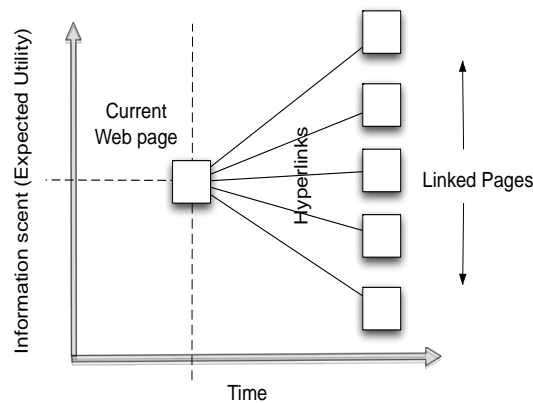


Figure 6.1. Information scent and the WWW. Adapted from (P. Pirolli, 2009)

Chapter 5 had two goals. The first goal was to investigate using information scent for predicting document accesses in the biomedical domain. The second goal was to provide an updated mathematical interpretation of information scent that is more consistent with the Bayesian theory of the ACT-R and the Information Foraging Theory. In implementation, the ACT-R and Information Foraging Theory make assumptions that result in what is essentially a TF-IDF model. In Chapter 5, I presented an update to information scent based on insights from language models, which enabled a probabilistic interpretation that is more consistent with the theoretical foundations of the ACT-R and Information Foraging theories.

For a full discussion on language models see the overview in Chapter 2.3.1. For an in-depth discussion of information scent and the interpretation of information scent based on language models see Chapter 5. Equation 6.3 presents the basic language model based on Dirichlet smoothing (MacKay & Peto, 1995; C. Zhai & Lafferty, 2002). The parameter  $w$  represents an element of the query  $Q$ . For the purpose of information scent,  $D$  represents the proximal cue. The proximal cue that is used in this work is the document title since this information is visible to the user and influences whether or not a document is clicked. Additional information is available to the user such as the journal in which the article is published and the authors of the paper. These additional cues are not investigated in this dissertation and will be the focus of future research. The maximum likelihood estimate in Equation 6.4 calculates the probability  $p(w|D)$  based on the number times  $w$  occurs in the proximal cue  $D$ . Equation 6.5 presents the maximum likelihood estimate for the



background language model. This estimate is based on the frequency of occurrence of  $w$  and the frequency of all terms in the collection  $C$ . The parameter  $\mu$  is the pseudo count parameter, which controls the amount of smoothing.

$$P(w|D) = \frac{P(w|D) + \mu P(w|C)}{|D| + \mu} \quad (6.3)$$

$$P(w|D) \approx P_{ML}(w|D) = \frac{\text{frequency}(w, D)}{\text{length}(D)} \quad (6.4)$$

$$P(w|C) \approx P_{ML}(w|C) = \frac{\text{frequency}(w)}{|C|} \quad (6.5)$$

In Equation 6.6, the language model in Equation 6.3 is updated with evidence from semantic relatedness scores, which enables partial matching. Conceptually, one can view this as combining a score, which reflects the likelihood of an element of the query  $w$  given the proximal cue (e.g., document title) with the likelihood of the neighbors of  $w$  given the proximal cue. Equation 6.7 presents the degree centrality, which is equivalent to the generalized measure for computing degree centrality in weighted networks (Barrat, et al., 2004). The  $P(w|D)$  for the connected term  $v$  is calculated using Equation 6.3.

$$P(w|D) = (1 - \lambda)P(w|D) + \lambda \sum_{v \in V} \frac{w(w, v)}{\text{Deg}(v)} P(v|D) \quad (6.6)$$

$$\text{Deg}(v) = \sum_{u \in V} w(u, v) \quad (6.7)$$

Chapter 5 described experiments with several semantic relatedness models to compute  $w(w, v)$  in Equation 6.6. The best performing model from the experiments utilized Equation 6.6 and used a topic model to compute the semantic relatedness score  $w(w, v)$ . I will refer to this model as LM\_TM. For reference, Table 6.1 presents the results for the LM\_TM model for predicting document clicks and document downloads.

Table 6.1

*Desirability results from combining multiple data sources*

<b>Pairwise judgments</b>	<b>Accuracy</b>
Document clicks	68.14%
Document downloads	73.18%

## 6.2 Motivation and discussion of experiments

The goal of these experiments is to evaluate the combination of the desirability and information scent models, which were presented separately in Chapters 4 and 5 of this dissertation. Additionally, these models are compared to existing state-of-the-art IR models.

The desirability model was discussed in detail in Chapter 4. The desirability model in these experiments used the CiteULike, HAM-TMC, Mendeley, and Scopus data sources. I used the LM\_TM model for calculating information scent, which was presented in Chapter 5. Recall that the desirability model is independent of the query and that the information scent model is dependent on the query. The information scent model ranks the documents using only the information that is visible to the user (i.e., the terms in the title of the document).

For comparison, I present the results for four existing IR models, which were presented previously in Chapter 2. These models include TF-IDF, BM25, divergence from randomness Bose-Einstein (DFR\_BE), and divergence from randomness TF-IDF (DFR\_IDF). I selected what can be considered the state of the art in IR models and used them in the traditional IR context where, at least for document ranking using MEDLINE abstracts, the models matched the query against the title and abstract of the document for ranking. The TF-IDF model is the oldest IR model in these experiments and it originated in the 1970s (K. S. Jones, 1972). Despite its age, TF-IDF remains widely used (Public Websites using Solr). The BM25 and DFR are probabilistic models that emerged in the 1990s (S. E. Robertson, Walker, Beaulieu, Gatford, & Paynet, 1996) and 2000s (Amati & Rijsbergen, 2002) respectively.

## **6.3 Methods and Data Sets**

### **6.3.1 Data sets**

The pairwise judgments used for evaluation were extracted from the query logs of HAM-TMC users, which captured their interactions with the PubMed IR system from October 18, 2012 to November 4, 2012 (19 days). I will refer to this data set as the HAM-TMC pairwise judgment data set. An additional data set was extracted from the HAM-TMC users. This data set contained PubMed accesses for 1,112 days (September 30, 2009 to October 17, 2012). The data set was comprised of 4,513,463 accesses over 2,107,806 documents (abstract views and full document downloads). I will refer to this data set as the

HAM-TMC document access data set. Only the number of document accesses was used from this data set and no pairwise judgments were extracted. In other words, this data set was used only for calculating desirability and not for evaluation.

I used the remaining data sets as additional evidence for calculating desirability. The CiteULike data set contained the number of CiteULike users who have a given document saved in their reading list (CiteULike). I obtained the data from the CiteULike website (<http://www.citeulike.org/>). The Mendeley application allows scientists to manage their reference library, rate articles, and discuss articles (Curran, 2011; Henning & Reichelt, 2008; Zaugg, et al., 2011). The Mendeley data set contained the number of users that have a given document in their personal library. I obtained this data source using the Mendeley API, which allows the download of article metrics such as the number of readers (Mendeley-API, 2013). Scopus is a bibliographic database that contains citations for scientific articles from over 19,000 journals (Archambault, et al., 2009; Burnham, 2006). The Scopus data source contained the number of citations for a given document. I obtained the Scopus citation counts through manual download.

### **6.3.2 Methods**

Figure 6.2 presents an overview of an experiment conducted for one query from the HAM-TMC pairwise judgment data set using the combination of information scent and desirability. Figure 6.3 presents an overview of the experiments conducted for one query from the HAM-TMC pairwise judgment data set that is representative for the TF-IDF, BM25, and DFR models. The experiments in Figure 6.2 and Figure 6.3 both rely upon the HAM-TMC pairwise judgment data set, which contains the query that was issued, the documents clicked by HAM-TMC PubMed users, and the pairwise judgments extracted

for the documents that were clicked. I previously discussed the method for extracting pairwise judgments and using them to evaluate IR algorithms in Section 2.4.2. As a review, I extracted two sets of pairwise judgments for evaluating the models in this work. I extracted the first set of pairwise judgments between documents that were clicked and documents that were not clicked. The goal of this set of pairwise judgments is to evaluate how well a given model can predict the documents that receive clicks and is referred to as the document click pairwise judgments in the remainder of this chapter. The second set of pairwise judgments was extracted between documents that were downloaded and documents that were not downloaded. The goal of this set of pairwise judgments is to evaluate how well a given model can predict document clicks that resulted in a download and is referred to as the document download pairwise judgments in the remainder of this chapter. For all of the approaches, the documents that the HAM-TMC users viewed were assigned a ranking score by one of the models in the experiments. The ranking scores were used to determine how many of the extracted pairwise judgments from the document download pairwise judgment data set or document click pairwise judgment data set were correctly ordered (e.g., documentA is preferred over documentB) based on the ranking scores.

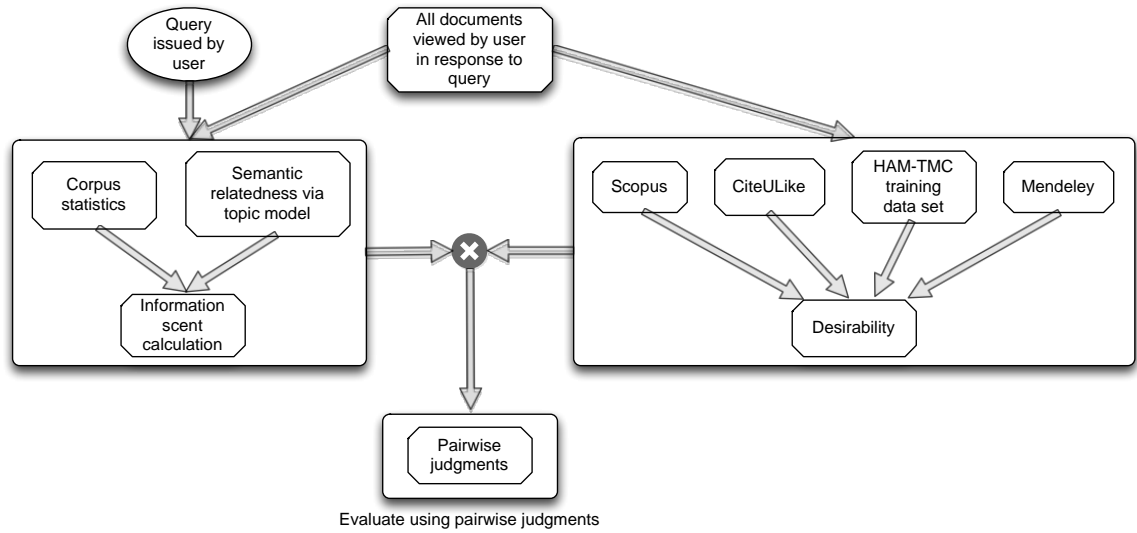


Figure 6.2. Overview of experiments for the combination of the information scent and desirability models

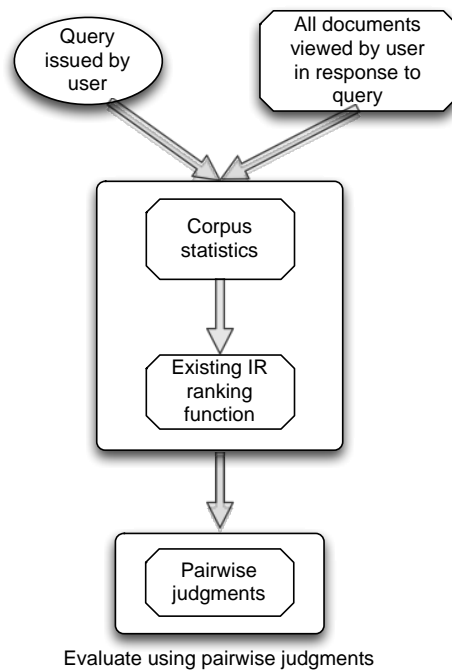


Figure 6.3. Overview of experiments for the existing IR models

I used a sliding window for evaluating all of the models in this work. In each window, I divided the data into training and test sets. The training set contained the data used to calculate the click prediction scores, which were subsequently evaluated using the pairwise judgments extracted from the test sets. Each test set was comprised of one of the days from the HAM-TMC pairwise judgment data set. If the test window was day  $N$  then the training set was comprised of data on day  $N - 1$  or earlier. Figure 6.4 presents an example of how the sliding window could be used to evaluate the combination of the information scent and desirability models over each day in the HAM-TMC pairwise judgment data set.

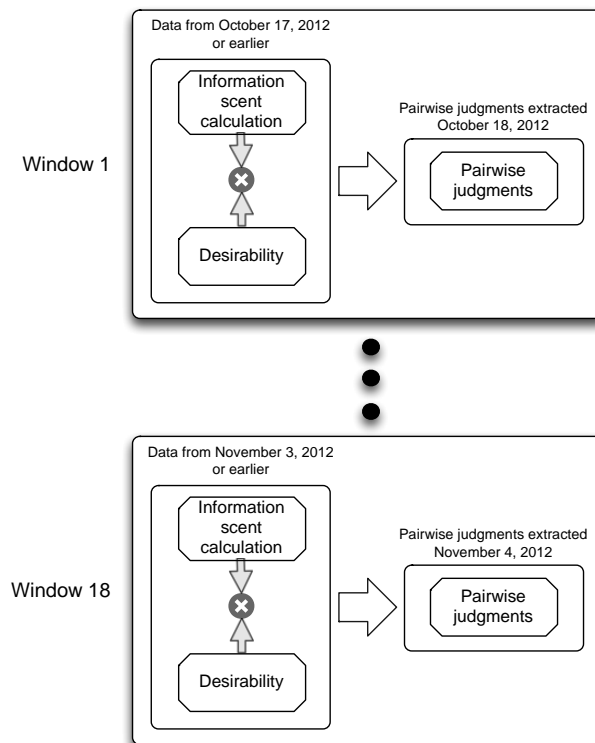


Figure 6.4. Example of sliding window evaluation for the combination of information scent and desirability

The algorithms in these experiments used different data from the training set to rank the documents in the test set. For desirability, the training set for each iteration was composed of document access information from Mendeley, Scopus, HAM-TMC document access data set, and CiteULike. All of the IR models including information scent, TF-IDF, BM25, etc. required some form of corpus statistics such as IDF. For the IR models, the training window was used to compute the corpus statistics. The corpus statistics for all of the models was computed using the titles and abstracts from the PubMed corpus. I utilized the stop word list generated by Salton and Buckley for the SMART information retrieval system for calculating the corpus statistics (Salton, 1971).

The information scent model required the use of a semantic relatedness measure that was not required for the TF-IDF, BM25, and DFR models. For the topic model used in this work, I excluded terms that occurred less than 10 times in the entire corpus. I used the MALLET package to generate the LDA topic model (McCallum, 2002). I selected 500 topics for this model. The MALLET package infers the topic distribution for new pieces of text using Gibb's sampling.

The BM25 and the LM\_TM model had free parameters that required tuning for a particular data set. To tune the parameters, I divided the HAM-TMC pairwise judgment data set into 50% for parameter tuning and 50% for evaluation. This resulted in nine windows for parameter tuning (October 18, 2012-October 26, 2012) experiments and nine windows for evaluation (October 27, 2012-November 4, 2012). Table 6.2 presents an overview of the number of pairwise judgments for the parameter tuning experiments and evaluation.

Table 6.2



#### *Number of pairwise judgments for parameter tuning and evaluation*

<b>Data set</b>	<b>Number of pairwise judgments</b>
Parameter tuning experiments – document click pairwise judgments	85,435
Evaluation – document click pairwise judgments	68,716
Evaluation – document download pairwise judgments	8,435

Section 2.4.2 presented in detail the methods used in this experiment to extract pairwise judgments and use them for evaluation. For reader convenience, Equation 6.8 presents the accuracy metric used to evaluate the performance of the models in these experiments. If a given algorithm resulted in a tie for the two documents in the pairwise judgment, the tie was broken at random.

$$Accuracy = \frac{\text{Number of correct pairwise judgments}}{\text{Correct} + \text{incorrect pairwise judgments}} \quad (6.8)$$

## **6.4 Results**

### **6.4.1 Parameter tuning experiments**

The BM25 model has two free parameters:  $b$  and  $k1$ . The  $b$  parameter controls the document length normalization and the  $k1$  parameter controls the influence of the  $tf$  component. The BM25 model was tested in the range  $[0.5, 1.0]$  with increments of 0.1 for the  $b$  parameter and in the range  $[0.5, 2.5]$  with increments of 0.5 for the  $k1$  parameter. Figure 6.5 presents the results of the parameter tuning experiments. The model obtained the best performance where  $b = 0.6$  and  $k1 = 1.0$  with an accuracy of 62.48%.

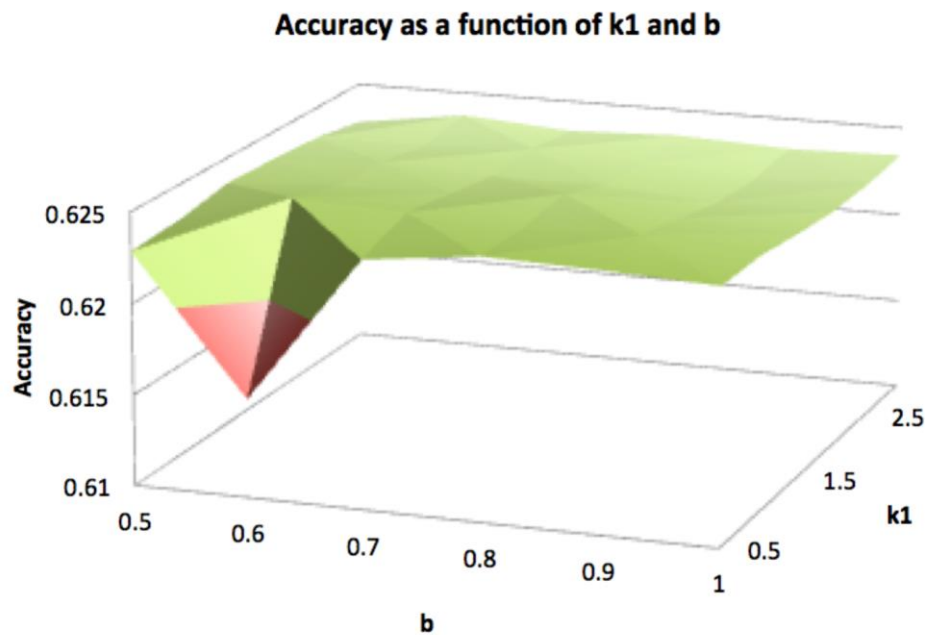


Figure 6.5. Results for parameter tuning for BM25

Figure 6.6 presents the parameter fitting results for the LM\_TM model. I tested the model in the range  $[0, 5000]$ <sup>8</sup>. I attained the best performance with a smoothing value of 10 with an accuracy of 66.49%, but smoothing had a very small impact on performance ( $<0.5\%$ ).

---

<sup>8</sup> The result for smoothing value of 0 is not shown in Figure 6.6. The accuracy at smoothing value 0 was 60.23% and adding the data point to Figure 6.6 rendered the remaining values illegible.

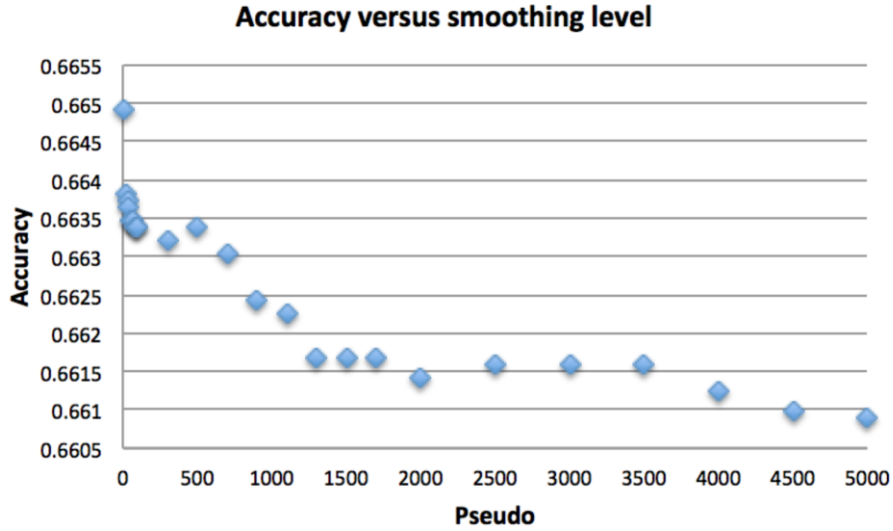


Figure 6.6. Smoothing level for language model with partial matching using a topic model

#### 6.4.2 Results for predicting document clicks

The experiments are divided according to the pairwise judgments used for evaluation. Section 6.4.2.1 presents the evaluation of all of the models using the document click pairwise judgments. Section 6.4.2.2 presents the evaluation of all of the models using the document download pairwise judgments.

##### 6.4.2.1 Results for document click pairwise judgments

Table 6.3 presents the results for all of the models for the document click pairwise judgments. One interesting finding was that, when evaluated separately, the difference in performance of the desirability and information scent models was not statistically significant. That is, the desirability model was able to attain comparable performance to the LM\_TM model based on past accesses alone without considering the query terms. In combination, the LM\_TM and desirability model (henceforth known as LM\_TM +

Desirability) had an accuracy of 74.01%. This is an improvement of 5.87% over the information scent model and 6.47% over the desirability model (t-test;  $p < 0.05$ ). Additionally, the LM\_TM + Desirability model outperformed TF-IDF, BM25, and DFR models. Of the existing IR models used as a base-line, TF-IDF performed the best. The LM\_TM + Desirability model outperformed TF-IDF by 9.81% (t-test;  $p < 0.05$ ).

Table 6.3

*Information scent and desirability results for all pairwise judgments*

<b>Model</b>	<b>Accuracy</b>
LM_TM	68.14%
Desirability	67.70%
LM_TM + Desirability	<b>74.01%</b>
TF-IDF	64.20%
BM25	63.57%
DFR_BE	63.90%
DFR_IDF	63.90%

#### **6.4.2.2 Results for document download pairwise judgments**

Table 6.4 presents the results for all of the models for the document download pairwise judgments. In this experiment, the desirability model showed significant performance degradation for predicting downloads. The results show that the LM\_TM model outperformed the desirability model by 8.92% (t-test;  $p < 0.05$ ). The LM\_TM + Desirability model resulted in a small 0.95% performance increase over the LM\_TM model (t-test;  $p > 0.05$ ). Of the existing IR models used as a base-line, TF-IDF performed the best. The LM\_TM + Desirability model attained a 6.9% performance improvement over TF-IDF (t-test;  $p < 0.05$ ).

Table 6.4

*Information scent and desirability results for predicting downloads*

<b>Model</b>	<b>Accuracy</b>
LM_TM	73.18%
Desirability	64.26%
LM_TM + Desirability	<b>74.13%</b>
TF-IDF	67.23%
BM25	66.69%
DFR_BE	66.53%
DFR_IDF	66.47%

## 6.5 Discussion

To summarize the results of this chapter, for document clicks, the LM\_TM + Desirability model improved performance over either LM\_TM (+5.86%) or desirability (+6.47%) alone. For downloads, the performance gain from LM\_TM + Desirability was small (+0.95% for LM\_TM). Holistically, when comparing to LM\_TM or desirability alone, the LM\_TM + Desirability model resulted in higher accuracy (5.87%) for document clicks while retaining comparable performances for downloads (74.13%). The performance of the LM\_TM + Desirability model was particularly dramatic when compared to the existing IR models. Table 6.5 summarizes the performance improvement of the LM\_TM + Desirability model as compared to the existing IR models. The TF-IDF model had the best performance of the existing IR models. The LM\_TM + Desirability improved performance by 9.81% for document clicks and 6.9% for document downloads.

Table 6.5

*Summary of performance increase of LM\_TM + Desirability compared to existing IR models*

<b>Algorithm</b>	<b>Performance increase by LM_TM + Desirability for document clicks</b>	<b>Performance increase by LM_TM + Desirability document downloads</b>
TF-IDF	9.81%	6.90%
BM25	10.44%	7.44%
DFR_BE	10.11%	7.60%
DFR_IDF	10.11%	7.66%

The contributions of this chapter cannot be isolated from those of Chapters 4 and 5. Previous research in information foraging theory used small user populations (Chi, Pirolli, Chen, et al., 2001; Chi, Pirolli, & Pitkow, 2001) and were conducted entirely outside of the biomedical domain. Additionally, the numerous studies conducted by Pirolli were investigated from a usability viewpoint. For example, these studies focused primarily on how well information scent could predict user browsing behavior in small laboratory experiments. These studies did not investigate if insights from ACT-R and the Information Foraging Theory could ultimately result in improved IR systems, which was the focus of this chapter.

All of the previous experiments by Pirolli assumed that the documents had a uniform prior probability of being accessed, which essentially ignored this component of the ACT-R model. Chapter 4 investigated desirability in-depth and showed that this property was present for documents accessed through PubMed. This chapter demonstrated that desirability has utility for predicting document accesses. In fact, the performance of the information scent model and the desirability model was not statistically significant for predicting document clicks. This finding shows that both models performed equally well

for predicting document clicks despite the desirability model being independent of the query. More importantly, these experiments demonstrated that desirability is an important component, which can improve IR performance. Thus, the prior probability of a document being accessed should not be assumed to be uniform.

The final contribution of this chapter is that it combined the information scent model and desirability model and showed that together these components improved performance over either component alone. Once again, this is the first study to investigate these components together for predicting document accesses. In addition, I showed that the combination of these components greatly outperformed the existing IR models (9.81% improvement for all document clicks and 6.9% improvement for document downloads).

An additional result of note from these experiments was that the LM\_TM model outperformed all of the existing IR models. A summary the performance gains of LM\_TM over the existing IR models is show in Table 6.6. For the document click data set the performance improvement was statistically significant ( $p < 0.05$ ) for BM25, DFR\_BE, and DFR\_IDF. For the document downloads all of the performance gains were statistically significant. These results are interesting as the LM\_TM model relies only upon information that the user can see, which was the document title in this case. The existing IR models rely upon the document title and abstract text. This is a somewhat counterintuitive result. Conventional wisdom would assert that the abstract text would provide a better representation of the document than the title alone. However, users are influenced only by what is visible on their screen and including the abstract text may not help if the goal is to predict document clicks or to model user information seeking behavior. From the results presented here I cannot make that claim. This would require a different experiment that

compared the click accuracy for each model using the abstract and title text or just using the title text. What can be asserted from this study is that using only the title resulted in a competitive model.

Table 6.6

*Summary of performance increase of LM\_TM compared to existing IR models*

<b>Algorithm</b>	<b>Performance increase by LM_TM for document clicks</b>	<b>Performance increase by LM_TM document downloads</b>
TF-IDF	3.94%	5.95%
BM25	4.57%	6.49%
DFR_BE	4.24%	6.65%
DFR_IDF	4.24%	6.71%

An additional interesting result from these experiments is that all of the existing IR models performed essentially the same (i.e. none of their results were statistically significant from one another). This is in contrast to much of the published literature. In fact, the literature is full of examples where these models have been claimed to outperform one another on various data sets. The results of a brief literature review is shown in Table 6.7. The results of the literature review highlight the contradictory findings that are prevalent in the literature. The results of this chapter showed that the performance results for BM25, DFR, and TF-IDF are essentially the same when compared to the preferences of the users. Interestingly, in this study, TF-IDF performed the best overall for the existing IR models, which is the defacto straw man for any new method. These findings, though alarming, is not entirely surprising. One of the drawbacks of Cranfield inspired experiments is that numerous studies have shown that the performance gains of IR systems using this protocol



do not necessarily translate to real-world user satisfaction (Al-Maskari, et al., 2008; Allan, et al., 2005; W. Hersh, et al., 2001; Jarvelin, 2009; Macdonald & Ounis, 2009; Sanderson, et al., 2010; Smith & Kantor, 2008; Smucker & Jethani, 2010; Su, 1992; Turpin & Scholer, 2001, 2006; Urbano, et al., 2012). It is quite possible that the lack of performance differences between BM25, TF-IDF, and DFR is that these models were evaluated based on preferences of real-world users and the performance gains in laboratory experiments often vanish in this scenario.

Table 6.7

*Summary of findings for different studies*

<b>Study</b>	<b>Finding</b>
(Zhao, Huang, Ye, & Zhu, 2009)	BM25 outperformed DFR
(Amati, 2003)	DFR outperformed BM25
(Kraaij, 2004; Trotman, Puurula, & Burgess, 2014; Urbain, Goharian, & Frieder, 2005)	BM25 outperformed language models
(Zhu, Song, & Ruger, 2009)	Language models outperformed BM25
(Bache, 2011)	BM25 outperformed TF-IDF
(de Almeida, Goncalves, Cristo, & Calado, 2007)	TF-IDF outperformed BM25
(Ye, He, Huang, & Lin, 2010)	Language models outperformed DFR

## Chapter 7: Conclusion and Future Research

This dissertation has presented numerous experiments to advance computational cognitive modeling applied to IR. The goal of this chapter is to summarize the research conducted in this dissertation and to discuss areas for future research. This chapter is organized as follows. Section 7.1 presents a summary and discussion of the work in this dissertation. Section 7.2 outlines areas for future research on the topics presented in this dissertation.

### **7.1 Summary and Discussion of Research in This Dissertation**

The theme of this dissertation is the application of cognitive science to information retrieval. Specifically, there are two main topics: desirability and information scent, which are both components of the overarching theoretical frameworks of the Information Foraging Theory and ACT-R. Chapter 4 provided an in-depth investigation into desirability. Chapter 5 focused on the information scent calculation. Finally, Chapter 6 unified the research threads in Chapter 4 and Chapter 5 by evaluating the components together. The remainder of this section will summarize and discuss the main results of this dissertation.

In this dissertation, desirability was computed based on the research of Anderson & Schooler (J. R. Anderson & Schooler, 1991) in which they showed that the past frequency (number of times an item appeared in the past) and recency (how recently a given item last appeared) influenced the probability that the item would appear in the future. However, the observations made by Anderson & Schooler lacked a mechanistic theory to explain the underlying phenomena. According to (J. R. Anderson & Milson, 1989), these results provide evidence of a universal law which governs the ebb and flow of information. (J. R. Anderson & Milson, 1989) summarize this hypothesis as follows.

Should we really believe that information retrieval by humans has the same form as library borrowings and file accesses? The fact that two very different systems display the same statistics suggests that there are “universals” of information retrieval that transcend device (library, file system, or human memory) and that these systems all obey the same form but differ only in parameterization.

(J. R. Anderson & Milson, 1989)

In this dissertation, I proposed the hypothesis that the recency-frequency effect is produced by the preferential attachment network growth mechanism which has shown in numerous experiments to give rise to scale free networks (Albert, Jeong, & Barabasi, 2000; Barabasi, 2003, 2005; Barabasi & Albert, 1999; Barabasi, et al., 2002; Dezso, et al., 2006; Jeong, et al., 2000; D. S. Lee, et al., 2008; Oliveira & Barabasi, 2005). As a review, the preferential attachment growth mechanism asserts that the probability of a vertex in a graph receiving a new connection is proportional to its current degree centrality (Barabasi & Albert, 1999). In a series of experiments, I showed that the recency-frequency effect was present only when the preferential attachment growth mechanism was present. This finding offers a potential mechanistic explanation for why Anderson & Schooler observed the recency-frequency effect in a wide variety of different domains.

The remainder of the research on desirability in this dissertation focused on whether the recency-frequency effect was present for document accesses and whether this information could be leveraged to improve click prediction performance. I performed a series of experiments on the HAM-TMC and Scopus datasets and found that the recency-frequency effect was present for both populations.

After verifying that the recency-frequency effect held for document accesses, I performed a series of experiments to determine if the recency-frequency effect could predict document accesses. In these experiments, I calculated desirability on several document access data sets and evaluated the performance of each data set individually and in combination with the other datasets. The most interesting finding from these experiments was that desirability outperformed existing IR models including TF-IDF and two instantiations of divergence from randomness. That is, the query-independent desirability function outperformed widely-used query-dependent ranking approaches that computed similarity between the document and the query. This finding provides strong support that desirability has utility for document ranking.

Chapter 5 focused entirely on information scent. The primary contribution is that it is the first exploration of applying the Information Foraging Theory in the medical domain. The previous applications of the Information Foraging Theory were applied entirely outside of the biomedical domain (Budi, et al., 2009; Card, et al., 2001; Chi, Pirolli, Chen, et al., 2001; Chi, Pirolli, & Pitkow, 2001; Hong, et al., 2008; Huberman, et al., 1998; P. Pirolli, 2005, 2009; P. Pirolli & Card, 1995, 1999b; P. Pirolli & W-T., 2006; P. L. Pirolli & Anderson, 1985; P. L. Pirolli & Pitkow, 2000). Additionally, this chapter presented an updated mathematical framework that was more consistent with the underlying Bayesian theory of ACT-R and the Information Foraging Theory. As discussed in detail in Chapter 5, the actual implementation of ACT-R and Information Foraging Theory made many simplifying assumptions that reduced it to what is essentially a TF-IDF computation. In Chapter 5, I presented a new mathematical interpretation based on recent insights from

statistical language models (C. X. Zhai, 2008) that avoids the simplifications made in previous implementations of information scent.

The purpose of Chapter 6 was to investigate the performance of combining desirability and information scent. The primary contribution of this chapter is that all other applications of information scent assumed a uniform prior probability. In this chapter, I combined information scent with the prior probability estimate discussed in detail in Chapter 4. I found that the combination of information scent improved click prediction by 6.31% over desirability alone and 5.87% over information scent alone. Additionally, the combination outperformed the existing IR models (TF-IDF, divergence from randomness, and BM-25) by over 9.0% in each experiment. These results provide compelling evidence to support the assertion that prior probabilities should not be assumed uniform.

## **7.2 Future Work**

This section outlines potential areas for future research in the topics covered in this dissertation. The remainder of this section is organized as follows. Section 7.2.1 discusses modeling additional features for click prediction. Section 7.2.2 discusses the development of a personal information scent model. Finally, Section 7.2.3 discusses modeling desirability at different levels of granularity.

### **7.2.1 Modeling additional information visible to the user in PubMed search results**

A primary weakness of the information scent experiments in this chapter is that it did not include all of the information visible to a user in the search results that could influence a document access. Figure 7.1 presents an example of the search results from PubMed. In this dissertation, I used only the title of the document for calculating information scent.

However, additional information such as the authors and journal in which the article was published were not used for calculating information scent.


-  [Predicting biomedical document access as a function of past use.](#)
1. Goodwin JC, Johnson TR, Cohen T, Herskovic JR, Bernstam EV.  
J Am Med Inform Assoc. 2012 May-Jun;19(3):473-8. doi: 10.1136/amiajnl-2011-000325.  
Epub 2011 Sep 13.  
PMID: 21917645 [PubMed - indexed for MEDLINE] **Free PMC Article**  
[Related citations](#)

Figure 7.1. Example result from PubMed

The author names and journal names can be modeled in terms of desirability (i.e. query-independent prior probability) and likelihood (query-dependent) components. The desirability for an author corresponds to document access patterns for their authored documents. Similarly, the desirability for the journal corresponds to document access patterns for all of the documents published by a given journal. For each model, experimentation will be required to model the underlying probability distribution. I showed that the recency-frequency effect held for document accesses in general, but it does not necessarily follow that the recency-frequency effect will hold for document clicks for authors or journals. Thus, additional experiments are required to determine the underlying distribution for estimating desirability of authors and journals. Once the underlying distribution is established, the desirability scores for the document, authors, and journal can be integrated using linear integration as shown in Equation 7.1.

$$B_i = \lambda_1 B_{document} + \lambda_2 B_{journal} + \lambda_3 B_{authors} \quad (7.1)$$

The information scent calculation for authors and journals follows from the information scent equations presented in Chapter 5 and Chapter 6 of this dissertation. Instead of calculating the likelihood of the terms in the query given the title of the document, the new components focus on calculating the likelihood of the terms in the query given the authors of the document or the journal in which the article was published.

For brevity, I focus on how the information scent calculation would apply to journals, but the application to authors would be nearly identical. I previously discussed in Chapter 5 and Chapter 6 the equations presented in the remainder of this section as they applied to modeling information scent for document titles. Equation 7.2 presents the information scent model for computing the likelihood of the query given the journal. Equation 7.3 and Equation 7.4 present the maximum likelihood estimate for the document language model and the background language model respectively. Here a journal  $J$  would be represented as all of the terms in the documents which  $J$  published. In other words, a journal is treated as a large document. The maximum likelihood estimate in Equation 7.3 calculates the probability  $p(w|J)$  based on the number times a term  $w$  occurs in the abstracts published by a journal  $J$  and the total number of terms in the abstracts published by  $J$  ( $|J|$ ). The background language model shown in Equation 7.4 is based on the frequency of occurrence of  $w$  and the frequency of all terms in the collection  $C$ . The parameter  $\mu$  is the pseudo count parameter, which controls the amount of smoothing.

$$P(w|J) = \frac{P(w|J) + \mu P(w|C)}{|J| + \mu} \quad (7.2)$$

$$P(w|J) \approx P_{ML}(w|J) = \frac{frequency(w, J)}{|J|} \quad (7.3)$$

$$P(w|J) \approx P_{ML}(w|C) = \frac{frequency(w)}{|C|} \quad (7.4)$$

Next, I describe how partial matching could work within the context of modeling information scent for journals. Equation 7.5 presents the information scent model with partial matching for calculating the likelihood of the query given the journal. Equation 7.5 updates the output of Equation 7.2 with evidence from semantic relatedness scores, which enables partial matching. Conceptually, one can view this as combining a score, which reflects the likelihood of an element of the query  $w$  given the proximal cue (e.g. journal name) with the likelihood of the neighbors of  $w$  given the proximal cue. The semantic relatedness between the term  $w$  and a term  $v$  in the document is represented by  $w(w, v)$ . The semantic relatedness score can be computed using a topic model as was done in Chapter 5 and Chapter 6. Equation 7.6 presents the degree centrality metric used in this work, which is equivalent to the generalized measure for computing degree centrality in weighted networks (Barrat, et al., 2004). The  $P(v|D)$  for the connected term  $v$  is calculated using Equation 7.6.

$$P(w|J) = (1 - \lambda)P(w|J) + \lambda \sum_{v \in V} \frac{w(w, v)}{Deg(v)} P(v|J) \quad (7.5)$$



$$Deg(v) = \sum_{u \in V} w(u, v) \quad (7.6)$$

### 7.2.2 Personalized information scent model

The information scent model in this dissertation uses a “one-size-fits-all” model to predict document accesses. However, this is a simplifying assumption. Numerous studies have shown that individual users have different relevance judgments for the same set of documents returned by a query (Teevan, Dumais, & Horvitz, 2005; White & Drucker, 2007; Wu, Turpin, & Zobel, 2008). For example, (Teevan, et al., 2005) compared the relevance judgments for identical documents returned by identical queries and found a low inter-agreement of 56%. One proposed reason for the low agreement is that queries are often ambiguous. For example, for the query term “cancer”, it was observed by (Teevan, et al., 2005) that some of the users were looking for information about cancer treatments and some users were looking for information about the astrological sign cancer. One method for dealing with ambiguous information needs to develop user models to enable personalized ranking. Towards this aim, I propose a personalized scent model and hypothesize that the “one-size-fits-all” information scent calculation can be improved by incorporating background information about the user who issued the query.

The notion of utilizing user background information in information scent calculation is closely related to personalized ranking and collaborative filtering. That is, each requires some notion of a user profile. One method for constructing a user profile is to have the user manually express their interests (Google Personal, 2013). Another method is to have users provide feedback or rate items to generate a profile. An example of this approach is the

Netflix movie recommendation engine, which relies, in part, upon user feedback when making personalized movie recommendations (Bennet & Lanning, 2007). A drawback to methods that require humans to either manually create a profile or provide explicit feedback is that users are reluctant to invest the time, which severely limits the accuracy of these methods (Bennet & Lanning, 2007). Given the reluctance of users to provide feedback, researchers have focused on automatically constructing user profiles based on implicit feedback (Dou, Song, & Wen, 2007; Matthijs & Radlinski, 2011; Shen, Tan, & Zhai, 2005). In this proposed model, I would automatically construct the user profile based on the past accesses of the user.

The proposed personal information scent model, like the standard information scent model, computes the likelihood of a document access based, in part, on the terms in the query and the terms in the title. In addition, the personal scent model includes evidence from the past document accesses of the user. That is, when the user issues the query, the documents with information scent values above a fixed threshold<sup>9</sup> are selected and used to smooth the likelihood score from the original information scent score that is based on the document title and query. Figure 7.2 presents an overview of the processing involved in the personal scent model.

---

<sup>9</sup> Proper threshold will have to be determined experimentally.

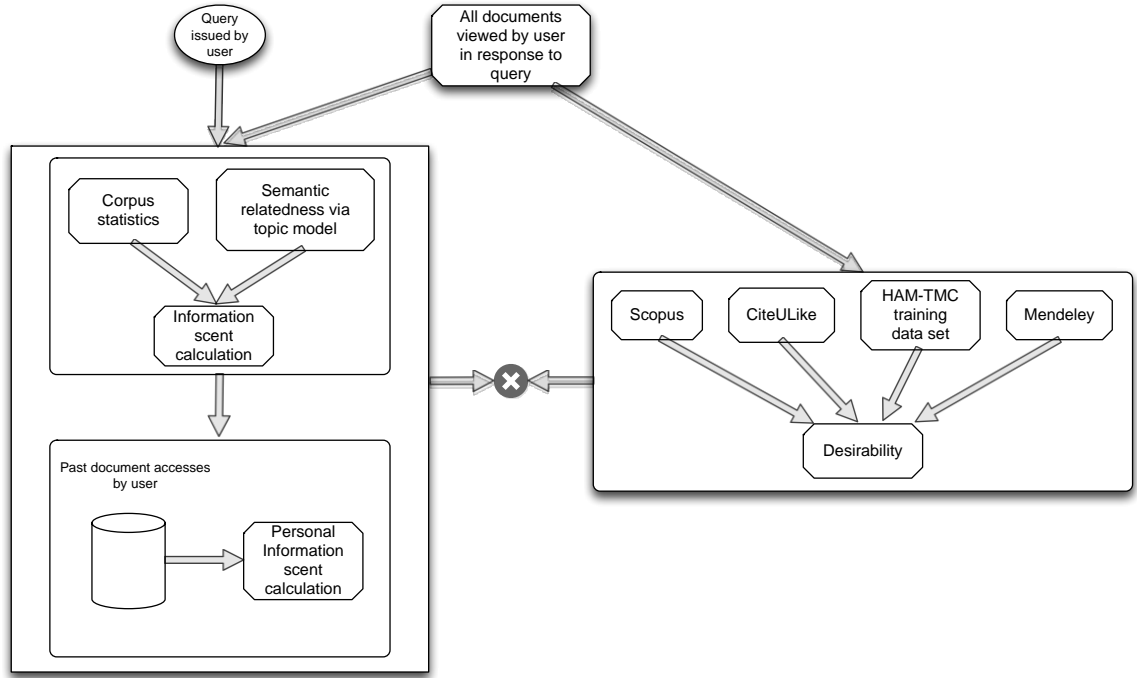


Figure 7.2. Personal information scent model

The remainder of this section will describe the mathematical framework of the personal information scent model, which is based on the general optimization framework for smoothing language models presented in (Mei, et al., 2008). For reader convenience, Equation 7.7 repeats the information scent model discussed in-depth in Chapter 5 and Chapter 6. Equation 7.7 calculates the information scent based on the terms in the query and the terms in the title of the document.

$$P(w|D) = (1 - \lambda)\dot{P}(w|D) + \lambda \sum_{v \in V} \frac{w(w, v)}{Deg(v)} P(v|D) \quad (7.7)$$

After applying Equation 7.7, the score is then smoothed using the past access of the users. Additionally, Equation 7.7 is applied to the past accesses of the user and the documents above a set threshold are selected for smoothing. The intended effect of smoothing (within the framework of language models) in this case is to essentially re-rank the documents based on the past accesses of a given user. I propose using Equation 7.8 for smoothing the likelihood score from Equation 7.7 based on the evidence from past accesses. In Equation 7.8, the likelihood score from Equation 7.7 is represented as  $\bar{P}(q|d_u)$ . The past documents accessed by the user that are similar to the query are represented by  $V$  and the likelihood score is represented by  $P(q|d_v)$ . The relatedness between a document in the result set ( $u$ ) and a document that a user has previously accessed ( $v$ ) is represented as  $w(u, v)$ . The semantic relatedness score  $w(u, v)$  can be computed using a variety of methods such as LDA (Blei & Lafferty, 2007). Equation 7.9 presents the degree centrality metric used in this work, which is equivalent to the generalized measure for computing degree centrality in weighted networks (Barrat, et al., 2004).

$$s(q, d_u) = (1 - \lambda)\bar{P}(q|d_u) + \lambda \sum_{v \in V} \frac{w(u, v)}{Deg(u)} P(q|d_v) \quad (7.8)$$

$$Deg(v) = \sum_{u \in V} w(u, v) \quad (7.9)$$

### **7.2.3 Desirability at multiple levels of granularity**

In this work, I modeled desirability using aggregate document accesses from multiple institutions. However, from this dataset alone, it is possible to model desirability at multiple levels of granularity with supplementation from institutional records. Figure 7.3 presents an example of potential levels where desirability could be modeled. The motivation behind modeling desirability at multiple levels of granularity is that it could theoretically provide prior probability estimates that are in higher agreement with the user. For example, interest in electronic health records could vary between the medical school and an informatics school. That is, the informatics school may be more interested in research in the latest scientific advancements in electronic health records whereas the medical school may be more interested in case studies that discuss the outcome of installing electronic health records in hospitals. The evidence from these multiple desirability scores could be combined using linear integration as shown in Equation 7.10.

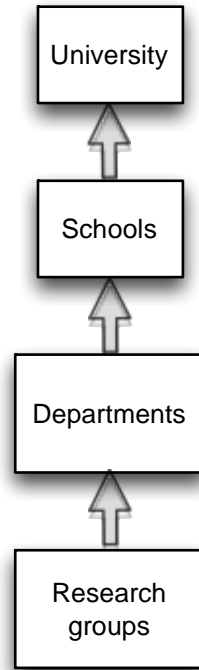


Figure 7.3. Potential levels for modeling desirability

$$B_i = \lambda_1 B_{University} + \lambda_2 B_{School} + \lambda_3 B_{Department} + \lambda_4 B_{Research\ group} \quad (7.10)$$

An obvious weakness of this approach is that modeling desirability at finer levels of granularity will result in a data sparsity problem. At the university level, desirability is computed using document accesses from thousands of individuals. At the research group level, there is generally less than one hundred people and frequently fewer than a dozen people. Overcoming this limitation would require supplementation from multiple institutions whereby similar research groups and similar departments are used in calculating desirability.

## REFERENCES

- Aerts, D., Aerts, S., & Gabora, L. (2009). *Experimental evidence for quantum structure in cognition*. Paper presented at the In Quantum Interaction.
- Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1), 45-65.
- Al-Maskari, A., Sanderson, M., Clough, P., & Airio, E. (2008). *The good and the bad system: does the test collection predict users' effectiveness?* Paper presented at the ACM SIGIR.
- Albert, R., Jeong, H., & Barabasi, A. L. (2000). Error and attack tolerance of complex networks. *Nature*, 406, 378-382.
- Allan, J., Carterette, B., & Lewis, J. (2005). *When will information retrieval be "good enough"*. Paper presented at the ACM SIGIR.
- Amati, G. (2003). *Probability models for information retrieval based on divergence from randomness*. PhD, University of Glasgow.
- Amati, G., & Rijsbergen, C. J. V. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM transactions on information systems*, 20(4), 357-389.
- Amati, G., & van Rijsbergen, C. (2002). Probabilistic modes of information retrieval based on measuring the divergence from randomness. *ACM transactions on information systems*, 20(4), 357-389.
- Aminul, I., & Inkpen, D. (2006). *Second order co-occurrence PMI for determining the semantic similarity of words*. Paper presented at the Proceedings of the International Conference on Language Resources and Evaluation, Genoa, Italy.

- Anderson, J. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press.
- Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*.
- Anderson, J. R. (1989). *The place of cognitive architectures in a rational analysis*. Paper presented at the 22nd Annu. Symp. Cognit., Carnegie-Mellon University.
- Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14, 411-517.
- Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Washington, DC: Winston & Sons.
- Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1120-1136.
- Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96, 703-719.
- Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology*, 128, 186-197.
- Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, 30(3), 221-256.
- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*(2), 396-408.



- Archambault, E., Campbell, D., Gingras, Y., & Lariviere, V. (2009). Comparing of science bibliometric statistics obtained from the web and Scopus. *J. Am. Soc. Inf. Sci.*, 60, 1320-1326.
- Bache, R. (2011). Measuring and improving access to the corpus *Current Challenges in Patent Information Retrieval* (pp. 147-165): Springer Berline Heidelberg.
- Barabasi, A. L. (2003). *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*: Basic Books.
- Barabasi, A. L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435, 207-211.
- Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509-512.
- Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, 311, 590-614.
- Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of The National Academy of Sciences*, 101(11), 3747-3752.
- Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An open source software for exploring and manipulating networks*. Paper presented at the In Proc. 3rd International AAAI Conference on Weblogs and Social Media.
- Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interfaces. *Online Review*, 13(5), 407-424.

- Belew, R. K. (1989). *Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents*. Paper presented at the Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Cambridge, MA.
- Belkin, N. J. (1993). *Interaction with texts: Information retrieval as information-seeking behavior*. Paper presented at the In Proc. of Information Retrieval.
- Belkin, N. J. (2008). *Some (what) grand challenges for information retrieval*. Paper presented at the ACM SIGIR Forum.
- Bell, C. G., & Newell, A. (1971). *Computer Structures: Readings and Examples*. New York: McGraw-Hill.
- Bennet, J., & Lanning, S. (2007). *The Netflix prize*. Paper presented at the KDD Cup and Workshop.
- Berthold, M. R., Brandes, U., Kotter, T., Mader, M., Nagel, U., & Thiel, K. (2009). *Pure spreading activation is pointless*. Paper presented at the In Proceedings of the 18th ACM conference on information and knowledge management.
- Blank, D. S. (1997). *Learning to see analogies: A connectionist exploration*. PhD, Indiana University.
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The annals of applied statistics*, 1(1), 17-35.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Bollobas, B. (1985). *Random Graphs*. London: Academic Press.

- Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2, 113-120.
- Bordogna, G., & Pasi, G. (1996). A user-adaptive neural network supporting a rule-based relevance feedback. *Fuzzy Sets and Systems*, 82, 201-211.
- Broder, A. (2002). *A taxonomy of web search*. Paper presented at the ACM Sigir forum.
- Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., . . .
- Roosin, P. S. (1990). A statistical approach to machine translation. *Computation Linguistics*, 16(2), 79-85.
- Brumby, D. P., Salvucci, D. D., & Howes, A. (2007). *Dialing while driving? A bounded rational analysis of concurrent multi-task behavior*. Paper presented at the Proc. 8th International Conf. on Cognitive Modeling.
- Bruza, P. D., Kitto, K., Ramm, B., Sitbon, L., Song, D., & Blomberg, S. (2012). Quantum-like non-separability of concept combinations, emergent associates, and abduction. *Logi Journal of IGPL*, 20(2), 445-457.
- Buccio, E. D., Melucci, M., & Song, D. (2011). *Towards predicting relevance using a quantum-like framework*. Paper presented at the Proceedings of the 33rd european conference on advances in information retrieval.
- Buchanan, M. (2011). Quantum minds: Why we think like quarks. *New Scientist*, 05.
- Budiu, R. (2001). *The role of background knowledge in sentence processing*. Doctoral Dissertation, Carnegie Mellon University, Pittsburgh, P.A. (CMU-CS-01-148)
- Budiu, R., & Anderson, J. R. (2000). *Integration of background knowledge in sentence processing: A unified theory of metaphor understanding, semantic illusions, and*

- text memory*. Paper presented at the In Proceedings of the Third International Conference on Cognitive Modelling, Groningen, Netherlands.
- Budiu, R., & Anderson, J. R. (2002). Comprehending anaphoric metaphors. *Memory & Cognition*, 30, 158-165.
- Budiu, R., & Anderson, J. R. (2004). Interpretation-based processing: A unified theory of semantic sentence processing. *Cognitive Science*, 28, 1-44.
- Budiu, R., & Anderson, J. R. (2006). Word learning in context: Metaphors and neologisms. In M. Vishnevsky (Ed.), *Focus on Cognitive Psychology Research*. New York, NY: Nova Science Publishers.
- Budiu, R., Pirolli, P., & Hong, L. (2009). *Remembrance of things tagged: How tagging effort effects tag production and human memory*. Paper presented at the In CHI 2009 Conference on Human Factors in Computing Systems, Boston, MA.
- Budiu, R., Royer, C., & Pirolli, P. L. (2007). *Modeling information scent: A comparison of LSA, PMI and GLSA similarity measures on common tests and corpora*. Paper presented at the In Proceedings of the 8th Annual Conference of the Recherche d'Information Assistée par Ordinateur (RIAO).
- Burnham, J. F. (2006). Scopus database: A review. *Biomedical Digital Libraries*, 3(1).
- Burrell, Q. (1980). A simple stochastic model for library loans. *Journal of Documentation*, 36(2), 115-132.
- Burrell, Q. (1985). A note on the aging in a library circulation model. *Journal of Documentation*, 41(2), 100-115.
- Burrell, Q., & Cane, V. R. (1982). The analysis of library data. *Journal of the Royal Statistical Society (Series A)*, 145, 439-471.

- Burrell, Q., & Fenton, M. R. (1994). A model of library book circulations incorporating loan periods. *Journal of the American Society for Information Science*, 45(2), 101-116.
- Bussemeyer, J. R., & Bruza, P. D. (2012). *Quantum Models of Cognition and Decision*: Cambridge University Press.
- Bussemeyer, J. R., Pothos, E., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment 'errors'. *Psychological Review*, 118(2), 193-218.
- Campbell, I., & van Rijsbergen, C. J. (1996). *The ostensive model of developing information needs*. Paper presented at the In Proceedings of CoLIS, Copenhagen, Denmark.
- Canfora, G., & Cerulo, L. (2004). A taxonomy of information retrieval models and tools. *J. Comput. Inf. Technol.*, 12, 175-194.
- Cao, G., Nie, J. Y., Gao, J., & Robertson, S. (2008). *Selecting good expansion terms for pseudo-relevance feedback*. Paper presented at the In SIGIR '08, New York, New York.
- Capocci, A., Servedio, V. D. P., Colaiori, F., Burrol, L. S., Donato, D., Leonardi, S., & Caldarelli, G. (2006). Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E*, 74(3).
- Card, S. K., Pirolli, P., Van Der Wege, M., Morrisison, J. B., Reeder, R. W., Schraedley, P. K., & Boshart, J. (2001). *Information scent as a driver of web behavior graphs: Results of a protocol analysis method for web usability*. Paper presented at the In Proc. CHI 2001.

- Carillo, M., Villatoro-Tello, E., Lopez-Lopez, A., Eliasmith, C., Montes-y-Gomez, M., & Villasenor-Pineda, L. (2009). Representing context information for document retrieval *Flexible Query Answering Systems* (pp. 239-250): Springer.
- Carrillo, M., Eliasmith, C., & Lopez-Lopez, A. (2009). *Combining text vector representations for information retrieval*. Paper presented at the In Proc. of the 12th International Conference on Text.
- Carrillo, M., Villatoro, E., Lopez, A., Eliasmith, C., Pineda, L., & Gomez, M. (2010). *Concept based representations for ranking in geographic information retrieval*. Paper presented at the In Proceedings of the 7th International Conference on Advances in Natural Language Processing.
- Center, T. M. (2013). Texas Medical Center Member Institutions Retrieved 8-1-13, from <http://www.texasmedicalcenter.org/all-institutions/>
- Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*: Oxford University Press.
- Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3, 191-209.
- Chi, E. H., Gumbrecht, M., & Hong, L. (2007). Visual foraging of highlighted text: An eye-tracking study *In Human-Computer Interaction*: Springer Berlin Heidelberg.
- Chi, E. H., Hong, L., Gumbrecht, M., & Card, S. K. (2005). *ScentHighlights: highlighting conceptually-related sentences during reading*. Paper presented at the In Proceedings of the 10th international conference on intelligent user interfaces.

- Chi, E. H., Hong, L., Heiser, J., & Card, S. K. (2004). *eBooks with indexes that reorganize conceptually*. Paper presented at the In CHI'04 extended abstracts on human factors in computing systems.
- Chi, E. H., Hong, L., Heiser, J., & Card, S. K. (2006). *ScentIndex: Conceptually reorganizing subject indexes for reading*. Paper presented at the IEEE Symposium on Visual Analytics Science and Technology.
- Chi, E. H., Hong, L., Heiser, J., Card, S. K., & Gumbrecht, M. (2007). ScentIndex and ScentHighlights: Productive reading techniques for conceptually reorganizing subject indexes and highlighting passages. *Information Visualization*, 6(1), 32-47.
- Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. E. (2001). *Using information scent to model user information needs and actions and the web*. Paper presented at the Conference on Human Factors in Computing Systems.
- Chi, E. H., Pirolli, P., & Pitkow, J. E. (2001). *The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a site*. Proceedings of the CHI 2001.
- Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.
- . CiteULike. Retrieved 7-1-13, from <http://www.citeulike.org/>
- Clauset, A., & Shalizi, C. R. (2009). Power-law distributions in empirical data. *SIAM Rev.*, 51.
- Cleverdon, C. W. (1960). *ASLIB Cranfield research project on the comparative efficiency of indexing systems*. Paper presented at the ASLIB Proceedings.
- Cleverdon, C. W. (1967). *The Cranfield test on index language devices*. Paper presented at the Aslib proceedings.

- Cleverdon, C. W., & Keen, E. M. (1966). *Factors determining the performance of indexing systems*. Aslib Cranfield Research Project. Cranfield, UK.
- Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, 82(6), 407-428.
- Conte, E., Khrennikov, A. Y., Todarello, O., Federici, A., Mendolicchio, L., & Zbilut, J. P. (2009). Mental states follow quantum mechanics during perception and cognition of ambiguous figures. *Open Systems & Information Dynamics*, 16(1), 85-100.
- Copeland, B. J., & Proudfoot, D. (1996). On Alan Turing's anticipation of connectionism. *Synthese*, 108(3), 361-377.
- Corominas-Murta, B., Valverde, S., & Sole, R. (2009). The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. *Advances in Complex Systems*, 12(3), 371-392.
- Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11, 453-482.
- Crestani, F. (1999). *WebSCSA: Web search by constrained spreading activation*. Paper presented at the IEEE Forum on Research and Technology Advances in Digital Libraries, Baltimore, MD.
- Crestani, F., & Lee, P. L. (2000). Searching the web by constrained spreading activation. *Information Processing & Management*, 36(4), 585-605.
- Croft, W. B. (2002). Combining approaches to information retrieval. *Advances in Information Retrieval*, 1-36.



- Crouch, C. C., Crouch, D. B., & Nareddy, K. (1994). Associative and adaptive retrieval in a connectionist system. *International Journal of Expert Systems*, 7(2), 193-202.
- Csanyi, G., & Szendroi, B. (2004). Structure of a large social network. *Physical Review*, 69, 1-5.
- Cunningham, S. J., Holmes, G., Littin, J., Beale, R., & Witten, I. H. (1997). Applying connectionist models to information retrieval. In S. Amari & N. Kasobov (Eds.), *Brain-like Computing and Intelligent Information Systems* (pp. 435-457): Springer-Verlag.
- Curran, M. (2011). Mendeley. *Technical Services Quarterly*, 28(4), 454-456.
- Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd & A. Group (Eds.), *Simple Heuristics That Make Us Smart*. New York: Oxford UP.
- Danker, J. F., & Anderson, J. R. (2007). *Using cognitive modeling to understand the roles of prefrontal and posterior parietal cortex in algebra problem solving*. Paper presented at the Proceedings of the Society for Cognitive Neuroscience, New York, New York.
- Dawson, M. R., & Shamanski, K. S. (1994). Connectionism, confusion, and cognitive science. *Journal of Intelligent Systems*, 4, 215-262.
- de Almeida, H. M., Goncalves, M. A., Cristo, M., & Calado, P. (2007). *A combined component approach for finding collection-adapted ranking functions based on genetic programming*. Paper presented at the In Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval.

- de Blasio, B. F., Svensson, A., & Liljeros, F. (2006). Preferential attachment in sexual networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(26).
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.
- Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, 93, 283-321.
- Deyne, S. D., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, 40(1), 213-231.
- Dezso, Z., Almaas, E., Lukacs, A., Racz, B., Szakadat, I., & Barabasi, A. L. (2006). Dynamics of information access on the web. *Physical Review*, 73.
- Doidge, N. (2007). *The brain that changes itself*. New York: Viking Books.
- Dou, Z., Song, R., & Wen, J. R. (2007). *A large-scale evaluation and analysis of personalized search strategies*. Paper presented at the In Proceedings of WWW'07.
- Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. Mineola, NY: Dover Publications.
- Ebel, H., Mielsch, L. I., & Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Phys. Rev. E*, 66.
- Eisenberg, E., & Levanon, E. Y. (2003). Preferential attachment in the protein network evolution. *Physical Review Letters*, 91(13).

- Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: a distributed model of analogical mapping. *Cognitive Science*, 25, 245-286.
- Erdos, P., & Renyi, A. (1959). On random graphs. *Publicationes Mathematicae*(6), 290-297.
- Erdos, P., & Renyi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5, 17-16`.
- Erdos, P., & Renyi, A. (1961). On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary*(12), 261-267.
- Falagas, M., Pitsouni, E., Malietzis, G., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal*, 22, 338-342.
- Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, 73, 421-435.
- Fishbein, J. M., & Eliasmith, C. (2008). *Integrating structure and meaning: A new method for encoding structure for text classification*. Paper presented at the Procs. of the 30th European Conf. on IR Research.
- Fodor, J. (1997). Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work. *Cognition*, 62, 109-119.
- Fodor, J. A., & MCLAughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35, 183-204.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), 3-71.

- Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3), 285-307.
- Fraser, A. G. (2010). On the impossibility of being expert. *BMJ*, 341.
- Freeman, L. C. (1977). A set of measures of centrality based upon betweenness. *Sociometry*, 40, 35-41.
- Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P., & van Rijsbergen, K. (2010). *Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework*. Paper presented at the Proceedings of the 2010 Information Interaction in Context Symposium, New Brunswick.
- Fu, W. T. (2008). *The microstructures of social tagging: A rational model*. Paper presented at the In Proceedings of the ACM 2008 conference on Computer Supported Cooperative Work, San Diego, CA.
- Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, 295, 90-93.
- Gavanski, I., & Roskos-Ewoldsen, D. R. (1991). Representativeness and conjoint probability. *Journal of Personality and Social Psychology*, 61, 181-194.
- Gayler, R. (2003). *Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience*. Paper presented at the International Conference on Cognitive Science (ICCS/ASCS), Sydney, Australia.
- Gayler, R. W., & Levy, S. D. (2009). *A distributed basis for analogical mapping*. Paper presented at the In new frontiers in analogy research; Proceedings of the second international analogy conference.

- Gayler, R. W., & Sandin, F. (2013). *Analogical mapping and inference with binary spatter codes and sparse distributed memory*. Paper presented at the Proceedings of International Joint Conference on Neural Networks 2013, Dallas, TX.
- Gehrke, J., Ginsparg, P., & Kleinberg, J. M. (2003). Overview of the 2003 KDD Cup. *SIGKDD Explorations*, 5(2), 149-151.
- Gentner, D., & Markman, A. B. (1992). Analogy - watershed or waterloo? Structural alignment and the development of connectionist models of analogy. In C. L. Giles, S. J. Hanson & J. D. Cowan (Eds.), *Advances in Neural Information Processing Systems* (pp. 855-862). San Mateo, CA: Morgan Kaufmann.
- Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: the adaptive toolbox*. Boston, MA: The MIT Press.
- Glucksberg, S. (1998). Understanding metaphors. *Current directions in psychological science*, 7, 39-43.
- Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5), 403-420.
- Goodwin, J. C., Cohen, T., & Rindflesch, T. (2012). *Discovery by scent: Discovery browsing system based on the Information Foraging Theory*. Paper presented at the The First International Workshop on the role of Semantic Web in Literature-Based Discovery (SWLBD2012), Philadelphia PA.
- Goodwin, J. C., Johnson, T. R., Cohen, T., Herskovic, J. R., & Bernstam, E. V. (2011). Predicting biomedical document access as a function of past use. *J Am Med Inform Assoc*.
- . Google Personal. (2013), from <http://labs.google.com/personalized>

- Granka, L. A., Joachims, T., & Gay, G. (2004). *Eye-tracking analysis of user behavior in WWW search*. Paper presented at the Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval.
- Green, R. F. (1980). Bayesian birds: A simple example of Oaten's stochastic model of optimal foraging. *Theor. Pop. Biol.*, 18, 244-256.
- Griffiths, T. L., & Steyvers, M. (2004). *Finding scientific topics*. Paper presented at the Proceedings of the National Academy of Science.
- Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science*, 18(12), 1069-1076.
- Guhe, M., Smaill, A., & Pease, A. (2010). *Towards a cognitive model of conceptual blending*. Paper presented at the Proceedings of the 10th international conference on cognitive modeling, Philadelphia, PA: Drexel University.
- Halford, G. S., Wiles, J., Humphreys, M. S., & Wilson, W. H. (1993). Parallel distributed processing approaches to creative reasoning: Tensor models of memory and analogy. *AI and Creativity*, 57-60.
- Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42, 335-346.
- Harris, Z. (1954). Distributional structure. *Word*, 10, 146-162.
- Hatcher, E., & Gospondnetic, O. (2004). *Lucene in action*. Greenwich, CT: Manning Publications Co.
- Hebb, D. O. (1940). Human behavior after extensive bilateral removal from the frontal lobes. *Archives of Neurology and Psychiatry*, 44(2), 421-438.

- Hebb, D. o. (1961). Distinctive features of learning in the higher animal. In J. F. Delafresnaye (Ed.), *Brain mechanisms and learning*. London: Oxford University Press.
- Henning, V., & Reichelt, J. (2008). *Mendeley - A Last.fm for research?* Paper presented at the IEEE fourth international conference on in eScience.
- Hersh, W., Turpin, J., Price, S., Kraemer, D., Olson, D., Chan, B., & Sachereck, L. (2001). Challenging conventional saaumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Information processing & management*, 37(3), 383-402.
- Hersh, W. R., Bhupatiraju, R. T., Ross, L., Roberts, P., Cohen, A. M., & Kraemer, D. F. (2006). Enhancing access to the Bibliome: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration*, 1(1).
- Hiemstra, D. (2000a). A probabilistic justification for using tf x idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3, 131-139.
- Hiemstra, D. (2000b). A probabilistic justification for using tf-idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2), 131-139.
- Hiemstra, D. (2000c). A probabilistic justification for using tf\*idf term weighting in infomation retrieval. *International Journal on Digital Libraries*, 3(2), 131-139.
- Hiemstra, D., & de Vries, A. (2000). *Relating the new language models of information retrieval to the traditional retrieval models*. University of Twente. Enschede, Netherlands.

- Hiemstra, D., & Kraaij, W. (1998). *Twenty-one at TREC-7: Ad-hoc and cross language track*. Paper presented at the In Proceedings of Seventh Text Retrieval Conference (TREC-7).
- Hills, T. T., Todd, P. M., & Jones, M. (2009). *Optimal foraging in semantic memory*. Paper presented at the Proceedings of the 31st annual conference of the cognitive science society.
- Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98, 74-95.
- Holme, P., & Kim, B. J. (2002). Growing scale-free networks with tunable clustering. *Phys. Rev.*, 65.
- Hong, L., Chi, E. H., Budiu, R., Pirolli, P., & Nelson, L. (2008). *SparTag. us: A low cost tagging system for foraging of web content*. Paper presented at the In Proceedings of the working conference on advanced visual interfaces.
- Huberman, B., Pirolli, P., Pitkow, J. E., & Lukose, R. (1998). Strong regularities in world wide web surfing. *Science*, 280.
- Huey, R. B., Bennett, A. F., John-Alder, H., & Nagy, K. A. (1984). Locomotor capacity and foraging behavior of Kalahari lacertid lizards. *Anim Behav*, 32, 41-50.
- Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham.
- Ingwersen, P., & Jarvelin, K. (2010). *The turn: Integration of information seeking and retrieval in context*: Kluver.
- Islam, M. M., & Hoque, A. L. (2010). Automated essay scoring using generalized latent semantic analysis. *13th International Conference on Computer and Information Technology*, 358-363.



- Islamaj Dogan, R., Murray, C. G., Neveol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database*. doi: 10.1093/database/bap018
- Jarvelin, K. (2009). *Explaining user performance in information retrieval: Challenges to IR evaluation*. Paper presented at the Proceedings of the 2nd international conference on the theory of information retrieval.
- Jelinek, F. (1997). *Statistical methods for speech recognition*: MIT Press.
- Jennings, A., & Higuchi, H. (1992). A browser with a neural network user model. *Library Hi Tech*, 10(1-2), 77-93.
- Jeong, H., Tombor, B., Albert, R., Oltval, Z. N., & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(5), 651-654.
- Joachims, T. (2003). Evaluating retrieval performance using clickthrough data. In J. Franke, G. Nakhaeizadeh & I. Renz (Eds.), *Text Mining* (pp. 79-96): Physica/Springer Verlag.
- Joachims, T., Granka, L., Bing Pan, H., Hembrooke, H., Randlinski, F., & Gay, G. (2007). *Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search*. Paper presented at the ACM Transactions on Information Systems.
- Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). *Accurately interpreting clickthrough data as implicit feedback*. Paper presented at the In proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval.

- Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2), 7.
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-21.
- Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and Comparative Experiments. *Information Processing and Management*, 36(6), 779-840.
- Jones, W. P. (1986a). *The memory extender personal filing system*. Paper presented at the CHI '86 Proceedings of the SIGCHI Conference on Human Factors in Computer Systems.
- Jones, W. P. (1986b). On the applied use of human memory models: the memory extender personal filing system. *International Journal of Man-Machine Studies*, 25(2), 191-228.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47, 263-291.
- Kanerva, P. (1994). *The spatter code for encoding concepts at many levels*. Paper presented at the ICANN '94 Proc. Int'l Conference on Artificial Neural Networks, Sorrento, Italy.
- Kanerva, P. (2010). *What we mean when we say "What's the dollar of mexico?"*: Prototypes and mapping in concept space. Paper presented at the In Proc. AAAI

Fall Symp. on Quantum Informatics for Cognitive, Social, and Semantic Processes.

- Kelly, T. D. (2003). Symbolic and sub-symbolic representations in computational models of human cognition: What can be learned from biology? *Theory & Psychology*, 13(6), 847-860.
- Khrennikov, A. Y., & Haven, E. (2009). Quantum mechanics and violations of the sure-thing principle: the use of probability interference and other concepts. *Journal of Mathematical Psychology*, 53(5), 378-388.
- Kie, J. G., Evans, C. J., Loft, E. R., & Menke, J. W. (1991). Foraging behavior by mule deer: the influence of cattle grazing. *Journal of Wildlife Management*, 55, 665-674.
- Kimoto, H., & Iwadera, T. (1989). *Construction of a dynamic thesaurus and its use for associated information retrieval*. Paper presented at the In proceedings of the 13th annual international ACM SIGIR conference on research and development in information retrieval.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*: Cambridge University Press.
- Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7.
- Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services, and Agents on the World Wide Web*, 2(2), 49-79.

- Kowk, K. L. (1989). *A neural network for probabilistic information retrieval*. Paper presented at the Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Cambridge, MA.
- Kowk, K. L. (1991). *Query modification and expansion in a network with adaptive architecture*. Paper presented at the Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago, Illinois.
- Kraaij, W. (2004). *Variations on language modeling for information retrieval*.
- Kurland, O., & Lee, L. (2005). *PageRank without hyperlinks: Structural re-ranking using links induced by language models*. Paper presented at the Proceedings of SIGIR 2005.
- Kwok, K. L. (1989). *A neural network for probabilistic information retrieval*. Paper presented at the Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Cambridge, MA.
- Lafferty, J., & Zhai, C. (2001). *Document language models, query models, and risk minimization for information retrieval*, In Proceedings of SIGIR'01.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2).
- Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Intelligent Systems*, 15(5), 27-31.
- Lawrance, J., Bellamy, R., & Burnett, M. (2007a). *Scents in programs: Does information foraging theory apply to program maintenance?* Paper presented at the IEEE Symp. Visual Languages and Human-Centric Computing (VL/HCC).

- Lawrance, J., Bellamy, R., & Burnett, M. (2007b). *Scents in programs: Does information foraging theory apply to program maintenance?* Paper presented at the IEEE Symposium on Visual Languages and Human-Centric Computing.
- Lawrance, J., Bellamy, R., Burnett, M., & Recker, M. M. (2008). *Using information scent to model the dynamic foraging behavior of programmers in maintenance tasks.* Paper presented at the In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.
- Lawrance, J., Bellamy, R., Burnett, M., & Rector, K. (2008a). *Can information foraging pick the fix? A field study.* Paper presented at the IEEE Symp. Visual Languages and Human-Centric Computing (VL/HCC).
- Lawrance, J., Bellamy, R., Burnett, M., & Rector, K. (2008b). *Using information scent to model the dynamic foraging behavior of programmers in maintenance tasks.* Paper presented at the ACM Conf. Human Factors in Computing Systems (CHI).
- Lawrance, J., Bogart, C., Burnett, M., Bellamy, R., Rector, K., & Flemming, S. D. (2013). How programmers debug, revisited: An information foraging theory perspective. *IEEE Transactions on Software Engineering*, 39(2), 197-215.
- Lebiere, C., Anderson, J. R., & Reder, L. M. (1994). *Error modeling in the ACT-R production system.* Paper presented at the Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.
- LEE, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. *IEEE Software*, 14(2), 67-75.
- Lee, D. S., Park, J., Kay, K. A., Christakis, N. A., Oltvai, Z. N., & Barabasi, A. L. (2008). The implications of human metabolic network topology for disease comorbidity.

*Proceedings of the National Academy of Sciences of The United States of America*, 105(29).

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). *Meme-tracking and the dynamics of the news cycle*. Paper presented at the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM KDD).

Li, R., Wang, S., & Chang, K. C. C. (2012). Multiple location profiling for users and relationships from social network and content. *PVLDB*, 5(11), 1603-1614.

Li, R., Wang, S., Deng, H., Wang, R., & Chang, K. C. C. (2012). *Towards social user profiling: unified and discriminative influence model for inferring home locations*. Paper presented at the KDD.

Light, S., Kraulis, P., & Elofsson, A. (2005). Preferential attachment in the evolution of metabolic networks. *BMC Genomics*, 6(159).

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., & Aberg, Y. (2001). Human sexual contacts. *Nature*, 411, 907-908.

Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society*, 270. doi: 10.1098/rsbl.2003.0057

Macdonald, C., & Ounis, I. (2009). *The influence of the document ranking in expert search*. Paper presented at the In proceedings of the 18th international conference on information and knowledge management, Hong Kong, China.

MacKay, D., & Peto, L. (1995). A hierarchical dirichlet language model. *Natural Language Engineering*, 1, 289-307.

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, 17, 457-472.

- Manning, C. D., & Schtze, H. (1999). *Foundations of statistical natural language processing*. Cambridge Massachusetts: MIT Press.
- Marchionini, G. (1995). *Information seeking in electronic environments*: Cambridge University Press.
- Matthijs, N., & Radlinski, F. (2011). *Personalizing web search using long term browsing history*. Paper presented at the In Proceedings of the fourth ACM international conference on Web Search and Data Mining (WSDM 11), New York, NY.
- McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Retrieved from <http://mallet.cs.umass.edu>
- McCarthy, J., & Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & P. Michie (Eds.), *Machine Intelligence* (Vol. 4). Edinburgh: Edinburgh University Press.
- McClelland, J., Rumelhart, D., & Group, P. R. (1986). *Parallel distributed processing: Explorations in the microstructures of cognition*. Cambridge, MA: MIT Press.
- Mei, Q., Zhang, D., & Zhai, C. X. (2008). *A general optimization framework for smoothing language models on graph structures*. Paper presented at the In SIGIR '08: Proceedings of the 31st annaul international ACM SIGIR conference on research and development in information retrieval, USA, New York, NY.
- Melucci, M. (2007). *Exploring a mechanics for context-aware in information retrieval*. Paper presented at the In Proc. of the AAAI Spring Symposium on Quantum Interaction, Stanford, CA.

- Melucci, M., & White, R. W. (2007a). *Discovering hidden contextual factors for implicit feedback*. Paper presented at the In Proceedings of the 2nd Workshop on Context-based Information Retrieval (CIR).
- Melucci, M., & White, R. W. (2007b). *Utilizing a geometry of context for enhanced implicit feedback*. Paper presented at the In Proceedings of the Conference on Information and Knowledge Management (CIKM).
- Mendeley-API. (2013). Mendeley API, 8-1-2013, from <http://dev.mendeley.com/>
- Miller, D. H., Leek, T., & Schwartz, R. (1999). *A hidden Markov model information retrieval system*. Paper presented at the In proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval.
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41.
- Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research*, 29(4), 495-512.
- Moore, D. W. (2002). Measuring new types of question-order effects. *Public Opinion Quarterly*, 66, 80-91.
- Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cognitive Science*, 37, 125-145.
- Nagy, W., & Anderson, R. (1984). The number of words in printed school English. *Reading Research Quarterly*, 20, 223-253.
- Nagy, W. E., & Herman, P. A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. C. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 19-35). Hillsdale, NJ: Erlbaum.



- Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.
- Newell, A., Shaw, J. C., & Simon, H. A. (1959). *Report on a general problem-solving program*. Paper presented at the Proceedings of the International Conference on Information Processing.
- Newell, A., & Simon, H. (1976a). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3).
- Newell, A., & Simon, H. (1976b). Computer science as empirical inquiry: symbols and search. *Communications of the ACM*, 19, 113-126.
- Newell, A., & Simon, H. A. (1956). The Logic Theory Machine: A complex information processing system. *The Rand Corporation*.
- Newman, T. (1994). Memories of the Pilot ACE. *Computer Resurrection*, 9.
- Ngo, V. M., & Cao, T. H. (2011). *Discovering latent concepts and exploiting ontological features for semantic text search*. Paper presented at the In IJCNLP.
- NLM. (2013). MBR Files (<http://mbr.nlm.nih.gov/Download/>) Retrieved 9-11-13
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*: Oxford University Press.
- Oliveira, J. G., & Barabasi, A. L. (2005). Darwin and Einstein correspondence patterns. *Nature*, 437(27).
- Olston, C., & Chi, E. H. (2003). *ScentTrails: Integrating browsing and searching on the Web*. Paper presented at the ACM Transactions on Computer-Human Interaction (TOCHI).

- Opsahi, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3), 245-251.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: bringing order to the web *Technical report*: Standord Digital Library Technologies Project.
- Pannu, A. S., & Sycara, K. (1996). *Learning text filtering preferences*. Paper presented at the Proceedings of the AAAI Syposium on Machine Learning and Information Access, Standord, CA.
- Papieni, K. (2001). *Why inverse document frequency?* Paper presented at the Proceedings of the North American Association for Computational Linguistics (NAACL).
- Pareto, V. (1964). *Cours d'economie politique*: Librairie Droz.
- Pavlik, P. I. J., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29, 559-586.
- Pereira, H. B. B., Fadigas, I. S., Senna, V., & Moret, M. A. (2011). Semantic networks based on titles of scientific papers. *Physica A: Statistical Mechanics and its Applications*, 390(6), 1192-1197.
- Petrov, A. A. (2006). *Computationally efficient approximation of the baselevel learning equation in ACT-R*. Paper presented at the Proceedings of the 7th International Conference on Cognitive Modeling, Trieste, ITA.
- Pirolli, P. (2005). Rational analyses of information foraging on the Web. *Cognitive Science*, 29, 343-373.

- Pirolli, P. (2009). *Information foraging theory: Adaptive interaction with information*: Oxford University Press.
- Pirolli, P., & Card, S. (1995). *Information foraging in information access environments*. Proceedings of the ACM Conference on Human Factors in Computing Systems.
- Pirolli, P., & Card, S. (1998). *Information foraging models of browsers for very large document spaces*. Paper presented at the Proceedings of the working conference on advanced visual interfaces.
- Pirolli, P., & Card, S. (1999a). Information Foraging. *Psychological review*, 106(4), 643.
- Pirolli, P., & Card, S. (1999b). *Information foraging*: Oxford University Press.
- Pirolli, P., Chi, W. T., & Farahat, A. (2005). *Information scent and web navigation: Theory, models and automated usability evaluation*. Paper presented at the HCI International.
- Pirolli, P., & W-T., F. (2006). *SNIF-ACT: a model of information foraging on the World Wide Web*. Ninth International Conference on Using Modeling.
- Pirolli, P. L., & Anderson, J. R. (1985). The role of practice in fact retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 11, 136-153.
- Pirolli, P. L., & Pitkow, J. E. (2000). USA Patent No. U.S. Patent No. 6,098,064. U. S. P. a. T. Office.
- Pitkow, J. E. (1997). *Characterizing world wide web ecologies*. PhD thesis, Georgia Institute of Technology, Georgia.
- Pitkow, J. E., & Recker, M. M. (1994). *A simple yet robust caching algorithm based on dynamic access patterns*. Paper presented at the In Electron. Proc. 2nd WWW Conf.

- Piwowarski, B., Frommholz, I., Lalmas, M., & Rijsbergen, K. V. (2010). *What can quantum theory bring to information retrieval*. Paper presented at the Proceedings of the 19th ACM International Conference on Information and Knowledge Management.
- Piwowarski, B., Frommholz, I., Lalmas, M., & van Rijsbergen, K. (2010). *Exploring a multidimensional representation of documents and queries*. Paper presented at the In Proceedings of the 9th RIAO Conferences (IRAO 2010), Paris, France.
- Piwowarski, B., & Lalmas, M. (2009). A quantum-based model for interactive information retrieval. *ArXiv e-prints*.
- Plate, T. A. (1994). Estimating analogical similarity by dot-products of Holographic Reduced Representations. *Advances in neural information processing systems*.
- Plate, T. A. (1995). *Holographic reduced representations*. Paper presented at the IEEE Trans. Neural Networks.
- Plate, T. A. (2000). Analogy retrieval and processing with distributed vector representations. *Expert Systems*, 17(1), 29-40.
- PLOS. (2013). Technology Retrieved 8-1-2013, from <http://www.plos.org/about/technology/>
- Pohl, S., Radlinski, F., & Joachims, T. (2007). *Recommending related papers based on digital library access records*. Paper presented at the In Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries.
- Ponte, J., & Croft, W. B. (1998). *A language modeling approach to information retrieval*. Paper presented at the In Proceedings of the ACM SIGIR 98.

- Pothos, E. M., & Busemeyer, J. R. (2009a). A quantum probability explanation for violations of "rational" decision theory. *Proc. R. Soc. B: Biological Sciences*, 276(1665), 2171-2178.
- Pothos, E. M., & Busemeyer, J. R. (2009b). *A quantum probability explanation for violations of "rational" decision theory*. Paper presented at the Proceedings of the Royal Society of Biological Sciences.
- Pothos, E. M., & Busemeyer, J. R. (2011). *A quantum probability explanation for violations of symmetry in similarity judgments*. Paper presented at the In Proceedings of the 32nd Annual Conference of the Cognitive Science Society, Mahwah, NJ.
- Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, 36, 255-327.
- Pothos, E. M., Busemeyer, J. R., Shiffring, R. M., Trueblood, J. S., Wang, Z., Reinhard, K. B., & Atamanspacher, H. (2011). *The potential of quantum probability for modelling cognitive processes*. Paper presented at the Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Austin, TX.
- Preece, S. E. (1981). *A spreading activation network model for information retrieval*. . Public Websites using Solr. from <http://wiki.apache.org/solr/PublicServers>
- Radiohead. (2011). Lotus Flower. *The King of Limbs*.
- Radlinski, F., & Joachims, T. (2005). *Evaluating the robustness of learning from implicit feedback*. Paper presented at the ICML Workshop on Learning in Web Search.

- Radlinski, F., & Joachims, T. (2006). *Minimally invasive randomization for collecting unbiased preferences from clickthrough logs*. Paper presented at the In proceedings of the national conference on artificial intelligence, Menlo Park, CA.
- Radlinski, F., & Joachims, T. (2007). *Active exploration for learning rankings from clickthrough data*. Paper presented at the In proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining.
- Radlinski, F., Kurup, M., & Joachims, T. (2008). *How does clicktrhough data reflect retrieval quality?* Paper presented at the In proceedings of the 17th ACM conference on information and knowledge management.
- Recchia, G. L., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information to latent semantic analysis. *Behavior Research Methods*, 41, 657-663.
- Recker, M. M., & Pitkow, J. E. (1996). Predicting document access in large multimedia repositories. *ACM transactions on Computer-Human Interaction (TOCHI)*, 3(4).
- Reuters, T. (2013). Web of Science Web Retrieved 7-1-13, from <http://thomsonreuters.com/web-of-science/>
- Rhodes, T., & Turvey, M. T. (2007). Human memory retrieval as L'evy foraging. *Physica A*, 385.
- Rieffel, E. G., & Polak, W. (2000). An introduction to quantum computing for non-physicists. *ACM Comput. Surveys.*, 32, 300-335.
- Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic

- propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6), 462-477.
- Risvik, K. M., Mikolajewski, T., & Boros, P. (2003). *Query segmentation for web search*. Paper presented at the In Poster Session in the Twelfth International World Wide Web Conference.
- Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5), 503-520.
- Robertson, S. E., Walker, S., Beaulieu, M., Gatford, M., & Paynet, A. (1996). *Okapi at TREC-4*. Paper presented at the In Proc. 4th Text Retrieval Conference (TREC-\$).
- Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press/Bradford Books.
- Sahlgren, M. (2005). *An introduction to random indexing*. Paper presented at the In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering.
- Salton, G. (1971). *The SMART retrieval system: experiments in automatic document processing*: Prentice-Hall.
- Salton, G., & Buckley, C. (1988). *On the use of spreading activation methods in automatic information retrieval*. Paper presented at the In Proceedings of ACM SIGIR, Grenoble, France.
- Sanderson, M., Paramita, M. L., Clough, P., & Kanoulas, E. (2010). *Do user preferences and evaluation measures line up?* Paper presented at the In proceedings of SIGIR.

- Saracevic, T. (1997). *Users lost: reflections of the past, future, and limits of information science*. Paper presented at the SIGIR Forum.
- Savage, L. J. (1954). *The foundations of statistics*. New York, New York: Wiley.
- Sayers, E., & Wheeler, D. (2004). Building customized data pipelines using the Entrez Programming Utilities (eUtils) *NCBI Short Courses*. Bethesda, MD: National Library of Medicine.
- Schooler, L. J. (1993). *Memory and the statistical structure of the environment*. PhD, Carnegie Mellon University, Pittsburgh, PA.
- Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of human memory. *Cognitive Psychology*, 32, 219-250.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: experiments on question form, wording, and content*. NY, NY: Academic Press.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-457.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523-568.
- . Semantic Vectors Package. (2012) Retrieved 3-1-12, from <http://code.google.com/p/semanticvectors/>
- Shafir, E. (1994). Uncertainty and the difficulty of thinking through disjunctions. *Cognition*, 50, 403-430.
- Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24, 449-474.



- Shen, X., Tan, B., & Zhai, C. (2005). *Implicit user modeling for personalized search*.  
Paper presented at the In Proceedings of CIKM '05.
- Shoval, P. (1981). *Expert/consultation system for a retrieval data-base with semantic network of concepts*. Paper presented at the In ACM SIGIR Forum.
- Sides, A., Oshershon, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory and Cognition*, 30, 191-198.
- Sidiropoulos, A., & Manolopoulos, Y. (2005). A new perspective to automatically rank scientific conferences using digital libraries. *Information processing & management*, 41(2), 289-312.
- Siegler, M., & Witbrock, M. (1999). *Improving the suitability of imperfect transcriptions for information retrieval from spoken documents*. Paper presented at the Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
- Simen, P., & Polk, T. (2010). A symbolic/subsymbolic interface protocol for cognitive modeling. *Log J IGPL*, 18(5), 705-761.
- Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42, 425-440.
- Simon, H. A. (1956). Rational choice and the structure of environments. *Psych. Rev.*, 63, 129-138.
- Singhal, A., Buckley, C., & Mitra, M. (1996). *Pivoted document length normalization*.  
Paper presented at the In Proceedings of the 1996 ACM SIGIR Conference on Research and Development in Information Retrieval.
- Smith, C. L., & Kantor, P. B. (2008). *User adaptation: good results from poor systems*.  
Paper presented at the ACM SIGIR.

- Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern J. Philos.*, 26, 137-163.
- Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46, 159-216.
- Smucker, M. D., & Allan, J. (2005). *An investigation of dirichlet prior smoothing's performance advantage*. The University of Massachusetts, The Center for Intelligent Information Retrieval.
- Smucker, M. D., & Jethani, C. (2010). *Human performance and retrieval precision revisited*. Paper presented at the SIGIR.
- Sole, R. V., Murta, B., Valverde, S., & Steels, L. (2006). Language networks: Their structure, function, and evolution. *Trends in Cognitive Sciences*.
- Sordoni, A., Nie, J., & Bengio, Y. (2013). *Modeling term dependencies with quantum language models for IR*. Paper presented at the Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval (SIGIR '13).
- Sousa, A. O. (2005). Consensus formation on a triad scale-free network. *Physica A: Statistical Mechanics and its Applications*, 348, 701-710.
- Sparack Jones, K. (1972). A statistical interpretations of term specificity and its application in retrieval. *Journal of Documentation*, 28, 11-21.
- Sparck Jones, K. (1988). *A look back and a look forward*. Paper presented at the Proceedings of the 11th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval.

- Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.
- Sternberg, S. (1966). High-speed scanning in human memory. *Science*, 153, 421-457.
- Steyvers, M., & Griffiths, T. L. (2010). Rational analysis as a link between human memory and information retrieval *The Probabilistic Mind: Prospects from Rational Models of Cognition*: Oxford University Press.
- Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29, 41-78.
- Stigler, G. J. (1961). The economics of information. *J. Pol. Econ.*, 69, 213-225.
- Stolarz-Fantino, S., Fantion, E., Zizzo, D. J., & Wen, J. (2003). The conjunction effect: New evidence for robustness. *American Journal of Psychology*, 116(1), 15-34.
- Su, L. T. (1992). Evaluation measures for interactive information retrieval. *IP&M*, 28(4), 503-516.
- Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). *A core of semantic knowledge: Unifying WordNet and Wikipedia*. Paper presented at the In WWW-2007.
- Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). YAGO - A large ontology from Wikipedia and WordNet. *Journal of Semantic Web*, 6(3), 203-217.
- Sun, R. (2001). Artificial intelligence: Connectionist and symbolic approaches. *International Encyclopedia of the Social and Behavioral Sciences*, 783-789.
- Symonds, M. (2013). *A tensor encoding model of word meaning: Theory and application to Information Retrieval*. PhD, Queensland University of Technology.

- Teevan, J., Dumais, S. T., & Horvitz, E. (2005). *Beyond the commons: Investigating the value of personalizing Web search*. Paper presented at the Workshop on New Technologies for Personalized Information Access (PIA).
- Tentori, K., & Crupi, V. (2012). On the conjunction fallacy and the meaning of and, yet again: A reply to Hertwig, Benz, and Krauss (2008). *Cognition*, 122, 123-134.
- Terra, E., & Clarke, C. L. (2004). *Scoring missing terms in information retrieval tasks*. Paper presented at the Proceedings of the thirteenth ACM international conference on information and knowledge management.
- Till, R. E., Mross, E. F., & Kintsch, W. (1988). Time course of priming for associate and inference words in discourse context. *Memory and Cognition*, 16, 283-299.
- Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*: Cambridge University Press.
- Trotman, A., Puurula, A., & Burgess, B. (2014). *Improvements to BM25 and Language Models Examined*. Paper presented at the In Proceedings of the 2014 Australasian Document Computing Symposium.
- Turing, A. M. (1950a). Computing machinery and intelligence. *Mind*, 49, 433-460.
- Turing, A. M. (1950b). Computing machinery and intelligence. *Mind*, 59(236), 433-460.
- Turing, A. M. (1956). Can a machine think? *The world of mathematics*, 4.
- Turing, A. M. (1999). Can Digital Computers Think? In K. Furukawa, D. Michie & S. Muggleton (Eds.), *Machine Intelligence*: Oxford University Press.
- Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL.
- Turney, P. D. (2001a). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, 2167, 491-502.

- Turney, P. D. (2001b). *Mining the Web for synonyms: PMI-IR versus LSA on TOEFL*. Paper presented at the In proceedings of the twelfth european conference on machine learning (ECML-2001), Freiburg, Germany.
- Turpin, A., & Scholer, F. (2001). *Why batch and user evaluations do not give the same results*. Paper presented at the ACM SIGIR.
- Turpin, A., & Scholer, F. (2006). *User performance versus precision measures for simple search tasks*. Paper presented at the In SIGIR.
- Turtle, H. R., & Croft, W. B. (1990). *Inference networks for document retrieval*. Paper presented at the Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Brussels, Belgium.
- Turtle, H. R., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3), 187-222.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science*, 3, 305-309.
- Urbain, J., Goharian, N., & Frieder, O. (2005). *IIT TREC 2005: Genomics Track*. Paper presented at the TREC.
- Urbano, J., McFee, B., Downie, J. S., & Schedl, M. (2012). *How significant is statistically significant? The case of audio music similarity and retrieval*. Paper presented at the 13th international society for music information retrieval conference (ISMIR 2012).

- Van Maanen, L., & Marewski, J. N. (2009). *Recommender systems for literature selection: A competition between decision making and memory models*. Paper presented at the Proceedings of the 31st Annual Meeting of the Cognitive Science Society.
- Van Maanen, L., Van Rijn, H., Van Grootel, M., Kemna, S., Klomp, M., & Scholtens, E. (2009). Personal publication assistant: abstract recommendation by a cognitive model. *Cognitive Systems Research*.
- van Rijsbergen, C. J. (2004). *The Geometry of Information Retrieval*: Cambridge University Press.
- Vesely, M., Rajman, M., & Meur, J. L. (2008). *Using bibliographic knowledge for ranking in scientific publication databases*. Paper presented at the Proceedings of the 2008 conference on knowledge-based software engineering.
- Villacorta, A., & Jammalamadaka, S. R. (2009). Optimal text space representation of student essays using latent semantic analysis *Advances in Multivariate Statistical Methods*.
- Volz, E. (2004). Random networks with tunable degree distribution and clustering. *Physical Review E*, 70(5).
- von Neumann, J. (1958). *The computer and the brain*. New Haven, CT: Yale University Press.
- Waddington, K., & Holden, L. (1979). Optimal foraging: On flower selection by bees. *American Naturalist*, 114(179-196).
- Wang, Z., Solloway, T., & Busemeyer, J. (2013). *Empirical tests of a quantum probability model for question order effects found in survey research*. Paper presented at the The Annual Meeting of the Cognitive Science Society.

- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393, 440-442.
- Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, 107, 105-136.
- White, R. W., & Drucker, S. M. (2007). *Investigating behavioral variability in Web Search*. Paper presented at the In Proc. WWW.
- Widdows, D. (2003). *Orthogonal negation in vector spaces for modelling word meanings and document retrieval*. Paper presented at the In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan.
- Widdows, D., & Cohen, T. (2010). *The semantic vectors package: New algorithms and public tools for distributional semantics*. Paper presented at the Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010), Carnegie Mellon University, Pittsburgh, Pennsylvania.
- Widdows, D., & Ferraro, K. (2008). *Semantic vectors: a scalable open source package and online technology management application*. Paper presented at the Sixth International Conference on Language Resources and Evaluation.
- Widdows, D., & Peters, S. (2003). Word vectors and quantum logic: Experiments with negation and disjunction. *Mathematics of Language*, 8, 141-154.
- Wilkinson, R., & Hingston, P. (1991). *Using the cosine measure in a neural network for document retrieval*. Paper presented at the Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago, Illinois.

- Wilson, A., & Hendler, J. (1993). Linking symbolic and subsymbolic computing. *Connection Science*, 5, 395-414.
- Wilson, E. O. (1992). *Consilience*. New York: Knoph.
- Wilson, T. D. (2000). Human information behavior. *Informing Science*, 3(2), 49-55.
- Wilson, W. H., Street, D. J., & Halford, G. S. (1995). *Solving proportional analogy problems using tensor product networks with random representations*. Paper presented at the In Proceedings of the IEEE International Conference on Neural NETworks (ICNN' 95).
- Wittek, P., Koopman, B., Zuccon, G., & Daranyi, S. (2013). *Combining word semantics within complex hilbert space for information retrieval*. Paper presented at the 7th International Quantum Iteration Conference, Leicester, United Kingdom.
- Wong, S. K. M., Cai, Y. J., & Yao, Y. Y. (1993). *Computation of term associations by a neural network*. Paper presented at the Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA.
- Woodruff, A., Gossweiler, R., Pitkow, J. E., Chi, E. H., & Card, S. K. (2000). *Enhancing a digital book with a reading recommender*. Paper presented at the In proceedings of the SIGCHI conference on human factors in computing systems.
- Wu, M., Turpin, A., & Zobel, J. (2008). *An investigation on a community's web search variability*. Paper presented at the In Proceedings of Australian Computer Science Conference, Woollongong, Australia.



- Yan, K. K., & Gerstein, M. (2011). The spread of scientific information: insights from the Web usage statistics in PLOS article-level metrics. *PLOS One*, 6. doi: 10.1371/journal.pone.0019917
- Yaoyong, L., & Cunningham, H. (2008). *Geometric and quantum methods for information retrieval*. Paper presented at the ACM SIGIR Forum.
- Ye, Z., He, B., Huang, X., & Lin, H. (2010). Revisiting rocchio's relevance feedback algorithm for probabilistic models *Information Retrieval Technology*: Springer Berlin Heidelberg.
- Zaugg, B. H., West, R. E., Tateishi, I., & Randall, D. L. (2011). Mendeley: Creating communities of scholarly inquiry through research collaboration. *TechTrends*, 55(1), 32-36.
- Zellhofer, D., Frommholz, I., Schmitt, I., Lalmas, M., & van Rijsbergen, K. (2011). Towards quantum-based DB+ IR processing based on the principle of polyrepresentation *In Advances in Information Retrieval*: Springer Berlin Heidelberg.
- Zhai, C., & Lafferty, J. (2001a). *Model-based feedback in the language modeling approach to information retrieval*. Paper presented at the In Tenth International Conference on Information and Knowledge Magagement (CIKM 2001).
- Zhai, C., & Lafferty, J. (2001b). *A study of smoothing methods for language models applied to ad hoc information retrieval*. Paper presented at the In Proceedings of ACM SIGIR'01.
- Zhai, C., & Lafferty, J. (2001c). *A study of smoothing methods for language models applied to ad hoc information retrieval*. Paper presented at the In Proceedings of

the 24th annual international ACM SIGIR conference on research and development in information retrieval, ACM.

Zhai, C., & Lafferty, J. (2002). *Two-stage language models for information retrieval*.

Paper presented at the Proceedings of ACM SIGIR.

Zhai, C. X. (2008). Statistical language models for information retrieval: a critical review. *Foundations and trends in information retrieval*, 2(3), 137-213.

Zhang, P., Song, D., Zhao, X., & Hou, Y. (2011). *Investigating query-drift problem from a novel perspective of photon polarization*. Paper presented at the ICTIR.

Zhang, S., Farooq, U., & Carroll, J. M. (2009). *Enhancing information scent: Identifying and recommending quality tags*. Paper presented at the Proc ACM Support Group Work.

Zhao, J., Huang, X., Ye, Z., & Zhu, J. (2009). *York University at TREC 2009: Chemical track*. Paper presented at the Proc. of TREC 2009.

Zhu, J., Song, D., & Ruger, S. (2009). Integrating multiple windows and document features for expert finding. *Journal of the American Society for Information Science and Technology*, 60(4), 694-715.

Zighele, L., & Kurland, O. (2008). *Query-drift prevention for robust query expansion*.

Paper presented at the SIGIR.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*: Addison-Wesley.

Zucco, G., & Azzopardi, L. (2010). Using the quantum probability ranking principle to rank interdependent documents. *Lecture notes in computer science: Advances in information retrieval*, 5993, 357-369.

Zuccon, G., Azzopardi, L. A., & van Rijsbergen, K. (2009). The quantum probability ranking principle for information retrieval. *Lecture notes in computer science: Advances in information retrieval theory*, 5766, 232-240.

Zuccon, G., Piwowarski, B., & Azzopardi, L. (2011). *On the use of complex numbers in quantum models for information retrieval*. Paper presented at the In Proc. of ICTIR, Bertinoro, Italy.