ONLINE FORUM THREAD RETRIEVAL USING DATA FUSION

AMEER TAWFIK ABDULLAH ALBAHEM

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

SEPTEMBER 2013

To my wife and parents

ACKNOWLEDGEMENT

First of all, all praise to Allah for giving me the strength and the patience to complete this task. My supervisor, Prof Naomie Salim, thanks for being a family member rather than an academic advisor. Your unlimited support in various aspects of my study has been a corner stone on the success of this research.

Part of the thesis would have not been completed without the valuable advice from Jangwon Seo at Center for Intelligent Information Retrieval, University of Massachusetts, Amherst. Using the corpus that developed by Sumit Bhatia from the Pennsylvania State University has enabled conducting the thesis experiments. Thank you Sumit, your collaboration is very much appreciated..

This work would have not seen the light without the scholarship provided by the Yemeni Ministry of High Education and Scientific Research. In addition, I would like to thank the Malaysian Ministry of Higher Education (MOHE) and the Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) for sponsoring the publication of this research.

It just has been a routine to make the family members the last ones to thank; however, my parents, family members and friends, thank you for your support.

My wife, I am speechless when it comes to thanking you. Your support, caring and sacrifice have been beyond what an ordinary person would do for his partner. What you have given me is just immeasurable, thank you!

Ameer Tawfik Albaham, Malaysia

ABSTRACT

Online forums empower people to seek and share information via discussion threads. However, finding threads satisfying a user information need is a daunting task due to information overload. In addition, traditional retrieval techniques do not suit the unique structure of threads because thread retrieval returns threads, whereas traditional retrieval techniques return text messages. A few representations have been proposed to address this problem; and, in some representations aggregating query relevance evidence is an essential step. This thesis proposes several data fusion techniques to aggregate evidence of relevance within and across thread representations. In that regard, this thesis has three contributions. Firstly, this work adapts the Voting Model from the expert finding task to thread retrieval. The adapted Voting Model approaches thread retrieval as a voting process. It ranks a list of messages, then it groups messages based on their parent threads; also, it treats each ranked message as a vote supporting the relevance of its parent thread. To rank parent threads, a data fusion technique aggregates evidence from threads' ranked messages. Secondly, this study proposes two extensions of the voting model: Top K and Balanced Top K voting models. The Top K model aggregates evidence from only the top K ranked messages from each thread. The Balanced Top K model adds a number of artificial ranked messages to compensate the difference if a thread has less than K ranked messages (a padding step). Experiments with these voting models and thirteen data fusion methods reveal that summing relevance scores of the top K ranked messages from each thread with the padding step outperforms the state of the art on all measures on two datasets. The third contribution of this thesis is a multi-representation thread retrieval using data fusion techniques. In contrast to the Voting Model, data fusion methods were used to fuse several ranked lists of threads instead of a single ranked list of messages. The thread lists were generated by five retrieval methods based on various thread representations; the Voting Model is one of them. The first three methods assume a message to be the unit of indexing, while the latter two assume the title and the concatenation of the thread message texts to be the units of indexing respectively. A thorough evaluation of the performance of data fusion techniques in fusing various combinations of thread representations was conducted. The experimental results show that using the sum of relevance scores or the sum of relevance scores multiplied by the number of retrieving methods to develop multi-representation thread retrieval improves performance and outperforms all individual representations.

ABSTRAK

Forum dalam talian membolehkan pengguna mencari dan berkongsi maklumat melalui benang perbincangan. Walau bagaimanapun, pencarian benang perbincangan adalah satu tugas yang bukan mudah disebabkan oleh beban maklumat. Disamping itu, teknik dapatan semula tradisional tidak sesuai dengan struktur unik benang perbincangan kerana dapatan semula benang mengembalikan benang, sementara teknik dapatan semula tradisional mengembalikan mesej teks. Beberapa perwakilan telah dicadangkan; dan mengagregat bukti relevansi maklumat carian merupakan satu langkah penting. Tesis ini mencadangkan beberapa teknik gabungan data untuk mengagregat bukti relevansi perwakilan benang perbincangan. Tesis ini mempunyai tiga sumbangan. Pertama, kerja ini mengadaptasi model undian dari tugas carian pakar kepada dapatan semula benang perbincangan. Kesesuaian Model Undian mendekati dapatan semula benang perbincangan sebagai satu proses undian. Ia memberi susunan kedudukan kepada senarai mesej, dan kemudian mengumpulkan mesej berdasarkan benang perbincangan induk mereka; ia juga bertindak pada setiap susunan mesej perbincangan sebagai undi yang menyokong kaitan benang induk. Untuk mendapatkan susunan kedudukan benang perbincangan induk, teknik gabungan data mengagregat bukti dari mesej benang perbincangan. Kedua, kajian ini mencadangkan dua lanjutan model undian: K-Teratas dan K-Teratas Seimbang model undian. Model K-Teratas mengagregat bukti hanya daripada K mesej tertinggi. Model K-Teratas Seimbang menambah sesuatu susunan mesej nombor untuk mengimbangi perbezaan jika benang perbincangan mempunyai kurang daripada K mesej tertinggi (langkah tambahan). Melalui kajian dengan Model Undian dan 13 kaedah gabungan data, keputusan menunjukkan bahawa penjumlahan skor dari K mesej tertinggi dari setiap benang perbincangan dengan langkah tambahan mengatasi kaedah semasa dalam semua penilaian ke atas dua set data. Sumbangan ketiga tesis ini adalah dapatan multi-perwakilan benang perbincangan menggunakan teknik gabungan data. Berbeza dengan Model Undian, kaedah gabungan data telah digunakan untuk menggabungkan beberapa senarai benang perbincangan dan bukannya satu senarai mesej. Senarai benang perbincangan telah dihasilkan oleh lima model dapatan semula berdasarkan pelbagai perwakilan, antaranya Model Undian. Tiga kaedah yang pertama menganggap mesej sebagai unit pengindeksan, manakala dua kaedah yang terakhir menggunakan tajuk dan gabungan teks mesej benang perbincangannya. Penilaian yang menyeluruh ke atas gabungan pelbagai kombinasi perwakilan benang perbincangan telah dijalankan. Keputusan ujikaji menunjukkan bahawa menggunakan jumlah skor relevan atau jumlah skor relevan didarab dengan bilangan kaedah dapatan untuk membangunkan multi-perwakilan dapatan semula benang perbincangan boleh meningkatkan prestasi dan mengatasi semua perwakilan individu.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MLE	–	Maximum Likelihood Estimation
P@10	–	Precision at 10 retrieved documents
MAP	–	Mean Average Precision
MRR	–	Mean Reciprocal Rank
NDCG@10	–	Normalized Discounted Cumulative Gain at 10 retrieved documents.
VD	–	Virtual document
PCS	–	Pseudo Cluster Selection method

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Online forums are platforms that facilitate discussion and knowledge sharing in the Web. In forums, a discussion starts when a user posts an initial message requesting for a help or initiating a conversation in a particular matter. Afterwards, the other users read and reply to the initial message. These replies are called reply messages. Each pair of an initial message and its replies forms a thread. Threads that address similar themes or topics are grouped into a sub forum. A collection of sub forums builds the entire forums.

Online forums are rich knowledge communities for several reasons. First, the asynchronous nature and the public accessibility of forums enable communication between community members regardless of the physical and the temporal boundaries. That empowers users with various areas and levels of expertise to share and seek knowledge through in depth discussions. Second, forums have accumulated a huge amount of content for a long time. For instance, Table 1.1 presents several examples of forums that have accumulated content over at least five years. Third, the archived content contains not only factual information but also detailed solutions and troubleshooting content (Bhatia and Mitra, 2010; Seo et al., 2011). In addition, the content is more comprehensive and objective than what might be found in web pages (Seo et al., 2011).

Nevertheless, that knowledge is not fully utilized. Forums' presentation and browsing tools are limited. Every day, forums members contribute new threads.

Table 1.1: Examples of forums and their content types

Forum	Type	Statistics
Ubuntu Forums	Official Ubuntu Forums	1.8 million threads, 1.7 million users, started on 2004 (Ubuntu Forums, 2013)
Lowyat.net	Malaysia's Largest Online Community	1.3 million threads, 0.5 million users, started on 2003 (Lowyat.net, 2013)
Body Building.com	Boding building related content	4.8 million threads, 5.2 million users, started on 2000 (Body Building.com, 2013)
Breast Cancer	Breast Cancer dedicated forums	0.95 million threads, 0.11 million users, started prior to 2008 (Breast Cancer Community, 2013)
Christian Forums	Faith and beliefs related content	7.69 million threads, 0.31 million users, started on 2003 (Christian Forums, 2013)

Unfortunately, the new threads are visible only for a short duration. Online forums list threads in a reverse and chronological order; newer threads push older threads down in the list. Figure 1.1 shows an example of a thread listing consists of ten thread entries. As shown in the upper right corner of the figure, the number of pages is huge. Therefore, it is impossible to navigate through threads to find a particular information. Although a user might manage to examine the thread listing in small forums, threads' titles do not convey threads' contents (Bhatia and Mitra, 2010). In addition, as shown in the figure, the information displayed for each thread entry is rather limited. Therefore, users need a search facility to find content satisfying their information needs.

Forums provide three methods to search content. The first method is a keywords matching. This method is limited because it is based on the Boolean model (Elsas, 2011b). The second method is a message search using a database backend full text search. This method is not adequate because it causes confusion. The reason is that a database engine indexes messages rather than threads. Therefore, when the database search engine matches a query against messages, it calculates the relevance score between all messages in the message table and the query; then, it returns the highest scoring messages to users. Returning a single message might confuse the user because it might be taken out of the discussion context, especially if it addresses other

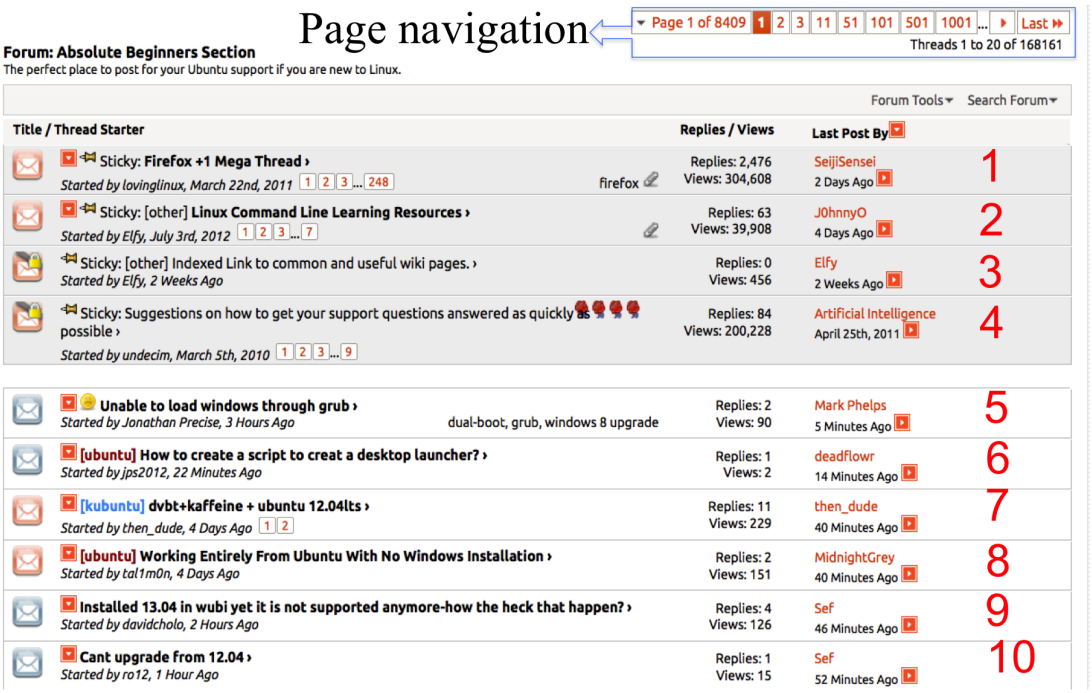

Figure 1.1: An example of a thread listing

reply messages in a thread. Lastly, the third method is searching using commercial web search engines. However, this method is not adequate due to the different structure and nature of retrieval in online forums (Seo et al., 2009; Bhatia and Mitra, 2010). In addition to that, web search engines index only publicly accessible content. Sometimes, some contents are provided only to registered users. Therefore, a built-in search engine is needed to leverage such contents.

Thread retrieval is one tool that enables searching content on forums. However, thread retrieval is not a trivial task because a thread is not the unit of text; the unit of text is the thread message. Researches in thread retrieval (Elsas and Carbonell, 2009; Seo et al., 2009; Bhatia and Mitra, 2010; Seo, 2011; Elsas, 2011b) have proposed several solutions to tackle this problem. These solutions leverage different thread structures and ways to fuse evidence within and across these structures. For instance, Elsas and Carbonell (2009) represented threads as collections of messages. Then, to rank threads, the authors fused the individual message relevance scores. Each individual message relevance score is an evidence of thread relevance. Aggregating evidence from messages is an example of fusing evidence within a structure or a representation. An example of aggregating evidence across structures or representations is proposed by (Seo et al., 2009). Here, the authors proposed two representations of threads. The

first is representing a thread as concatenation of its message texts. The second is representing a thread as a set of messages. To rank threads, a weighted product of the thread relevance scores based on the individual representations was used. Each representation based relevance score is considered as an evidence. This thesis focus is using data fusion techniques to combine relevance evidence in the above scenarios.

Data fusion is related to the concept of fusing several inputs related to a certain object into a useful output (Wikipedia, 2013b). In the context of information retrieval, the terms "data fusion" and "meta search" have been used interchangeably to refer to the techniques that fuse several ranked lists of documents (Aslam and Montague, 2001). Based on what a data fusion technique aggregates about a document in the ranked lists, it can be classified as a rank or a score based technique.

A score based method aggregates the relevance scores of a document. Some seminal score based methods are CombMAX, CombSUM and CombANZ (Shaw and Fox, 1994). CombMAX assigns the maximum of relevance scores as the document final score. CombSUM, and CombANZ assign the sum and the average of scores as the document final score.

A rank based method fuse the ranking positions of documents on the ranked lists such as the Reciprocal Rank (RR) (Zhang et al., 2002) method. For instance, if there are two ranked lists of documents, and a document was ranked first in the first list and fourth in the second list, then the document's final score using RR is equal to $1/1 + 1/4 = 1.25$.

Many data fusion inspired solutions have been used to combine evidence of relevance in many search tasks that share resemblance to thread retrieval such as ranking aggregates tasks (Macdonald and Ounis, 2008a,b, 2011) and multi-representation search tasks (Ogilvie and Callan, 2003; Wu et al., 2012). Because of that, this thesis develops several evidence combination methods for thread retrieval based on the data fusion techniques.

1.2 Problem Background

One of the problems in thread retrieval is that the retrieval, and the text units are different. In thread retrieval, the text units are the messages, whereas a user expects a ranked list of threads. This section reviews solutions to this problem.

1.2.1 Text and Retrieval Units Mismatch

Elsas and Carbonell (2009) proposed two strategies to tackle the mismatch problem: inclusive and selective. The inclusive strategy utilizes all messages ranking parent threads. Two models from previous work on blog feed retrieval (Elsas et al., 2008) were adapted to thread search: the large and the small document models. The large document model creates a virtual document for each thread by concatenating the thread message texts, then it scores threads based on their virtual document relevance to the query. In contrast, the small document model defines a thread as a collection of text units (messages). Then, it scores threads by averaging their message relevance scores. In the selective strategy, Elsas and Carbonell (2009) treated threads as collections of messages and used only few messages to rank threads. Three selective methods were used. The first one is scoring threads using only the initial message relevance score. The second method scores threads by taking the maximum score of their message relevance scores. The third method is based on the Pseudo Cluster Selection method (PCS) (Seo and Croft, 2008). PCS scores threads in two steps: it retrieves an initial ranked list of messages, then it ranks threads by taking the geometric mean of the top K ranked messages' scores from each thread. If a thread has less than K messages, PCS adds a padding message whose value is the relevance value of the minimum scoring message on the initial ranked list. Generally, it was found that the selective models are statistically superior to the inclusive models (Elsas and Carbonell, 2009; Elsas, 2011a). In addition, PCS is superior to all methods.

Nevertheless, PCS was not only applied to fuse message relevance scores, but also to fuse relevance scores of other types of text units. For instance, Seo et al. (2009) treated a thread as a collection of several local contexts. A local context is a self-contained text unit. Seo et al. (2009) proposed four contexts: posts, pairs, dialogues and the entire thread. The thread and the post contexts are identical to the virtual

document, and the message based representations (Elsas and Carbonell, 2009). In the pair and the dialogue contexts, the conversational relationship between messages is exploited to build text units. To rank threads using a particular local context, the authors then ranked a list of local contexts, then they ranked threads by applying PCS to the threads' ranked contexts' relevance scores. It was observed that the retrieval using the dialogue context outperformed retrieval using other contexts. In summary, regardless the definition of text unit, two steps were involved to rank threads: rank a list of text units based on their relevance to the user query, and then score threads by aggregating their ranked text units' relevance scores.

Ranking text units in order to rank their associated objects is well studied in ranking aggregates tasks (Macdonald and Ounis, 2011). Examples of these tasks are: experts finding (Balog et al., 2012) and blog feed topic distillation (Santos et al., 2012). The expert finding task is defined as retrieving a list of people who are experts on the topic of the user query (Macdonald and Ounis, 2008b). To estimate the expertise of a person, methods in this task leverage the documents that are associated with, or written by that person (Macdonald and Ounis, 2008b). The blog topic distillation task focuses on finding blogs that have recurring interest on the topic of the user query (Santos et al., 2012). To estimate this aspect, methods in this task leverage the blogs' postings.

Macdonald and Ounis (2008b) (Voting Model) modeled the problem of expert finding as a voting process. Motivated by the success of the Voting Model in the expert finding task, Macdonald and Ounis (2008a) modeled the blog topic distillation task as a voting process as well. In both tasks, the Voting Model first ranks a list of documents (the people's documents or the blogs' postings) with respect to a user query using an underlined text retrieval model, then it considers each ranked document as a vote supporting the relevance of its associated object (the person or the blog). Then, data fusion techniques were used to aggregate these votes. However, instead of fusing several ranked lists, there was only one list. In addition, the relevance scores or the ranking positions of the documents are not fused to rank documents but to rank the associated expert or blog. Indeed, in both tasks, the voting approach was found to be statistically superior to baseline methods (Macdonald and Ounis, 2008b,a). However, the performance of each voting technique was not consistent across tasks and datasets (Macdonald and Ounis, 2011): the CombMAX method, which performed well on the expert finding setting, was significantly worse than the other voting methods on the blog distillation setting (Macdonald and Ounis, 2008a).

This inconsistency of CombMAX's performance is due to the differences on the nature of documents and the definition of relevance on both tasks (Macdonald and Ounis, 2008a, 2011; Santos et al., 2012). Such factors are also different in thread retrieval as well. For instance, the average length of a thread message is much shorter than the average length of a blog posting (Elsas, 2011b). In addition, a thread message is topically related to its parent thread's messages; hence, understanding the message content might not be possible without looking at the thread discussion context. In contrast, in a blog, the postings are written independently from each other; the purpose of a blog posting is to share an experience or a thought the blog owner has. In fact, a major issue in blog topic distillation is how to model the topic diversity of postings while ranking blogs (Seo and Croft, 2008; Santos et al., 2012). In addition, blog postings are written mostly by a single author. However, in threads, the messages are written by several authors who vary on their writing styles, skills and expertise. Furthermore, in thread retrieval, a thread is considered relevant if it is topically relevant to the user query. However, in blog topic distillation, whereas a blog is relevant only if it has a recurring interest though it might have a posting that is relevant (Macdonald and Ounis, 2008a; Seo and Croft, 2008).

As compared to expert finding, a thread message belongs only to a single thread, whereas a document might be associated with several people. In addition, the association of documents to people is a challenge in expert finding (Balog et al., 2012), but this is not the case on thread retrieval.

In summary, thread retrieval resembles expert finding and blog topic distillation in the general problem, but it differs from them on several aspects such as the thread physical and conversational structure and the notion of a relevant thread. Furthermore, although the maximum score method (Elsas and Carbonell, 2009) is similar to the CombMAX technique from the Voting Model (Macdonald and Ounis, 2008b,a), it is only one method from the twelve aggregation methods employed by the Voting Model. In addition, this particular method has shown inconsistent performance as discussed above. Therefore, in thread retrieval, the Voting Model will provide a general approach to tackle the mismatch between the text and the retrieval units. That is it ranks an initial list of messages, then it considers each ranked message as a vote supporting the relevance of its parent thread. However, the performance of the aggregation (voting) techniques is still an open question.

In connection to previous works in thread retrieval, Elsas (2011a) reported that ranking threads using the maximum score of their message relevance scores is superior to using the average; and, ranking threads using the Pseudo Cluster Selection method (PCS) (Seo and Croft, 2008; Elsas and Carbonell, 2009) is statistically superior to both methods. The average score method will be affected by the messages with low scores, whereas the maximum score method favors threads with highly ranked messages. In addition, the maximum score method is a special case of PCS(K = 1) (Elsas and Carbonell, 2009; Elsas, 2011a). In other words, focusing on highly ranked messages improve the retrieval performance. In the ranking aggregates tasks introduced previously (Macdonald and Ounis, 2008b,a), voting techniques, other than CombMAX, which promote aggregates with highly ranked documents, were found to perform consistently good. Therefore, applying the Voting Model to fuse evidence from only the top K ranked documents might improve performance.

1.2.2 Leveraging Multiple Representations

Researches in thread retrieval have focused on combining evidence of relevance not only within the same representation but also across representations. For instance, Seo et al. (2009) proposed a weighted product (non-linear interpolation) between the relevance score using the thread context and the relevance score using the message, the pair or the dialogue context. It was reported that the weighted product was better than the individual contexts. Seo (2011) used the learning to rank approach (Liu, 2011) to combine evidence of relevance from the thread, the message and the dialogue contexts. Bhatia and Mitra (2010) defined a thread as a structured document: a document that consists of three small text units, which are the title, the initial message and the reply messages set; then, the query terms were given different weights based on where these terms appear. Generally, these models (Seo et al., 2009; Seo, 2011; Bhatia and Mitra, 2010) are examples of retrieval using multiple representations. In addition, they require training in order to obtain good results.

The representation proposed by Elsas and Carbonell (2009) were subsumed by the non-linear interpolation of the virtual document representation and the message context representation introduced in (Seo et al., 2009, 2011). However, none of the thread representations proposed in Elsas and Carbonell (2009) and Seo et al. (2009) were integrated by (Bhatia and Mitra, 2010). Similarly, the structural components

of Bhatia and Mitra (2010) were not considered by (Seo, 2011). One reason for the exclusion of some methods across these studies is the lacking of a general framework that combines all representations. Many of these representations used different approaches on exploiting thread structure. As a result, this thesis proposes to use meta search techniques to combine these representations. A meta search technique consumes the outputs of these representations, which are ranked lists of threads, hence combining these representations becomes feasible. Furthermore, many meta search techniques do not require training. Considering that online forums vary in their abilities to construct training models, using meta search techniques to combine various thread representations is a better approach.

In the literature of information retrieval, Ogilvie and Callan (2003) and Wu et al. (2012) identified two approaches to retrieval using multiple representations: within search and after search. To estimate the relevance of a document, the within search based methods use a supervised approach on fusing the relevance scores coming from each representation. In other words, the within search methods require training to work effectively. The after search methods used meta search techniques to fuse ranked lists of documents that were generated by retrieval models based on the individual representations. It was found that both approaches are competitive to each other on two search tasks: the known item search (Ogilvie and Callan, 2003; Wu et al., 2012) and the web topic distillation (Wu et al., 2012).

In summary, this thesis proposes several data fusion inspired techniques to improve thread retrieval. A group of these techniques focus on the combination of evidence within the same representation such as fusing the message relevance scores when treating a thread as a collection of messages. The other group of techniques address the combination of relevance evidence from different thread representations such as combining evidence from the message and the virtual document based representations. Throughout this thesis, the term ”Voting Model” refers to methods from the first group, while the term ”meta search techniques” refers to methods from the second group. In addition, a voting technique fuses evidence from a single ranked list of messages, whereas a meta search technique aggregates evidence from multiple ranked lists of threads.

1.3 Problem Statement

Based on the discussion presented in the previous section, this thesis hypothesizes that data fusion inspired techniques such as the voting and the meta search techniques can improve evidence combination on thread retrieval. To test this hypothesis, the thesis investigates the following questions:

RQ 1: Will the Voting Model improves thread retrieval? Which voting technique is the best?

RQ 2: Will the combination of the Voting Model with the Pseudo Cluster Selection method improve the performance of the voting techniques? If yes, is the improvement consistent on all techniques?

RQ 3: Will the combination of multiple thread representations using meta search techniques improve retrieval? Which meta search technique is the best?

RQ 4: Is it possible to improve retrieval using a few representations? If yes, what are these representations?

1.4 Objectives

The following are the objectives of this thesis:

1. To adapt the Voting Model to thread retrieval and evaluate its performance.

2. To combine the Voting Model with the Pseudo Cluster Selection method.

3. To develop multi-representation thread retrieval using meta search techniques.

1.5 Scope

The following are the scopes of this thesis:

- The major focus of this work is fusing relevance evidence within and across structures.
- Only query dependent relevance estimates are fused. No document priors are assumed.
- In developing thread retrieval methods, only methods that use thread content are used. In addition, the methods that require recovering of thread structure are not considered due to the instability of retrieval using inaccurate thread structure (Seo et al., 2009; Seo, 2011).
- Only the data fusion methods that do not require training are used to enable the implementation of these methods on web forums; where training resources are not available.
- All experiments are conducted using a corpus that was used by (Bhatia and Mitra, 2010).
- All retrieval methods are benchmarked using the standard Ad Hoc retrieval evaluation measures. These measures are Precision at the top 10 retrieved documents (P@10), Normalized Discounted Cumulative Gain at 10 (NDCG@10), Mean Reciprocal Rank(MRR) and Mean Average Precision (MAP) (Mark, 2010).

1.6 Thesis Structure

The remaining of this thesis consists of five chapters. Chapter 2 presents the academic literature in which this thesis is built upon. In particular, it reviews related concepts to forums, information retrieval and data fusion applications on information retrieval. It then reviews related works to thread retrieval and identifies research gaps in which data fusion could fill in.

Chapter 3 details the thesis methodology in addressing the research questions and achieving the objectives. Chapter 4 presents the first contribution of the thesis, which is adapting the Voting Model to thread retrieval and evaluating its performance. Chapter 5 presents the thesis second contribution which is extending the Voting Model by borrowing ideas from the Pseudo Cluster Selection method. Chapter 6 presents the third contribution: a meta search approach to rank threads using multiple representations. Chapter 7 reasserts the main findings of this thesis and outlines the potential future works.

REFERENCES

G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, Oct. 2002. ISSN 1046-8188. doi: 10.1145/582415.582416. URL http://doi.acm.org/10.1145/582415.582416.

J. A. Aslam and M. Montague. Models for metasearch. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 276–284, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6. doi: 10.1145/383952.384007. URL http://doi.acm.org/10.1145/383952.384007.

K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 43–50, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148181. URL http://doi.acm.org/10.1145/1148170.1148181.

K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *Information Processing & Management*, 45(1):1 – 19, 2009. ISSN 0306-4573. doi: 10.1016/j.ipm.2008.06.003. URL http://www.sciencedirect.com/science/article/pii/S0306457308000678.

K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *Found. Trends Inf. Retr.*, 6(2):127–256, Feb. 2012. ISSN 1554-0669. doi: 10.1561/1500000024. URL http://dx.doi.org/10.1561/1500000024.

N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. The effect multiple query representations on information retrieval system performance. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 339–346, New York, NY, USA, 1993. ACM. ISBN 0-89791-605-0. doi: 10.1145/160688.160760. URL http://doi.acm.org/10.1145/160688.160760.

M. Bendersky and O. Kurland. Utilizing passage-based language models for ad hoc document retrieval. *Information Retrieval*, 13:157–187, 2010. ISSN 1386-4564. URL http://dx.doi.org/10.1007/s10791-009-9118-8.

10.1007/s10791-009-9118-8.

M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 95–104, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935849. URL http://doi.acm.org/10.1145/1935826.1935849.

S. Bhatia and P. Mitra. Adopting inference networks for online thread retrieval. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1300–1305, Atlanta, Georgia, USA,, July 11–15 2010.

Body Building.com. Bodybuilding.com forums - bodybuilding and fitness board. http://forum.bodybuilding.com, 2013. [Online; accessed 25-January-2013].

Breast Cancer Community. Breast cancer discussion forums - access the shared knowledge of thousands of people affected by breast cancer. http://community.breastcancer.org, 2013. [Online; accessed 25-January-2013].

C. Buckley. Text retrieval conference (trec) trec eval. http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz, 2013. [Online; accessed 25-January-2013].

J. Callan. Distributed information retrieval. In *In: Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, 2000.

J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 302–310, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X. URL http://dl.acm.org/citation.cfm?id=188490.188589.

Cancun Travel Forums. Cancun forum, travel discussion for cancun, mexico – tripadvisor. http://www.tripadvisor.com/ShowForum-g150807-i8-Cancun_Yucatan_Peninsula.html, 2013. [Online; accessed 25-January-2013].

Christian Forums. Christian forums - where christian community meets faith. http://www.christianforums.com, 2013. [Online; accessed 25-January-2013].

W. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Alternative Etext Formats. ADDISON WESLEY Publishing Company Incorporated, 2010. ISBN 9780136072249. URL http://books.google.com.my/books?id=VVYAPgAACAAJ.

W. B. Croft. Knowledge-based and statistical approaches to text retrieval. *IEEE Expert: Intelligent Systems and Their Applications*, 8(2):8–12, Apr. 1993. ISSN 0885-9000. doi: 10.1109/64.207424. URL http://dx.doi.org/10.1109/64.207424.

P. Das-Gupta and J. Katzer. A study of the overlap among document representations. In *Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '83, pages 106–114, New York, NY, USA, 1983. ACM. ISBN 0-89791-107-5. doi: 10.1145/511793.511809. URL http://doi.acm.org/10.1145/511793.511809.

J. L. Elsas. Ancestry.com online forum test collection. Technical Report CMU-LTI-017, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2011a.

J. L. Elsas. *Leveraging Collection Structure in Information Retrieval With Applications to Search in Conversational Social Media*. PhD thesis, Carnegie Mellon University, 2011b.

J. L. Elsas and J. G. Carbonell. It pays to be picky: an evaluation of thread retrieval in online forums. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 714–715, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572092. URL http://doi.acm.org/10.1145/1571941.1572092.

J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog distillation. In *TREC*, 2007.

J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 347–354, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4. doi: 10.1145/1390334.1390394. URL http://doi.acm.org/10.1145/1390334.1390394.

Indri. Lemur project components: Indri. http://sourceforge.net/projects/lemur/files/lemur/indri-5.4, 2013. [Online; accessed 25-January-2013].

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL http://doi.acm.org/10.1145/582415.582418.

K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 36 (6):779–808, Nov. 2000. ISSN 0306-4573. doi: 10.1016/S0306-4573(00)00015-7. URL http://dx.doi.org/10.1016/S0306-4573(00)00015-7.

P. B. Kantor and E. M. Voorhees. The trec-5 confusion track: Comparing retrieval methods for scanned text. *Information Retrieval*, 2:165–176, 2000. ISSN 1386-4564. URL http://dx.doi.org/10.1023/A:1009902609570. 10.1023/A:1009902609570.

M. Kaszkiel and J. Zobel. Passage retrieval revisited. In *Proceedings of the 20th annual*

international ACM SIGIR conference on Research and development in information retrieval, SIGIR '97, pages 178–185, New York, NY, USA, 1997. ACM. ISBN 0-89791-836-3. doi: 10.1145/258525.258561. URL http://doi.acm.org/10.1145/258525.258561.

S. N. Kim, L. Wang, and T. Baldwin. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 192–202, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-83-1. URL http://dl.acm.org/citation.cfm?id=1870568.1870591.

R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 191–202, New York, NY, USA, 1993. ACM. ISBN 0-89791-605-0. doi: 10.1145/160688.160718. URL http://doi.acm.org/10.1145/160688.160718.

J. H. Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '97, pages 267–276, New York, NY, USA, 1997. ACM. ISBN 0-89791-836-3. doi: 10.1145/258525.258587. URL http://doi.acm.org/10.1145/258525.258587.

Y. Lee, S.-H. Na, and J.-H. Lee. Utilizing local evidence for blog feed search. *Information Retrieval*, 15:157–177, 2012. ISSN 1386-4564. URL http://dx.doi.org/10.1007/s10791-011-9176-6. 10.1007/s10791-011-9176-6.

T.-Y. Liu. *Learning to Rank for Information Retrieval*. Springer Berlin Heidelberg, 2011.

X. Liu and W. B. Croft. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 375–382, New York, NY, USA, 2002. ACM. ISBN 1-58113-492-4. doi: 10.1145/584792.584854. URL http://doi.acm.org/10.1145/584792.584854.

X. Liu and W. B. Croft. Evaluating text representations for retrieval of the best group of documents. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 454–462, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 3-540-78645-7, 978-3-540-78645-0. URL http://dl.acm.org/citation.cfm?id=1793274.1793330.

Lowyat.net. Lowyat.net - insanely addictive malaysia forum. http://forum.lowyat.net, 2013. [Online; accessed 25-January-2013].

C. Macdonald. *The Voting Model for People Search*. PhD thesis, University of Glasgow, 2009.

C. Macdonald and I. Ounis. Key blog distillation: ranking aggregates. In *Proceedings*

of the 17th ACM conference on Information and knowledge management, CIKM '08, pages 1043–1052, New York, NY, USA, 2008a. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458221. URL http://doi.acm.org/10.1145/1458082.1458221.

C. Macdonald and I. Ounis. Voting techniques for expert search. *Knowl. Inf. Syst.*, 16(3):259–280, Aug. 2008b. ISSN 0219-1377. doi: 10.1007/s10115-007-0105-3. URL http://dx.doi.org/10.1007/s10115-007-0105-3.

C. Macdonald and I. Ounis. Learning models for ranking aggregates. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 517–529, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-20160-8. URL http://dl.acm.org/citation.cfm?id=1996889.1996957.

C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

S. Mark. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4, 2010.

D. Metzler and O. Kurland. Experimental methods for information retrieval. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1185–1186, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348534. URL http://doi.acm.org/10.1145/2348283.2348534.

M. Montague and J. A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 538–548, New York, NY, USA, 2002. ACM. ISBN 1-58113-492-4. doi: 10.1145/584792.584881. URL http://doi.acm.org/10.1145/584792.584881.

P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 143–150, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3. doi: 10.1145/860435.860463. URL http://doi.acm.org/10.1145/860435.860463.

phpBB. phpbb - free and open source forum software. https://github.com/phpbb/phpbb3/blob/develop/phpBB/includes/search/fulltext_mysql.php, 2013. [Online; accessed 25-January-2013].

J. M. Ponte. *A language modeling approach to information retrieval*. PhD thesis, UNIVERSITY OF MASSACHUSETTS AMHERST, 1998.

J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York,

NY, USA, 1998. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.291008. URL http://doi.acm.org/10.1145/290941.291008.

M. F. Porter. An algorithm for suffix stripping. In K. Sparck Jones and P. Willett, editors, *Readings in information retrieval*, chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1-55860-454-5. URL http://dl.acm.org/citation.cfm?id=275537.275705.

S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 42–49, New York, NY, USA, 2004. ACM. ISBN 1-58113-874-1. doi: 10.1145/1031171.1031181. URL http://doi.acm.org/10.1145/1031171.1031181.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL http://doi.acm.org/10.1145/361219.361220.

G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 49–58, New York, NY, USA, 1993. ACM. ISBN 0-89791-605-0. doi: 10.1145/160688.160693. URL http://doi.acm.org/10.1145/160688.160693.

R. L. T. Santos, C. Macdonald, R. McCreadie, I. Ounis, and I. Soboroff. Information retrieval on the blogosphere. *Found. Trends Inf. Retr.*, 6(1):1–125, Jan. 2012. ISSN 1554-0669. doi: 10.1561/1500000026. URL http://dx.doi.org/10.1561/1500000026.

J. Seo. *Search Using Social Media Structures*. PhD thesis, UNIVERSITY OF MASSACHUSETTS AMHERST, 2011.

J. Seo and W. B. Croft. Homepage search in blog collections. IR IR-570, UNIVERSITY OF MASSACHUSETTS AMHERST, 2007.

J. Seo and W. B. Croft. Blog site search using resource selection. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1053–1062, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458222. URL http://doi.acm.org/10.1145/1458082.1458222.

J. Seo, W. B. Croft, and D. A. Smith. Online community search using thread structure. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1907–1910, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646262. URL http://doi.acm.org/10.1145/1645953.1646262.

J. Seo, W. Bruce Croft, and D. Smith. Online community search using conversational

structures. *Information Retrieval*, 14:547–571, 2011. ISSN 1386-4564. URL http://dx.doi.org/10.1007/s10791-011-9166-8. 10.1007/s10791-011-9166-8.

J. A. Shaw and E. A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.

M. Shokouhi and L. Si. Federated search. *Found. Trends Inf. Retr.*, 5(1):1–102, Jan. 2011. ISSN 1554-0669. doi: 10.1561/1500000010. URL http://dx.doi.org/10.1561/1500000010.

L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 298–305, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3. doi: 10.1145/860435.860490. URL http://doi.acm.org/10.1145/860435.860490.

L. Si and J. Callan. Unified utility maximization framework for resource selection. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 32–41, New York, NY, USA, 2004. ACM. ISBN 1-58113-874-1. doi: 10.1145/1031171.1031180. URL http://doi.acm.org/10.1145/1031171.1031180.

M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 623–632, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. doi: 10.1145/1321440.1321528. URL http://doi.acm.org/10.1145/1321440.1321528.

A. Spoerri. Authority and ranking effects in data fusion. *J. Am. Soc. Inf. Sci. Technol.*, 59(3):450–460, Feb. 2008. ISSN 1532-2882. doi: 10.1002/asi.v59:3. URL http://dx.doi.org/10.1002/asi.v59:3.

T. Strohman. Lemur project components: Galago. www.galagosearch.org/, 2013. [Online; accessed 25-January-2013].

Trip Advisor Forums. New york forum, travel discussion for new york, united states – tripadvisor. http://www.tripadvisor.com/ShowForum-g28953-i4-New_York.html, 2013. URL ubuntuforums.org. [Online; accessed 25-January-2013].

Ubuntu Forums. Ubuntu forums. ubuntuforums.org, 2013. [Online; accessed 25-January-2013].

H. Wang, C. Wang, C. Zhai, and J. Han. Learning online discussion structures by conditional random fields. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 435–444, New York, NY, USA, 2011a. ACM. ISBN 978-1-4503-0757-

4. doi: 10.1145/2009916.2009976. URL http://doi.acm.org/10.1145/2009916.2009976.

L. Wang, M. Lui, S. N. Kim, J. Nivre, and T. Baldwin. Predicting thread discourse structure over technical web forums. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 13–25, Stroudsburg, PA, USA, 2011b. Association for Computational Linguistics. ISBN 978-1-937284-11-4. URL http://dl.acm.org/citation.cfm?id=2145432.2145435.

Y.-C. Wang and C. P. Rosé. Making conversational structure explicit: identification of initiation-response pairs within online discussions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 673–676, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL http://dl.acm.org/citation.cfm?id=1857999.1858096.

W. Weerkamp, K. Balog, and M. Rijke. Blog feed search with a post index. *Information Retrieval*, 14:515–545, 2011. ISSN 1386-4564. doi: 10.1007/s10791-011-9165-9. URL http://dx.doi.org/10.1007/s10791-011-9165-9.

Wikipedia. Condorcet method — wikipedia, the free encyclopedia, 2012. URL http://en.wikipedia.org/w/index.php?title=Condorcet_method&oldid=528911950. [Online; accessed 26-January-2013].

Wikipedia. Borda count — wikipedia, the free encyclopedia, 2013a. URL "http://en.wikipedia.org/wiki/Borda_count". [Online; accessed 25-January-2013].

Wikipedia. Data fusion — wikipedia, the free encyclopedia, 2013b. URL http://en.wikipedia.org/w/index.php?title=Data_fusion&oldid=546618234. [Online; accessed 6-July-2013].

Wikipedia. Student's t-test — wikipedia, the free encyclopedia", 2013c. URL http://en.wikipedia.org/wiki/Student%27s_t-test. [Online; accessed 25-January-2013].

R. Wilkinson. Effective retrieval of structured documents. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 311–317, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X. URL http://dl.acm.org/citation.cfm?id=188490.188591.

I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition*. Morgan Kaufmann. Morgan Kaufmann Publishers, 1999. ISBN 9781558605701. URL http://books.google.com.my/books?id=2F74jyPl48EC.

World of Warcraft Forums. Forums - world of warcraft. http://us.battle.net/wow/en/forum/, 2013. [Online; accessed 25-January-2013].

M. Wu, D. Hawking, A. Turpin, and F. Scholer. Using anchor text for homepage and topic distillation search tasks. *Journal of the American Society for Information Science and Technology*, 63(6):1235–1255, 2012. ISSN 1532-2890. doi: 10.1002/asi.22639. URL http://dx.doi.org/10.1002/asi.22639.

S. Wu. *Data Fusion in Information Retrieval*. Springer Berlin Heidelberg, 2012a.

S. Wu. Linear combination of component results in information retrieval. *Data & Knowledge Engineering*, 71(1):114 – 126, 2012b. ISSN 0169-023X. doi: 10.1016/j.datak.2011.08.003. URL http://www.sciencedirect.com/science/article/pii/S0169023X11001236.

C. Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2(3):137–213, Mar. 2008. ISSN 1554-0669. doi: 10.1561/1500000008. URL http://dx.doi.org/10.1561/1500000008.

C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, Apr. 2004. ISSN 1046-8188. doi: 10.1145/984321.984322. URL http://doi.acm.org/10.1145/984321.984322.

M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, and L. Zhao. Expansion-based technologies in finding relevant and new information: Thu trec2002 novelty track experiments. In *the Proceedings of the Eleventh Text Retrieval Conference (TREC*, pages 586–590, 2002.