

國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

比較並結合基於內部和網路文件之查詢擴展方法

Comparing and Combining Query Expansion Approaches

based on Local and Web Documents



Chien - Hung Chen

指導教授：李秀惠 博士

Advisor: Hsiu-Hui Lee, Ph.D.

中華民國 98 年 6 月

June, 2009

國立臺灣大學電機資訊學院資訊工程所

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master thesis

比較並結合基於內部和網路文件之查詢擴展方法

Comparing and Combining Query Expansion Approaches

based on Local and Web Documents



Chien - Hung Chen

指導教授：李秀惠 博士

Advisor: Hsiu-Hui Lee, Ph. D.

中華民國 98 年 6 月

June, 2009

誌 謝

JavaLab 像是一個溫暖的家，不僅是因為有舒適的沙發床、存放食物的大冰箱，主要還是我們的大家長—李秀惠老師媽媽般的氣質，讓每次 meeting 不僅聊著我們的研究，也談論著我們的生活點滴、將來的工作以及人生觀的分享，謝謝老師對我們學生的用心與放心，讓我們碩士兩年的學習如此豐富，認識到做研究的方法，也學習到獨立自主的態度。願上帝祝福老師的身體健康，也希望以後有機會能去老師打算買地興建的透天厝參觀、玩耍。

還要感謝有著專業知識、和善態度的俊雄學長，你的建議常帶給我們研究上很大的實質幫助；感謝榮宏介紹我來到 JavaLab，大學同窗好友的建議果然是可信的；感謝阿土和香腸，你們是學長、也是朋友，或在工作、或在當兵都加油！

畢業也意謂著，我們這一群一起奮鬥的 JavaLab 閉門弟子們要各奔東西了。mojo，你總是帶來活力與歡笑，相信認真負責的你未來到了訊連打拼一定沒問題！常常一起守在實驗室的彭彭，你在資訊技術與網站經營的想法與創意總是帶給我們許多驚喜，如果哪天開公司、股票上市可別忘了我們歐！還有安東尼，雖然你都會給自己很大的壓力，但我發現你最後的表現總是比你想像中的好，送你一句聖經的話：「不要為明天而憂慮，一天的憂慮一天當就夠了」，如果真有心事也別忘了找我們這些同窗好友聊，祝福你在 ASUS 的成長與學習。

自己喜歡在台大讀書的一個很重要原因，是學校離家裡很近，回到家，總可以感受到家人對我的接納與支持，這也一直帶給我內心極大的安全感。

最後，要感謝愛我的上帝，因我知道祢是一切幸福的來源、恩典的源頭，讓我有機會認識好老師、好學長和好同學，並與大家一同相處，謝謝祢！

Contents

中文摘要	v
Abstract	vi
Chapter 1 Introduction	1
1.1 Background.....	1
1.2 Motivation and Objective	3
1.3 Organization.....	3
Chapter 2 Related Works	4
2.1 Query Expansion based on Local Documents.....	4
2.2 Query Expansion based on Web Documents	5
2.3 Combined Query Expansion Method.....	6
Chapter 3 Research Approach	8
3.1 Framework	8
3.2 Query Expansion Method.....	9
3.2.1 Co-occurrence Statistics Method.....	10
3.2.2 Frequency Statistics Method	11
3.2.3 Combined Method	12
3.3 Combined Query Expansion Method.....	13
3.3.1 Score Normalization Methods	14
3.3.2 Expansion Term Merging Methods	16
3.4 Term weighting of Expanded Query Terms.....	17
Chapter 4 Experiments	19
4.1 Experimental Environment	19
4.2 Blind Relevance Feedback (BRF) based on Local and Web Document.....	21
4.3 Combined Query Expansion Methods	25
4.4 Experiments Results Discussion	31
Chapter 5 Conclusions and Future Works	33
5.1 Conclusions	33

5.2 Future Works..... 33

References 35



List of Figures

Figure 2-1	Bo1 Model.....	5
Figure 2-2	Wiki-Link Algorithm.....	6
Figure 3-1	The framework of our approach.....	9
Figure 3-2	Co-Occurrence Statistic Method.....	11
Figure 3-3	Frequency Statistics Method.....	12
Figure 3-4	Combined Co-Occurrence and Frequency Statistics Method.....	13
Figure 3-5	Score-Normalization: Max-Min Method.....	15
Figure 3-6	Score Normalization: Z-Score Method.....	15
Figure 3-7	Merging: Adding Method.....	16
Figure 3-8	Merging: Average Method.....	17
Figure 3-9	Term Weighting Method.....	18



List of Tables

Table 4-1	TREC 2004 document collection	20
Table 4-2	TREC 2004 Topics	20
Table 4-3	TREC 2004 Measures	21
Table 4-4	Bo1 vs Local	22
Table 4-5	Wiki-Link vs Web	23
Table 4-6	Comparison: Init, Local and Web.....	24
Table 4-7	Based on Max-Min Normalization: Adding Method vs Average Method ...	26
Table 4-8	Based on Z-Score Normalization: Adding Method vs Average Method.....	27
Table 4-9	Max-Min vs Z-Score	28
Table 4-10	Comparison: Local, Web, Max-Min and Z-Score	30
Table 4-11	P-Value Computation	31
Table 4-12	Comparison of Top Expansion Terms(Topic 641 For Example).....	32



中文摘要

關鍵字擴展是增進資訊檢索效能的重要技術。依據擴展的文件集，可分為兩類，一類是基於內部文件集的關鍵字擴展，另一類則是基於網路文件集。

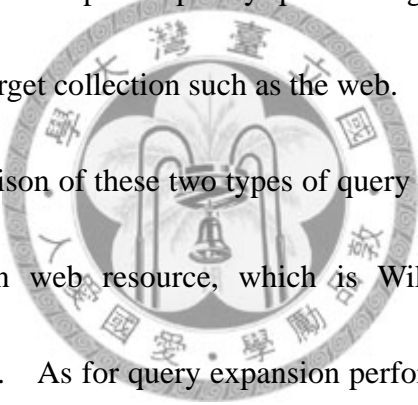
之前的研究顯示，基於內部文件集的擴展方法在增進檢索效能不佳的查詢上有瓶頸，然而，有人提出解決此問題之方法，是選用外部的文件集做為查詢擴展文件集，像是網路文件。

關於這兩類擴展方法的比較，我們的實驗結果顯示，的確在增進檢索效能不佳的查詢上，基於網路文件集－維基百科的關鍵字擴展有較好的效能，至於基於內部文件集而實行的關鍵字擴展，則在其他不同的查詢問題上有較好的表現。因此，我們提出將兩類不同文件集產生的擴展關鍵字列表作結合的方法，並評量其檢索效能的表現。大致上，我們擴展關鍵字的結合方法產生了較好的結果；但嚴謹的來說，我們只說我們的方式提供了平衡的結果。

關鍵字：虛擬相關回饋、關鍵字擴充、維基百科

Abstract

Query expansion is an important technique to improve search capability in information retrieval. According to expansion collection, there are two types of query expansion. One is query expansion performed on local documents and another is performed on web documents. The previous research found the method which is based on local documents has bottleneck on poorly performing topics, called hard topics. However, others propose to improve poorly performing topics is exploiting text collections other than the target collection such as the web.



Regarding our comparison of these two types of query expansion, our result shows query expansion based on web resource, which is Wikipedia, indeed has better performance on hard topics. As for query expansion performed on local documents, it has better performance on other topics. Therefore, we propose a combined method to integrate two ranked lists of terms expanded by these two types of query expansion, and evaluate the corresponding search performance. Roughly speaking, our combined query expansion methods produce better performance. However, to view it in a strict way, our methods provide balanced results.

Keywords: Blind Relevance Feedback, Query Expansion, Wikipedia

Chaper 1 Introduction

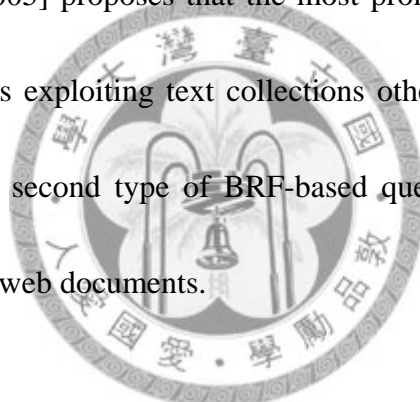
In this chapter, we make a brief introduction for this thesis. We introduce the research background in section 1.1 and describe the motivation of our study in section 1.2. Finally, the organization of this thesis is shown in section 1.3.

1.1 Background

Nowadays, tons of data and information spread on internet and the amount of them keep growing on an unpredictable way. In this situation, it is critical to have good performance on search capability and then return relevant documents and information to users. In order to improve retrieval performance, whether in web retrieval or information retrieval, query expansion is an important technique to achieve the goal. It can provide related terms or synonyms to expand the original query and return more numbers of relevant documents, as well improve the precision of top retrieval documents.

As for query expansion, Blind Relevance Feedback (BRF) was proposed in [ER 1994] and has been demonstrated to be an effective method for improving retrieval results. BRF expands original query by selecting relevant terms from top-ranked

retrieval documents. There are two types of BRF-based query expansion. One is the original version, where query expansion is performed on local documents, but there is bottleneck on poorly performing topics, called hard topics. In hard topics case, most of top-ranked retrieval documents may be irrelevant to the original query and it may result in worse performance. Therefore, more research and studies begin to focus on this issue. For example, Text Retrieval Conference (TREC), which is a novel conference in information retrieval, held robust retrieval track in 2003,2004 and 2005. The TREC report in [V 2005] proposes that the most promising approach to improve poorly performing topics is exploiting text collections other than the target collection such as the web. It is the second type of BRF-based query expansion, where query expansion is performed on web documents.



With the amount of electronic resources available on internet, it is often possible that query expansion based on local documents and web documents can be performed as means to improve retrieval results. Though query expansion based on web documents generally has better performance on hard topics, the approach based on local documents may exhibit better on other topics. It shows the possibility to combine these two methods to have better performance.

1.2 Motivation and Objective

Query expansion based on local and web documents have their own advantages. They may have better performance on different query topics. It is possible to combine these two methods and improve the retrieval performance.

In our thesis, we implement BRF method on local and web documents. Moreover, we compare the result of two methods and combine two ranked lists of expansion terms. Hopefully, two methods are complementary with each other and the combination improves the retrieval performance.



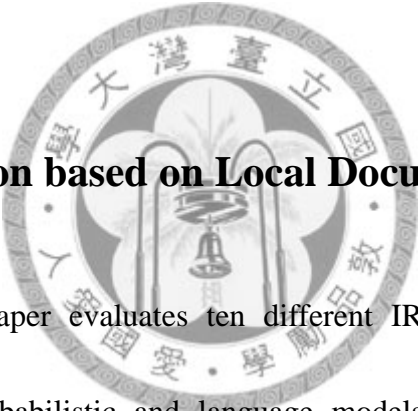
1.3 Organization

The remainder of this thesis is organized as follows. Related works are summarized in chapter 2. The detailed description of our research approach is in chapter 3. In chapter 4, we design some experiments to evaluate our proposed methods and present their evaluation results. Finally, conclusions and future works are presented in chapter 5.

Chapter 2 Related Works

In this chapter, we introduce related works to our study. In section 2.1, we focus on state of art query expansion method based on local documents, which is Bo1. As for query expansion based on web documents, we adopt Wikipedia as our web resource. The related researches are described in section 2.2. Finally, we mention works concerning combined query expansion method in section 2.3.

2.1 Query Expansion based on Local Documents



In [AS 2008], the paper evaluates ten different IR models, including recent developments in both probabilistic and language models. It shows that the best performing IR model is a probabilistic model developed within the Divergence from Randomness framework (DFR) .

Moreover, another paper [RA 2008] mentions the most effective DFR term weighting model is the Bo1 model that uses the Bose-Einstein statistics. Therefore, we choose Bo1 model as our competitive query expansion method based on local documents to prove high performance of our proposed approach.

The simple idea of DFR is that “The more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word”. We can see the equation of Bo1 in Figure 2-1.

$$w(t) = tf_x \log_2 \left(\frac{1 + P_n}{P_n} \right) + \log_2 (1 + P_n)$$

tf_x : frequency of the query term in the x top – ranked documents
 P_n : given by $\frac{F}{N}$
 F is the frequency of query term in collection
 N is the number of documents

Figure 2-1 Bo1 Model

2.2 Query Expansion based on Web Documents

Wikipedia is the largest and widely-used encyclopedia nowadays. Because it freely provides good quality and quantity articles for everyone, not only the general public benefit from Wikipedia, but also researchers apply it to different problems in last few years.

In [MLMH 2008], the paper provides first comprehensive summary of Wikipedia related research. It introduces several fields which Wikipedia applies to. One of them is query expansion. It cites several papers, including [LLHC 2007], [MWN 2007],

[AECC 2008] and [EGM 2008]. Among these query expansion methods based on Wikipedia, we choose [AECC 2008] to be our comparison method. The algorithm makes use of anchor phrase to generate expansion terms. The detailed algorithm is in Figure 2-2.

$$Score(a_i) = \sum_{a_{ij} \in S_w} (I(\text{target}(a_{ij}) \in S_R) \times (R - \text{rank}(\text{target}(a_{ij}))))$$

a_i : unique anchor phrase
 a_{ij} : an occurrence of anchor phrase a_i
 $\text{target}(a_{ij})$: return target article linked to by occurrence j of anchor phrase a_i
 $\text{rank}(\cdot)$: return the rank of Wikipedia article
 $I(\cdot)$: identity function, which equals 1 if its argument is true and 0 otherwise
 S_R : top - ranked R documents
 S_w : top - ranked W documents

Figure 2-2 Wiki-Link Algorithm

2.3 Combined Query Expansion Method

In [HP 2006], it discusses these two blind relevance feedback techniques. The result of the paper shows the expansion terms obtained from two methods have only a few overlap and has potential for combining these two methods. In [LLHC 2007], the paper evaluates 50 hard topics on two methods. The result shows that sometimes query expansion based on web documents performs better, and sometimes it is just opposite. It also shows the possibility to combine these two methods to have better performance.

Regarding above two papers, we know there is potential to combine two ranked lists of expansion terms, but we can't find representative combined query expansion method to be our competitive target until now.



Chapter 3 Research Approach

In this chapter we describe not only the procedures of our research, but also our main research approaches in detail. Our framework shows in section 3.1, and main research methods are in the rest of chapter.

3.1 Framework

Figure 3-1 shows the procedures of our research. First of all, we issue the original query to local and web documents individually for query expansion. After generating candidate expansion terms with BRF(Blind Relevance Feedback) method on two different document resources, we adopt our proposed method to combine candidate expansion terms which are from local and web documents.

Candidate terms are combined with the original query to form the expanded query, and it is sent to an IR system for further evaluation of retrieved result.

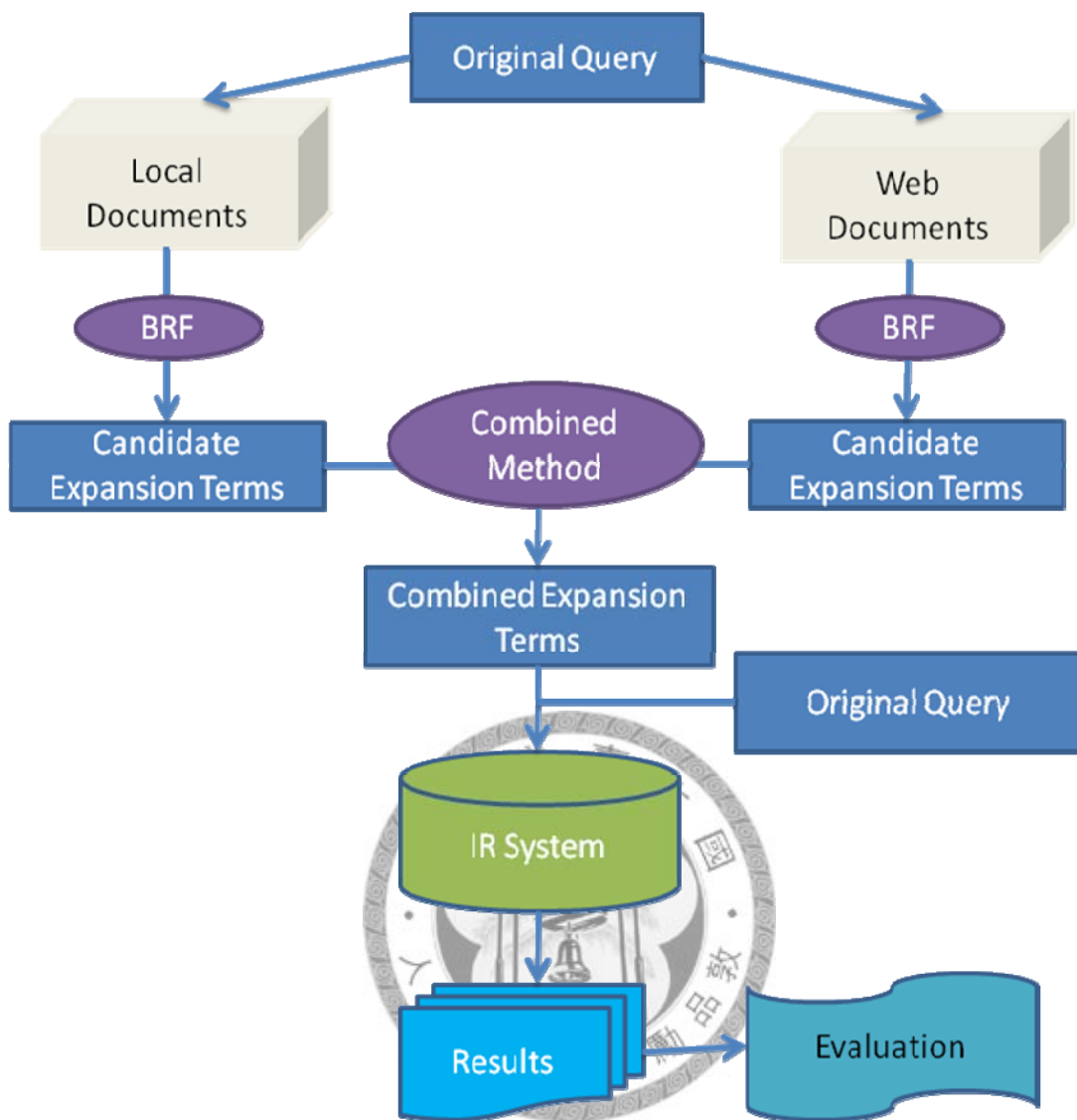


Figure 3-1 The framework of our approach

3.2 Query Expansion Method

Based on BRF, we propose our query expansion method, which is the combination of co-occurrence and frequency statistics method. Co-occurrence statistics method is introduced in section 3.2.1, and the following is frequency statistics method in section 3.2.2. As for combined approach of these two methods, we can see the detail in section, 3.2.3.

3.2.1 Co-occurrence Statistics Method

Co-occurrence statistics method is used to detect some kind of semantic similarity between terms and exploit it to query expansion. There are many co-occurrence statistics methods available nowadays. We choose the well-known Tanimoto-coefficient.

We use the terms on top-N retrieval documents as expansion candidates and then sum the Tanimoto-coefficient between original query terms and expansion candidates. The most useful terms are at the top score of CC value (co-occurrence statistic value).

The equation is in Figure 3-2.

$$Tanimoto(t_i, t_j) = \frac{c_{ij}}{c_i + c_j - c_{ij}}$$

$$\begin{cases} t_i, t_j : \text{terms } i, j \\ c_i, c_j : \text{the number of documents in which terms } t_i, t_j \text{ occurs} \\ c_{ij} : \text{the number of documents in which } t_i \text{ and } t_j \text{ cooccur} \end{cases}$$

$$CC(q, t_e) = \sum_{t_i \subset q} Tanimoto(t_i, t_e)$$

$$\begin{cases} q : \text{original query} \\ t_i : \text{term } i \text{ in the original query} \\ t_e : \text{candidate term } e \end{cases}$$

Figure 3-2 Co-Occurrence Statistic Method

3.2.2 Frequency Statistics Method

The traditional and well-known query expansion method based on frequency statistics is TF*IDF.

TF means term frequency. The term frequency of a term is defined by the number of times a term appears in a document and can be viewed as local or document specific information. We calculate TF value of terms on top-N retrieval documents and also take into account the rank of retrieval documents at the same time. Higher TF value shows that the term is more related with the original query and more suitable to be expansion candidate.

As for IDF, it means inverse document frequency. The inverse document frequency of a term is defined by total number of documents dividing the number of documents in which a term appears and can be viewed as global information. Higher IDF value shows that the term appears less times in whole collections and is better to be expansion candidate.

After calculating TF and IDF score, we multiply these two values to have our TFIDF score. The equation is in Figure 3-3.

$$TF(t_i) = \sum_{t_{ij} \in D} (N - Rank(t_{ij})) \times \frac{tf(t_{ij})}{DL_j}$$

$$\left\{ \begin{array}{l} t_i : \text{term } i \\ t_{ij} : \text{document } j \text{ containing term } i \\ D : \text{retrieval documents} \\ N : \text{numbers of top documents to retrieve} \\ Rank(t_{ij}) : \text{document rank of } t_{ij} \\ DL_j : \text{length of document } j \\ tf(t_{ij}) : \text{term frequency of } t_{ij} \end{array} \right.$$

$$IDF(t_i) = \log \frac{TD}{df(t_i)}$$

$$\left\{ \begin{array}{l} t : \text{term } i \\ df(t_i) : \text{document frequency of term } i \\ TD : \text{numbers of total documents} \end{array} \right.$$

$$TFIDF(t_i) = TF(t_i) \times IDF(t_i)$$

Figure 3-3 Frequency Statistics Method

3.2.3 Combined Method

Referring to [RAA 2008], we know it is workable to combine co-occurrence statistics method and frequency statistics method. Two approaches can complement

each other because they rely on different information. The drawback of the co-occurrence statistics method is that the performance is reduced by words which are not stop-words but very frequent in the collection. These words, which are a kind of noise, may have a high co-occurrence statistics score and result in bad performance after query expansion. However, these words have low frequency statistics score because they appear in any set of document collection and result in low IDF score.

But the next question is how we integrate both of these two methods at the same time. Regarding to [RAA 2008], they obtain two lists of candidate terms by each method separately and intersect the lists to result in final expansion terms. As for us, we adopt another way to apply both methods together, which is to multiply two scores to generate final BRF scores. The equation is in Figure 3-4.

$$BRF(t_i) = CC(t_i) \times TFIDF(t_i)$$

$$\begin{cases} t_i : \text{term } i \\ CC(t_i) : \text{score calculated by co-occurrence statistics method} \\ TFIDF(t_i) : \text{score calculated by frequency statistics method} \end{cases}$$

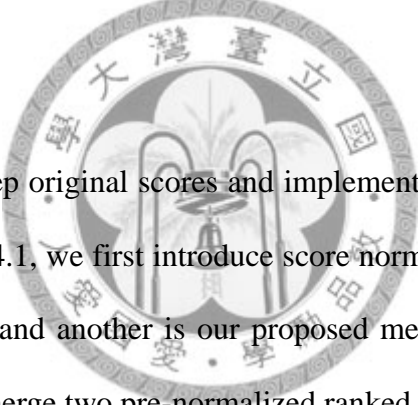
Figure 3-4 Combined Co-Occurrence and Frequency Statistics Method

3.3 Combined Query Expansion Method

The main goal of our thesis is to propose another combined query expansion approach based on local and web documents. After we implement BRF method to expand original query on local and web documents separately, we derive two ranked

lists of new terms.

To integrate two ranked lists of expansion candidates, [CR 1999] adopts the method which ignores the original scores of each ranked terms, but consider the relative position of each term on ranked lists. For example, the score of the term ranked as first as N and the score of the second term was $N-1$. After assigning new scores according to ranked position, the next step is to compute an average score for each term from the scores assigned to that term by individual lists. The most useful terms are at the top score. This method is easy to implement but loses the precise scores of each term.



Our approach is to keep original scores and implement score normalization in two ranked lists. In section 3.4.1, we first introduce score normalization methods. One is proposed by [MTT 1999], and another is our proposed method. In section 3.4.2, we introduce two methods to merge two pre-normalized ranked lists.

3.3.1 Score Normalization Methods

Before merging two score-based lists, we have to implement normalization in order that terms in two lists will have similar absolute scores. First method is proposed by [MTT 1999], we call this score normalization method as “Max-Min” in the following thesis. The normalization equation is in Figure 3-5.

$$Score_{new} = \frac{Score_{old} - Score_{min}}{Score_{max} - Score_{min}}$$

$$\left\{ \begin{array}{l} Score_{new} : \text{score after normalizing} \\ Score_{old} : \text{original score} \\ Score_{max} : \text{highest score in ranked list} \\ Score_{min} : \text{lowest score in ranked list} \end{array} \right.$$

Figure 3-5 Score-Normalization: Max-Min Method

Using this normalization strategy bring new score into range [0,1] and have similar absolute scores.

Another approach we propose is based on normalization in statistics, called z-score or standard score. The score is derived by subtracting the population mean from an individual old score and then dividing the difference by the population standard deviation. Z-Score equation is in Figure 3-6.

$$Score = \frac{x - \mu}{\sigma}$$

$$\left\{ \begin{array}{l} x : \text{original score to be normalized} \\ \mu : \text{the mean of population} \\ \sigma : \text{the deviation of the population} \end{array} \right.$$

Figure 3-6 Score Normalization: Z-Score Method

In statistics, z-score allows data on different scales to be compared, by bringing them into a common scale. By using z-score method, we can bring our two ranked

lists of terms into a common scale and implement further expansion term merging.

3.3.2 Expansion Term Merging Methods

After score normalization, the next step is to combine two ranked lists of terms in a reasonable way. There are two merging methods we propose in this section. One is adding method, and another is average method. The difference between these two methods is the calculating of overlapping terms in two lists.

Adding method handles overlapping terms by adding two normalized scores together. As for non-overlapping terms, we only keep its' original scores. The result of this merging method emphasize in overlapping terms and higher scores' terms of original ranked list. The equation is in Figure 3-7.

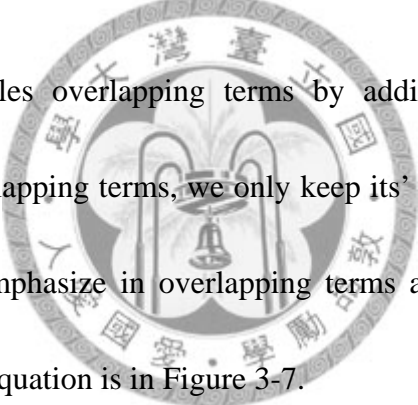

$$\begin{aligned} \mathit{Score}_{\mathit{merge}\text{-}\mathit{adding}} &= \mathit{Score}_{\mathit{local}} + \mathit{Score}_{\mathit{web}} \\ \left\{ \begin{array}{l} \mathit{Score}_{\mathit{merge}\text{-}\mathit{adding}} : \mathit{score\ after\ adding\ method} \\ \mathit{Score}_{\mathit{local}} : \mathit{term\ score\ based\ on\ local\ document} \\ \mathit{Score}_{\mathit{web}} : \mathit{term\ score\ based\ on\ web\ document} \end{array} \right. \end{aligned}$$

Figure 3-7 Merging: Adding Method

The second approach is average method. The score of overlapping terms is defined as the average of their normalized scores. This merging method doesn't not

give overlapping terms more scores but adjust the scores to be more objective. See the equation in Figure 3-8.:

$$Score_{merge-average} = \frac{Score_{local} + Score_{web}}{2}$$

$$\left\{ \begin{array}{l} Score_{merge-average} : \text{score after average method} \\ Score_{local} : \text{term score based on local document} \\ Score_{web} : \text{term score based on web document} \end{array} \right.$$

Figure 3-8 Merging: Average Method

3.4 Term weighting of Expanded Query Terms

After expansion terms have been generated, we have to assign term weighting to expansion terms and add them into original query.

Because we believe the original score of the term means how important the term is, we propose our term weighting method depending on the data, which is collected during the selection process of expansion terms. That is we calculate the division of original score of the term and the amount of original scores, the quotient is weighting of the term.

Figure 3-9 is our term weighting (TW) equation:

$$TW(t_i) = \frac{Score(t_i)}{\sum_{t_j \in ST} Score(t_j)}$$

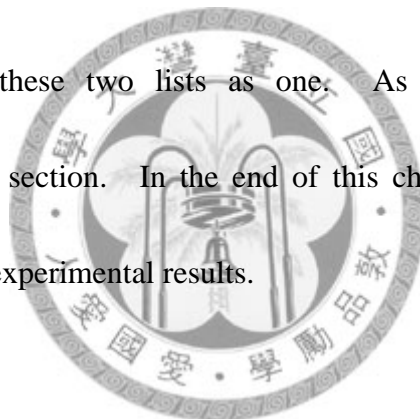
$\left\{ \begin{array}{l} t_i, t_j : \text{term } i, j \\ ST : \text{all selected terms} \\ Score(t_i) : \text{original score of } t_i \end{array} \right.$

Figure 3-9 Term Weighting Method



Chaper 4 Experiments

In this chapter, we design some experiments to evaluate our proposed methods and present evaluation results. In section 4.1, we describe our experimental environment, including our fundamental information retrieval system, evaluation corpus, and testing topic sets. In next section 4.2, we implement combined BRF methods to expand query over local documents and web documents, and then analyze the performance individually. Furthermore, in section 4.3, we normalize the terms' scores of two ranked lists and merge these two lists as one. As well there are evaluation performances in the same section. In the end of this chapter, section 4.4, we have deeper discussions on our experimental results.



4.1 Experimental Environment

We have used the Vector Space Model implementation provided by [Lucene] to build our information retrieval system. Stemming and stop-word removing has been applied in indexing and expansion process.

Evaluation is carried out on the TREC 2004 Robust Retrieval Track corpus. The summary of the corpus is on the following table.

Table 4-1 TREC 2004 document collection

Source	# Docs	Size (MB)
Financial Times	210,158	564
Federal Register 94	55,630	395
FBIS, disk 5	130,471	470
LA Times	131,896	475
Total Collection:	528,155	1904

With these corpus, we use a set of 250 topics (one of which was subsequently dropped due to having no relevant documents) to evaluate. The following is the description of topic sets.

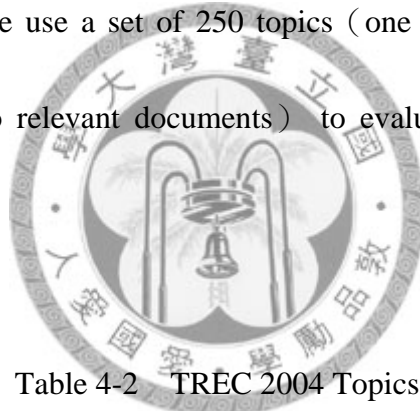


Table 4-2 TREC 2004 Topics

Topic Set	Number of Topics	Topics Description
Old	200	Topics From TRECs 6-8 (301-450) TREC 2003 Robust Track (601-650)
New	49	TREC 2004 Robust Track (651-700)
Hard	50	50 particularly difficult topics from 301-450 set
Combined	249	Old+New Topics

To evaluate the performance on different aspects, we use four measures, which are MAP, P10, NO%, and AREA. Each of them provides a different estimation of the retrieval documents. We explain in the following table.

Table 4-3 TREC 2004 Measures

Measure	Explanation	Estimation of System	Better
MAP	mean average precision	average performance of system	↑
P10	precision after 10 docs retrieved	top retrieval performance of system	↑
NO%	number of topics with no relevant in top 10	top retrieval performance of system	↓
AREA	the area underneath the MAP(X) vs X curve for worst topics	worst topics performance of system	↑

4.2 Blind Relevance Feedback (BRF) based on Local and Web Document

We implement combined BRF method, which is mentioned in section 3.2.3, based on local and web documents individually to evaluate the performance on different kinds of topics.

First, we discuss the evaluation on query expansion based on local documents. We compare Bo1 method and our approach at the same time. We have best

performance of Bo1 when top 10 documents are picked up and top 40 expansion terms are selected. As for our method, the experiment with top 10 documents and top 50 expansion terms has best data. The following table is the comparison of representative data.

Table 4-4 Bo1 vs Local

System	Old Topic Set (200)				New Topic Set (49)			
	Map	P10	%no	area(50)	Map	P10	%no	area(12)
Bo1	0.2511	0.4	15.50%	0.0091	0.2925	0.4102	12.20%	0.0245
Local	0.2511	0.416	15.50%	0.0097	0.3017	0.4122	12.20%	0.0243
System	Hard Topic Set (50)				Combined Topic Set (249)			
	Map	P10	%no	area(12)	Map	P10	%no	area(62)
Bo1	0.1043	0.21	26.00%	0.0034	0.2592	0.402	14.90%	0.0106
Local	0.1055	0.244	28.00%	0.0035	0.261	0.4153	14.90%	0.0111

We mark higher score with red. Even though some of measures are equal, we can still see our approach with better performance on most of measures. Considering Bo1 as state of art query expansion method, the result shows that our query expansion approach based on local documents has outstanding performance.

As for web documents, Wikipedia is our document resource for query expansion. The competitive method which we call it “wiki-link” in brief is the algorithm in [AECC 2008]. We have best performance of wiki-link when SR set to 500 and Sw set to 1000, with top 20 expansion phrases selected. As for our method, the experiment with top 50 documents and top 30 expansion terms has best data.

Table 4-5 Wiki-Link vs Web

System	Old Topic Set (200)				New Topic Set (49)			
	Map	P10	%no	area(50)	Map	P10	%no	area(12)
Wiki-Link	0.1991	0.3505	15.5%	0.0095	0.2031	0.3184	20.4%	0.0135
Web	0.2596	0.4405	12.00%	0.0169	0.2833	0.4061	12.20	0.0274
System	Hard Topic Set (50)				Combined Topic Set (249)			
	Map	P10	%no	area(12)	Map	P10	%no	area(62)
Wiki-Link	0.0848	0.1980	28%	0.0049	0.1999	0.3442	16.5%	0.0098
Web	0.1243	0.282	24%	0.0101	0.2642	0.4337	12.00%	0.0181

It is obvious that our approach is better because of higher scores in each measure.

After picking up representative data from two query expansion experiments based

on different target resource. We summarize the experimental results in the following table, which contains the initial data without query expansion, the representative data of BRF based on local documents and representative data of BRF based on web documents.

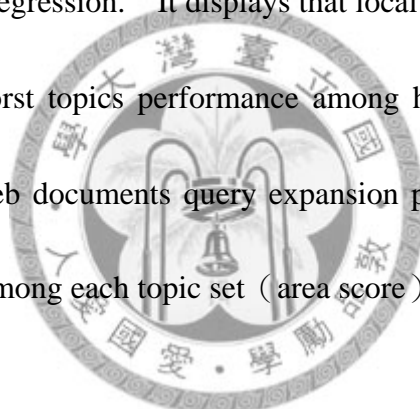
Table 4-6 Comparison: Init, Local and Web

System	Old Topic Set (200)				New Topic Set (49)			
	Map	P10	%no	area(50)	Map	P10	%no	area(12)
Init	0.1908	0.3495	18.00%	0.0083	0.2303	0.3449	12.20%	0.0142
Local	0.2511 (32%)	0.416 (19%)	15.50% (14%)	0.0097 (17%)	0.3017 (31%)	0.4122 (20%)	12.20% (0%)	0.0243 (71%)
Web	0.2596 (36%)	0.4405 (26%)	12.00% (33%)	0.0169 (104%)	0.2833 (23%)	0.4061 (18%)	12.20 (0%)	0.0274 (93%)
System	Hard Topic Set (50)				Combined Topic Set (249)			
	Map	P10	%no	area(12)	Map	P10	%no	area(62)
Init	0.0756	0.19	30.00%	0.0042	0.1986	0.3486	16.90%	0.0089
Local	0.1055 (40%)	0.244 (28%)	28.00% (6%)	0.0035 (-17%)	0.261 (31%)	0.4153 (19%)	14.90% (12%)	0.0111 (25%)
Web	0.1243 (64%)	0.282 (48%)	24% (20%)	0.0101 (140%)	0.2642 (33%)	0.4337 (24%)	12.00% (29%)	0.0181 (103%)

We mark the highest performance in each measure with red, and the regression

data with blue. Out of these experimental data, we can observe some interesting point:

1. Expansion based on local and web documents have their own advantage. In old topic set, expansion based on web documents outperforms expansion based on local documents. As for new topic set, it goes on the opposite way.
2. Expansion based on local documents exhibits bad performance on hard and worst topics, but expansion based on web documents is just opposite. The data marked with blue is the only regression. It displays that local documents query expansion can't improve the worst topics performance among hard topic set, but makes it worse. However, web documents query expansion performs well on hard topic set and worst topics among each topic set (area score).



Regarding what are mentioned above, we believe combining the expansion terms based on local and web documents can take the advantage on each other and improve the overall performance.

4.3 Combined Query Expansion Methods

There are two steps to combine two ranked lists of expansion terms. First one is score normalization. We mentioned two approaches, Max-Min and Z-Score, in section

3.4.1. The second is merging process. There are two methods, adding method and average method, in section 3.4.2.

We implement all these methods to evaluate the performance. First of all, we display the comparison between adding and average method based on Max-Min normalization.

Table 4-7 Based on Max-Min Normalization: Adding Method vs Average Method

Max-Min	Old Topic Set (200)				New Topic Set (49)			
	Map	P10	%no	area(50)	Map	P10	%no	area(12)
Adding Method	0.2626	0.4405	12.50%	0.0153	0.31	0.4224	10.20%	0.0327
Average Method	0.258	0.4395	13%	0.015	0.3025	0.4204	10.20%	0.0295
Max-Min	Hard Topic Set (50)				Combined Topic Set (249)			
	Map	P10	%no	area(12)	Map	P10	%no	area(62)
Adding Method	0.1117	0.248	28.00%	0.007	0.2719	0.4369	12.00%	0.0173
Average Method	0.1137	0.254	32.00%	0.0066	0.2688	0.4357	12.40%	0.0166

We mark higher scores on each measure with red. Obviously, Adding method exhibits better performance on most of measures and topic set.

As for normalization, not only Max-Min but also Z-Score we would like to know its performance. Therefore, we display the comparison between adding and average method based on z-score normalization in the following table.

Table 4-8 Based on Z-Score Normalization: Adding Method vs Average Method

Z-Score	Old Topic Set (200)				New Topic Set (49)			
	Map	P10	%no	area(50)	Map	P10	%no	area(12)
Adding Method	0.2636	0.4465	11.50%	0.0171	0.2967	0.4163	6.10%	0.0311
Average Method	0.2562	0.4365	11.5%	0.0167	0.2866	0.4163	10.2%	0.0293
Z-Score	Hard Topic Set (50)				Combined Topic Set (249)			
	Map	P10	%no	area(12)	Map	P10	%no	area(62)
Adding Method	0.1219	0.278	26.00%	0.01	0.2701	0.4406	10.40%	0.0186
Average Method	0.1214	0.2920	22.00%	0.0103	0.2622	0.4325	11.2%	0.0182

Although average method has good performance on hard topic, we can observe that adding method still have better scores on overall measures and topic set, especially on measures of combined topic set.

Regarding above two experiments on adding method and average method, we both conclude that adding method performs better based on whether Max-Min normalization or Z-Score normalization.

Further, with adding method, we would compare the performance of Max-Min and Z-Score normalization. The evaluation result is in the following.

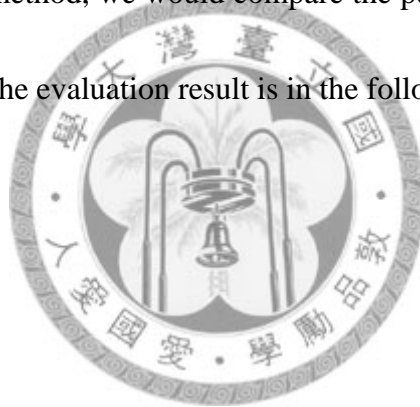


Table 4-9 Max-Min vs Z-Score

Adding Method	Old Topic Set (200)				New Topic Set (49)			
	Map	P10	%no	area(50)	Map	P10	%no	area(12)
Max-Min	0.2626	0.4405	12.50%	0.0153	0.31	0.4224	10.20%	0.0327
Z-Score	0.2636	0.4465	11.50%	0.0171	0.2967	0.4163	6.10%	0.0311

Adding Method	Hard Topic Set (50)				Combined Topic Set (249)			
	Map	P10	%no	area(12)	Map	P10	%no	area(62)
Max-Min	0.1117	0.248	28.00%	0.007	0.2719	0.4369	12.00%	0.0173
Z-Score	0.1219	0.278	26.00%	0.01	0.2701	0.4406	10.40%	0.0186

According to evaluation figures, we are hard to judge which normalization method is better because each of them displays better performance on different topic set. For example, Max-Min method performs well on new topic set but Z-Score method wins all measures on both old and hard topic set. As for combined topic set, Z-Score method derives higher figures on P10, %no and area than Max-Min method but lose the competition on Map measure.

At last, we would list the overall comparison, which compares Max-Min and Z-Score method with initial data without query expansion, local documents query expansion and web documents query expansion.

Table 4-10 Comparison: Local, Web, Max-Min and Z-Score

System	Old Topic Set (200)				New Topic Set (49)			
	Map	P10	%no	area(50)	Map	P10	%no	area(12)
Local	0.2511 (32%)	0.416 (19%)	15.50% (14%)	0.0097 (17%)	0.3017 (31%)	0.4122 (20%)	12.20% (0%)	0.0243 (71%)
Web	0.2596 (36%)	0.4405 (26%)	12.00% (33%)	0.0169 (104%)	0.2833 (23%)	0.4061 (18%)	12.20 (0%)	0.0274 (93%)
Max-Min	0.2626 (38%)	0.4405 (26%)	12.50% (31%)	0.0153 (84%)	0.31 (34%)	0.4224 (22%)	10.20% (16%)	0.0327 (130%)
Z-Score	0.2636 (38%)	0.4465 (28%)	11.50% (36%)	0.0171 (106%)	0.2967 (29%)	0.4163 (21%)	6.10% (50%)	0.0311 (119%)
System	Hard Topic Set (50)				Combined Topic Set (249)			
	Map	P10	%no	area(12)	Map	P10	%no	area(62)
Local	0.1055 (40%)	0.244 (28%)	28.00% (6%)	0.0035 (-17%)	0.261 (31%)	0.4153 (19%)	14.90% (12%)	0.0111 (25%)
Web	0.1243 (64%)	0.282 (48%)	24% (20%)	0.0101 (140%)	0.2642 (33%)	0.4337 (24%)	12.00% (29%)	0.0181 (103%)
Max-Min	0.1117 (48%)	0.248 (31%)	28.00% (7%)	0.007 (67%)	0.2719 (37%)	0.4369 (25%)	12.00% (29%)	0.0173 (94%)
Z-Score	0.1219 (61%)	0.278 (46%)	26.00% (13%)	0.01 (138%)	0.2701 (36%)	0.4406 (26%)	10.40% (38%)	0.0186 (109%)

It is like other tables, data marked with red represents highest score in each

measure. We can see combined query expansion method, including Max-Min and Z-Score, really improve overall system performance because the highest score of each measure on old, new and combined topic set belongs to combined query expansion methods, but we also observe that both methods have lower performance on hard topic set than expansion based on web documents.

In addition, we would like to prove the statistical significance of our combined approaches based on adding method. The following table is p-value computation of average precision in all 249 topics, and we can see that each p-value is less than 0.05, in other words, we have more than 95% confidence interval of mean.

Table 4-11 P-Value Computation

P-value	Local	Web
Max-Min	0.001679	0.045228
Z-Score	0.033721	0.016005

4.4 Experiments Results Discussion

Regarding section 4.3, Expansion based on local and web documents have their own advantage. Although expansion based on web documents performs better on

most topic sets, especially on hard topic set, but there still exists certain topic set, such as new topic set, expansion based on local documents has better performance. We believe the merging of two ranked lists of expansion terms can complement with each other and have better evaluation performance.

In order to have deeper understanding of combined query expansion method, we not only have relative experiments in section 4.3 but also specifically take topic 641 which is in new topic set for an example. The following is the list of combined expansion terms and we can figure out how it works.

Table 4-12 Comparison of Top Expansion Terms(Topic 641 For Example)

Topic 641 : Valdez wildlife marine life		
Local Only	Web Only	Combined
<i>Exxon</i>	Otter	Otter
<i>Valdez</i>	<i>sea</i>	oil
oil	<i>Alaska</i>	Alaska
spill	<i>oil</i>	sea
wildlif	<i>marine</i>	Valdez
<i>Alaska</i>	spill	spill
Bird	<i>ship</i>	Exxon

Regarding above table, we can observe Web Only query expansion provides more general and common terms, such as sea, oil and ship. As for Local Only query expansion provides additional details on topic, such as Exxon, Valdez and Alaska which are people or place.

Chaper 5 Conclusions and Future Works

5.1 Conclusions

In this thesis, we have compared the performance of query expansion based on local and web documents. Moreover, we also combine the expansion terms from two ranked lists and provide the comparison with initial data without query expansion, local documents query expansion and web documents query expansion.

Roughly speaking, our combined query expansion methods produce better performance. However, to view it in a strict way, the methods provide balanced results. If the combination of local and web documents expansion terms works appropriately, it can improve evaluation performance because of its complementary. However, we also observe the bad performance on hard topic set, and it is probably caused by expansion based on local documents pulling down the evaluation result of combined method.

5.2 Future Works

Future work includes more studies on the effectiveness of combining expansion terms based on local and web documents. Especially, to eliminate queries hurt by

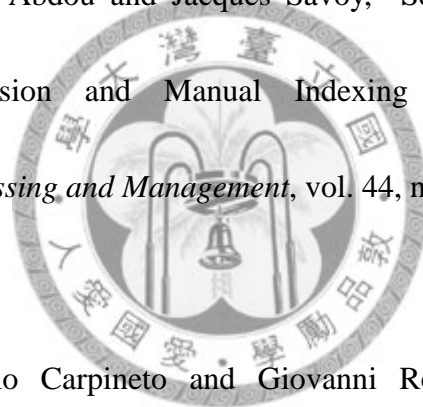
either, such as the result of hard topic set, is an important topic to discuss and explore.



References

[AECC 2008] Jaime Arguello, Jonathan L. Elsas, Jamie Callan and Jaime G. Carbonell, “Document Representation and Query Expansion Models for Blog Recommendation,” *Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM 2008)*, 2008.

[AS 2008] Samir Abdou and Jacques Savoy, “Searching in Medline: Query Expansion and Manual Indexing Evaluation,” *Information Processing and Management*, vol. 44, no. 2, pp781-789, 2008.



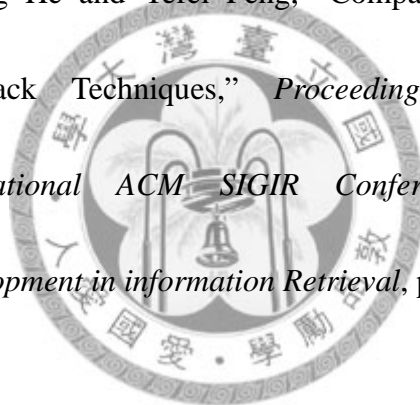
[CR 1999] Claudio Carpineto and Giovanni Romano. “Comparing and Combining Retrieval Feedback Methods: First Results,” *Proceedings of the Workshop on Machine Learning for Intelligent Information Access, Advanced Course on Artificial Intelligence (ACAI-99)*, 1999.

[EGM 2008] Ofer Egozi, Evgeniy Gabrilovich and Shaul Markovitch, “Concept-Based Feature Generation and Selection for Information

Retrieval,” *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008.

[EL 1994] D. Evans and R. Lefferts, “Design and Evaluation of the Clarit-Trec-2 System,” *Proceedings of the Second Text REtrieval Conference (TREC-2)*, 1994.

[HP 2006] Daqing He and Yefei Peng, “Comparing Two Blind Relevance Feedback Techniques,” *Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, pp. 649-650, 2006.

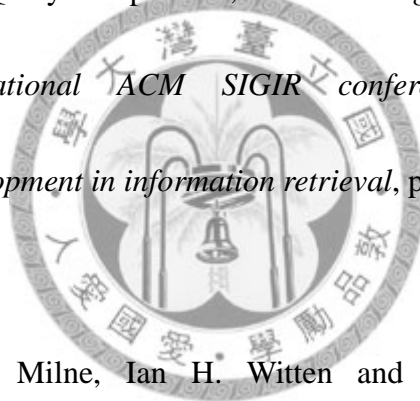


[LLHC 2007] Y. Li, R.W.P. Luk, E.K.S. Ho and F.L. Chung, “Improving Weak Ad-Hoc Queries Using Wikipedia as External Corpus,” *Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, pp. 797-798, 2007.

[Lucene] <http://lucene.apache.org>

[MLMH 2008] Olena Medelyan, Catherine Legg, David Milne and Ian H. “Mining Meaning From Wikipedia,” *Artificial Intelligence*, Submitted on 26 Sep 2008, last revised 10 May 2009.

[MTT 1999] Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. “Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion,” *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in information retrieval*, pp. 191-197, 1999.



[MWN 2007] David Milne, Ian H. Witten and David M. Nichols. “A Knowledge-Based Search Engine Powered by Wikipedia,” *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 445-454, 2007.

[RA 2008] Jose R. Perez-Aguera. and Lourdes Araujo. “Comparing and Combining Methods for Automatic Query Expansion,” *Advances in Natural Language Processing and Applications. Research in*

Computing Science 33, pp. 177-188, 2008.

[V 2005] Ellen M. Voorhees. “The TREC Robust Retrieval Track,” *SIGIR Forum* 39, pp. 11-20, 2005

