Computational Linguistics and Chinese Language Processing
Vol. 16, No. 1-2, March/June 2011, pp. 47-68

以籠統查詢評估查詢擴展方法與線上搜尋引擎之資訊檢索效能

Evaluating the Information Retrieval Performance of Query Expansion Method and On-line Search Engine on General Query

許志全*，吳世弘*

Chih-Chuan Hsu and Shih-Hung Wu

摘要

當資訊檢索系統的使用者在不確知正確關鍵字時，可能會使用不精確的查詢詞描述其所需的訊息，嘗試著尋找所需要的資訊。我們稱這些不精確的查詢關鍵字爲籠統查詢(general query)。本文將評估線上搜尋引擎與資訊檢索研究者所研發的檢索系統對籠統查詢的檢索效能。現有的資訊檢索測試集不適合用於評估線上系統，因爲一來面對的文件集不相同，其次是測試集並不包含籠統查詢詞與精確查詢詞。爲了提供線上搜尋引擎與一般資訊檢索系統一個一致性的評比環境，我們利用維基百科建立一個測試集，這樣一來每個系統都可以檢索同樣的文件集容，同時可以比較籠統查詢詞與精確查詢詞的查詢結果。

在我們的檢索系統中，我們利用此測試集的鏈結結構特性，提出了新的查詢擴展方法。使用維基百科作爲查詢擴展方法的同義辭典，並與虛擬關聯回饋的查詢擴展方法結合，我們稱此方法爲維基百科查詢擴展。

實驗結果表明，本篇論文所建構的基於籠統查詢的資訊檢索測試集，能夠合理的評估線上搜尋引擎，且在籠統查詢與精確的關鍵字檢索效能的比較，可以明

* 朝陽科技大學資工系
Chaoyang University of Technology, Taichung, Taiwan
E-mail: shwu@cyut.edu.tw
The author for correspondence is Shih-Hung Wu.

顯的觀察到，籠統查詢的檢索效能的確較差。並且我們發現，在使用籠統查詢下，虛擬關聯回饋的檢索系統會優於主流的線上搜尋引擎，如:Google, Alta Vista。

而在查詢擴展方法的部分，適當的使用維基百科查詢擴展方法的確是可以提升檢索效能，而且只使用維基百科查詢擴展與只使用虛擬關聯回饋查詢擴展間效能的比較，顯示利用維基百科作爲查詢擴展的同義辭典是很好的資源。

關鍵詞：資訊檢索、籠統查詢、測試集、維基百科、查詢擴展

Abstract

Users might use general terms to query the information in need, when the exact keyword is unknown. We treat these inexact query terms as general queries. In this paper, we consturct a test data set to evaluate the performance of online search engine on searching Wikipedia with general queries and exact queries.

We also proposed a new query expansion method that performs better on general queries. The Wikipedia query expansion method is regarding the Wikipedia as a thesaurus to find candidates of query expansion. The expanded queries are then combined with the pseudo relevance feedback. The performance of this method is better than online search engine on the general queries.

Keywords: Information Retrieval, General query, Test Collection, Wikipedia, Query Expansion.

1. 研究動機

在本篇論文中，我們將評估資訊檢索系統面對一個普遍發生的情況:「使用者不知道如何使用精確的關鍵字描述其資訊需求」。此時使用者只能使用籠統的關鍵字描述其資訊需求，並透過多次重新檢索之後，才找到精確的關鍵字，最後取得相關資訊。我們稱這些不精確的查詢關鍵字爲籠統查詢(general query)。本文將評估線上搜尋引擎與我們所研發的檢索系統對籠統查詢的檢索效能。

現今的資訊檢索測試集(Test Collection)不適合評估使用籠統查詢與精確關鍵字的檢索效能，因爲目前的各種大型測試集提供的是主題式的查詢資訊。如 TREC(Text REtrieval Conference)、CLEF(Cross Language Evaluation Forum)、NTCIR(NII Test Collection for IR Systems)等。各測試集典型的查詢主題(Topic)，都是使用主題當關鍵字描述其查詢的內容，而這些查詢主題是不區分爲精確或籠統的。

在本研究主要可以分爲兩個主題，第一部分是建置籠統查詢之資訊檢索測試集，主要是依據國際資訊檢索測試集機構(TREC、NTCIR)，以標準流程建構在 Web 文件上使用籠統查詢的資訊檢索測試集。第二部份是對查詢擴展作進一步的探討，其中我們提出

新的方法作查詢擴展，亦爲虛擬關聯回饋與同義辭典的結合。

在建置測試集的標準流程，如圖 1 所示。標準的測試集是由三個文件集合所構成，即查詢主題(Query set)、文件集(Document set)、相關判斷集(Relevance Judgment set)，其中查詢主題與文件集是事先蒐集建構的，而相關判斷集則由利用各個不同的檢索系統的檢索結果中，透過 Pooling Method 建構出相關判斷的 Pool，最後再透過參與判斷之相關判斷者，對 Pool 中的每篇文件判斷所建構而成的。我們利用其維基百科所釋出的資料(Wikipedia Dump Data)，作爲我們的測試集的內容。而查詢主題的建構，是由真實世界使用者對於維基百科全書知識需求所建構而成，其中包含了籠統查詢以及精確查詢等資訊。

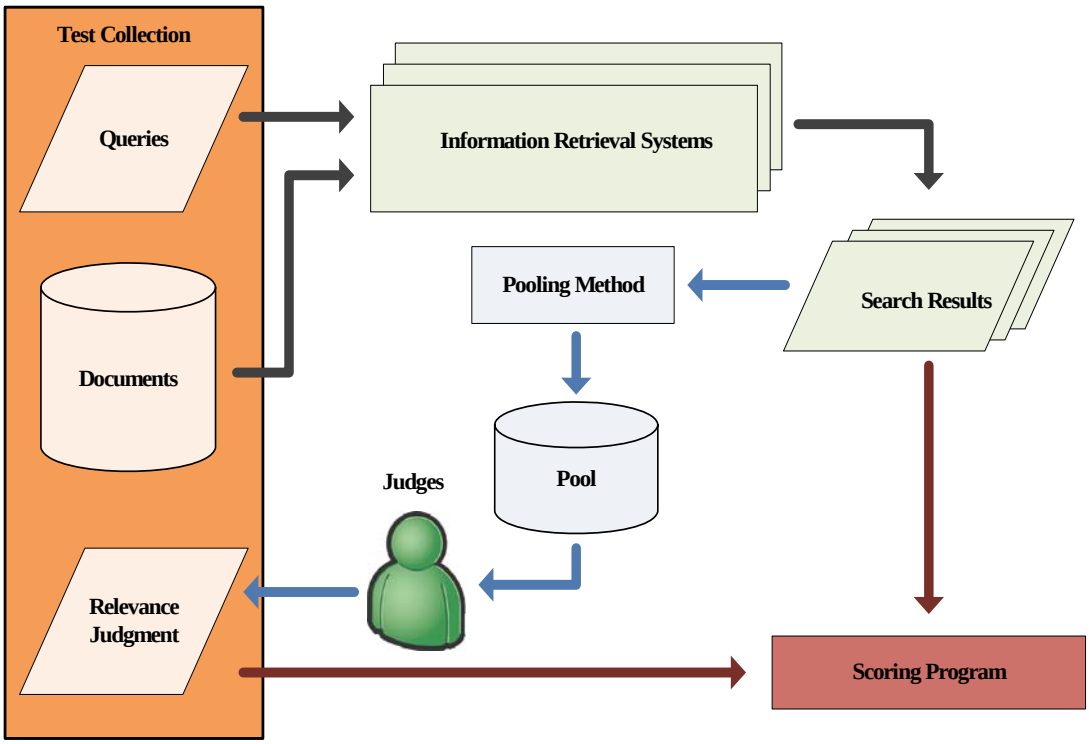

圖 1. 評估資訊檢索效能之流程圖

本研究的另一主題，探討查詢擴展檢索系統的效能。查詢擴展方法能夠有效提升資訊檢索系統的召回率(Recall)，而我們認爲在籠統查詢上，由於使用者缺乏相關的資訊詳細描述其精確的資訊需求，我們可以透過查詢擴展的方法，幫助使用者找尋到更多更相關的文件。在查詢擴展的研究，我們將維基百科視爲同義辭典，並結合虛擬關聯回饋的機制，自動的查詢擴展，以提升檢索的效能。維基百科擁有超鏈結的特性，每個條目(頁面)中都包含了許多超鏈結的鏈結文字(anchor text)，而這些字詞都是與條目擁有高度相關的字詞，我們將這些相關字詞視爲查詢擴展的候選詞，並與虛擬關聯回饋中的字詞共同競選，以挑出更多相關的詞彙作查詢擴展。

2. 文獻探討

2.1 資訊檢索測試集 (Information Retrieval Test Collection)

資訊檢索評估是透過在一致性的評比環境中進行測試，衡量不同的資訊檢索技術或各檢索系統間的效能評比。此方法最早是由 Cleverdo 在 Cranfield II(Cleverdon, 1967)時提出，主要以文件集(Document set)、查詢主題集(Query set)、相關判斷集(Relevance Judgment set)建構測試集，作爲評估各系統的基礎資料，並訂定一套效能評估準則，評估各資訊檢索技術及檢索系統間的效能。在 1992 年美國 Defense Advanced Research Projects Agency (DARPA)與 National Institute of Standards and Technology(NIST)共同舉辦 Text REtrieval Conference (TREC) (Harman, 1993)，TREC 提供了當時最龐大的測試集，使得資訊檢索測試的環境更接近於真實。

繼 TREC 之後，各界對於提供資訊檢索一致性的評比環境，許多機構亦開始提供不同語系，相似於 TREC 的大型測試集，例如：Cross Language Evaluation Forum (CLEF) (Braschler, 2001)、NACSIS Test Collection for IR Systems (NTCIR) (Manning , Raghavan & Schütze, 2008)，這些機構與 TREC 每屆都會舉行各種不同的資訊檢索任務(Harman, 2005) (Ferro & Peters, 2008) (Kando, 2007)，如：單語言資訊檢索(Single Language IR, SLIR)、跨語言資訊檢索(Cross Language IR, CLIR)、跨多語言資訊檢索(Multi-Lingual IR, MLIR)。

在文件集中 TREC、CLEF、NTCIR 都會將收集的新聞文件加上各種不同的標記(Tag)，詳細的區分文件的特性，以利於系統進行剖析各種不同資訊，並有效的將其應用，表 1 爲 NTCIR 文件標記之說明(陳光華, 2001) (陳光華, 2004) (Chen & Chiang, 2000)，圖 2 爲 NTCIR 所使用的中文文件範例，其資訊是在陳光華教授所收集的 CIRB040r 文件集，本研究是透過參與 NTCIR-6 CLIR 與 NTCIR-7 IR4QA 之任務所取得。

表1. NTCIR使用文件標記說明

Tag		
<DOC>	</DOC>	The tag for each document
<DOCNO>	</DOCNO>	Document identifier
<LANG>	</LANG>	Language code: CH, EN, JA, K
<HEADLINE>	</HEADLINE>	Title of this news article
<DATE>	</DATE>	Issue date
<TEXT>	</TEXT>	Text of news article

近幾年 TREC、CLEF、NTCIR 等在不同的任務也使用了不同的文件集，如：TREC 的 Web Track(Craswell & Hawking, 2004)、Terabyte Track (Büttcher , Clarke & Soboroff 2006)、Blog Track (Ounis & Soboroff, 2008)，CLEF 的 WebCLEF (Jijkoun & Rijke, 2007)，NTCIR 的 WEB Task 等 (Eguchi , Oyama, Aizawa & Ishikawa, 2004)，其文件集的型態不同於以往的新聞文件，而是網際網路上的網路文本，即每一篇文件皆爲網頁。

```
<DOC>
<DOCNO>edn_xxx_20000101_0265056</DOCNO>
<LANG>CH</LANG>
<HEADLINE>總統府前升旗　喜迎千禧曙光</HEADLINE>
<DATE>2000-01-01</DATE>
<TEXT>
<P>記者蕭君暉／台北報導</P>
<P>全台陷入迎接千禧的狂熱！台東太麻里的海邊，數以萬計的人潮共同迎接台灣第一道千禧曙光；...。</P>
</TEXT>
</DOC>
```

圖 2. NTCIR 中文文件範例

在查詢問題集中使用者描述其資訊需求稱之爲查詢問題(Query)，查詢問題中包含使用者所需的查詢關鍵字。TREC 提出使用查詢主題集(Topic Set)取代查詢問題集(Query Set)，其中的不同，在於查詢主題集以多欄位的方式陳述各種不同層次的查詢需求，而後 NTCIR、CLEF 等亦使用查詢主題集作爲各種查詢需求的陳述。表 2 爲 NTCIR 所使用 Topic 的標記及其意義(陳光華, 2001)(陳光華, 2004)(Chen & Chiang, 2000)。

TREC 在早期是使用模擬的方式建構查詢主題集，而 NTCIR 的查詢主題集則是來自於真實使用者的需求(CIRB010)，之後透過人工以及全文檢索工具的輔助，濾除敘述不清、不夠詳盡，或者主題涵蓋範圍太廣泛、主題不符合等(陳光華, 2001)。

表2. NTCIR 查詢主題標記及說明

<TOPOIC>	</TOPOIC>	The tag for each topic
<NUM>	</NUM>	Topic identifier
<SLANG>	</SLANG>	Source language code: CH, EN, JA, KR
<TLANG>	</TLANG>	Target language code: CH, EN, JA, KR
<TITLE>	</TITLE>	The concise representation of information request, which is composed of noun or noun phrase.
<DESC>	</DESC>	A short description of the topic. The brief description of information need, which is composed of one or two sentences.
<NARR>	</NARR>	A much longer description of topic. The <NARR> has to be detailed, like the further interpretation to the request and proper nouns, the list of relevant or irrelevant items, the specific requirements or limitations of relevant documents, and so on.
<CONC>	</CONC>	The keywords relevant to whole topic.

相關判斷即爲由判斷者(人)判斷查詢主題與文件集中的每篇文件之相關程度。現今測試集的規模都相當龐大，無法閱讀所有文件，因此發展出 Pooling Method (Kageura & others, 1997)，此方法是假設真正相關的文件，會被多數的資訊檢索系統所檢索出，將所有資訊檢索系統檢索的結果，建構一個相關文件候選的 Pool，評估者只需要判斷相關候選 Pool 的文件，以此可降低建構相關判斷集的時間與人力。

NTCIR 在進行相關判斷時(陳光華, 2001)，每位判斷者必須詳細閱讀並瞭解查詢主題，並以查詢主題中<NARR>欄位作爲主要的判斷依據，將文件分到判斷者認爲最適當的相關類別。NTCIR 的相關判斷集分爲四個層級，如表 3.所示，然而 TREC 的相關判斷集是採取二元分層的方式(Harman, Braschler, Hess, Kluck, Peters & Schäuble, 2001)，TREC 的作法被視爲資訊檢索評估的標準流程，所以 NTCIR 亦採取二元分層的方式產生兩組相關判斷集，即嚴謹相關(Rigid Relevance)以及寬鬆相關(Relaxed Relevant)，嚴謹相關視"S"與"A"爲相關，寬鬆相關視"S"、"A"、"B"爲相關。

表3. NTCIR相關判斷層級

Label of Relevance	Sign	Score
Highly Relevant	S	3
Relevant	A	2
Partially Relevant	B	1
Irrelevant	C	0

NTCIR 中每一個查詢主題由 3 位判斷者做判斷，判斷者必須在一段連續的時間內完成一個查詢主題的判斷工作，以儘量確保判斷標準前後的一致性。之後透過以下公式結合三位判斷者的判斷:

$$R = \frac{avg(X_A + X_B + X_C)}{Z} \quad (1)$$

其中 X 爲各判斷者對文件所給的類別等級，A, B, C 則爲三位判斷者之代號，Z 爲正規化參數(爲最高分數)。所得的值 R 介於 0 與 1 之間，若 R 愈接近 1，則表示二者愈相關。 其結合相關判斷分數時是視每位判斷者對於相關判斷的整體貢獻是相同的，所以不作特別加權，並且每個判斷都是獨立的。

嚴謹相關判斷集以及寬鬆相關判斷集的區分，則是透過訂立兩個門檻值 0.6667 與 0.3333，區分嚴謹以及寬鬆，如前述，嚴謹爲 B 以上之分數(2)，所以嚴謹的門檻值是透過以下的運算所取得，寬鬆則是 C 以上之分數(1)，亦透過相同的運算取得寬鬆之門檻值。

2.2 查詢擴展 (Query Expansion)

查詢擴展爲資訊檢索系統中常見的技術，最早由(Robertson & Sparck Jones, 1976) 所提出，主要的概念爲將原始的查詢詞擴展，將其加入至原始的查詢中，再使用擴展後的查詢作進一步的檢索，此種方法能夠有效提升資訊檢索系統的召回率(Recall)。

關聯回饋(Relevance Feedback)是一種藉由反覆查詢提高檢索精準度技術，其概念爲透過第一次檢索出來的文件，取得其中與原始查詢的關聯程度，並回饋給檢索系統，而系統可以利用這些相關或是不相關的文件，修改檢索系統中的各種參數值或是修改原使的查詢(Query)，之後進行下一次的檢索時，即可得到較精準的檢索結果。

傳統關聯回饋方法以 Rocchio 等所提出的演算法(Joachims, 1997)最具代表性，其公式如下：

$$Q_{new} = \alpha Q_{current} + \frac{\beta}{|R'|}\sum_{D\in R'} D - \frac{\gamma}{|NR'|}\sum_{D\in NR'} D \tag{2}$$

Q_{new} 爲經過關聯回饋後產生的新查詢，$Q_{current}$ 爲舊有的查詢，R'表示與查詢相關的文件，NR'表示與查詢不相關的文件，α、β 及 γ 爲參數比值，α+β+γ=1。其中 γ 通常設爲 0，因爲在真實世界中比較少會去區分文件爲不相關，通常使用者最多只標注哪些文件爲相關。

關聯回饋分爲三種：顯性回饋（Explicit feedback）、隱性回饋（Implicit feedback）、和隱蔽的回饋（blind feedback）或 "虛擬"關聯回饋(Buckley, Salton, & Allan, 1994)(Harman, 1992)(Salton & Buckley, 1990)(Saracevic, 1970)(Sparck Jones & Rijsbergen, 1976)。顯性回饋指使用者主動標記哪些文件是相關或不相關。隱性回饋指系統監視使用者的行爲，像是使用者有點選或沒點選哪些網頁、觀看網頁多久時間，收集這些資訊可以讓系統個人化。

隱蔽的回饋又稱爲虛擬關聯回饋(Fan, Luo, Wang, Xi, &Fox, 2004)，由於使用者自己提供關聯回饋的意願不高，因此需要系統自動產生出模擬使用者所做的關聯回饋。系統會先進行一次檢索，擷取出 Top N 篇的文件當做虛擬關聯回饋文件，用來新增查詢字詞，讓最終檢索效能提高。

Okapi BM25 是一種排序的公式，搜索引擎在接受查詢句後，使用此公式排序相符合文件的高低，藉此找出相關文件出來。此公式是在 1970 年所發展出來屬於機率模式的演算法(Robertson & Sparck Jones, 1976)。現今許多資訊檢索的方程式都是改進自 BM25(Robertson, Walker, Sparck Jones, Hancock-Beaulieu & Gatford, 1995)。

Okapi 的公式如下：

$$Sim\left(Q, D_n\right) = \sum_{T\in Q} w^1 \frac{(k_1+1)tf\,(k_3+1)qtf}{(K+tf)(k_3+qtf)} \tag{3}$$

$$w^1 = \log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{4}$$

$$K = k_1((1-b)+b\frac{dl}{avdl}) \tag{5}$$

Okapi BM25 演算法將會進行兩次的檢索，第一次的檢索結果當成虛擬關聯回饋的文件，再從其中虛擬關聯回饋文件中，挑選出 n 個字詞加入查詢中，達到查詢擴展的作

用，再進行第二次的檢索。而在第一次的檢索時，其中的其小 R、r 的數值爲 0，第二次檢索時透過第一次檢索後取得 R、r 的數值。

第一次挑選出 n 個字詞加入原始的查詢中，挑選是透過 TF-IDF 的公式計算每個字詞的權重值：

$$w(i) = tf(i) \times idf(i) \tag{6}$$

$$idf(i) = \log \frac{N}{df(i)} \tag{7}$$

其中 tf(i)爲 i 字詞所出現的頻率(次數)，df(i)爲 i 字詞出現在多少篇文章的頻率，N 爲所有文章數量。

3. 建置簡統查詢的資訊檢索測試集

我們介紹如何利用維基百科全書的釋出資料(dump data)建置我們的資訊檢索測試集，以及建置測試集的成果。我們依據前述所介紹的測試集建構之標準流程建置測試集，在以下分別介紹我們所建構的文件集、查詢主題集、相關判斷集。

3.1 文件集 (Document Set)

我們希望建構一個繁體中文的文件集，所以必須將維基百科的內容作簡體中文轉換爲繁體中文的處理。在本研究中我們是透過 MediaWiki (http://www.mediawiki.org/)建置本地端的維基百科網站，將簡體中文與繁體中文參雜的內容，轉換爲只有繁體中文的內容。

文件收集流程如圖 3 所示，我們會將維基百科釋出資料中命名空間條目(Namespace Articles)以及重定向條目(Redirect Articles)濾除，這是因爲命名空間條目以及重定向條目，並沒有百科全書實質上的資訊，所以必須將其濾除。我們由維基百科釋出資料所擷取條目真正有內容的文件集 211,147 篇。

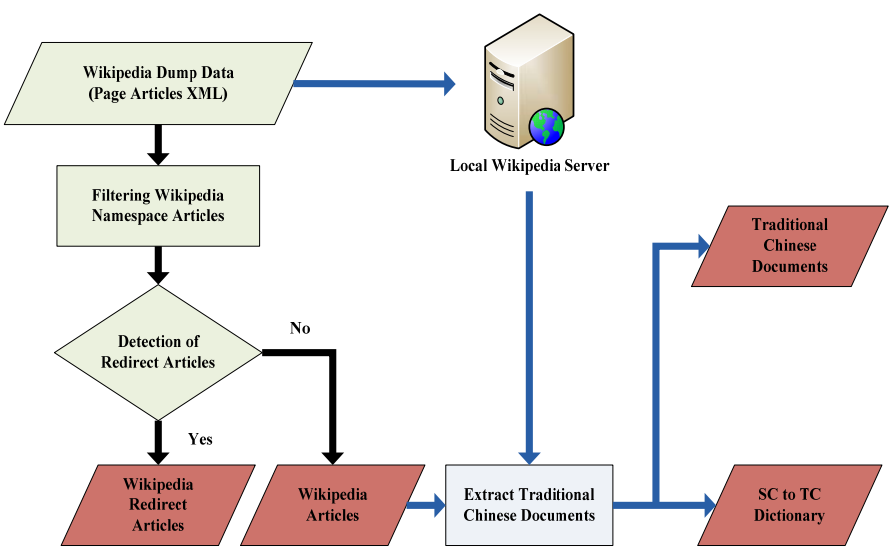

圖 3. 文件集收集流程

3.2 查詢主題集 (Topic set)

爲了使測試集的評比更接近真實的資訊檢索環境，查詢主題的建立通常都是透過真實使用者對於資訊的需求，並且須涵蓋多個不同的主題。所以我們收集了真實使用者對於維基百科的資訊需求。爲了評估現有的資訊檢索系統，是否能夠滿足使用者作籠統查詢，我們針對有缺乏精確關鍵字知識經驗的使用者收集其查詢 。我們透過訪談收集使用者在尋找資訊時，知道自己第一次查詢時使用的關鍵字是不精確的，並且經過幾次查詢可以找到可以滿足的資訊需求的文件。使用者修改其原始的查詢關鍵字最終找到需要的文件，這些修正前後的查詢就是我們所要蒐集的查詢主題。

挑選查詢總共有三個階段的篩選，第一階段，依據以上的假設蒐集，我們共收集了 84 個籠統查詢。在第二階段，使用搜尋引擎輔助，使用籠統查詢到維基百科官方搜尋引擎作檢索，在前 Top 40 筆搜尋結果中必須有三篇以上使用者認爲非常相關或相關的文件，如果沒有則將此查詢濾除。最後再由使用者經由多次的檢索與閱讀其所需的資訊，修改籠統查詢成精確查詢，將這些資訊作更詳細的描述。如果發現此查詢主題不適用這個流程則將其刪除。

經由三階段篩選後，我們總共建立 34 個查詢主題。我們所建構查詢主題集各種使用的標記(Tag)及其定義如表 4 所示。

表4. 我們建置的查詢主題標記說明

Tags		Description
<TOPIC>	</TOPIC>	The tag for each topic
<NUM>	</NUM>	Topic identifier
<SLANG>	</SLANG>	Source language code: CH
<TLANG>	</TLANG>	Source language code: CH
< C-Query>	</ C-Query>	The concise representation of the general query
<DESC>	</DESC>	Description of this topic with one sentence
<NARR>	</NARR>	Length description of this topic, which contains two more tags: <BACK>, and<REL>
<BACK>	</BACK>	The background knowledge of the topic
<REL>	</REL>	How to judge the relevance
< EXACT >	</ EXACT >	The concise representation of information request, which is composed of noun or noun phrase.

3.3 相關判斷集 (Relevance Judgment set)

爲了降低建構相關判斷集所花費的時間以及人力，必須使用 Pooling Method 建構相關判斷候選集。我們的 Pool 是由 Google (http://www.google.com/) 、 Altavista (http://www.altavista.com/) 、 Wikipedia (http://zh.wikipedia.org/) 、 Wikigazer

(http://wil.csie.cyut.edu.tw/Wikigazer/)等線上搜尋引擎，分別使用籠統查詢以及精確查詢作查詢，每次查詢取得最多 1000 筆的搜尋結果，查詢時間爲 2009 年 1 月，所以每個查詢主題會由 8 個搜尋結果建構成一個 Pool。其中 Wikipedia 及 Wikigazer 爲維基百科專屬的搜尋引擎，而 Google 及 Altavista 則是透過此兩種搜尋引擎指定網域搜尋中文維基百科，例如:兵法 site:zh.wikipedia.org。

當完成上述之前處理後，我們將建構每個查詢主題的 Pool(相關候選文件)，以提供相關判斷者作判斷。建構 Pool 之後，每個查詢主題皆有兩人參與相關判斷，每篇文件與查詢主題相關程度是依據 NTCIR 所訂定的四個層級(2.1 小節)，之後結合兩位相關判斷者之判斷分數，是透過公式(8)計算，並且參照 NTCIR 之寬鬆及嚴謹門檻值，0.3333 與 0.6667，產生兩組相關判斷集，即寬鬆相關判斷集以及嚴謹相關判斷集。

$$R = \frac{avg(X_A + X_B)}{Z} \tag{8}$$

我們利用 kappa 統計量，統計每個查詢主題的兩位判斷者所做判斷之一致性分析，其公式如(9)所示，kappa 的假設爲:判斷者在有意識的情況下所做的判斷，其一致性的結果應該大於隨機判斷的結果，其中 P(A)爲兩位判斷者所做的判斷中，觀測到一致性判斷的機率，P(E)則代表爲兩位判斷者偶然一致性判斷的機率。根據 An Introduction to Information Retrieval(Manning, Raghavan, & Schütze,2008)第八章所說明，kappa 值大於 0.8 爲屬於好的一致性判斷，若借於 0.67~0.8 之間則屬於能認可的一致性判斷，如果小於 0.67 則屬於不好的判斷。

$$kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{9}$$

圖 4 爲我們各個查詢主題的 kappa 值統計分析，其總平均 kappa 值爲 0.91，最差爲查詢主題編號 22，其 kappa 爲 0.63，而編號 5、17、20，是介於 0.67~0.8 之間，其餘 30 個查詢主題皆高於 0.8。基於 kappa 的統計分析，我們將捨棄查詢主題 022，其餘 33 個查詢主題則用於評估線上搜尋引擎與資訊檢索系統之效能。

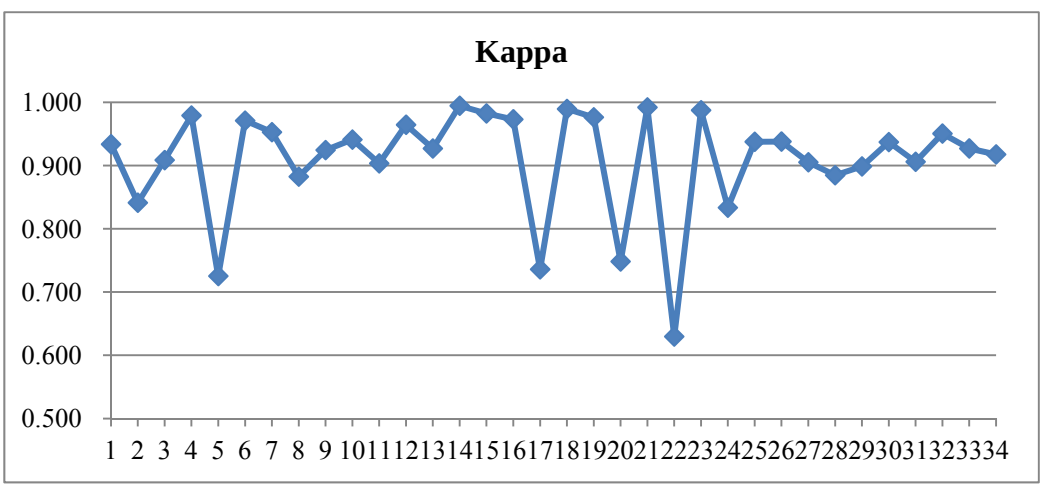

圖 4. 相關判斷的 Kappa 統計分析

4. 查詢擴展方法與系統

4.1 維基百科查詢擴展 (Wikipedia Query Expansion)

維基百科查詢擴展的概念，是利用維基百科擁有高品質、更新迅速的特性，當作是額外的同義辭典，以此輔助虛擬關聯回饋機制的查詢擴展方法。維基百科中每個頁面(條目)都有一個唯一的標題，而在每個頁面中都包含了各種超鏈結以及其鏈結文字連繫到其它相關的條目，而這些鏈結文字即爲與此條目高度相關且重要的字詞，利用原始的查詢中的關鍵字，與維基百科中的條目名稱比對，並收集此條目中的鏈結文字作爲查詢擴展的候選字詞。此方法在 NTCIR-7 IR4QA 的任務中，已證明適當的使用維基百科查詢擴展，可以有效的增進檢索效能(Hsu, Li, Chen & Wu, 2008)。

在以上的維基百科查詢擴展方法，只考慮到條目中的超鏈結，亦即只使用到鏈結結構中的鏈出鏈結，而在本篇論文中我們還使用了鏈入鏈結的鏈結文字的資訊，因爲此資訊與鏈出鏈結相同，很可能也是有高度相關且重要的資訊，所以我們將其加入到維基百科查詢擴展方法中，以期能夠找出更多相關的字詞，增進檢索的效能。

圖 5 爲我們查詢擴展檢索系統的系統流程圖，我們使用 OKAPI BM25 作爲排序的演算法，BM25 參數設定爲：k_1=1.2、k_3=7、b=0.75。我們系統的檢索可以分爲三個步驟，首先由第一次檢索的結果中取出 Top 100 篇文件作爲虛擬關聯回饋的文件，第二，由維基百科中擷取出與查詢字詞相關的字詞，第三，使用 TF-IDF 計算查詢擴展候選字的權重，最後由這 Top 100 篇虛擬關聯回饋文件和維基百科網站中，挑選出 n 個查詢擴展字詞加入原始查詢之中，作查詢擴展，最後再進行第二次的檢索

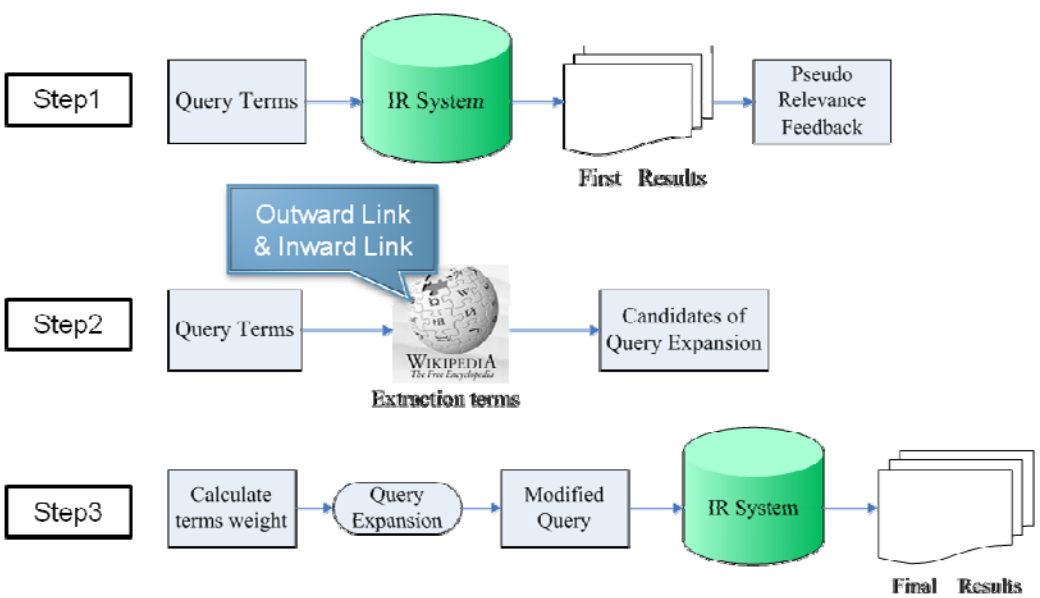

圖 5. 維基百科查詢擴展檢索系統流程圖

5. 實驗結果與分析

我們總共有三種實驗，這三種實驗所使用的資料爲我們所建置的籠統查詢的 Web 資訊檢索測試集。實驗一爲使用此測試集評估線上搜尋引擎以及查詢擴展檢索系統的檢索效能，而我們所使用的查詢擴展檢索系統爲 2.3 節所介紹的 Okapi BM25。實驗二爲 4.1 節所介紹的經由維基百科查詢擴展的效能評估。

5.1 實驗的目的與限制

在本文中，主要使用的資料爲 2009 年 1 月 16 日中文維基百科的釋出資料所建構的測試集，而我們所使用的評分方式爲 MAP，我們實驗的主要目的爲評估線上搜尋引擎的檢索效能，以及各種查詢擴展方法在檢索效能上的評估。在檢索時所使用的查詢有兩種，一爲籠統查詢，我們稱使用籠統查詢作檢索爲 C-run，另爲精確查詢，而使用精確查詢作檢索則爲 E-run。在評分時所使用的答案也分爲寬鬆(Relax)與嚴謹(Rigid)兩個答案集。所以在每個評估上會取得 4 種結果。我們的檢索系統主要是使用 Okapi BM25，其 Okapi 參數設定爲:K1=1.2、K3=7、B=0.75。

5.2 實驗的前處理

由於中文語系單字與單字之間並沒有以空格隔開，因此在中文語言的資訊檢索中，將句子斷詞這個前處理的步驟是相當重要的，在中文斷詞的部份，我們主要使用的是中研院 CKIP 小組開發的斷詞工具，其斷詞的平均準確度能夠達到 95%（http://ckipsvr.iis.sinica.edu.tw/），另外在處理爲中文維基百科中文斷詞時，我們避免將維基百科的條目名稱斷開，以提升檢索系統以及查詢擴展的效能。

5.3 評估方法

在本研究中主要以 MAP(Mean Average Precision) 評分公式，評估線上搜尋引擎以及我們所提出的新的 Okapi BM25 檢索系統的檢索效能 。線上搜尋引擎包含: Google、Wikipedia、Altavista、Wikigazer。MAP 公式如下:

$$MAP = \frac{1}{T}\sum_{j=1}^{T}\frac{\sum_{i=1}^{r_j}\frac{i}{Doc\ (i)}}{r_j} \tag{10}$$

T 爲總共查詢主題數，*r* 表示在檢索文件中相關文件的數量，*Doc(i)*表示第 *i* 篇相關文件被檢索出來時，檢索文件的數量。

5.4 實驗1：評估線上搜尋引擎與查詢擴展演算法

5.4.1 實驗說明

本實驗主要評估的目標爲線上搜尋引擎與查詢擴展演算法，線上搜尋引擎有：Google、Wikipedia 官方搜尋引擎、Altavist、Wikigazer，查詢擴展演算法爲原始的 Okapi BM25，其 Okapi 參數設定爲:K1=1.2、K3=7、B=0.75，虛擬關聯回饋文件是第一次檢索的 Top 100 篇文件，並使用標準的 TFIDF 計算 Top 100 篇文件中的查詢擴展候選字詞，最後挑選 50 個字詞作查詢擴展。

5.4.2 實驗1結果

表5. C-run：使用寬鬆評估各搜尋引擎以及Okapi之MAP

	Okapi*	Google	Wikigazer*	Altavista	Wikipedia
Relax-MAP	**0.184**	0.145	0.133	0.087	0.082

表6. C-run：使用嚴謹評估各搜尋引擎以及Okapi之MAP

	Okapi*	Google	Wikigazer	Altavista	Wikipedia
Rigid-MAP	**0.185**	0.126	0.119	0.078	0.078

表7. E-run：使用寬鬆評估各搜尋引擎以及Okapi之MAP

	Google	Wikigazer	Okapi	Wikipedia	Altavista
Relax-MAP	**0.289**	0.287	0.259	0.198	0.189

表8. E-run：使用嚴謹評估各搜尋引擎以及Okapi之MAP

	Google	Wikigazer*	Okapi	Altavista	Wikipedia
Rigid-MAP	**0.326**	0.303	0.224	0.224	0.190

其中*代表此項數值與下一項(右側)數值，經由 T-test 統計檢定，其 P 值小於 0.05，則代表有顯著的差異。而由表 5~表 8 我們可以觀察到以下幾個重點：

(1)無論是使用寬鬆或嚴謹評估這些搜尋引擎或 Okapi，都可以發現 C-run 的檢索效能明顯的低於 E-run，這代表在現實生活中，在我們常使用的關鍵字搜尋引擎，如果使用了籠統查詢，其效能是不太能滿足使用者的需求，即使是著名的搜尋引擎 Google，其 C-run 的檢索效能亦不甚理想。

(2)在 C-run 中，我們發現基於虛擬關聯回饋的查詢擴展檢索系統 Okapi BM25 的檢索效能是最好的(高於 Google)，這代表在使用者使用籠統查詢時，這些相關但卻模糊的關鍵字，可以透過虛擬關聯回饋的機制，由其中挑選出正確的字詞，幫助使用者找到更多相關的文件。

(3)另外在寬鬆與嚴謹的結果中，會發現有幾個搜尋引擎的嚴謹會高於寬鬆的 MAP，經由我們分析之後發現，這是因爲嚴謹的相關文件數量低於寬鬆相關文件數量許多，在此情形下，使用 MAP 評估時其分母會縮小許多，而這些系統排序的方式又把嚴謹的答

案排序在很前面，所以嚴謹所得到的 MAP 值才會高於寬鬆的 MAP。

5.5 實驗2：維基百科查詢擴展

5.5.1 實驗說明

在本實驗中，我們將進行兩種實驗，實驗 2-1 爲使用維基百科中的鏈出鏈結與鏈入鏈結的鏈結文字同時作爲查詢擴展的候選字詞，在此稱其查詢擴展方式爲 $WikiQE_{out+in}$，並與 Okapi BM25 的 PRF 中的字詞共同競選，在此稱由 Okapi BM25 的 PRF 查詢擴展方式爲 OkapiQE，此實驗中會呈現使用 $WikiQE_{out+in}$ 與 OkapiQE 不同比重的查詢擴展效能。實驗 2-2 爲更詳細區分維基百科查詢擴展方法，將其區分爲鏈出鏈結 ($WikiQE_{out}$)、鏈入鏈結 ($WikiQE_{in}$)與 OkapiQE，針對不同比重的查詢擴展比較。

在這兩個實驗，我們將進行使用不同的查詢擴展字詞數，由 10、20 到 500，分別使用來自於不同查詢擴展方法的比重作查詢擴展，實驗中分別使用籠統查詢與精確查詢，且評分時也區分爲寬鬆與嚴謹。由於實驗的數據太過龐大，所以我們只針對 10~50 的查詢擴展字詞數的特別數據作呈現。

5.5.2 實驗2-1結果

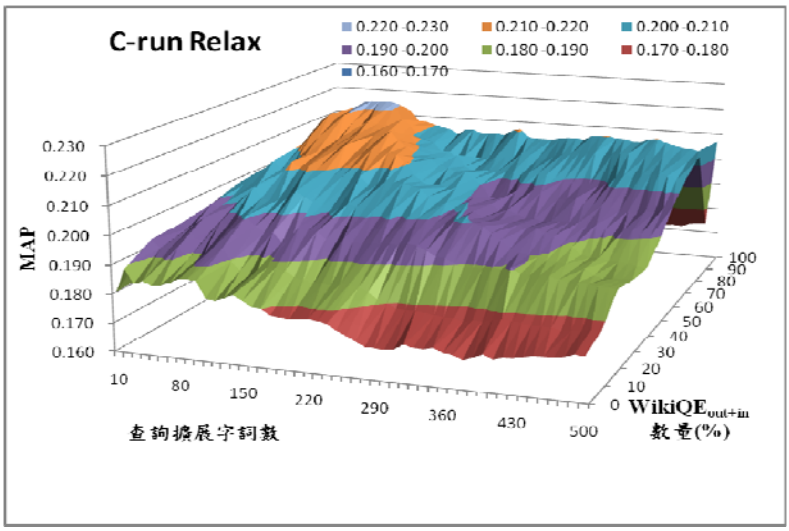

圖 6. C-run：不同查詢擴展方法比重與不同擴展字詞數之寬鬆 MAP

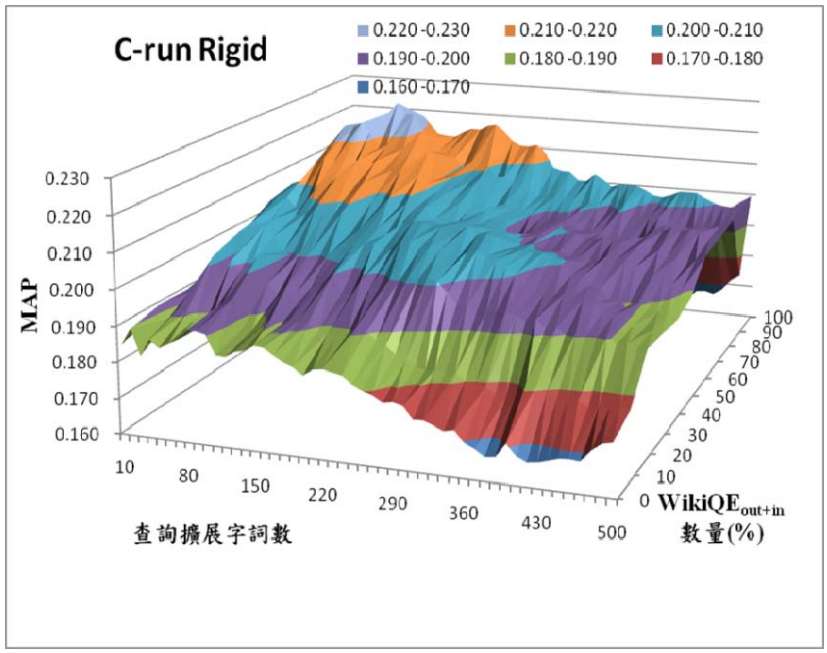

圖 7. C-run：不同查詢擴展方法比重與不同擴展字詞數之嚴謹 MAP

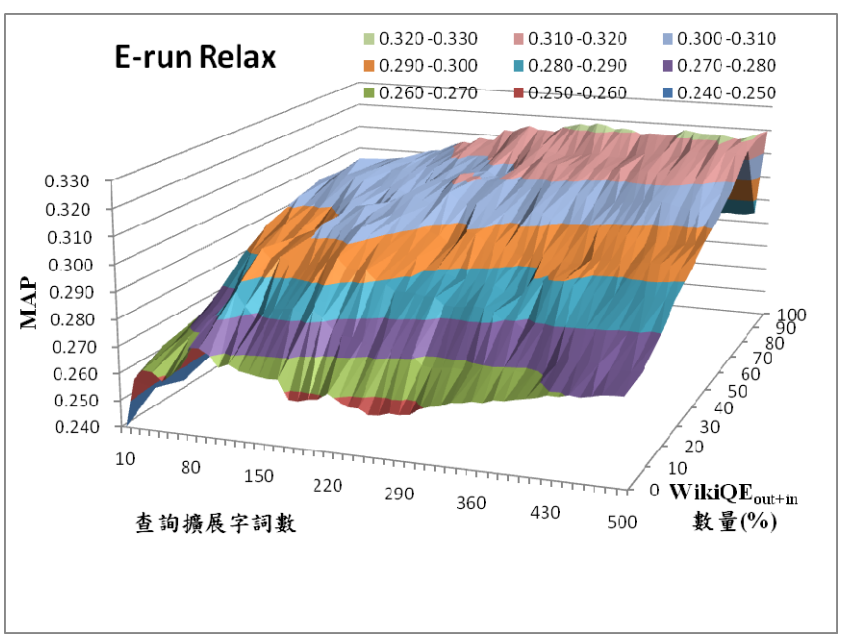

圖 8. E-run：不同查詢擴展方法比重與不同擴展字詞數之寬鬆 MAP

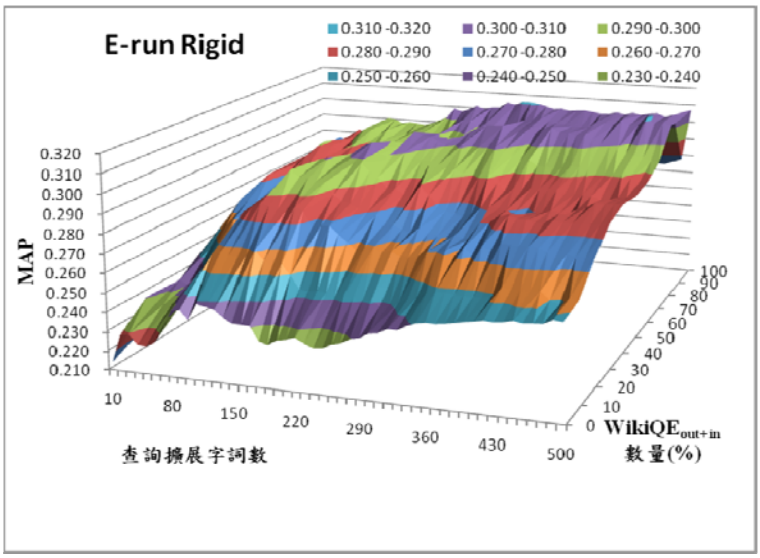

圖 9. E-run：不同查詢擴展方法比重與不同擴展字詞數之嚴謹 MAP

由圖 6~圖 9 的實驗結果中，我們歸納以下幾個重點：

(1)維基百科查詢擴展的確可以改進檢索的效能，無論是在使用籠統查詢或精確查詢，或是在寬鬆、嚴謹的評估下，適度的使用 OkapiQE 與 $WikiQE_{out+in}$ 的結合，能夠幫助挑選出更相關的查詢擴展字詞。在 C-run 中最佳查詢擴展字詞的比重是，80%來自於 $WikiQE_{out+in}$，20%來自於 OkapiQE。而在 E-run 中最佳查詢擴展字詞的比重是，60%來自於 $WikiQE_{out+in}$，40%來自於 OkapiQE。

(2)在不同查詢擴展字詞數的分析上，在有使用 $WikiQE_{out+in}$ 查詢擴展時，在使用較多的查詢擴展字詞上會得到比較好的 MAP，而單純使用 OkapiQE 查詢擴展時，對於使用不同字詞數的影響並不大。

(3)表 9 爲 Topic 編號 009，C-run 檢索，而各種不同查詢擴展字詞的例子。首先，此查詢主題其精確關鍵字是"孫子兵法、孫武兵法、孫子、孫武"等字詞，而使用"風林火山、兵法、謀略"等籠統查詢，可以取得的字詞可由各個欄位作詳細檢視。由於我們所使用的 Okapi BM25 的演算法，因有 PRF 的查詢擴展，所以其演算法容許有重複的字詞出現，而我們的查詢擴展字詞來源總共有三種:PRF、維基百科的 Outward Link 與 Inward Link，所以同一個字詞其容許出現最多三次，由表中我們可以看到查詢擴展的效果，的確是可以有效的找出與籠統關鍵字高度相關的字詞。

表9. 查詢主題009；C-run；OkapiQE、$WikiQE_{out+in}$查詢擴展字詞

	OkapiQE:$WikiQE_{out+in}$		
Topic	100:0	30:70	0:100

General Query 風林火山、兵法、謀略	噴發、**孫子兵法**、武田、風林火山、**孫子**、聖海倫火山、爆發、山鵑、林峯、武田信玄、火星、岩漿、司馬伷、聖托里尼、火山爆發、海倫、helens、武者、地質、司馬法、公園、將軍、司馬、mount、加勒比板塊、火山碎屑、板塊、碎屑、活火山、軍事、大□、灰濛、拉森火山、戰略、孫臏兵法、孫臏、軍團、accessed、凌日、mandarin、mountains、里茲、板垣信方、次髎、國家、川中島、泥火山、形成、harris、九降風	噴發、**孫子兵法**、武田、風林火山、**孫子**、聖海倫火山、爆發、山鵑、林峯、武田信玄、火星、岩漿、司馬伷、聖托里尼、火山爆發、風林火山、(大河劇)、日本戰國、武田信玄、中國、**孫子兵法**、**孫子**、南北朝時代、奧州、北畠顯家、足利尊氏、村上源氏、井上靖、軍師、山本勘助、風林火山、(大河劇)、中國、兵家、中國、學術、知識份子、儒、法、晏武、修文、政治、文事、制度、宋、明、**孫子兵法**、中國、六韜、張良	風林火山、(大河劇)、日本戰國、武田信玄、中國、**孫子兵法**、**孫子**、南北朝時代、奧州、北畠顯家、足利尊氏、村上源氏、井上靖、軍師、山本勘助、風林火山、(大河劇)、中國、兵家、中國、學術、知識份子、儒、法、晏武、修文、政治、文事、制度、宋、明、**孫子兵法**、中國、六韜、張良、中國人民解放軍、中國共產黨、美國、前512年、吳楚之戰、黃帝、風後、握奇經、中國共產黨、抗日戰爭、八路軍
Exact Query 孫子兵法、孫武兵法、孫子、孫武			

5.5.3 實驗2-2結果

表10. C-run : OkapiQE 、$WikiQE_{out}$ 、$WikiQE_{in}$之查詢擴展比較

	C-run									
	Only OkapiQE		Only $WikiQE_{out}$		Only $WikiQE_{in}$		Best		Worst	
	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid
50	0.184	0.185	0.191	0.197	0.158	0.167	0.212 (2:8:0)	0.212 (2:8:0)	0.158 (0:0:10)	0.167 (0:0:10)
40	0.188	0.188	0.188	0.197	0.159	0.169	0.207 (4:4:2 4:6:2)	0.211 (3:4:3 4:5:1)	0.159 (0:0:10)	0.169 (0:0:10)
30	0.187	0.184	0.173	0.184	0.161	0.167	0.204 (5:3:2 5:5:0)	0.210 (4:5:1)	0.161 (0:0:10)	0.167 (0:0:10)
20	0.191	0.190	0.173	0.187	0.157	0.165	0.199 (5:4:1 5:5:0)	0.198 (6:1:3)	0.157 (0:0:10)	0.162 (1:0:9)
10	0.183	0.186	0.164	0.176	0.155	0.165	0.184 (5:5:0 9:1:0)	0.188 (9:1:0)	0.152 (1:0:9)	0.157 (2:1:7)

表11. E-run : OkapiQE 、$WikiQE_{out}$ 、$WikiQE_{in}$之查詢擴展比較

	E-run									
	Only OkapiQE		Only $WikiQE_{out}$		Only $WikiQE_{in}$		Best		Worst	
	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid
50	0.259	0.224	0.253	0.256	0.250	0.243	0.300 (4:4:2*)	0.288 (0:4:6*)	0.250 (0:0:10)	0.226 (10:0:0)
40	0.262	0.225	0.256	0.260	0.245	0.237	0.292 (4:6:0)	0.278 (1:4:5*)	0.242 (0:1:9)	0.225 (10:0:0)
30	0.263	0.231	0.264	0.266	0.243	0.232	0.284 (3:7:0 4:6:0)	0.277 (0:6:4)	0.235 (0:1:9)	0.231 (8:0:2 10:0:0)
20	0.259	0.229	0.257	0.264	0.229	0.223	0.271 (3:7:0)	0.267 (1:9:0)	0.224 (0:1:9)	0.223 (0:0:10 0:1:9)
10	0.240	0.216	0.230	0.242	0.216	0.204	0.249 (5:5:0)	0.249 (1:6:3)	0.212 (0:4:6)	0.203 (1:0:9)

在表內*代表，此數值與只使用 OkapiQE 的 Relax 或 Rigid，其經由 T-test 統計檢定，其 P 值小於 0.05，亦即有顯著的差異。由表 10、表 11 的實驗結果中，我們歸納以下幾個重點：

(1)在 C-run 或 E-run 的結果顯示，只使用 $WikiQE_{out}$ 的效果與只經由虛擬關聯回饋 OkapiQE 的效果是很相近的，$WikiQE_{in}$ 的效果會比較差，這代表 $WikiQE_{out}$ 所找到的字詞是有很高度關連的字詞，而 $WikiQE_{in}$ 的輔助效果比較沒有 $WikiQE_{out}$ 好。在鏈結結構中，鏈出鏈結的鏈結文字通常是與此網頁非常相關的字詞，而鏈入鏈結對此網頁來說，其相關性並不一定很高，所以使用其鏈結文字查詢擴展的效能並不如使用鏈出鏈結的鏈結文字之效能好。

(2)在 Worst 的欄位中，可以看到最差的三種比重，以 OkapiQE 與 $WikiQE_{in}$ 占的比例最高，這代表此兩種擴展方式的效果是很接近的。在 C-run 中，Best 的欄位中，明顯的使用 $WikiQE_{out}$ 比重是偏高的，這代表在籠統查詢檢索中，透過維基百科內容中鏈出鏈結的鏈結文字，可以幫助我們由模糊、籠統的字詞來找到精確的字詞，以提升檢索效果。

(3)經由使用不同查詢擴展字詞數量的比較中，只使用 OkapiQE 的方式並沒有對效能有太大的影響，而只使用 $WikiQE_{out}$ 或 $WikiQE_{in}$ 的方式最佳的擴展字詞數約是 30~50 個擴展字詞。另外，我們可以看到在使用 50 個字詞擴展，E-run 使用 40%的 OkapiQE、40%$WikiQE_{out}$ 與 20%的 $WikiQE_{in}$ 效果會最好，寬鬆的 MAP 有 0.300，已經有 Google 的水平(見表 7)。

6. 結論

在本篇論文中，我們探討在真實世界的使用者，使用籠統查詢描述其資訊需求的情形。本研究依據建置測試集的標準流程，建構了一套基於籠統查詢的資訊檢索測試集。並且能夠使用在評估真實世界中線上的搜尋引擎，如 Google、Wikipedia、Altavista、Wikigazer 等。

由實驗結果證明，目前線上的搜尋引擎並不能滿足使用者使用籠統查詢，雖然 Google 使用精確查詢所檢索的效能的確很好，但是在籠統查詢上的表現並不是很好。另外我們發現基於虛擬關聯回饋的 Okapi BM25 演算法在籠統查詢的效果比 Google 還好，這證明基於 PRF 的查詢擴展方法能夠有效的幫助使用者在使用籠統查詢下的檢索效能。

在本文研究中，我們研究了維基百科鏈結結構的資訊，提出了維基百科查詢擴展方法，利用維基百科頁面中鏈出鏈結與鏈入鏈結的鏈結文字作爲查詢擴展的候選字詞，並與 PRF 的集合結合，提升檢索的效能。由實驗結果證明，在使用籠統查詢的檢索中，透過維基百科鏈結資訊的鏈結文字，可以有效提升單純基於 PRF 查詢擴展方法(Okapi)的檢索效能；而在使用精確查詢的檢索中，此方法也能提升檢索的準確度，甚至可以達到 Google 的檢索水平。

參考文獻

陳光華(2001)。資訊檢索系統的評估 - NTCIR 會議。國立台灣大學圖書資訊學系四十週年系慶學術研討會論文集, 67-86, 台北：台灣大學。

陳光華(2004)。中文資訊檢索標竿測試集之建置, In *The Association for Computing Linguistics and Chinese Language Processing*, 15(4), 4-12.

Braschler, M. (2001). CLEF - Overview of Results. In Cross-Language Information Retrieval and Evaluation. *Lecture Notes in Computer Science 2069*, 89-101, Springer Verlag.

Buckley, C., Salton, G., & Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment, In *Proceedings of SIGIR 17.*, 292-300.

Büttcher, S., Clarke, C. L. A., & Soboroff, I. (2006). The TREC 2006 Terabyte track. In *TREC 2006*, Gaithersburg, Maryland USA.

Chen, K.H., & Chiang, Y.T. (2000). The Design and Implementation of the Chinese IR Benchmark. *Journal of Information,Communication, and Library Science*, 6(3), 61-80.

Cleverdon, C.W. (1967). The Cranfield Tests on Index Language Devices. In *Aslib Proceedings*, 19(6), 173-194.

Craswell, N., & Hawking, D. (2004). Overview of the trec-2004 web track. In *Proceedings of TREC-2004*, Gaithersburg, Maryland USA.

Eguchi, K., Oyama, K., Aizawa, A., & Ishikawa, H. (2004). Overview of the WEB task at the fourth NTCIR workshop. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization.*

Fan, W., Luo, M., Wang, L., Xi, W. & Fox, E. A. (2004). Tuning Before Feedback：Combining Ranking Discovery and Blind Feedback for Robust Retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval,* 138-145.

Ferro, N., & Peters, C. (2008). From CLEF to TrebleCLEF: the Evolution of the Cross-Language Evaluation Forum. In *Proceedings of NTCIR-7 Workshop Meeting*, December 16-19, Tokyo, Japan.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipediabased explicit semantic analysis. In *IJCAI 2007. Proc. of International Joint Conference on Artificial Intelligence*, 1606-1611.

Harman, D.K. (1992). Relevance feedback revisited, In *Proceedings of SIGIR 15*, 1-10.

Harman, D.K. (1993). The First Text REtrieval Conference (TREC-1). *Information Processing and Management,* 29(4), 411-414.

Harman, D.K., Braschler, M., Hess, M., Kluck, M., Peters, C., & Schäuble, P. (2001). CLIR Evaluation at TREC. In *Peters(2001)*, S. 7-23.

Harman, D.K. (2005). *The TREC Test Collections*, Voorhees, E. M. and Harman, D. K. (eds.), TREC: Experiment and Evaluation in Information Retrieval, 21-52.

Hsu, C.C., Li, Y.T., Chen, Y.W., & Wu, S.H. (2008). Query Expansion via Link Analysis of Wikipedia for CLIR. In *Proceedings of NTCIR-7*.

Jijkoun, V., & Rijke, M. (2007). The University of Amsterdam at WebCLEF 2007: Using Centrality to Rank Web Snippets. In *CLEF 2007*, Budapest, Hungary, 2007.

Joachims, T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *Proceedings of 14th International Conference on Machine Learning (ICML-97)*, 143-151.

Kageura, K. & others, eds. (1997). NACSIS Corpus Project for IR and Terminological Research. In *Natural Language Processing Pacific Rim Symposium '97*, 493-496, December 2-5, Phuket, Thailand.

Kando, N (2007). Overview of the Sixth NTCIR Workshop. In *Proceedings of the Sixth NTCIR Workshop*, May 15-18, NII, Tokyo.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*, Cambridge University.

Ounis, I., & Soboroff, C.M.I. (2008). Overview of the TREC-2008 Blog Track. In *TREC 2008.*

Robertson, S.E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *The American Society for Information Science*, 27(3), 129-146.

Robertson, S.E., Walker, S., Sparck Jones, K., Hancock-Beaulieu, M., & Gatford, M. (1995). Okapi at TREC-3. In *Proceedings of the Third Text Retrieval Conference (TREC-3).*

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 288-297.

Saracevic, T. (1970). The Concept of 'Relevance' in Information Science: A Historical Review. In *Introduction to Information Science*, 111-151, New York, USA: R.R.Bowker.

Turmo, J., Comas, P.R., Rosset, S., Lamel, L., Moureau, N., & Mostefa, D. (2008). Overview of QAST 2008. In *Proceedings of the CLEF 2008 Workshop on Cross-Language Information Retrieval and Evaluation*.