國立臺灣大學「台大工程」學刊　第九十期　民國九十三年二月　第 35–42 頁
Bulletin of the College of Engineering, N.T.U., No. 90, February 2004, pp. 35–42

正面和反面資訊需求對資訊檢索效能之影響

EFFECTS OF POSITIVE AND NEGATIVE INFORMATION NEEDS ON INFORMATION RETRIEVAL

馮 廣 明[*]　　陳 信 希[†]

Kuang-Ming Feng[*]　　*Hsin-Hsi Chen*[†]

[*]碩士班研究生　　[†]教授兼系主任
[*]國立台灣大學資訊工程學研究所
[†]國立台灣大學資訊工程學系

[*]Graduate student　　[†]Professor and Chairman
[*]Graduate Institute of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan 10617, R.O.C.
[†]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan 10617, R.O.C.

Abstract

The paper studies positive and negative information needs in information retrieval. The negative information needs are classified into three types. Three methods, including elimination, parametric, and re-ranking approaches, are proposed. NTCIR Chinese test collection is used for evaluation. The experimental results show that parametric approach is effective for information needs of type 1, and re-ranking approach can be adopted for information needs of types 2 and 3 when the base line is poor.

Keywords: information retrieval, negative information need, performance evaluation.

摘　要

本文研究資訊檢索正面和反面資訊需求，我們把反面資訊需求分成三類，並提出三種方法，包括移除法、係數法、和重新排序法。同時採用 NTCIR 中文測試集評估這三種方法的效能，實驗顯示係數法對第 1 種型態的反面資訊需求有效，當以基本法處理第 2 種和第 3 種型態的反面資訊的效能不好時，可採用重新排序法。

關鍵詞：資訊檢索、反面資訊需求、效能評估。

1. 緒　論

使用者的資訊需求除了正面描述外，也包括反面描述，表達不希望檢索出的文件的條件。大部分資訊檢索系統將查詢內的關鍵詞，都視爲相關文件的正面描述，並用以計算與文件的相似度，因此檢索的準確性必然受到影響。

Bierner [1] 分析英文 alternative phrases，例如「such as …」、「other than …」等，以獲得正反面資訊，並轉換成檢索語言。以下是個範例：

What are some web browsers *other than* Netscape?

查詢處理後的結果爲：

What are some web browsers ? :|: ANSWER NOT NEAR (| netscape)

系統先找出符合查詢前半部分條件 (What are some web browsers ?) 的候選答案，再運用後半部分的額外條件 (ANSWER NOT NEAR (| netscape)) 自候選答案中挑出正確答案。

查詢正反面描述之檢索包括兩個議題：查詢正反面資訊的辨識，以及對資訊檢索效能之影響。本文重點在於後者，我們採用 NTCIR [2] 資訊檢索測試集，對正、反面描述之辨識採用人工方式區分。本文之架構如下：第 2 節說明查詢之正反面資訊需求，並將反面資訊需求區分爲三類；第 3 節介紹採用之資訊檢索模型，以及測試用之文件集；第 4 節列出實驗結果；第 5 節總結可能改進檢索效能之做法，以及可能之延伸議題。

2. 正反面資訊需求

NTCIR 測試集 [2] 共有 50 個主題，主要欄位包

括：

<title>：資訊需求之主題 (名詞或名詞片語)。

<question>：對資訊需求之簡要說明 (一至兩個句子)。

<narrative>：對資訊需求之詳細說明，包含相關或不相關之資訊，及其他限制等 (數個句子)。

<concept>：與查詢主題有關之關鍵詞 (數個關鍵詞)。

查詢內容依其欄位及性質，可區分爲三部分：

1. 查詢中描述相關文件應包含哪些資訊，構成正面資訊需求。本文採用的正面資訊，包括查詢中之<question>, <narrative>去除不相關資訊後的描述、以及<concepts>欄位。對正面資訊需求部分之描述，以 P 表示。
2. 查詢中註明包含某些資訊的文件，視爲不相關，該資訊構成反面資訊需求。所有查詢一定包含正面資訊需求，但不一定含有反面資訊需求。實驗的 50 個測試查詢，有 33 個含有反面資訊。<narrative> 欄位中所描述的不相關資訊，構成查詢反面資訊需求，以 N 表示。
3. 對含有反面資訊需求之查詢而言，符合正面資訊需求之描述，且不符合反面資訊需求說明之文件，才算是相關文件。NTCIR 反面資訊需求可分爲以下三類：

(a) 若含有 N 是不相關的 (第 1 類)

若某文件包含符合正面資訊需求的關鍵詞，也包含了符合反面資訊需求的關鍵詞，則該文件被視爲不相關，如圖 1 所示。以下是個範例：

<title>國民卡</title>
<question>
查詢國民卡所可能引發之安全性議題與因應防範之道。
</question>
<narrative>
相關文件內容應包括使用國民卡所可能產生的安全危機，規劃單位對安全問題的說明，所使用的防範措施，依循之安全標準、安全性驗證方法，對承包廠商的技術監督，現有之破碼能力等。推動國民卡的實際措施、議約程序、及遭受的其他阻礙，均是不相關的。
</narrative>

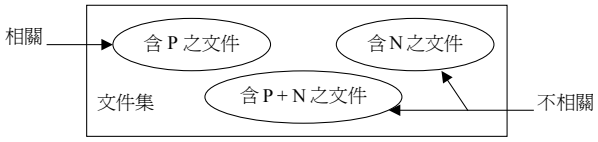

圖 1 第 1 類反面資訊需求

此例檢索國民卡的文件，若包含「推動國民卡的實際措施、議約程序、及遭受的其他阻礙」等資訊 (即 N)，則應視爲不相關文件。NTCIR 測試集共有五個此類型之查詢。

(b) 若僅有 N 是不相關的 (第 2 類)

若某文件包含符合正面資訊需求的關鍵詞，則不論該文件是否出現符合反面資訊需求的關鍵詞，皆可被視爲相關文件。但若某文件中只包含符合反面資訊需求的關鍵詞，而無符合正面資訊需求的關鍵詞，則該文件被視爲不相關文件，如圖 2 所示。以下是個範例：

<title>數位神經系統</title>
<question>
查詢數位神經系統的意義與應用。
</question>
<narrative>
相關文件內容應敘述數位神經系統的意義、內涵、理論架構，可能的應用層面及實施模式，對各項產業的支援影響，以及國內外目前的實施現況與前景。文件中若僅敘述比爾蓋茲出版新書的消息，是不相關的。
</narrative>

文件如果提到數位神經系統，但僅討論「比爾蓋茲出版新書的消息」(N)，而未敘述數位神經系統的意義、內涵等 (P) 之文件應視爲不相關。NTCIR 測試集共有十八個此類型之查詢。

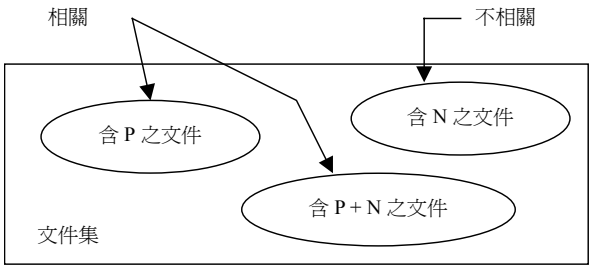

圖 2 第 2 類反面資訊需求

(c) …只有 N_1 而無 N_2 是不相關的 (第 3 類)

圖 3 顯示此類查詢，以下是個範例。

<title>美容瘦身產品查驗</title>
<question>
查詢衛生署、消基會、消保會等單位對各類美容瘦身產品與業者之查驗報告。
</question>
<narrative>
相關文件應有衛生署藥物食品檢驗局、消費者文教基金會、行政院消費者保護委員會等相關單位，對市售之各類化妝、保養、瘦身美容產品及服務業者之任一稽查檢驗報告結果，包括檢驗標準、違規不合格商品之名稱與廠牌、違規原因、對不合格產品的處置等。文件中若只敘述產品管理條例或消費者申訴之案件，而無對特定產品做檢測，是不相關的。
</narrative>

在此例有關美容瘦身產品查驗之文件，若只敘述 「產品管理條例或消費者申訴之案件」，而無「對特定產品做檢測」之資訊，則視爲不相關文件。本次實驗此類型之查詢共有十個。

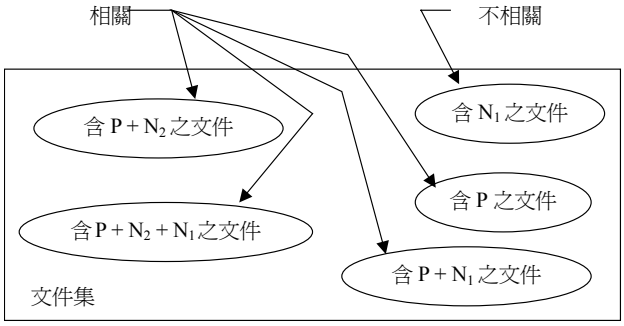

圖 3 第 3 類反面資訊需求

3. 處理策略

本文採用向量空間模型，以 Cosine 公式計算查詢與文件間之相似度 [3]。實驗採用 NTCIR CIRB010 測試集，該文件集共包含 132,173 篇選自台灣五家主要報紙的新聞，且爲 XML 文件格式。

實驗比較之基準爲以 <title>、<question>、<narrative> 等三個欄位作爲查詢之關鍵詞，不特別區分反面資訊需求，而直接進行檢索所得之結果。

本文提出三個方法，詳述如下：

1. 移除法 (elimination approach)

假設反面資訊需求對檢索效能不利，因此將 <Narrative> 欄位中敘述反面資訊需求之文字刪除。

2. 係數法 (parametric approach)

假設反面資訊需求之關鍵詞 N 在計算相似度時，貢獻應是反面而非正面，因此將傳統計算相似度之公式修改爲：

$$sim(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times (\beta_{i,q}\, w_{i,q})}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

其中

$w_{i,q}$ ：查詢中關鍵詞 k_i 之權重，

$w_{i,j}$ ：第 j 篇文件中關鍵詞 k_i 之權重，

$\beta_{i,q} = 1.0$ ：關鍵詞 k_i 在查詢中爲正面資訊需求，

$\beta_{i,q} \neq 1.0$ ：關鍵詞 k_i 在查詢中爲反面資訊需求。$\beta_{i,q}$ 之數值，將以多組數據進行實驗。

3. 重新排序法 (re-ranking approach)

對第 2 類反面資訊需求而言，文件若同時含有 P 及 N 即爲相關，因此含有 N 不一定就是不相關，但若只含 N，則一定不相關。重新排序法先作一次檢索，找出只含 N 之文件，再將這些文件的排名向後調整。詳細步驟如下：

(a) 首先依照移除法，將反面資訊需求移除，並進行第一次檢索，取回排名前一千名之文件。

(b) 將查詢中正面及反面資訊需求反轉作如下之變動：

<title>：仍爲查詢主題。

<question>：原爲正面資訊需求，改爲反面資訊需求。

<narrative>：原爲正面者改爲反面，原爲反面者改爲正面。

(c) 查詢變動後，依係數法再進行第二次檢索，並嘗試多組係數。如此含較多 N、含較少 P，且仍爲同一查詢主題之文件 (即可能之不相關文件)，將會被排在前面。

(d) 由第二次檢索結果中取前一千篇，自第一名開始與第一次檢索之前一千篇作比對。若有相同文章，則將該文章從第一次檢索之排名下降至最後一名 (第一千名)。繼續比對之調整，若有相同者，下降至前次調整排名之前。全部比對調整完畢後，即爲最後之排名。

4. 實驗結果

4.1 基準法

表 1 列出全部 50 個查詢之平均結果，其中查詢之關鍵詞若只包含 <title> 欄位，則 Query Type 爲 **T**。若包含 <title>、<question> 欄位，則 Query Type 爲 **T + Q**。若包含 <title>、<question>、<narrative> 欄位，則 Query Type 爲 **T + Q + N**。若包含 <title>、<question>、<narrative>、<concepts> 欄位，則 Query Type 爲 **T + Q + N + C**。

表 1 顯示實驗之結果隨查詢所含之關鍵詞數目之增加而提昇，以含有 <concepts> 欄位查詢之效果爲最佳。Chen 和 Chen (2001) 曾指出 <concepts> 在 NTCIR 2 的效果太好，無法反應眞正的檢索效能，因此爲了實驗正反面資訊所帶來之影響，後面的研究將僅採用 Query Type 爲 **T + Q + N** 之查詢結果。

表 1 基本法之實驗結果

	T	T+Q	T+Q+N	T+Q+N+C
平均準確率	0.3072	0.3246	0.3919	0.4943

4.2 移除法

4.2.1 第 1 類反面資訊

第 1 類反面資訊需求指若含有 N 便是不相關的文件，基本法在建構查詢時，將反面資訊需求之關鍵詞視爲正面，相似度計算排名較前面之文件，可能同時包含正反面資訊 (P + N) 之不相關文件。

相較之下，移除法在建構查詢時，已將將反面資訊需求之關鍵詞移除，因此相似度排名較前面之文件中，含 P + N 之不相關文件之篇數，必然較基本法減少。圖 4 顯示三種方法對於 5 個第 1 類反面資訊需求實驗的結果。

移除法將平均準確度提升 9.1%。此外，反面資訊需求關鍵詞之鑑別率，對平均準確度之提升亦有相當之影響。以 Topic 21 (國民卡) 爲例，其反面資訊需求之敘述如下：

推動國民卡的實際措施、議約程序、及遭受的其他阻礙

經斷詞後之關鍵詞爲：

及、其他、阻礙、推動、程序、實際、遭受、議約。

上列關鍵詞對使用者希望排除之文件內容描述的相當清楚，因此只要將這些關鍵詞自查詢中移除，平均準確度便可獲得相當的提升 (71.8%)。若以 Topic 8 (經濟痛苦指數) 爲例，其反面資訊需求之敘述如下：

對於環境痛苦指數之議論

經斷詞後之關鍵詞爲：

對於、環境、議論。

上列關鍵詞與 Topic 21 相比則較爲模糊，在許多文件中皆有可能出現。移除這些關鍵詞對相似度排名之影響甚小，因此平均準確度與基本法相比相差甚小。

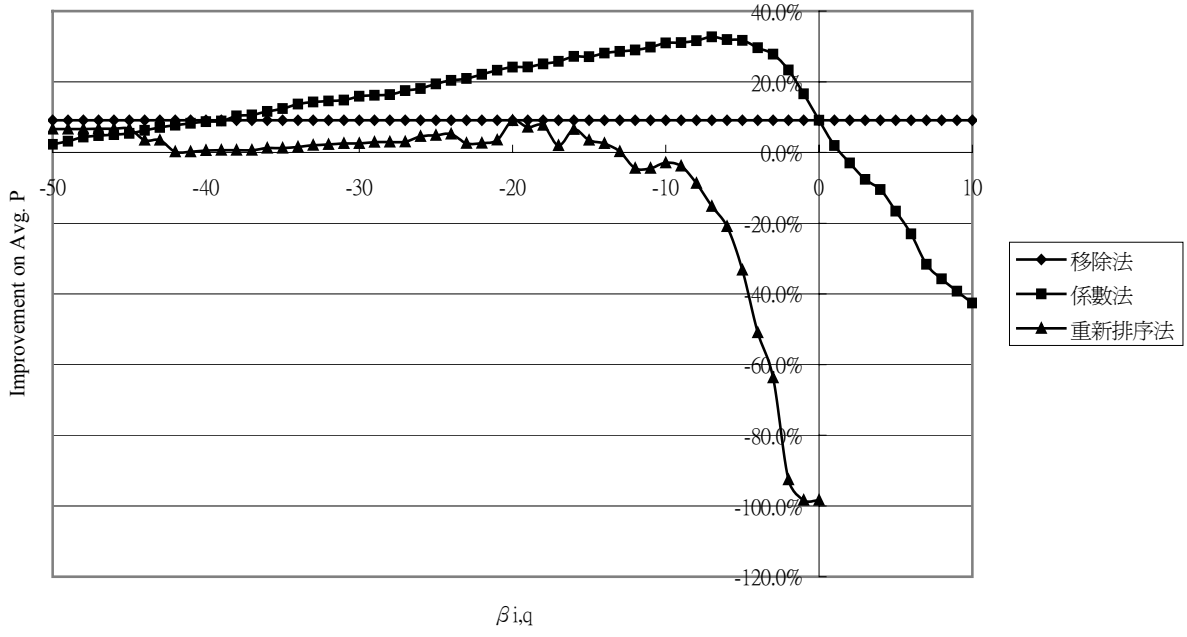

圖 4 對第 1 類查詢之平均準確度改進比較

4.2.2 第 2 類反面資訊

若某文件只含有 N，但不含 P，則視爲不相關。移除法在建構查詢時，已將反面資訊需求之關鍵詞移除。只含 N 之不相關文件之排名有可能下降 (有利)，但含 P + N 之相關文件之排名也可能下降 (不利)，因此移除法對平均準確度之效果有好有壞，而非如第 1 類般將近全面性的提升，圖 5 顯示實驗結果。對 18 個第 2 類查詢而言，平均準確度可提升 2.6%。

4.2.3 第 3 類反面資訊

屬於本類型 10 個查詢，平均準確度提升 0.6%，變化有限，與第 2 類查詢之結果類似。

4.3 係數法

4.3.1 第 1 類反面資訊

移除法雖然可減少排名前面的文件中之不相關文件數，但只運用正面資訊需求關鍵詞計算相似度，並未針對反面資訊需求之關鍵詞作剔除的動作，排名高之文件中仍有可能含 P + N 之不相關文件。圖 4 顯示係數法將查詢中反面資訊需求關鍵詞之 $\beta_{i,q}$ 取爲負數，則文件中若含有 N，則其相似度分數必然降低，相似度之排名也下降。若取 $\beta_{i,q}$ 爲正數，直覺上平均準確度應較基本法爲差，實驗結果如預期。當 $\beta_{i,q}$ 由 10 開始遞減時，對平均準確度之改善率將會遞增，達到某一最大值後又開始遞減。平均準確度提升之最大值爲 32.7% ($\beta_{i,q} = -7$)。

對個別查詢而言，當 $\beta_{i,q}$ 介於 −5 ~ −15 時，對平均準確度之提升可達最大值。至於各查詢間之差異，則如同第 4.2.1 節所述，反面資訊需求之描述越明確，對檢索效能之提昇也越明顯 (Topic 10, 21, 29)，反之則提昇效果有限 (Topic 8, 31)。

4.3.2 第 2 類反面資訊

對本類型 18 個查詢而言，與第 1 類查詢不同的是，平均準確度之提昇百分比到達最大值之 $\beta_{i,q}$ 值介於 0 到 1 之間 (詳圖 5)，之後不論 $\beta_{i,q}$ 遞增或遞減均呈現下降之趨勢。平均準確度提升之最大值爲 2.9%，效果與忽視或剔除反面資訊需求關鍵詞之差異不大。

對個別查詢而言，$\beta_{i,q}$ 值對平均準確度影響之程度與反面資訊需求所提供之條件是否充分有相當大的關係。以 Topic 11 (金融機構合併) 爲例，其反面資訊需求之敘述如下：

陳述金融機構合併之個案

經斷詞後之關鍵詞爲：

個案、*陳述*。

發現平均準確度在 $\beta_{i,q}$ 值小於零時幾乎沒有變動，因爲陳述金融機構合併個案之文章內不一定會含有「個案」、「陳述」等字眼，即使有也不是重要的關鍵詞。

至於變動較大之查詢，我們可發現部分平均準確度達到最大值對應之 $\beta_{i,q}$ 值爲正值 (如 Topic 13「NBA 勞資糾紛」、Topic 15「白曉燕綁架撕票案」等)，而非如同第 1 類查詢皆爲負值，主要原因是「P + N」文件，對第 2 類查詢是相關，對第 1 類查詢而言，則是不相關。

4.3.3 第 3 類反面資訊

對本類型 10 個查詢而言，平均準確度提升之最大值爲 6.9% ($\beta_{i,q} = -6$)。觀察本法改善值之線形，並無第 2 類般明顯的山峰形，反而是隨 $\beta_{i,q}$ 值之遞減而漸趨平緩。對個別查詢而言，當 $\beta_{i,q}$ 值爲正時，除 Topic 16 (受虐兒童) 改善值會隨 $\beta_{i,q}$ 值之遞增，而逐漸上升外，其餘皆呈現下降之趨勢，此點與第 1 類查詢之現象類似。

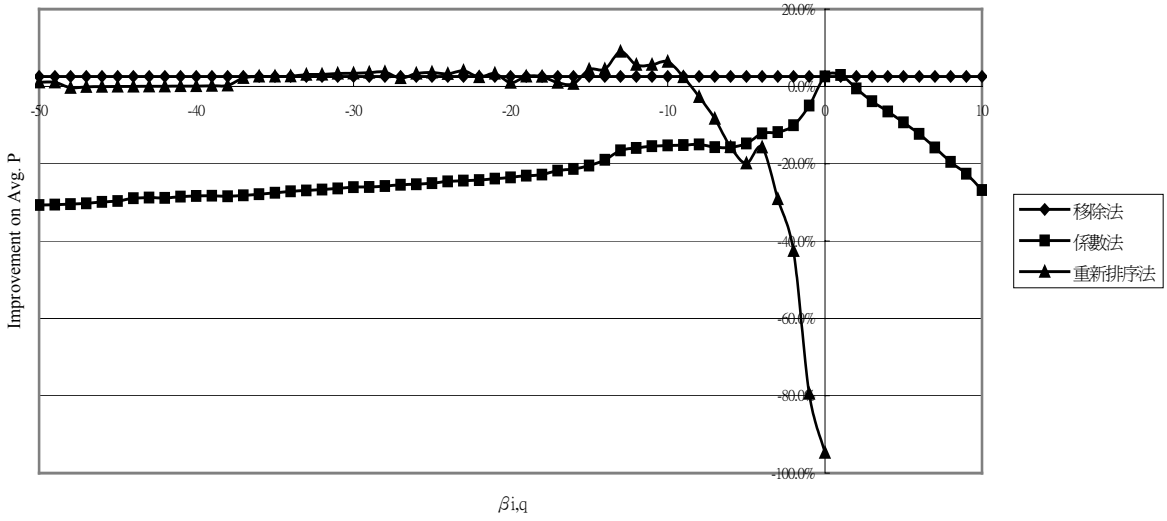

圖 5 對第 2 類查詢之平均準確度改進比較

4.4 重新排序法

4.4.1 第 1 類反面資訊

重新排序法將含反面資訊需求關鍵詞 N 較多、正面資訊需求關鍵詞 P 較少之文件挑出，並將之相似度排名向後調整。挑出這些文件的做法是將正、反面資訊需求對調，並以係數法計算相似度 (作第二次檢索)。

圖 4 顯示重新排序法對 5 個第 1 類查詢平均準確度之影響，$\beta_{i,q}$ 由 0 開始，由於第二次檢索相當於以 <title> 欄位及反面資訊需求 (在此處之角色爲正面) 作檢索，結果與第一次檢索有相當程度的一致性，因此重新排序必然造成平均準確度的大幅下降 (將近 –100%)。此一現象隨 $\beta_{i,q}$ 之遞減而逐漸改善，最後對平均準確度提升之最大值分別爲 9.0% ($\beta_{i,q}$ = –20)，整體效果尚不如移除法。

對個別查詢而言，如同前節所述，反面資訊需求之描述越明確，對檢索效能之提升也越明顯 (Topic 10, 21, 29，Topic 21 甚至可達 366%)，反之則提升效果有限 (Topic 8, 31)。由於各查詢達到平均準確度最大值時之 $\beta_{i,q}$ 值差異甚大，例如以 Topic 21 來說，當 $\beta_{i,q}$ = –5 時出現最大值 366%，此時 Topic 29 之值則爲 –30.5%，因此較難對 $\beta_{i,q}$ 值定出一個合理的範圍，以求得各查詢檢索效能之全面提升。

4.4.2 第 2 類反面資訊

重新排序法對平均準確度之影響詳見圖 5。對 18 個第 2 類查詢而言，平均準確度提升之最大值爲 9.0% ($\beta_{i,q}$ = –13)，整體效果較移除法及係數法明顯。

觀察各查詢之變化可發現，未對反面資訊需求作處理之檢索，成效不彰時，如 Topic 12 (麥可喬登退休)、Topic 27 (反美濃水庫興建)、Topic 43 (CIH 電腦病毒)、及 Topic 49 (海外旅遊保險) 等，重新排序法改善效果相當明顯，且達最大改善值時之 $\beta_{i,q}$ 值，變化不會太大。

4.4.3 第 3 類反面資訊

本法對平均準確度之影響詳圖 6，對 10 個第 3 類查詢而言，平均準確度提昇之最大值爲 17.9% ($\beta_{i,q}$ = –20)，整體效果較移除法及係數法明顯。

5. 結論及延伸研究議題

觀察實驗結果，可歸納以下結論：

1. 對第 1 類反面資訊需求而言、由於只要包含反面資訊需求之關鍵詞 (N) 的文件便是不相關的 (不論是否含有 P)，因此若移除查詢中之反面資訊需求 (移除法)，其效能便可得到提升。若將反面資訊需求對相似度計算之貢獻取爲負數 (係數法)，其效能之提昇將更爲顯著，且其對 $\beta_{i,q}$ 值之變化量較爲穩定。若採用第二次檢索後重新排名的方式 (重新排序法)，效能之提昇不若移除法來得顯著，且其改善量對 $\beta_{i,q}$ 值之變化量較爲敏感。因此整體而言，係數法對第 1 類查詢較爲合適，$\beta_{i,q}$ 可取一介於 –5 ~ –10 間之值。

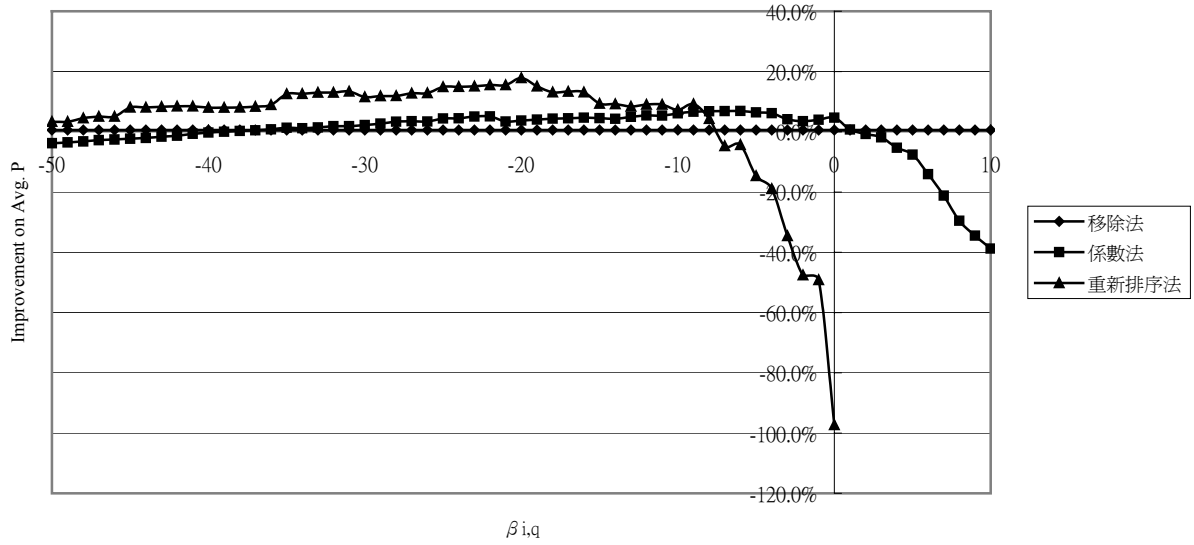

圖 6 對第 3 類查詢之平均準確度改進比較

2. 對第 2 類反面資訊需求而言、只包含反面資訊需求之關鍵詞 (N)，且不含 P 的文件才是不相關的，所以含有 N 不一定就是不相關 (如 P + N)，不若第 1 類查詢般明確。實驗結果證實整體而言，無論採用移除法、係數法或重新排序法，其效能之提升均不甚明顯。重新排序法不甚穩定，但對於部分原來檢索效能甚差的查詢而言則有相當大幅的改善。因此對第 2 類查詢仍可採用基本法，但可設計一介面，使用者可先瀏覽基本法傳回的結果，若其效果甚差 (例如前十篇皆爲不相關)，則可利用重新排序法再檢索一次。
3. 對第 3 類反面資訊需求而言，只包含反面資訊需求之關鍵詞 (N_1)，且不含 P 及 N_2 的文件是不相關的，類似於第 2 類查詢。實驗顯示各法對其效能之提升仍然有限，故可比照第 2 類以基本法爲主，但對效果較差之個別查詢則利用重新排序法再做檢索。

實驗發現反面需求用詞非常不明確時，三種做法之效用有限。因此，如何掌握使用者的資訊需求的語義，需進一步研究。另外，除了查詢內含有正反面資訊之外，文件中也帶有相關資訊。以下是個例子：

一九九五年下半年到九六年春天，中共在臺海製造飛彈危機，結果引來了北京最不願意看到的美國航空母艦艦隊。

在詞彙「不願意」後所跟隨的詞組 (美國航空母艦艦隊)，必然受到前面否定詞的影響，未來擬針對文件層次正反面資訊進一步探討。贊成和反對意見的擷取，是另一種正反面資訊的辨識，也是句子層次分析重要的應用之一，值得深入探討。

參考文獻

[1] G. Bierner, “Alternative phrases and natural language information retrieval,” *Proceedings of the 39th Meeting of the ACL*, Toulouse, France, 2001, pp. 58–65.

[2] K. H. Chen and H. H. Chen, “The Chinese text retrieval tasks of NTCIR workshop 2,” *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, 2001, pp. 51–72.

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

馮 廣 明 (Kuang-Ming Feng) 美國密蘇里大學土木工程系碩士，國立台灣大學資訊工程系碩士，結構工程技師，現任中鼎工程公司土木設計部工程師。

陳 信 希 **(Hsin-Hsi Chen)** 國立台灣大學電機工程學研究所博士，現職國立台灣大學資訊工程學系教授兼系主任，專長爲自然語言處理、資訊檢索與擷取、人工智慧、資料庫系統。

收稿日期 92 年 11 月 19 日、修訂日期 93 年 2 月 2 日、接受日期 93 年 2 月 12 日
Manuscript received November 19, 2003, revised February 2, 2004, accepted February 12, 2004