情報檢索시스팀과 語彙調整

현 은 정
〈KORSTIC 電子計算室〉

檢索시스팀의 性能을 支配하는 要素는 索引言語의 質과 索引作成에 의한 것으로 크게 나누어진다. 索引言語의 質은 再現과 精度에 影響을 미치는 用語의 特定性과 用語間의 關係를잘 나타내며 索引作成時 文獻當 몇개의 用語를 割當하느냐에 따라서 情報檢索시스팀의 性能이 좌우된다. 즉 索引言語와 索引作成過程에서 어휘(Vocabulary)를 效率的으로 調整하는 데 따라 시스팀의 性能을 向上시킬 수 있다.

1. 檢索시스팀에서의 檢索效率

必要한 情報를 檢索하여 要求情報가 全部 探索되고 不必要한 情報가 나타나지 않는다면 理想的이라 할 수 있지만 실제로 이러한 結果를 얻는다는 것은 不可能하다. 情報의 檢索結果가 그 近似値에 어느 정도 接近하느냐를 測定하는 尺度로는 精度率(Precision)과 再現率(recall)이 있다.

$$精度率=\frac{탐색된\ 適合情報}{탐색된\ 全情報}\times 100\,(\%)$$

$$再現率=\frac{탐색된\ 適合情報}{全適合情報}\times 100\,(\%)$$

精度率과 再現率은 높을수록 探索結果가 좋다고 할 수 있는데 精度率과 再現率은 한쪽이 높으면 한쪽이 저하되는 相反關係를 가지고 있다. 어떤 主題 A에 關係있는 文獻을 入手코자 할 때 同一主題의 文獻이라도 그 主題의 取扱程度는 各各 다르다. 예를 들어 主題A에 관한 文獻이 10個 있을 때 이 파일에서 무조건 A에 관한 文獻을 探索하면 10個의 文獻이 모두 探索되며 再現率은 100%가 되겠으나 이것이 모든 質問者의 要求에 滿足하는지는 알 수 없다. 즉 精度率이 低下될 경우 探索時 特定性의 範圍를 좁혀서 該當主題를 重點的으로 取扱하고 있는 것만 探索한다면 精度率은 向上되나 再現率은 低下된다. 이와 같이 兩者는 相反關係에 있다.

2. 語彙調整(Vocabulary Control)과 再現率

어떤 主題 A에 關係되는 用語(Term) A, B, G, H, L, N, R, V가 있다면 다음의 세가지 方法을 생각할 수 있다.

첫째로 同意關係(Synonymous Relation)가 B=X, L=D, R=C 또는 Y일 때 表 1과 같은

표 1.

A			
B	USE	FOR	X
C	USE		R
D	USE		L
G			
H			
L	USE	FOR	D
N			
R	USE	FOR	C
	AND	FOR	Y
V			
X	USE		B
Y	USE		R

표 2.

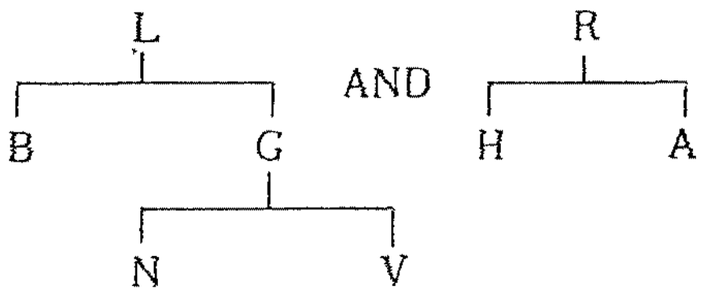

표 3.

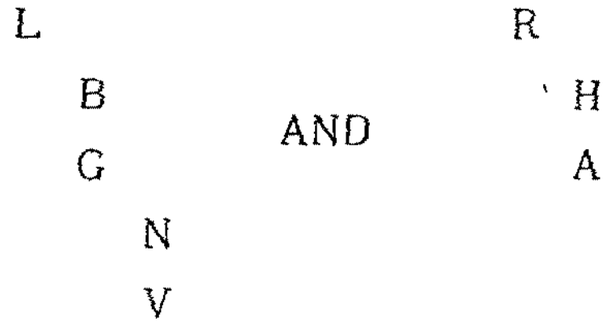

표 4

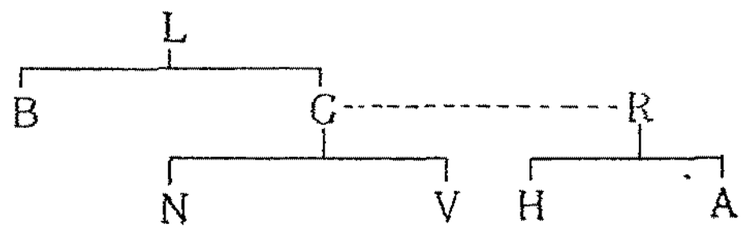

표 5.

A		H		R	
BT	R	BT	R	UF	C
RT	H	RT	A		Y
B		L		NT	A
UF	X	UF	D		G
BT	L	NT	B		H
RT	G		G		N
			N		V
			V		
C		N		V	
USE	R	BT	G	BT	G
			L		L
			R		R
		RT	V	RT	N
G		P		Y	
BT	L	USE	L	USE	R
	R				
NT	N				
	V				
RT	B				

리스트를 作成할 수 있다.

표 1은 探索時 選擇範圍를 확장시키기 위해 同意語를 採擇(어휘조정)함으로써 再現率(Recall)이 어떻게 增加되는가를 나타낸 것이다.

둘째로 이 여덟개의 用語는 어느 分類項目을 基準으로 하여 그에 대한 上位概念, 下位概念, 同位概念으로 關聯시키는 두개의 階層(Hierarchy) 또는 木(Tree) 形態로서 표 2와 같이 排列할 수도 있으며 또한 모두가 單一階層(monohierarchy)으로 되어 있으므로 표 3과 같이 직선형으로도 나타낼 수 있다.

특히 표 3에는 用語間에 聯關關係가 內包되어 있는데 G의 이웃(Neigbour), G의 主(Generic) L 또는 G의 種(Specific)인 N, V를 選擇함으로써 用語 G의 探索範圍를 넓힐 수 있다면 再現率을 向上시킬 수 있을 것이다.

물론 표 3에서도 用語間의 聯關關係가 쉽게 單純化되지는 않지만 만일 두 木사이에 어떤 하나의 關係가 存在한다면 表 4와 같이 復合階層(Polyhierarchy)도 構想할 수 있다.

표 4는 직선형태로 圖示할 수는 없지만 關係된 모든 用語들이 알파벳順序로 나타나는 디소오러스形式(Thesaurus Form)으로 排列할 수 있다.

또한 探索技法을 變更시킬 수 있는 可能性을 통해 再現率을 向上시키고 표 4와 같이 分類된 形態는 R-H-A, G-N-V 또는 L-B, G-N-V와 같은 그룹과 個個의 主題가 分類表에 全部 列擧되는 것이 아니다. 各 主題의 性質에 따라 作成된 標準位表(Unit Schedule)의 合成에 의하여 分類하는 파스트 分類(Faceted Classification : 일명 ; 分析合成型 分類法)의 形態로 여덟개의 用語를 再排列하여 블록 B-G-H-A, N-V를 探索時 使用함으로써 再現率을 增加시킬 수 있다.

세째로 표 4에서 어떤 特性要素에 따른 크럼프槪念(Clumping Concepts)에 의해 H-A, L-B, R-G-N-V로도 分類할 수 있으며, 다음과 같은 用語를 使用하여 표 5와 같이 디소오러스形態에 一致시킬 수 있다.

USE
UF　　USED FOR
BT　　BROADER TERM
NT　　NARROWER TERM
RT　　RELATED TERM

위에서 言及한 파스트分類法(Faceted Classification), 階層(Hierarchies)分類法, 分析合成型 分類表 외에도 用語의 切斷(Term Truncation)이나 래티스(Lattice) 등은 再現率을 向上시키기 위한 再現디바이스(Recall Devices)이다.

3. 語彙調整과 精度率

文獻의 內容은 여러가지 複雜한 槪念으로 構成되어 있는데 索引은 어떤 程度의 複雜함도 充分히 表現할 수 있어야 한다. 文獻의 內容을 單一槪念이나 複合主題로 索引하는 데는 前組合索引과 後組合索引 2種類가 있다.

前組合索引은 文獻의 內容을 單一標目으로 나타내는 것으로 原文을 찾고자 할 때도 그 單一標目으로만 探索하는 것이며, 後組合索引은 文獻의 內容을 여러 개의 單位槪念으로 索引하고 探索時에는 各 單位槪念을 서로 組合하여 必要文獻에 접하게 되는 것이다.

특히 어떤 形態의 索引에서 보다 後組合索引에서는 생략過程에서 누락되는 構文이 자주 發生되는데 組合(Coordination), 링크(Links), 로울(Roles) 등과 같은 精度디바이스(Precision Devices)는 이러한 누락을 防止하고 精度率을 向上시키기 위해 使用된다.

다음 표 6과 같이 用語를 索引했을 때 用語 R을 使用해서 探索했을 때 세개의 項目이 探索되고 L+R로 探索했다면 ②와 ③에서 探索된다.

표 6.

① B+H+R
② L+R
③ L+R+B+H

만일 文獻의 構文上 關係있는 L과 B, R과 H로써 連結된 別個의 組合으로 링크(Links)되어 있고 이것을 索引했다면 L+R을 使用한 探索은 ②에서 探索되고 L+B로써의 探索은 L+B에 관한 最適合 情報를 얻을 수 있다. 즉 링크의 使用으로 再現率은 低下되나 精度率은 向上시키게 된다.

또한 ①에서 用語들을 자세히 分析하여 H와 R 사이의 相互關係에 D가 미치는 影響을 細分시켜 特定性을 높이며 링크로서 그 影響과 關係를 索引할 수 있다면 精度率은 더욱 向上될 것이다.

4. 語彙調整에 관한 Cost-Effectiveness

索引作成者 및 探索者가 索引 또는 質問事項을 記述할 때 索引言語의 效率을 改善하는 데는 다음 두가지 方法이 있다.

① 現象能의 探索效率을 維持시키고 시스템 費用을 減少시키는 일환으로 索引言話를 変化시키는 方法

② 全體시스템 費用의 增加없이 探索效率을 높이기 위해 索引言語를 変化시키는 方法

一般的으로 情報檢索시스템에서 正確하고 細分된 索引言語를 維持하고 最新情報를 업데이트(update)하는 데는 費用이 많이 드는데 이것은 語彙의 量과 密接한 關係가 있다.

語彙數 즉 用語數가 많을수록 文獻의 數도 많아지며 語彙의 特定性이 높을수록 시스템의 精度率도 높아지고 이렇게 調整된 用語는 上位槪念을 나타내는 用語(Broad Term)로 構成된 것보다 探索過程에서나 그것을 索引化할 때 用語數에 많은 変化가 생기며 이러한 대부분의 変化는 比較的 用語階層의 下位水準에서 發生된다.

用語의 特定性은 시스템에서 要求하는 特定性과 直接的으로 關係가 있어야 하는데 이것은 시스템에서 必要한 特定性의 水準보다 높거나 낮을 때 費用面에서나 效率에서 不適合하기 때문이다.

用語의 特定性을 고려할 때 探索當 平均書誌

事項(citation)數에 대해 有意해야 하는데 만일 精度率이 20%이고 平均出力 書誌事項數가 12個라면 適當하고 할 수 있으나 精度率이 20%고 平均出力 書誌事項數가 120個라면 效率性이 좋지 않은 시스템인데 이 시스템의 效率을 높이기 위해 링크, 로울 등은 使用하여 잘못 選択된 同義關係의 結果로 인해 探索時 検索된 원치 않은 項目을 減少시켜 주며 精度率도 向上되나 索引費用과 人力處理用을 增加시킨다. 또한 索引者의 生産性도 링크나 로울을 使用했을 때 줄어든다.

Sinnett J.D는 로울은 探索時間과 시스템내에서 探索處理時間을 增加시키며 Montague B. A는 링크나 로울의 指定은 한개의 特許를 索引하는데 36分중 平均 15.6%를 消費하고 로울은 使用했을시 약 50%정도 파일의 量이 增加함을 찾아냈으며 숙련되지 않은 索引者가 索引할 경우 30% 이상 더 時間을 消費한다고 밝혔다.

機械化된 遡及探索(Restrospective Search) 시스템에서 로울이나 링크는 관련없는 書志事項 數를 減少시켜 주나 이것은 실지 探索者가 비슷한 探索結果를 얻을 수 있는 여러 方法보다 이들은 使用했을 때 費用이 적게 들 경우 비로소 經濟性이 있는 것이다. 바꾸어 말하면 費用과 效率分析은 情報要員에 의해 探索出力選択作業을 통해 探索된 不必要한 書誌事項을 除去시키고 發生되는 얼마간의 不正確한 用語關係를 받아들임으로써 索引時間이나 探索時間을 로울이나 링크를 使用했을 때 보다 節約하고 실지 探索者가 滿足을 느낀다면 이 方法이 經濟性이 있는 것으로 評價된다.

4.1 入·出力 費用関係

情報検索시스템의 또 다른 關點에서 費用과 效率 關係를 살펴볼 때 検索效率에 影響을 미치는 入力費用과 出力費用을 均衡있게 調整하여야 한다.

出力效率을 높이고 最上의 出力結果를 얻기 위해 入力處理를 어떤 形態로 하느냐에 따라 시스템의 經濟性과 性能이 좌우된다.

4.2 自由키 워어드와 固定키워어드

単語 또는 複合語로 表現되는 키워어드는 標目으로 定하지 않은 自由키워어드(Free Keyword)索引보다 標目으로 定한 즉 調整된 固定키워드(Fixed Keyword) 索引時 時間과 努力이 많이 들며 숙련된 索引者가 必要하지만 探索者가 自然語(Natural Language)나 自由키워어드로 探索했을 때 보다 負擔을 덜어준다.

즉 探索者가 매번 探索戦略을 準備할 때 自然語나 自由키워드를 調整하여야 하기 때문이다.

예를 들어 探索時 自然語 主題나 自由키워드로 表現할 수 있는 "Textile Industry" 또는 "Petrochemicals" 등 모든 可能한 方法을 생각하여야 한다.

그밖에 自由키워드의 使用은 平均探索精度率을 減少시키고 探索出力 選択作業때도 努力과 費用을 더 들인다.

로울이나 링크 또는 關係表示(Relational Indicators)등을 使用하여 特定性이 높은 語彙를 索引할 때 索引者나 情報探索者의 水準이 높아야 되며 效率的인 索引(Indexing Consistency)을 하는데도 어려움이 있지만 特定性이 높은 用語일수록 探索精度率을 높이며 探索出力의 選択作業 時間도 줄일 수 있다.

4.3 情報시스템의 Trade-Off比較

다음 표 7은 두개의 情報시스템에서 Trade-Off를 나타내고 있다.

시스템A에서는 出力結果와 費用에 經濟性을 두고 入力處理를 正確하고 詳細히 分析處理한 반면에 시스템B에서는 出力費用의 增加와 좋지 않은 出力結果가 나타나지만 入力費用을 적게 하는데 經濟性을 두고 있다.

시스템A와 B를 比較할 때 시스템A가 B보다 效用性있고 經濟的이지만은 않다. 그것은 시스템B에 接近해 나가는 方式이 실지 探索者가 滿足할 만한 水準의 結果를 얻을 수 있고 시스템A에 關係되는 全體費用보다 적게 든다면 시스템B가 效用性이 있다.

<시스템A>
<入力>
① 固定키워드使用
② 文献當 平均 10個의 用語를 索引
③ 숙련된 索引者
④ 索引을 校訂處理
⑤ 一日當 40個의 項目을 索引
⑥ 入力費用이 많다.
<出力>
① 探索者의 負擔을 減少
② 높은 精度率
③ 適當한 再現率
④ 探索出力의 選択作業이 不必要
⑤ 빠른 應答時間
⑥ 探索費用 減少

<시스템B>
<入力>
① 自由키워드 使用
② 文献當 平均 10個의 用語를 索引
③ 미숙련 索引者
④ 索引校訂 處理가 없다.
⑤ 一日當 100個의 項目을 索引
⑥ 入力費用이 적다.
<出力>
① 探索者의 負擔을 增加
② 낮은 精度率
③ 適當한 再現率
④ 探索出力의 選択作業이 必要
⑤ 늦은 應答時間
⑥ 探索費用 增加

5. 結 論

費用과 效率에 관한 分析은 情報檢索시스템에 索引作業, 索引言語, 探索方法 등의 副시스템(Subsystem)에 適用시킬 수 있지만 現實的이지못하다. 왜냐하면 이런 副시스템은 서로 密接하게 內部的으로 關係되어 있고 어떤 副시스템의 変化는 시스템 全體를 통해 間接的으로 影響을 미치기 때문이다.

情報檢索시스템의 評價에 있어 再現率과 精度率이 높아도 經濟性과 迅速性이 欠如되면 效用性이 없게 된다. 그러나 檢索效率, 經濟性, 迅速性은 一般的으로 両立할 수 없는 것이다. 그러므로 각 시스템에 따라 最適基準을 정하여야 한다.

또한 情報시스템은 하나의 複合的인 要素를 갖춘 組織體이므로 단지 국한된 效果만으로서 어떤 変化를 期待해선 안된다. 成功的인 檢索시스템을 完成하는 데는 많은 方法이 있지만 現在 情報檢索시스템의 흐름은 単純化, 単一化하는 傾向이 있다.

参 考 文 献

1) Lancaster F. W. Vocabulary Control for Information Retrieval. Information Resources Press Washington, D. C. 1972. pp. 121~134, 218~222.
2) A. Gilchrist The Thesaurus in Retrieval. London SW1X8PL. Aslib, 1971. pp. 4~17.
3) 司空 哲 : 情報檢索論. 亞細亞文化社, 1977. pp. 123~167.
4) 中井 浩 / 笹森勝之助 情報檢索システム. 東京, 日本經營出版會, 1971. pp. 159~163.
5) Sinnett, J. D. An Evaluation of Links and Roles Used in Information Retrieval. Dayton, Ohio : Airforce Materials Laboratory, Wright-Patterson Airforce Base, 1964. AD432198.
6) Hontague, B. A. Testing Comparison and Evaluation of Recall, Relevance and Cost of Coordinate Indexing with Links and Roles. Proceedings of the American Documentation Institute, 1964. pp. 357~367.
7) 坂本 徹 : 情報檢索. 東京, 雄山閣出版, 1976. pp. 173.