韓國 OA學會 論文誌
第3卷 第1號, 1998. 3.

論文 98-3-1-1-4

정보 검색 시스템의 적합성 피드백에 관한 연구

명 순 희*

Automatic Term Relevance Feedback in IRS

Soon-Hee, Myong*

요 약

비정형 데이터를 다루는 정보검색 시스템에서 검색의 정확도는 사용자의 인지에 의존하며 따라서 사용자의 검색 평가는 시스템의 효율을 측정하는 척도의 하나이다. 적합성 피드백은 사용자의 검색 평가를 시스템에 입력하여 질의의 수정, 재 검색을 반복함으로써 재현율과 정확도를 높이고자 하는 질의 확장 방법의 일종이다. 본고에서는 적합성 피드백의 이론적 배경과 구현 절차를 기술하였다.

Abstract

In the Information Retrieval System, the relevance of retrieved items is determined by the judgement of the user and thus the evaluation of the system efficiency counts on the cognizance of users to some extent. The relevance feedback mechanism provides a device allowing iterative searches during which the query can be modified and refined based on user input from the relevant documents. The feedback systems are generally reported to outperform non-feedback systems. The procedures and algorithms to implement the feedback mechanism are surveyed in this paper.

---

* 용인공업전문대학 사무자동화과 전임강사

논문접수 : 98.1.3. 심사완료 : 98.2.23.

I. 서 론

정형화된 데이터를 저장. 검색하는 데이터베이스 시스템(DBMS)과 달리 비정형 데이터를 다루는 정보검색시스템(IRS)의 사용자는 저장된 데이터의 구조나 내용에 관한 사전지식이 없고, 따라서 검색 결과를 예측하기 어렵다. DBMS의 검색은 매칭 여부나 검색결과가 결정적인데 비하여 정보검색시스템의 검색에서 정보요구의 충족은 다분히 확률적이다. 대부분의 경우 정보검색 시스템에서 한 번의 조작으로 질의에 적합한 문헌을 얻을 수가 없으며, 새로운 고성능 시스템이라도 정보검색 시스템의 특성상 재현율은 제한적일 수 밖에 없다. 더구나 일반 사용자에게 있어서 정보요구를 대변하는 질의 구성은 쉬운 작업이 아니다. 연구에 의하면 성공적이라고 사용자가 생각하는 검색에서도 정확률은 25%에 불과한 것으로 보고되고 있다. 또한 미 의회 도서관 THOMAS 시스템의 조사에서 사용자 전체의 88%이상이 3개 이하의 질의 용어를 사용한다는 연구 결과는 사용자 입장에서 질의 구성의 어려움을 나타낸다[7].

사용자 입장에서는 매우 단순한 질의 형식이 유리하고, 시스템에서는 질의가 정교할수록 검색 효율이 높은데 이 두 가지의 상반되는 목표를 수용하는 인터페이스가 필요한 것이다. 앞서 말한 탐색 상의 시행착오 또는 재 탐색은 질의어와 색인어가 일치하지 않는 데서 비롯된다. 따라서 많은 정보검색 시스템에서 질의어가 색인어와 정확히 일치되지 않아도 관련 문헌을 검색할 수 있도록 질의 확장 기능이 제공되는데, 시소러스, 문헌 클러스터링, 적합성 피드백 등이 이에 이용된다.

시소러스는 색인어와 관련된 용어를 구조화하여 탐색 중 자동 확장하거나, 혹은 사용자의 조회 요구에 따라 제시되는 통제어휘집이며, 클러스터링은 문헌들간의 거리, 즉 유사도를 측정하여 유사항목의 집합으로 분류함으로써 검색대상의 확대를 용이하게 한다. 이들 이론적 근거를 가진 방법에 비하여 적합성 피드백은 질의를 여러 번 수정하여 최적의 질의에 접근 해 가는, 실질적인 운용과정에서 개발된 장치이다. 수행 방식은 최초의 검색 결과에서 사용자가 수 작업으로 용어를 선정, 수정하는 수동식피드백 시스템과, 사용자가 시스템과의 대화형 조작을 통하여 질의를 개선해나가는 자동 방식, 그리고 최초의 질의 및 개념의 입력을 기반으로 검색을 확대하는 인공지능형 피드백 시스템이 있다.

수 작업을 통한 질의 수정은 사용자의 지적 노력을 요하고, 무엇보다도 효율적인 탐색을 위해 이용될 수 있는 용어확장이나 용어변경 등의 대안을 일관성 있게 적용하기를 기대할 수는 없다. 1960년대 중반부터 시

작된 대부분의 적합성 피드백 연구는 자동 피드백 시스템에 관한 것이고 일부 실용화 되었다. 자동 피드백 시스템은 사용자가 문헌의 적합성에 대한 판단을 내리고 시스템과의 상호작용을 통하여 원하는 자료의 범위를 좁히는 작업을 반복함으로써 최적의 탐색에 근접하게 하는 장치로서 SMART 시스템에 구현된 것이 대표적인 예이다[9].

이러한 시스템이 사용자의 판단을 요구하는데 비하여 완전자동 피드백은 이른바 '종자(seed)' 질의만을 요하는 것으로 첨단 정보검색 엔진인 INQUERY을 이용한 판례검색 시스템에 구현되어 있다[7].

텍스트 데이터베이스를 대상으로 하는 정보검색 기술은 도서관을 비롯하여 매뉴얼이나 EDI문서, 비즈니스 정보 등을 관리하는 기업 환경, 기술 문헌과 학술 정보의 공유, 소프트웨어 데이터베이스 관리 등을 위해 이용된다. 이들의 공통점은 대용량의 비정형 정보라는 점이다. 여기서 검색되는 자료의 '적합성'은 정보검색 분야의 핵심 개념으로 많은 연구의 주제가 되었다. 특히 용어가 문헌의 콘텐트를 대표하는 정도를 정량화하여 질의와의 유사도를 계산하기 위한 검색 모델 중에는 벡터 공간 모델과 확률 모델이 활발히 연구, 실험되었다. 자동 피드백의 방법 또한 시스템의 검색 모델에 따라 다른데 본고에서는 앞의 두 가지 전통적 검색모델에서 자동 피드백 시스템을 구현하기 위한 절차와 알고리즘을 기술한다.

II. 본 론

적합성 피드백은 사용자가 원시 질의의 바탕 위에 새로운 질의를 구성하는 방법이다. 탐색 결과에 대한 만족도가 낮을 때 사용자는 결과물 중 가장 적합하다고 판단되는 문헌을 시스템에 지정함으로서 피드백 기능은 작동된다. 시스템은 자동으로 질의를 수정하여 보다 정교한 재 탐색을 수행하게 된다. 질의 수정작업은 매우 단순한 가정에서 출발하는 것으로 적합한 문헌에 출현빈도가 높은 용어는 질의어와 관계가 밀접한 용어이므로 다른 용어보다 우선적으로 질의 수정에 이용한다는 것이다. 대부분의 적합성 피드백 기술은 이 밀접한 용어를 참고하여 용어 가중치를 변화시키는 방법과, 용어 자체를 변경하는 방법, 이 두 가지를 병행하거나 그 중 한 가지를 이용한다. 따라서 적합성 피드백 연구는 두 가지 분야를 다루어 왔는데 하나는 원시 질의를 바탕으로 한 질의어의 가중치 재 산정 문제이고, 다른 하나는 용어의 첨가, 삭제 등 질의 변경에 필요한 용어 선정의 문제이다.

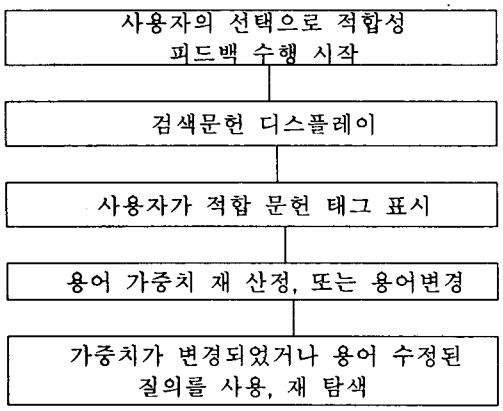
가중치 재산정시 용어 가중치를 이용하므로 적합성 피드백은 문헌의 적합성에 따른 순위 부여와 밀접한 연관이 있다. 여기서의 용어는 통제어 이거나 또는 불용어 처리, 스테밍, 등의 자연어 처리 과정을 거쳐 문헌의

콘텐트를 대표하는 색인어로 확정된 것을 말한다. 색인어의 가중치는 문헌내의 용어 빈도(tf: term frequency)와 전체 문헌 집합에 나타나는 빈도(df: document frequency)를 이용한 표준적 역빈도(inverse document frequency; df·idf) 방식으로 산출하여 역색인 파일에 색인어와 함께 저장된 정보를 이용한다.

이러한 방식의 근거는 소수 문헌에 집중적으로 나타나는 용어는 문헌의 대표성이 높다는 것으로, 색인어 가중치는 문헌 주제와의 밀접성과 문헌길이에 따라 값이 무한히 커지지만 문헌의 길이로 정규화 하여 0과 1사이의 값으로 산출된다.

적합성 피드백의 절차를 개괄하면 다음과 같다.

표 1 적합성 피드백
Table. 1 Relevance feedback



2.1 가중치 재산정

대표적인 순위부여 시스템인 벡터공간모델과 확률모델에 각기 적합성 피드백을 위한 알고리즘이 있다. 이 두 가지의 기본 알고리즘에는 여러 가지 변형이 있다. 예를 들어 INQUERY 검색 엔진의 적합성 피드백에는 Roccio 가중치로 불리는 변형된 벡터모델이 이용되고 있다[2]. 그러나 각 알고리즘의 우열은 가리기 어려운 것으로 연구되었다[4].

불리언 검색 시스템에서의 피드백 수행이 불가능한 것은 아니나 본고에서는 기본적으로 가중치에 근거한 순위 부여 시스템에서의 접합성 피드백만을 다루었다.

수행 방법은 사용자가 일차 검색 후 상위 문헌을 점검하여 적합한 문헌에 태깅을 하면 시스템은 이에 의거 용어의 가중치를 재계산하여 질의를 수정하게 된다. 여기서 가장 기초적인 근거는 적합한 문헌에 많이 등장하는 용어는 정보 요구에 적합한 '주요' 색인어이므로 가중치를 상향시키고 반대로 부적합 문헌에 나타나는 용어에는 가중치 하향 요인을 부여한다는 단순한 원리이다. 질의 용어의 상대적인 중요도를 결정하는 프로세스로서 이용자의 정보 요구가 충족될 때까지 반복된다.

2.1.1 벡터공간모델에서의 피드백

SMART 시스템에서 볼 수 있는 것과 같은 초기의 피드백 장치는 초기의 Roccio와 Ide의 피드백 모델을 개선하여 벡터 공간 모델의 검색 시스템을 위해 고안되었다[9].

사용자의 정보 요구는 질의어로 대표되고, 시스템에 저장된 문헌의 내용은 색인어로 대표되며 검색은 이들 질의 벡터와 문헌 벡터를 매치시켜 질의어를 공유하는 문헌을 찾아내는 과정이다. 벡터 모델의 순위 부여 기능은 이에 더하여 질의와 문헌간의 유사도 계수를 계산, 값이 높은 문헌부터 역순으로 보여준다.

질의와 문헌 벡터는 다음과 같이 표시된다.

$Q = (q_1, q_2, \ldots. q_t)$ (1)

$D = (d_1, d_2, \ldots. d_t)$ (2)

q_i : 질의 Q에 포함된 용어 i의 가중치

d_i : 문헌 D에 포함된 용어 i의 가중치

질의와 문헌간의 유사도는 기본적으로 상호 용어리스트를 비교하여 일치하는 용어 가중치의 곱을 합산하여 얻는다.

$$SIMILAR(D,Q) = \sum_{i=0}^{t} d_i \cdot q_i \quad (3)$$

피드백의 결과 변형된 질의는 다음의 벡터로 표현할 수 있으며 문헌 집합이라는 공간 내에서의 위치이동은 Q→Q′로 이동함으로써 적합 문헌에 더욱 근접하는 이미지로 표현할 수 있다.

$Q' = (q'_1, q'_2, \ldots. q'_t)$ (4)

적합 문헌과 부적합 문헌을 구분하기 위한 최적의 질의식은 적합 문헌 집합과 부적합 문헌 집합과의 차이의 합산으로 얻어진 벡터이다.

$\left(\frac{1}{R}\right)\left(\sum_{i\in D_R} TERM_{ik}\right)$: 적합한 문헌 집합에 들어있는 용어 k 가중치의 평균

$\left(\frac{1}{N-R}\right)\left(\sum_{i\in D_R} TERM_{ik}\right)$: 부적합한 문헌 집합에 들어있는 용어 k 가중치의 평균

최적의 질의에서 용어 k의 값은

$$(Q_{opt})_k = C\left(\frac{1}{R}\sum_{i\in D_R} TERM_{ik} - \frac{1}{N-R}\sum_{i\in D_{N-R}} TERM_{ik}\right) \quad (5)$$

C : 상수

$TERM_{ik}$: 문헌 i 에 포함되어 있는 용어 k 의 가중치

D_R : 적합한 문헌

D_{N-R} : 부적합한 문헌

그러나 실제로는 D_R (적합한 문헌), D_{N-R}(부적합한 문헌)을 처음 질의부터 알 수가 없으므로 질의 수정을 반복하면서 최적의 질의에 접근하는 것이다. 따라서 최초의 질의 벡터 Q로부터 향상된 질의벡터 Q′를 전개하여, $D_{N'}$과 $D_{N R'}$ 집합을 얻을 수 있다. 수정된 질의식은 아래와 같다.

$$Q' = C\left(\frac{1}{R}\sum_{i\in D_R} DOC_i - \frac{1}{N-R}\sum_{i\in D_{N-R}} DOC_i\right) \quad (6)$$

DOC_i : 문헌 i 의 용어 가중치 벡터

그러나 실상은 최초의 질의에 중요한 용어가 있을 수 있으므로 이것을 포함하여 개선된 질의를 만들 수 있다.

$$Q' = \alpha Q + \beta\left(\frac{1}{R}\sum_{i\in D_R} DOC_i\right) - \gamma\left(\frac{1}{N-R}\sum_{i\in D_{N-R}} DOC_i\right) \quad (7)$$

여기서 α, β, γ는 적절한 상수이다. 이 식에서 새로이 수정된 질의벡터는 원시 질의의 벡터합에 더한 적합 문헌의 평균과 부적합 문헌 가중치의 평균 차가 된다.

표 2 질의. 문헌 벡터 매트릭스
Table. 2 Query and document vector matrix

	T_1	T_2	T_3	T_4	T_5
Q =(	6	0	4	1	0)
D_R =(	4	2	4	0	1)
D_{N-R}=(	2	0	0	4	0)

예를 들어 표2의 데이터를 이용하여 식 (7)에 따라 가중치를 재 산정하면 다음의 수정된 가중치를 얻는다. 상수는 각각 1, ½, ¼로 하는데 양성피드백을 수행하려면 γ를 0으로 한다.

$$Q' = Q + 1/2\left(\sum_{D_R} D_i\right) - 1/4\left(\sum_{D_{N-R}} D_i\right)$$
$$= (6,0,4,1,0) + ½(4,2,4,0,1) - ¼(2,0,0,1,0)$$
$$= (7½,1,6,0,½)$$

결과는 적합 문헌에 포함된 용어 T1, T3의 가중치가 증가하고 T2, T5는 원시질의에 없던 것이 추가되었으며, T4는 배제되었다. 최초 질의와 문헌간의 유사도, 그리고 수정된 질의와 문헌간의 유사도를 각각 구하여 비교하면 결과적으로 수정된 질의와 문헌간의 유사도가 높아졌음을 알 수 있다.

$$SIMILAR(Q,D_R) = \sum_{j=1}^{t}(Q_j \cdot D_{ij}) = (6 \cdot 4) + (0 \cdot 2) + (4 \cdot 4) + (1 \cdot 0) + (0 \cdot 1) = 40$$

$$SIMILAR(Q',D_R) = \sum_{j=1}^{t}(Q'_j \cdot D_{ij}) = (7½ \cdot 4) + (1 \cdot 2) + (6 \cdot 4) + (0 \cdot 0) + (½ \cdot 1) = 56½$$

앞서 본 벡터 수정모델은 수정된 용어 가중치를 적합 또는 부적합 문헌내의 용어로부터 계산하는, 매우 간단한 개념에 기초하고 있다. 따라서 최초의 질의나 저장된 문

헌의 컨텐트를 정확히 반영할 수 있는 가중치가 유효하다면 벡터 수정 모델도 강력한 질의 구성 방법이다[8].

SMART시스템의 실험운용 결과 재현율이 높은 검색의 경우 50%의 정확도 향상이, 재현율이 낮은 탐색의 경우 약 20%정도의 정확율 향상을 보여주었다[9].

실험 결과에 의하면 1, 2회의 피드백은 적합도 향상에 매우 효과적이었으나 2회 이후에는 좀처럼 검색효과가 개선되지 않았다.

2.2 확률적 피드백 방법

확률 검색모델에 구현되는 적합성 피드백 또한 유용한 대안이다. 확률 모델에서의 질의 용어는 다음 계산식에 의해 계산된 가중치를 가지며, 검색된 문헌은 문헌. 질의간의 유사도에 따라 검색순위가 부여된다[3].

$$w_i=\log \frac{p_i(1-u_i)}{u_i(1-p_i)} \qquad (8)$$

p_i = P(x_i =1 | relevant) : 용어 i 를 포함하고 적합할 확률 (= r_i/R)

u_i = p(x_i =1 | nonrelevant) : 용어 i 를 포함하고 비적합할 확률(= $(n_i-r_i)/(N-R)$)

질의 q에 대한 적합 문헌과 부적합문헌에서 용어 i의 분포는 표 3과 같이 나타낼 수 있다.

표 3 용어 분포표
Table. 3 Occurences of Term i in a Collection

		적합성 + (relevant)	적합성 - (non-relevant)	
색인 유무	+ (d_i=1)	r_i	n_i-r_i	n_i
	- (d_i=0)	$R-r_i$	$N-n_i-R+r_i$	$N-n_i$
		R	N-R	N

따라서 식(3)을 원용하면 질의와 문헌간의 유사도는 다음과 같이 계산된다.

$$SIMILAR(D,Q)=\sum_{i=1}^{t} d_i \cdot \log \frac{p_i(1-u_i)}{u_i(1-p_i)}$$

그러나 실제로는 처음부터 pi 와 ui 의 값을 알 수가 없다. 적합성 정보가 없는 초기질의에서는 pi 의 값이 모든 용어에 대하여 0.5 정도로 일정하고 부적합 문헌에서의 용어분포는 전체 분포와 유사 하다고 본다. 따라서 최초의 검색에서 ui의 값은 ni/N으로 둘 수 있으므로 질의, 문헌 유사도는 다음과 같다.

$$InitialSIMILAR(D,Q)=\sum_{i=i}^{t} d_i \cdot \log \frac{N-n_i}{n_i} \qquad (10)$$

N = 총 문헌 수

n_i = 용어 i를 포함한 전체 문헌

검색이 계속되고 적합성에 관한 수치가 누적되면 p_i 와 u_i 는 각각

$$p_i = \frac{r_i}{R}, \quad u_i = \frac{n_i - r_i}{N - R} \qquad (11)$$

r_i = 용어 i 가 있는, 검색된 적합문헌 수
R = 질의 q에 대한 적합문헌 전체

피드백 검색을 위해서는 이전 검색 결과에서 적합한 문헌과 부적합한 문헌과 관련된 통계를 축적하여 식(9)를 평가한다. 이 경우 이전에 검색된 연관 문헌에서의 용어 분포는 적합 문헌 전체에 걸쳐서 같다고 간주한다. 표1에 나와있는 용어 분포를 검색 결과에 적용하면 얻는다. 이를 식(9)에 대입하면 식(12)를 얻는다.

여기서 R은 검색된 문헌 중에서 적합문헌의 수이고, ri는 용어 i 를 포함한 적합문헌, ni는 용어 i 를 포함한 문헌 수이다.

$$feedbackSIMILAR(D,Q) = \sum_{i=1}^{t} d_i \cdot \log\left(\frac{r_i}{R - r_i} \div \frac{n_i - r_i}{N - R - n_i + r_i}\right) \qquad (12)$$

이후의 피드백 탐색에서는 이전의 적합성, 비적합성 통계가 이용된다.

위의 계산에서 R=O, 즉 적합 문헌이 초기에 검색되지 않았을 때 적합 문헌에 용어 i 가 포함될 확률 p_i는 단순히 전체 문헌에서 나타날 확률 n_i/N과 같다. 따라서 조정상수를 이용하여 식(11)을 다음과 같이 변경할 수 있다[8].

$$p'_i = P(x_i = 1 \mid \text{relevant}) = \frac{r_i + n_i/N}{R+1}$$

$$u'_i = p(x_i = 1 \mid \text{nonrelevant}) = \frac{n_i - r_i + n_i/N}{N - R + 1} \qquad (13)$$

벡터 모델과 비교할 때 확률적 피드백 모델의 장점은 피드백 프로세스가 질의용어의 가중치 산정과 직접 연관되어 있다는 점이다. 즉, 질의용어 i가 문헌과 매치될 때마다 $\log[p_i(1-u_i)/u_i(1-p_i)]$로 계산되는 역문헌빈도가 유사도에 합산되어 값이 커지며 이 용어 가중치는 색인이 0,1로 표시되는 이진 색인일 때 극대화된다. 이 경우 단점으로는 문헌에 들어있는 용어들의 가중치라든가 최초의 질의에 들어있는 용어의 가중치 등은 무시된다는 점이다. 더구나 확률모델에서는 검색된 적합 문헌이 질의변경에 직접 반영되지 않고 용어분포에 관한 정보가 확률적 용어 가중치를 결정하는데 간접적으로 사용되는데 이러한 점들은 시스템의 효율에 영

향을 미치게 된다. 사용자의 질의어는 단지 이와 관련이 있을 확률이 임계값 이상으로 큰 모든 문헌을 추출하기 위한 출발점이 되는 것이다. 확률적 피드백 모델은 많이 실용화 되어왔으나 초기질의에 가중치를 쓸 수 없다는 점이 결점으로 지적되어왔다.

2.2 용어변경 피드백

원시 질의 용어에 한정하여 가중치를 재산정 하는 방법은 문헌의 순위에만 영향을 주므로 질의 수정에 있어 매우 제한적이며, 용어의 삽입, 삭제를 이용한 질의 재구성을 통하여 보다 융통성 있게 개선할 수 있다. 파일 탐색 이전에 제공되는 질의어의 경우 시소러스에 의존하거나, 사용자가 대표적인 적합 문헌(source document)을 이미 가지고 있는 경우 질의는 단순히 이 문헌에 포함된 용어를 이용한다. 용어변경 피드백 장치는 공통적으로 원시질의에 의한 최초 검색 이후 사용자에 맞는 새로운 개념을 대표하는 색인어를 찾아서 보다 정교한 질의어로 재구성하는 절차를 제시하고 있다[9].

원시 질의어의 광의어나 협의어, 또는 같은 레벨에 있는 관련어, 유사어 등으로 대체하는 방법이다. 이 단계에서도 시소러스가 이용되기도 하는데 질의어의 동의어나 유사어 등을 포함하여 질의 수정하므로 기본적인 질의 구조는 같다고 보아야한다[1]. 다른 연구에 의하면 적합성 피드백 단계에서 이미 검색된 문헌에서 특정 필요에 의해 선택한 용어가 시소러스나 사용자가 알고 있는 일반 용어보다 검색효율이 좋았다[11].

다른 방법은 적합한 문헌을 분석하여 중요한 용어와 가장 밀접히 연관되어 있는 용어를 더하거나 다른 용어로 대체하는 것이다. 즉 최초의 질의에 없던 새로운 용어의 편입은 적합 문헌에 포함된 색인어의 가중치를 검사하여 가중치가 가장 큰 용어를 결정하는 것이다. 또는 공동발생 빈도가 높은 단어를 대상으로 할 수 있다. '전화'라는 용어를 확장, 일반화하는 경우 '통신', '전보', 등의 용어가 공동발생 빈도가 높은 관련어로 나타날 것이다. 용어 변경 결과 원시질의에서 0의 값을 갖는 용어가 수정된 질의에서는 0과 1사이의 값을 갖는 것이며, 부적합 용어의 경우 그 반대가 된다. 물론 빈번한 용어의 삽입, 삭제는 탐색에 혼란을 가져올 수 있으므로 삽입, 삭제에 관한 결정은 신중하게 이루어져야한다.

III. 결 론

적합성 피드백은 검색의 개선을 위해 널리 이용되는 방법이며 실제 상당한 검색효과의 개선이 있는 것으로 연구되었다. 그러나 벡터모델이나 확률모델과 같이 텍스트

처리에 있어서 문헌의 용어분산과 빈도에 근거한 통계적 해석은 구현의 용이함과 이에 따른 경제적 이점이 있으나 그 성능에 대해서는 비판이 적지 않다. 방대한 문헌 집합 가운데 요구에 적합한 자료가 들어 있다 하더라도 앞서 말한 정보검색 모델에는 추론을 통하여 도달할 수 있는 체제가 없다는 것이다[1].

이러한 텍스트 처리에 바탕을 둔 적합성 피드백 방법 또한 그 효과가 제한적일 수밖에 없다. 기존의 텍스트 데이터베이스에 인공 지능적인 지식 기반 시스템이나, 의미론적 해석 기술이 발전함에 따라 텍스트의 주제 분석 및 적합성 피드백 방법 또한 보완될 것이다.

적합성 피드백을 실제 적용하는데는 두 가지 문제가 발생한다. 하나는 검색성능 향상을 위해 얼마만큼의 적합성 정보가 필요하며, 둘째, 질의확장은 어디까지 해야하는가. 측정된 용어가중치가 신뢰성이 높은 것은 아니지만 소량의 적합한 데이터는 질의 전개에 도움이 되는 것으로 나타나 있다[10]. 질의 확장에 대해서는 요구사항이 충분히 질의에 반영되었는가, 문헌의 색인은 잘 되어있는가, 그리고 이미 확보된 적합성 정보는 충분한가 등에 달려있다. 실제 운용중인 대화형 검색 시스템에서 실제 검색을 통한 연구결과 적합성 피드백은 시소러스나 정보 검색사의 도움과 같은 다른 용어확장 방법에 비하여 많이 사용되지는 않았으나 일단 이용된 경우에는 2/3 이상의 경우에 양성 검색이 되는 등, 효과적인 것으로 나타났다[12].

이상의 효율적인 가중치 재산정과 용어변경 방법 이외에도 성공적인 피드백 시스템은 몇 가지 일반적인 요건을 갖추어야한다[6].

첫째, 사용자가 질의수정을 무한정 반복하지 않도록 3, 4회 이내의 수정을 통해 최대의 검색효과를 볼 수 있어야 한다.

둘째, 한 두개의 문헌을 찾고자 할 때 사용자는 탐색 중간에라도 탐색을 끝낼 수 있어야한다. 셋째, 질의변경 탐색결과는 항상 새로운 문헌을 보여줘야 한다. 즉, 앞서 찾아낸 문헌은 질의 변경후 검색에서 나타나지 않아야 한다. 그러면서도 필요하면 이전에 탐색된 적합한 문헌을 한꺼번에 열람할 수 있어야한다.

자동 피드백 시스템의 장점은 사용자가 질의 구성에 관한 지식이나 문헌집합의 구조나 저장 내용에 관한 지식이 없어도 시스템의 통제된 절차를 통하여 정교한 질의를 전개하는 효과를 얻는 것이다. 적합성 판단의 주관성과 피드백의 효율성과 관련하여 문제가 있음에도 적합성 피드백의 사용은 사용자가 질의를 직접 구성해야하는 검색시

스템의 문제를 해결하는데 도움이 된다. 또한 적합성 피드백 기술을 이용하여 전통적인 데이터베이스의 검색을 강화할 수 있는 가능성을 제시하였다는 점에서 정보검색 시스템의 개선에 기여를 한 것으로 평가된다 [5].

참 고 문 헌

[1] Bordogna, Gloria and Pasi, Gabriella (1996) "A user-adaptive neural network supporting a rule-based relevance feed back,"Fuzzy Sets and Systems, pp. 82, 201-211.

[2] Croft, W. Bruce (1995) "Effective Text Retrieval Based on Combining Evidence from the Corpus and Users", IEEE Expert, Dec. 1995, pp. 59-63.

[3] Croft, W. B. and Harper,D. J. (1979) "Using probabilistic models of document retrieval without relevance information."Journal of Documentation, Vol. 35, pp. 285-295.

[4] Efthimiadis, Efthimis N. (1995) "User Choices: a New Yardstick for the Evaluation of Ranking Algorithms for Inte -ractive Query Expansion", Information Processing&Management, Vol. 31, No. 4, pp. 605-620.

[5] Ellis, David, Progress and Problems in Information Retrieval. London, Library Association, 1996. pp. 30-36.

[6] Korphage, Robert R. (1997) Information Storage and Retrieval. New York, John Wiley, 1997. pp. 221-224.

[7] Rissland, Edwina A., and Damiels, Jody, J. (1996) "The Synergetic Application of CBR to IR", Artificial Intelligence Review, Vol. 10, pp. 441-475.

[8] Salton, Gerard and Buckley, Chris, 1990, "Improving Retrieval Performance by Relevance Feedback", Journal of American Society for Information Science, Vol. 42, pp. 288-297.

[9] Salton, Gerard and McGill, Michael J. Introduction to Modern Information Retrieval. New York, McGraw-Hill, 1983.

[10] Spark Jones, K. (1979) "Search Term Relevance Weighting Given Little Relevance Information", Journal of Documentation, Vol. 35, pp. 30-48.

[11] Spink, Amanda and Losee, Robert M. (1996) "Feedback in Information Retrieval", Annual Review fo Information Science and Technology, Vol. 31, pp. 33-78.

[12] Spink, Amanda and Saracevic, Tefko (1997) "Interaction in Information Retrieval: Selection and Effectiveness of S-

earch Terms", Journal of the American Society for Information Science, Vol.48, No.8, pp. 741-761.

□ 筆者紹介

명순희

韓國 OA學會 論文誌 VOL. 2. NO. 3, DEC. 1997