1997년도

제9회 한글 및 한국어 정보처리 학술대회

인간과 기계와 언어

한글 및 한국어정보처리

■ 일 시 : 1997년 10월 10일(금) ~ 11일(토)

■ 장 소 : 부 산 대 학 교

공동주최 : 한 국 정 보 과 학 회
한 국 인 지 과 학 회

주 관 : 부산대학교 컴퓨터 및 정보통신연구소

후 원 : 전자통신연구원 시스템공학연구소
연구개발정보센터(KORDIC)

"미리내" 정보검색 시스템에서 Relevance Feedback 구현[1]

*박수현, **박세진, ***권혁철
*동서대학교 컴퓨터공학과, **부산대학교 인지과학협동과정, ***부산대학교 전자계산학과

Implement of Relevance Feedback in "MIRINE" Information Retrieval System

*Su-Hyun Park, **Se-Jin Park, ***Hyuk-Chul Kwon
*Department of Computer Engineering, Dongseo University.
**Interdisciplinary Research Program of Cognitive Science, Pusan National University
***Department of Computer Science, Pusan National University

subak@kowon.dongseo.ac.kr
sejin @solge.cs.pusan.ac.kr
hckwon@hyowon.pusan.ac.kr

요 약

이 논문은 부산대학교 전자계산학과 인공지능 연구실에서 개발한 정보검색 시스템 "미리내"의 적합성 피드백 방법을 분석하고, 그 방법들의 검색 효율을 비교 분석하였다. "미리내"에서 질의문은 자연언어 질의문을 사용하고 재검색을 위한 적합성 피드백은 원질의문에서 검색된 문서 중 이용자가 직접 선택한 적합 문서에서 추출한다. 적합성 피드백은 크게 단어 확장(Term Expansion)을 위한 단어 선택 방법과 추가될 단어에 가중치를 부여하는 단어 가중치 부여(Term Weighting)의 2 가지 요소로 이루어 진다. 단어 선택을 위해서는 적합 문서에 나타난 단어 빈도합(tf), 역문헌빈도(idf), 적합 문서 중에서 해당 단어가 있는 적합 문서의 비율(r/R) 등의 정보를 이용한다. 단어 가중치 부여 방법으로는 정규화 또는 코사인 함수를 이용하여 부여하였다. 단어확장에는 tfidf 가 tfidf(r/R)보다 정확도 면에서 나은 향상율을 보였으나, 30 위 내 검색된 적합문서의 수를 비교해 보았을 때 tfidf(r/R)의 정확도가 높았다. 단어 선택 방법에서 계산된 값을 정규화 하여 가중치를 부여하였을 때 보다 코사인 함수를 이용하여 가중치를 부여하였을 때 정확도가 높았다. 실험은 KT-Set 2.0 (4391 건), 동아일보 96 년 신문기사(70459 건)를 대상으로 수행하였다.

Keyword
정보검색, 적합성 피드백, relevance feedback, 질의문 확장

1. 서론

정보검색시스템이란 여러 가지 정보를 수집하여 분석한 뒤, 찾기 쉬운 형태로 조직하여 두었다가 정보에 대한 요구가 발생할 때, 이용자의 질의문과 데이터베이스에 저장되어 있는 문서를 비교하여 이용자의 질의문과 가장 유사한 문서를 보여준다. 대부분의 정보검색시스템은 검색된 문서에 순위를 매겨서 이용자에게 가장 적합한 문서부터 먼저 보여준다.

그러나 원질의문에 의한 검색만으로는 부적합 문서의 검색율(정확도)이 높고, 적합 문서의 검색율(재현율) 또한 낮다. 즉, 정보검색 분야에서 정확도와 재현율을 동시에 높이는 문제는 매우 어려운 문제점으로 인식되어 왔다. 최근의 연구들은 원질의문만을 이용하여 정확도와 재현율을 높이는 것이 불가능하므로, 이를 위해서는 원질의문을 수정하여 재검색 해야 한다는 결과들을 내놓았다 .

지난 20 여년간 연구결과를 보면 질의문을 수정하는 방법으로 적합성 피드백(Relevance Feedback)이 가장 적절하다[1][2][6]. SMART 시스템을 사용한 초기 연구

[1] 본 연구는 한국과학재단 산학협력연구비 (962-0100-001-2) 지원으로 수행되었으며 지원에 감사를 드립니다.

(1971)와 확률적 가중치 모델 (Robertson & sparck Jones 1976)을 사용한 실험은 적합성 피드백을 이용하여 검색 효율이 매우 향상되었다[1][6].

그러나 기존 정보검색 시스템에서 적합성 피드백을 이용하여 검색 효율이 높아지기는 하였지만, 정확도면에서는 약 20~30% 정도로 더욱 향상시켜야 한다. 대부분의 검색 시스템에서 원질의문 검색과 적합성 피드백에 의한 재검색의 재현율은 약 80% 이상이다. 즉,.검색 대상이 되는 문서의 수가 많은 경우에는 적합 문서의 수도 많으므로 검색 효율을 높이기 위해서는 재현율을 높이기 보다는 정확도를 높이는 것이 더욱 의미가 있다. 또한 검색된 문서의 수가 많은 경우에는 상위 순위에 적합 문서가 많은 것이 더욱 효과적이다.

따라서 이 논문에서는 정보검색 시스템의 재현율과 정확도를 높이기 위해 "미리내"에 구현한 적합성 피드백 방법들을 분석하고, 각 방법들의 검색 효율을 재현율-정확도를 이용하여 비교 분석한다. 또한 신문기사 실험에서는 적합 문서 총수를 알 수 없으므로, 상위 30 내의 적합 문서를 이용한 정확도를 구했다.

2 장에서는 미리내 시스템과 미리내 시스템에서 적합성 피드백이 처리되는 과정에 대해 알아보고, 3 장에서는 단어 선택 방법과 단어 가중치 부여 방법에 대해 자세히 설명한다. 실험 결과와 분석은 4 장에 있다. 마지막으로 결론이다.

2. 미리내 시스템

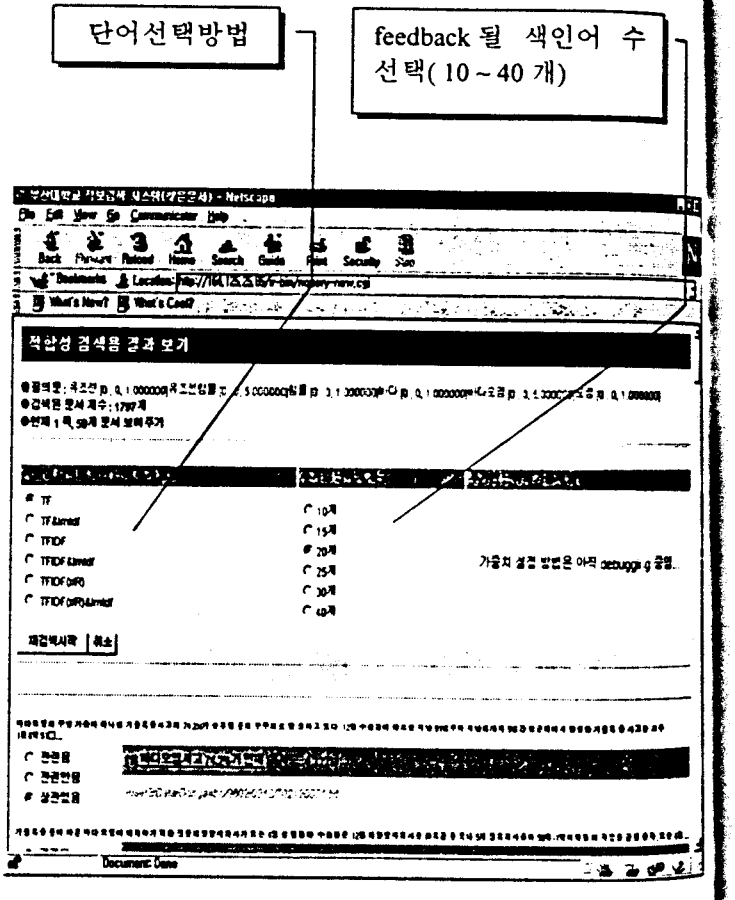
이 논문에서 적합성 피드백 실험에 사용하는 시스템은 부산대학교 전자계산학과 인공지능연구실에서 개발한 한국어 정보검색시스템 "미리내(MIRINE)"이다.

"미리내" 시스템은 검색기, 색인기, 등록기로 구성되어 있다. 등록기는 수집된 문서를 처리하여 검색에 필요한 형태로 조직하여 검색용 사전을 구축한다. 색인기는 자연언어 문장으로부터 색인어를 추출하여 주는 시스템이다. 검색기는 이용자 질의문을 입력 받아 색인기를 이용하여 질의어 리스트 추출하여 검색을 수행한다. 이때 등록기의 출력물인 검색용 사전을 사용한다

"미리내"는 벡터스페이스 모델에 기반한 검색시스템으로 자연언어 질의어를 기본으로 처리한다. 검색된 문서는 코사인 유사도 측정 방법을 이용하여 순위가 부여된다. 또한, 이용자가 직접 적합 문서를 선택하여 재검색할 수 있는 기능을 제공한다.

또한 한국어 문서에서 복합명사가 단일명사보다 해당 문서의 특성을 잘 나타내어 줄 뿐만 아니라 이용자 질의문에서도 이용자의 요구를 더욱 잘 나타낸다. 그러므로 "미리내" 시스템은 한국어의 특성을 고려하여 질의문의 복합명사를 단일명사에 비해 더 높은 가중치를 부여하여 준다.

다음은 "미리내" 시스템에서 적합성 피드백을 이용하여 정보를 검색하는 사용자 인터페이스이다. 그림 1의 사용자 인터페이스에서는 문서를 재검색하기 위해서 원 질의문에 대한 검색을 수행한 뒤, 이용자가 적합 문서와 단어 선택 방법을 선택하여 재검색을 수행한다. "미리내"에서는 검색 효율을 높이기 위해 여러 가지 질의문 확장과 가중치 조정 방법을 제공한다. 이용자는 이들 방법 중에서 한 개를 선택하여 적합성 피드백을 수행한 후 재검색할 수 있다.



[그림 1] "미리내" 정보검색시스템

"미리내" 시스템에서 적합성 피드백 과정은 다음과 같다.

1. 색인기를 이용해 원질의문에서 질의어 추출
2. 질의어와 검색용 사전의 색인어를 유사도 함수를 통해 비교하여 문서 순위 결정
3. 이용자가 1 차 검색 문서 중 적합문서를 판단
4. 단어 선택 방법을 이용하여 적합문서에서 단어를 자

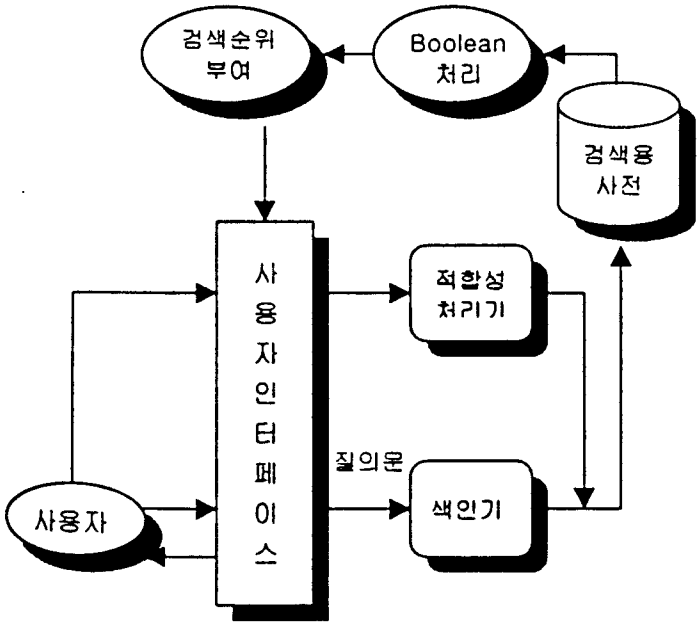
동으로 선택 (상위 20 개 단어)

5. 가중치 재산정 방법에 따라 선택한 단어에 가중치를 부여
6. 원질의문을 수정하여 재검색

3. 적합성 피드백

적합성 피드백에는 크게 두 가지 요소가 있다.[1] 첫째, 단어 선택 방법(term expansion method)은 검색된 적합문서에 있는 단어의 중요도에 따라 상위 n 개의 단어를 선택하여 질의문을 확장한다. 전체문서 집단 내에서의 중요도와 적합문서 집단 내에서의 중요도를 이용하여 단어의 중요도를 측정한다. 각 단어 선택 방법은 서로 다른 단어를 선택하고 따라서 검색 효율도 다르다. 둘째, 단어 가중치 부여 방법(term weighting method)은 선택한 단어에 가중치를 부여하여 검색의 정확도를 높인다.

미리내 시스템의 적합성 피드백을 이용한 검색시스템의 구성도는 그림 1 과 같다



[그림 2] 미리내 시스템 적합성 피드백 구성도

3.1 단어 선택 방법

이용자가 적합하다고 판단한 문서의 내용을 가장 잘 나타내는 단어를 선택하기 위해 각 단어가 가질 수 있는 모든 정보를 이용해야 한다. 먼저 단어가 적합 문서 내에서 가지는 중요도(local 정보)를 측정하기 위해 tf(단어빈도)와 (r/R)을 이용한다. 각 단어가 전체 문서 집단 내에서 가지는 중요도(global 정보)를 측정하기 위해 역문헌빈도(idf)를 사용한다. 역문헌빈도는 Spark Jones 가 제안한 (log(N) _ log(n) + 1)값을 사용한다. r 은 단어 k 가 있는 적합문서의 수이고, R 은 적합성 피드백에 이용하는 적합문서의 수이다.

- *Tf* - R 에서 단어 k 의 총 출현빈도
- *r/R* _ 적합문서 내에서 단어 k 가 나타난 적합문서의 비율
- *idf* - 전체 문서 집단 내에서 단어 k 가 나타난 문서의 비율

이 논문에서는 위에서 언급한 3 가지 요소들을 결합하여 사용한다. 그리고 각 방법들의 검색 효율을 정확도-재현율을 이용하여 비교한다. Local 정보와 Global 정보를 결합하는 방법은 다음과 같다.

(1) Tf _ 적합 문서 내에서 총 단어 빈도가 높은 단어를 우선으로 선택한다.
(2) Tfidf - 전체 문서에서 적은 수의 문서에 나타나면서 적합 문서 내에서 총 단어 빈도가 높은 단어를 선택한다.
(3) Tfidf*(r/R) - 단어 빈도가 높고 전체 문서 집단 내에서는 적은 수의 문서에 나타나면서 검색된 적합 문서에서 중에서 많은 수의 적합 문서에 나타나는 단어를 선택한다.
(4) Lmtdf _ 위의 3 가지 방법에 의해서 선택된 단어들 중에서 문헌 빈도가 높은 단어를 제외한 뒤, 단어를 선택한다.

"미리내" 시스템에서는 단어 선택 방법으로 위의 1-3 과 4 를 결합하여 6 가지 방법을 제공한다. 적합 문서에서 나타난 모든 문서에서 추출된 단어 중에서 값이 큰 상위 20 개 단어를 원질의문에 추가한다.

3.2 단어 가중치 부여 방법

"미리내"에서는 질의문에서 단어들의 중요도를 고려한 검색을 통해 검색 효율을 높이기 위해서 질의문의 단어에 가중치를 부여하여 검색을 수행한다. 재검색을 위해 추가되는 단어는 원질의문에서 검색된 문서들로부터 추출한다. 그러나 추가될 단어들이 적합 문서들을 대표하는 정도가 다르므로 그 단어의 중요도를 재검색 시에 사용해야 한다. 그러므로 "미리내" 시스템은 앞의 단어 선택 방법에서 계산된 값에 다음의 2 가지 함수를 이용하여 각 단어들의 가중치를 계산한다.

- 정규화 방법(Normalization Function)

단어 선택 함수에서 얻은 가장 높은 값으로 나머지 값들을 나누어 가중치를 0 에서 1 사이로 조정하고, 이 결

과를 f 라 한다.

- Cosine Function

위의 f 값에 코사인 함수를 적용하여 가중치를 계산한다. 코사인 그래프의 특징을 이용하여 f 값이 높은 중요한 단어들의 중요도는 더욱 높여주고, f 값이 낮은 단어들의 중요도는 낮춰준다.

$$W = \cos\left(\frac{\pi}{2} * (1 - f) \right)$$

4. 실험 및 결과

4.1 실험 방법

실험 데이터는 KT-Set 2.0과 동아일보 96년 신문기사 이다. KT-Set에 포함된 문서는 전자와 전산분야에 관련된 내용으로 KT 논문초록, 전자 신문 그리고 잡지 기사 4,391건으로 구성되어 있으며, 테스트 문서와 함께 50개의 자연어와 그에 해당하는 불리언 질의문이 제공된다. 각 질의문에는 적합문서가 미리 정의되어 있으며, 질의문은 평균 4.27개 단어이고, 질의문 하나의 평균 적합문서 수는 29개이다.

동아일보 데이터는 96년도 신문기사 70459건으로, 20개의 자연언어 질의문을 직접 개발하였다. 여러 종류의 다양한 질의문을 개발하기 위하여 학부생 100명을 대상으로 설문지 조사를 통해 질의문 500개를 수집하였다. 학생들에게 모든 질의문에 상세한 설명을 달도록 요구하여 검색 문서의 적합성을 판단할 때 판단기준으로 이용하였다. 500개의 질의문 중 96년 동아일보 신문기사 내용에 적합한 질의문을 정치, 경제, 사회, 체육, 정보통신 등의 각 분야에서 17개 선택하였다.

대부분의 이용자는 검색된 문서 전부를 이용하여 적합성을 판단하기 보다는 상위 순위의 문서만을 이용하는 경향이 있다. 그러므로 이 논문에서 수행한 실험은 적합문서 판단을 상위 30위 또는 50위에 대해서만 수행하였다.

부산대학교 전자계산학과 인공지능 연구실에서는 "미리내" 시스템의 적합성 피드백 실험에서 원질의문에 추가되는 단어의 수를 10, 20, 40, 60, 80, 100개로 변화하여 수행하였다. 그 결과 20, 40, 60개일 때 다른 경우에 비해서 검색율이 좋았다. 또한, Donna Harman의 실험에서는 20개 단어를 추가했을 때 가장 좋은 결과를 얻을 수 있었다[6]. 그러므로 이 실험에서는 적합성 피드백에 사용하는 단어의 수를 20개로 하여 실험을 수행한다. KT-Set 2.0에서는 50개의 자연어 질의문으로 상위 30위 이내에 검색된 적합 문서를 이용하여 적합성 피드백 실험을 하였다.

동아일보 실험은 한 사람에 의해서 실험을 수행한 1차 실험과 이 실험의 결과를 토대로 여러 사람이 토의한 결과를 이용한 2차 실험으로 구성된다. 먼저 1차 실험은 학부생 2명이 각각 질의문 8개, 9개씩 분배하여 개별적인 실험을 수행하였다. 이때 적합성 판단은 실험자 개인의 판단에 따라 수행하였다. 그러나 질의문에 대한 적합문서가 제공되지 않으므로, 2차 실험에서는 1차 실험에서의 검색 문서의 적합성을 컴퓨터공학과 학부생 4명과 전자계산학과 박사 1명, 인지과학 협동과정 석사 1명이 합의하여 3명 이상의 의견이 일치할 때 적합하다고 판단하였다.

피드백을 이용한 검색의 효율을 평가하기 위해 재현율-정확도를 이용한다. 그러나 동아일보의 경우 적합 문서를 알 수 없기 때문에 재현율-정확도 값을 구할 수 없기 때문에 검색된 문서 중 상위 50위(30위) 내에 검색된 문서의 적합성을 판단하여 50위(30위) 내의 정확도 값으로 검색효율을 평가한다. KT-Set은 동아일보 실험결과와 비교하기 위해 상위 30위 내의 정확도 값을 구하고 표준 재현율-정확도 값으로 검색 효율을 평가한다.

4.2 실험 결과 분석

4.2.1 단어선택방법

- *tf* - 이 방법은 문서 집단 내에서의 단어의 분포를 전혀 고려하지 않고 적합문서 내에서의 단어빈도 만을 이용하므로, 변별력이 떨어지는 단어를 추가할 수 있다. 따라서 뒤의 2가지 방법에 비해 부적합문서의

단어 선택 방법	KT-Set 2.0			동아일보	
	정확도	재현율	30위 내의 정확도	30위 내의 정확도	50위 내의 정확도
원질의문 검색	0.4192	0.82	0.3053	0.3867	0.3106
Tf	0.4720 (+12.6%)	0.97 (+18.3%)	0.3193(+4.6%)	--	--
Tfidf	0.5075 (+21.1%)	0.93 (+13.4%)	0.3247(+6.4%)	0.5490(+41.9%)	0.447(+43.92%)
Tfidf(r/R)	0.4850 (+15.7 %)	0.95 (+15.9%)	0.3253(+6.5%)	0.5510(+42.5%)	0.4518(+45.46%)

[표 1] 단어 선택 방법에 따른 검색 효율 비교 (Weight option = cosine)

검색율이 높아진다.

- *Tfidf* – 여러 문서에 골고루 분포되어 있는(변별력이 떨어지는) 단어를 역문헌빈도를 이용하여 제거함으로써 부적합문서의 검색율을 낮출 수 있었다. 따라서 매우 특정한 단어가 많이 선택되므로 이 단어가 나타난 문서는 순위가 상승하여 검색의 정확도가 높아진다.

- *Tfidf(r/R)* –. *Tfidf* 방법에 비해 r/R 의 영향으로 인해 변별력이 떨어지는 단어가 많이 선택되므로 정확도는 떨어지지만 재현율은 높아진다.

표 1 에서는 3 가지 방법의 검색효율을 정확도-재현율을 이용하여 나타내었다. KT-Set 에서는 tfidf 방법이 21.06%로 정확도가 가장 높이 증가했고, tfidf(r/R)은 15.85%로 재현율이 증가하였다. 동아일보는 KT-Set 에서 검색 효율이 높은 2 가지 방법을 적용하여 실험하였다. KT-Set 에서 30 위 내의 정확도를 보면 tfidf 는 6.4%%, tfidf(r/r)은 6.5% 향상하였다. 이에 반해 데이터의 크기가 큰 동아일보의 경우 30 위 내의 정확도를 보면, tfidf 는 41.9% tfidf(r/r)은 42.5% 증가하였다. 데이터의 크기가 큰 문서집단의 경우 질의문에 대한 적합문서의 수가 많으므로 질의문을 수정하여 재검색하였을 때 더 많은 수의 적합문서를 검색할 수 있었다. 데이터의 크기와 특성에 따른 적합성 피드백의 효과에 대한 연구와 실험이 필요하다.

또한 질의문에 따라 tfidf, tfidf(r/R) 방법의 결과에 차이가 있었다. tfidf 방법을 적용했을 때 더 많은 수의 적합문서를 검색한 경우와 tfidf(r/R)방법을 적용했을 때 더 많은 수의 적합문서를 검색한 경우가 있었다. 질의문의 특성에 따른 적절한 적합성 피드백 방법에 대한 연구와 실험이 필요하다.

4.2.2 문헌빈도 제한

질의문을 확장하기 위해 적합문서에 나타난 단어의 로컬 정보와 글로벌 정보를 이용하여 단어를 선택하였다. 그러나 선택된 단어들을 분석한 결과, 현재 이 논문에서 tf 의 값이 idf 값보다 크기 때문에 idf 값이 큰 단어보다 tf 값이 큰 단어가 우선적으로 선택된다. tf 가 큰 단어들은 많은 문서에 나타나는 일반적인 단어이기 때문에 이러한 단어가 질의문에 추가되었을 때 많은 수의 부적합 문서를 검색하게 되어 정확도가 떨어진다. 단어빈도의 경우 중간빈도의 단어가 중요하고, 문헌빈도의 경우 N/100 에서 N/10 사이에 오는 단어가 중요한 의미를 가진다[8]. 따라서 이 논문에서는 문헌빈도를 제한하여 너무 일반적인 단어가 질의문에 추가되어 부적합 문서의 검색을 억

(원질의문 : 0.4192(정확도), 0.82(재현율), 0.3053(정확도(30))

	Tf		Tfidf			Tfidf*(r/R)		
	정확도	재현율	정확도	재현율	정확도 (30)	정확도	재현율	정확도 (30)
No-lmtdf	0.4720 +12.59	0.97 +18.29%	0.5075 +21.06%	0.93 +13.41%	0.326 +6.8%	0.4850 +15.7%	0.95 +15.85%	0.326 +6.8%
N/3 (1464)	0.4740 +13.07%	0.96 +17.07%	0.5112 21.95%	0.92 +12.2%	0.324 +6.1%	0.4875 +16.29%	0.94 +14.63%	0.3267 +7%
N/5 (878)	0.4790 +14.27%	0.95 +15.85%	0.5131 +22.4%	0.91 +10.98%	0.3253 +6.6%	0.4917 +17.29%	0.93 +13.41%	0.3273 +7.2%
N/10 (439)	0.4936 +17.75%	0.93 +13.41%	0.5176 +23.47%	0.91 +13.98%	0.3267 +7%	0.4980 +18.8%	0.92 +12.2%	0.3293 +7.9%
N/15 (293)	0.5047 +20.4%	0.92 +12.2%	0.5232 +24.81%	0.90 +9.8%	0.3287 +7.6%	0.5073 +21.02%	0.91 +10.98%	0.3306 +8.3%
N/20 (220)	0.5650 +20.61%	0.90 +9.8%	0.5234 +24.86%	0.89 +8.5%	0.3247 +6.3%	0.5099 +21.64	0.90 +9.8%	0.3267 +7%
N/25 (176)	0.5084 +21.28%	0.90 +9.8%	0.5251 +25.26%	0.88 +7.3%	0.3253 +6.5%	0.5121 +22.16%	0.89 +8.5%	0.3313 +8.5%
N/30 (146)	0.5055 +20.59%	0.89 +8.5%	0.5245 25.12%	0.87 +6.1%	0.322 +5.5%	0.5087 +21.35%	0.88 7.3%	0.3273 +7.2%
N/35 (125)	0.5057 +20.63%	0.89 +8.5%	0.5250 25.24%	0.87 +6.1%	0.3233 +5.9%	0.5087 +21.35%	0.88 +7.3%	0.33 +8.1%

[표 2] 문헌빈도 제한(lmtdf)에 따른 검색 효율 비교

(정확도(30) : 30 위 내의 정확도)

(Weight Option = cosine)

제하고자 하였다. 문헌빈도 기준선을 N/3, N/5, N/10, N/15, N/20, N/25, N/30, N/35로 변화 시키면서 실험한 결과가 표 2이다.

문헌빈도를 제한하면, 변별력이 떨어지는 단어들을 제외되고 그 문서의 내용을 정확하게 나타낼 수 있는 특정한 단어들이 많이 선택된다. 따라서 문헌빈도 제한선을 1464에서 125까지 변화 시킨 결과 문헌빈도를 낮게 제한할수록 정확도는 계속 증가한다. 그러나 문헌빈도를 제한하지 않은 결과와 비교하면 재현율은 떨어진다. 따라서 30위 내의 적합문서의 수를 이용해서 비교한 결과 tfidf는 N/15(293)으로 제한하였을 때 30위 내에 검색된 적합문서의 수가 평균 9.86개로 7.6% 증가하였고, tfidf(r/R)는 N/25(176)으로 제한하였을 때 30위 내에 평균 10개의 적합문서를 검색하여 9% 증가하였다. tfidf에 비해 tfidf(r/R)에서 일반적인 단어가 많이 선택되었기 때문에 문헌빈도를 제한하였을 때 tfidf(r/R)에서 그 효과가 컸다.

4.2.3 가중치 설정 방법에 따른 검색 효율

단어선택방법으로 질의문을 확장한 다음, 검색의 정확도를 높이기 위해 가중치를 조정할 필요가 있다. 코사인 함수를 이용한 방법은 중요하다고 판단한 단어의 가중치는 더 높여주고, 중요하지 않다고 판단한 단어의 가중치는 낮춰주는 효과가 있다.

4.2.4 지식에 따른 적합성 판단

우리는 동아일보 데이터를 실험하면서, 검색하고자 하는 질의문에 대한 지식 정도에 따라 적합성 판단에 차이가 있음을 발견하였다. 1차 실험의 실험자를 A라 하고, 2차 실험의 실험자를 B라고 한다.

Q1. 2002년 월드컵을 한국과 일본에서 주최하게 된 배경

질의문의 검색의도는 "월드컵을 두 나라에서 공동으로 개최하는 것은 흔치 않은 일이다. 왜 공동개최결정이 이루어졌는지 와 일본과 한국에서 어떠한 방식으로 월드컵 경기를 할 것인가?"에 대한 기사를 찾는 것이다. 실험자 A는 월드컵에 대한 지식이 거의 없는 상태이고, 실험자 B는 월드컵에 대한 기본 지식이 많은 상태였다. 실험자 A는 검색 문서에 대한 적합성을 판단할 때, "월드컵, 공동개최, 한국, 일본"과 같은 키워드가 많은 문서를 적합하다고 판단하였다. 실험자 B는 "FIFA의 아벨란제 회장과 일본과의 관계, 98년에 있을 FIFA 회장 선거에 대한 전략"과 같이 월드컵 공동개최에 대한 구체적인 사실을 다루는 기사를 적합하다고 판단하였다.
A는 50위 내에 6개의 문서가 적합하다고 판단하였고, B는 8개의 문서가 적합하다고 판단하였는데 4개의 문서에 대한 적합성 판단이 일치하였다. 실험자의 지식과 검색 경험에 따라 적합성 판단에 많은 차이가 있음을 알 수 있었다.

5. 결론

단어 선택에 있어서는 각 문서들에 대한 정보와 문서 집단에 대한 정보를 결합하여 사용하는 경우에 검색 효율이 향상되었다. 그러나 단어 빈도(tf)가 상대적으로 너무 큰 경우에는 역문헌빈도에 의해서 단어 선택을 위한 값이 조정되지 않았다. 그러므로 "미리내"에서는 문헌빈도를 제한하여 문헌 빈도가 한계치 이상 되는 단어를 제거한 뒤, 실험한 결과 효율이 향상되었다.

단어 가중치부여 방법은 추가되는 단어의 가중치가 원질의문의 가중치를 넘지 못하게 조정하였고, 원질의문과 동일한 단어가 추가되면 재검색 시 그 단어의 가중치의 합은 1을 넘게 된다. 즉 이 방법은 원질의문의 단어가 추가되는 단어보다 이용자 요구에 더욱 적합하다는 가정하에서 이루어졌다. 이 논문에서 제시한 코사인 함수를 이용한 가중치부여 방법이 정규화 방법에 비하여 두 문서 집단 모두에서 보다 나은 효과를 보였다.

이 논문에서 제시한 적합성 피드백을 이용하여 재검색한 결과 방법들 모두 "미리내" 시스템의 검색 효율을 향상시킬 수 있었다. 그러나 방법에 따라서 검색 효율은 다소의 차이가 있었고, 문서 집단의 크기가 클수록 적합성 피드백의 효과가 현저하게 증가되었다.

향후 연구방향으로는 이 실험에서 사용한 문서 집단

가중치 부여 방법	Tf		Tfidf		Tfidf(r/R)	
	정확도	재현율	정확도	재현율	정확도	재현율
Normalize	0.4689 +11.85%	0.97 +18.29%	0.5038 +20.18%	0.93 +13.41%	0.4775 +13.91%	0.95 +15.85%
Cosine	0.4720 +12.59%	0.97 +18.29%	0.5075 +21.06%	0.93 +13.41%	0.4850 +15.7%	0.95 +15.85%

[표 3] 가중치 부여 방법에 따른 검색 효율 비교

의 크기가 너무 적으므로 보다 큰 문서 집단에 대한 실험이 필요하다. 또한 동아일보 실험은 17개 질의문만을 이용하여 실험하였는데, 더 많은 수의 질의문에 대한 실험이 필요하다.

정보검색 시스템에서 질의문의 특성에 따라 적절한 피드백방법에 대한 연구 실험이 필요하고, 문헌 집단의 크기와 특성에 따른 연구 실험이 필요하다. 또한 한국어 정보 검색을 위해서는 복합 명사의 가중치를 높여주는 것과 같은 한국어의 특성을 이용한 적합성 피드백 방법의 개발도 필요하다.

6. 참고 문헌

[1] W. B. Frakes and R.Baeza-Yates, *Information retrieval : data structures and algorithms*, 1992, Prentice Hall, New Jersey.

[2] G. Salton and C. Buckely, "Improving Retrieval Performance by Relevance Feedback", JASIS, 1990 41(4), 288-297.

[3] D.Hains and W.B. Croft, "Relevance Feedback and Inference Networks", http://ciir.ca.umass.edu/info/psfiles/irpub/ir.html.

[4] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGrawHill, 1983.

[5] C. Buckely and G. Salton, "The Effect of Adding Relevance Information in a Relevance Feedback Environment", SIGIR 94, 292-301.

[6] Donna Harman, "Relevance Retrieval Revisited", SIGIR 92, 1-10.

[7] Donna Harman, "An Experimental Study of Factors Important in Document Ranking", Information Retrieval, 1986, 8, 186-193.

[8] 정영미, "정보검색론" 구미무역㈜출판부, 1993.

[9] 박세진, 강상배 "Relevance Feedback을 이용한 정보검색 시스템의 검색 효율 향상", HCI 97, 1997, 3-8.

[10] 이준영, 강상배. "다중색인에 의한 정보검색 시스템 구현" 한글 및 한국어정보처리 학술발표 논문집, 1996, 63-37.

[11] 이준영. "다중색인과 압축저장에 의한 정보검색 시스템 개발에 관한 연구", 부산대학교 전자계산학과 석사학위논문, 1997.