한국어 웹 정보검색 시스템의 정확도 향상을 위한 연관 피드백 에이전트

백 준 호† · 최 준 혁†† · 이 정 현†††

요 약

기존의 한국어 웹 정보 검색 시스템은 대부분이 불리언 검색 시스템이므로 사용자가 원하는 정보를 한 번의 질의에 의해 얻기가 매우 어렵다. 또한 생략이 빈번하고 링크가 많은 웹 문서의 특성상 기존의 역문헌 빈도에 의한 키워드 선정은 중의성의 문제를 가중시켜 부적절한 키워드가 추출된다. 따라서 원하는 정보를 얻을 때까지 사용자는 질의어의 수정을 반복한다.

본 논문에서는 이러한 문제를 해결하기 위해 연관 피드백(Relevance Feedback) 에이전트 시스템을 설계하고 구현하였다. 연관 피드백 에이전트 시스템은 사용자의 선호 키워드에 대한 적합 문서를 추출하여 선호 키워드를 선호 DB 테이블로 저장하였다가 사용자가 추후에 검색할 때 사용자 질의에 연관 키워드를 추가하여 검색한다. 이 결과로 사용자의 질의 수정의 횟수를 줄이고 검색 효율을 향상시킬 수 있었다.

Relevance Feedback Agent for Improving Precision in Korean Web Information Retrieval System

Jun-Ho Back† · Jun-Hyeog Choi†† · Jung-Hyun Lee†††

ABSTRACT

Since the existed Korean Web IR systems generally use boolean system, it is difficult to retrieve the information to be wanted at one time. Also, because of the feature that web documents have the frequent abbreviation and many links, the keyword extraction using the inverted document frequency extracts the improper keywords for adding ambiguous meaning problem. Therefore, users must repeat the modification of the queries until they get the proper information.

In this paper, we design and implement the relevance feedback agent system for resolving the above problems. The relevance feedback agent system extracts the proper information in response to user's preferred keywords and stores these keywords in preference DB table. When users retrieve this information later, the relevance feedback agent system will search it adding relevant keywords to user's queries. As a result of this method, the system can reduce the number of modification of user's queries and improve the efficiency of the IR system.

1. 서 론

컴퓨터의 증가와 인터넷의 발달로 접할 수 있는 정보도 다양하고 그 양도 무수히 증가하고 있다. 이러한 많은 정보의 홍수 속에서 사용자가 원하는 정보를 빠르고 정확하게 찾아주는 시스템이 정보 검색 시스템이다. 또한 인터넷 사용 인구의 증가로 인하여 웹(Web)이라는 새로운 형식의 정보 구조가 등장하게 되었다.

대부분의 정보 검색 시스템 엔진들의 자료구조는 처리속도와 반응시간을 줄이기 위하여 데이터 베이스를

※ 본 연구는 인하대학교 98년도 연구비 지원에 의하여 수행되었음.
† 준 회 원 : 인하대학교 대학원 전자계산공학과
†† 종신회원 : 김포대학 컴퓨터계열 교수
††† 종신회원 : 인하대학교 전자계산공학과 교수
논문접수 : 1998년 11월 30일, 심사완료 : 1999년 4월 13일

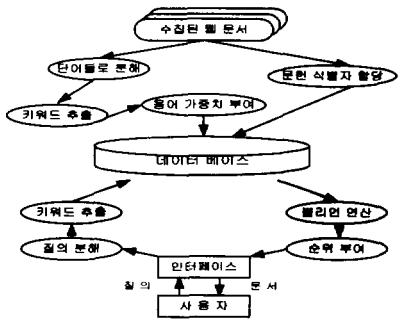
사용하지 않고 역화일 구조를 사용한다. 그러나 역화일 구조의 특성상 갱신 비용이 많이 들기 때문에 갱신을 하지 않고 역화일을 새로 생성한다[1]. 따라서 사용자의 질의에 의해 수집된 과거 데이터를 저장하지 않고 새로 수집해온 데이터에만 의존하여 검색하기 때문에 대부분의 사용자는 만족스런 결과를 얻지 못하고 결과에 만족할 때까지 질의문을 수정해 가면서 탐색을 반복한다[2,3].

이러한 문제를 해결하기 위한 노력으로 연관 피드백 방법을 사용하였으나 대부분이 사용자 질의어 분석에 대한 연구에 머무르고 있으며, 일차 검색한 문서에서 다시 키워드를 추출하려는 연구가 대부분이기 때문에 시스템의 응답시간을 중요시하는 웹 검색 시스템에서는 적절하지 못하다[4].

본 논문에서는 기존의 역화일에서 사용자가 검색에 자주 사용하는 질의를 가공하여 질의의 연관 키워드를 사용자 선호 DB 테이블로 저장한다. 그리고 사용자가 다시 질의하면 질의어에 기존의 적합 문서에서 추출한 연관 키워드를 추가하여 재검색시에 이용하는 연관 피드백 방법을 설계하고 구현하였으며, 이에 사용자의 질의를 적절한 질의로 수정하는 에이전트 개념을 적용하였다. 이러한 연관 피드백 에이전트 시스템을 사용하여 최근에 수집해온 역화일 자료에 사용자 질의어의 키워드와 연관된 키워드를 추가하여 검색함으로써 응답시간이 향상되었으며, 재현율은 유지하면서도 정확도가 향상됨을 알 수 있었다.

2. 기존 웹 정보 검색 시스템

정보 검색 시스템은 시스템의 사용자가 필요로 하는 정보를 수집하여, 정보 자료의 내용을 분석한 후, 검색하기 쉬운 형태로 저장하여 두었다가 정보에 대한 요구가 발생할 때 적합한 정보를 검색하여 제공하는 시스템이다[5]. 또한 정보 검색 시스템의 중요한 요소는 웹 상의 문서 형식으로 제공되는 정보를 찾아내어, 그 문서의 중요 내용을 요약하여 표현하는 것이다[6].



(그림 1) 기존의 웹 정보 검색 시스템

기존의 일반적인 정보 검색 시스템의 구조는 (그림 1)과 같다.

이러한 과정을 거쳐 데이터 베이스를 작성하기 위하여 웹 문서들을 읽어 들여 단어들로 분해하며, 웹 문서에서 출현하는 각 단어들의 빈도를 계산한다. 이 값은 전체 문서들에 대해서 표준화되고, 문서들과의 관련성을 계산하여 데이터 베이스에 키워드와 문서의 쌍으로 (그림 2)와 같이 역화일 형태로 저장된다[7,8].

키워드 1	URL(1), 항목
키워드 2	URL(3), 항목
키워드 3	URL(2), 항목
키워드 4	URL(2), 항목
키워드 5	URL(3), 항목
..	..
..	..
키워드 n	URL(N), 항목

(그림 2) 키워드와 문서식별쌍

키워드와 문서들로 만들어진 데이터 베이스를 탐색하기 위해서 사용자는 적절한 질의어를 입력한다. 그리고 사용자 질의어에서 추출된 키워드는 데이터 베이스 내의 역화일 검색을 통해 검색된 적합 문서에 대해 순위를 부여하여 사용자에게 검색 결과를 제공한다.

이러한 기존 검색 시스템에 대한 문제점은 역화일을 새롭게 생성하기 때문에 갱신 비용이 많이 든다는 점이다. 또한 기존의 자료에 대한 키워드를 보존하고 있지 않기 때문에 기존의 적합한 자료를 상실할 수 있다. 따라서 기존의 검색 시스템에서는 검색을 수행하였을 때 정확한 문서 위주의 검색보다는 재현율 위주의 검색결과를 제공하였다. 그러나 사용자는 실제로 재현율보다는 높은 정확도를 요구한다.

2.1 연관 피드백

연관 피드백은 첫 번째 반복 검색에서 초기 키워드로부터 검색된 문서와 검색되지 않은 문서들에 의하여

키워드들의 가중치를 재산정하는 것을 의미하며, 검색한 문서 중에서 적합하다고 생각되는 문서들에서 추출된 키워드를 사용자가 검색할 키워드에 추가하는 시스템이다[4].

기존의 연관 피드백을 이용하는 시스템으로는 사용자가 일차 검색 문서 중에서 적합 문서와 비적합 문서를 판단한 후 사용자와 관련이 있다고 생각되는 문서의 키워드를 질의한 키워드에 추가하여 사용하는 방법에 관한 연구가 있었다[2].

또 다른 연구로는 사용자 키워드를 수정하는 과정에서 불리언 연산을 수행하는 연구가 대부분이었다. 이는 사용자 질의에 대해 검색된 문서에서 적합하거나 비적합한 문서를 분류하고, 각각의 문서에서 추출한 키워드에 대해서 사용자 질의와 불리언 연산을 수행하는 시스템이다[4,9]. 이러한 시스템은 사용자가 정보를 검색하려고 질의를 수행하였을 때 곧바로 결과를 제시하지 못하는 단점이 있고, 또한 원 질의어와 수정된 질의어가 같은 데이터 베이스를 사용하여 유사도를 계산하는 관계로 사용자 응답 시간이 길어지는 단점이 있다. 따라서 이러한 방법은 웹 정보 검색 시스템에는 적합하지 않다.

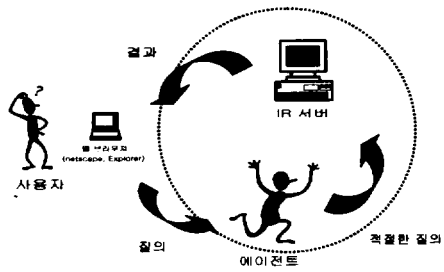
본 논문에서 사용하는 연관 피드백 에이전트는 사용자가 선호하는 키워드로 일차 검색한 후 적합 문서에서 추출한 키워드를 연관 키워드 테이블에 저장한다. 그리고 연관 키워드 테이블을 역화일 구조 형태로 저장한다.

3. 사용자 선호 DB를 이용한 연관 피드백 에이전트 시스템

3.1 웹 검색 에이전트

정보검색에서 사용되는 검색 엔진은 사용자의 습성에 대한 지식을 고려하지 않기 때문에 복잡한 질의를 반복하여야 한다. 따라서 검색 엔진 자체를 에이전트라고 간주하기는 어려우며, 사용자는 여러 단어를 통한 검색을 통해서 필요한 정보를 얻는다[10,11].

본 논문에서의 에이전트의 역할은 (그림 3)과 같이 사용자가 원하는 정보를 얻기 위하여 반복적으로 질의하는 문제를 해결하기 위하여 사용한다. 이때 에이전트는 사용자의 질의를 적절하게 수정하여 검색을 수행하기 때문에 검색 시스템은 보다 적합한 문서를 추출하여 사용자에게 제공할 수 있다.



(그림 3) 질의 수정 에이전트

3.2 연관 피드백 에이전트 시스템

본 논문에서 설계된 연관 피드백 에이전트는 사용자 선호 DB를 구축하여 사용한다.

사용자의 질의에 대해 역화일을 통해서 적합한 문서를 찾은 결과가 (그림 4)와 같이 생성되면 <Pri> </Pri>의 키워드 중에서 가중치가 높은 키워드를 연관 키워드에 대한 선호 DB로 구축한다.

```
<URL>1720</URL>
<Title>동아리</Title>
<Pri>의학과,간호학과,동아리,</Pri>
<Sec>학생,연구소,의학,</Sec>
<URL>1722</URL>
<Title>의과대학 부속병원</Title>
<Pri>진료,병원,의과대학,부속병원,</Pri>
<Sec>의학,의학과,예과,간호학과,</Sec>
<URL>2318</URL>
<Title>인하대학교를 소개합니다.</Title>
<Pri>인하대학교,</Pri>
<Sec></Sec>
```

(그림 4) 적합 문서의 식별자와 키워드

사용자 선호 DB의 테이블 구조는 <표 1>과 같으며, 테이블은 새로운 역화일이 생성될 때마다 사용자 연관 키워드에 대해 생성하게 된다.

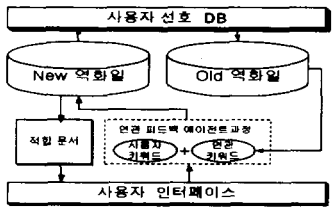
〈표 1〉 데이터 테이블

사용자 질의	연관키워드1	연관키워드2	연관키워드3
…	…	…	…
…	…	…	…

사용자 선호 DB를 이용함으로써 비용상의 문제로 갱신이 불가능하여 일정한 기간에 새로 생성해야 하는

역화일의 단점을 보완할 수 있으나 속도가 느리다는 문제는 여전히 남아있다. 그러나 이러한 문제는 사용자 선호 DB를 역화일로 다시 변환하고 사용자 연관 키워드와 사용자 키워드를 새로운 역화일에 추가하여 검색함으로써 해결할 수 있다.

(그림 5)는 삭제된 역화일에서 적합 문서를 추출하고 그 문서의 키워드를 사용자 선호 DB에 저장하는 과정을 나타낸다. 그리고 사용자가 검색을 요구할 때 사용자 키워드에 대한 적합 문서 키워드를 추가하여 검색하는 연관 피드백 에이전트 시스템이다.



(그림 5) 연관 피드백 에이전트 시스템

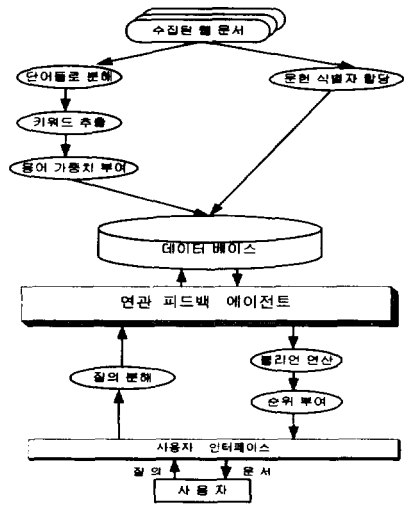
Old 역화일은 기존의 수집된 웹 문서에서 사용자가 자주 사용하는 질의어에 대한 선호 DB의 키워드로부터 적합 문서를 추출한다. 그리고 그 문서의 키워드를 연관 키워드와 문서의 쌍의 역화일 구조로 생성한 것이고, New 역화일은 최근에 수집된 웹 문서에서 추출한 키워드를 키워드와 문서의 쌍의 역화일 구조로 생성한 것이다.

본 논문에서의 연관 피드백 에이전트 시스템은 사용자 키워드와 몇 개의 유사한 키워드를 Old 역화일에서 추출하여 New 역화일에 추가하여 검색함으로써 새로운 역화일에서 보다 정확한 검색 결과를 얻을 수 있다. 또한 사용자가 선호하는 키워드를 지속적으로 갱신하여 선호 DB를 재작성한다면 검색 시스템의 정확도는 더욱 향상될 것이다.

4. 전체 시스템 설계

본 논문의 시스템 구성도는 (그림 6)과 같다.

전체 시스템은 먼저, 인터넷 상에서 웹 로봇을 이용하여 웹 문서를 수집한다. 그리고 역문헌 빈도를 이용하여 가중치를 부여하고 문서들의 키워드를 추출한다. 문서에서 새로운 키워드가 추출되면 기존의 역화일을 갱신하지 않고 새로운 역화일로 저장하며, 또한 문서 식별자도 저장한다.



(그림 6) 연관 피드백 에이전트를 이용한 시스템

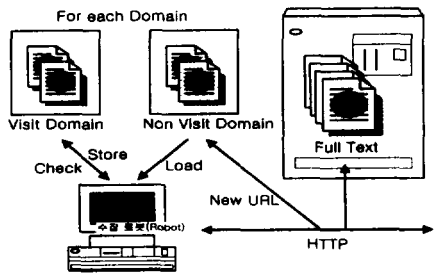
웹 문서의 특성상 소멸과 생성이 빈번히 반복되기 때문에 수집 작업과 저장 작업을 주기적으로 수행해야 한다. 또한 수집된 문서의 키워드는 문서의 빈도에 의해 계산되어지기 때문에 기존의 역화일에서 적절하게 추출된 키워드는 새로 수집한 역화일에서 키워드가 아닐 수 있다.

그러나 사용자는 검색 시스템이 언제 새롭게 키워드를 수집하였는지 알 수 없다. 그러므로 첫 번째 검색한 질의를 다음 번 검색할 때에도 그대로 이용하며, 원하는 정보를 얻지 못하였을 경우에는 원하는 정보를 얻을 때까지 계속하여 검색 시스템에 질의를 보내게 된다. 이렇게 사용자가 질의를 반복하여 검색하는 문제를 해결하기 위하여 본 논문에서는 기존 검색 시스템에 연관 피드백을 이용한 에이전트 시스템을 추가하였다.

4.1 웹 문서의 수집

방대한 양의 웹 문서를 수집하는 로봇(robot)은 키워드 추출과 문서를 처리하는 검색 시스템에 입력으로 제공하는 역할을 수행한다.

문서 수집 로봇은 사용자가 선정한 URL을 우선 처리하고, 프레임이나 이미지 맵, 구조체 등 웹 문서를 분석한다. 그리고 문서의 수정 상황을 검사하고, 수집된 문서들을 가지고 문서의 수와 내용, 제목, URL 등이 저장된 엔트리 파일을 생성하게 된다.



(그림 7) 웹 문서 수집 로봇

본 논문에서 사용하는 웹 문서 수집 로봇은 (그림 7)과 같이 하나의 URL을 먼저 검색하고, 다음 URL을 선정하는 깊이 우선 탐색 방법을 이용한다.

4.2 웹 문서 키워드 추출

키워드 추출은 방대한 개개의 인터넷 문서자료의 특성을 표현하는 데이터 요소를 뽑아 문서의 내용을 대표하도록 한 것으로 키워드 추출 작업을 통해 만들어진다. 키워드는 많은 문서 정보로부터 가장 적절한 문서를 선별해 주는 역할을 한다[5].

키워드를 검색 도구로 하는 검색 시스템에서 특정한 재현율 수준에서 높은 정확도를 가져오는 키워드일수록 효과적인 키워드라고 볼 수 있다. 정확도는 검색된 문서 속에 적합 문서가 많을수록 높아진다. 좋은 키워드일수록 문서들을 가능한 한 분리시킨다. 즉, 밀집도를 낮추고 유사도를 떨어뜨린다. 그리고 나쁜 키워드일수록 문서들을 가능한 한 무리짓게 한다[12,13].

본 논문에서는 기존의 검색 알고리듬을 적용하여 키워드를 추출함에 있어 (그림 6)과 같이 웹 문서를 내용형 문서와 내용이 없는 홍보형 문서로 분류하였다. 그리고 내용형 문서에 대해서만 기존의 키워드 추출 알고리듬을 적용하였으며, 내용이 없는 홍보형 문서에 대해서는 제목에서 출현하는 단어에 대하여 가중치를 배가시킴으로써 적절한 키워드를 추출하였다.

문서에서 출현하는 단어를 이용하여 웹 문서에 출현하는 단어가 일정 갯수 이상이면 이를 내용이 있는 문서로 추정하여 가중치를 준다. 이때 기준을 낮추면 재현율을 높일 수 있다. 내용형 문서를 선택해 가중치를 부여하는 알고리듬은 [알고리듬 1]과 같다.

다른 문서와 연결되어진 링크가 일정 갯수 이상이면 내용이 없는 문서로 추정한다. 또한 단어간의 결속력을 이용하는데 한 문서에 나타나는 단어들 사이에 유사도가 높으면 내용이 있는 내용형 문서로 추정한다. 즉, 단어들 사이의 결속력이 높으면 그 페이지는 내용형이고 결속력이 낮으면 내용이 없는 웹 문서로 간주하며, 템플릿을 이용해 내용형 문서를 선택한다.

[알고리듬 1] 가중치 부여 알고리듬

```
Begin
  Doc_num <-# //문서내의 단어수;
  Doc_Link <-# //문서내의 링크수;
  if(Doc_num ≥ Doc_N) //문서수가 N개 이상
    then(Entropy++);
  if(Doc_Link ≥ Doc_L) //링크수가 N개 이상
    then(Entropy--);
  for(i =1, N) { //단어간의 결속력 계산
    for(j =1, N) {
      CalcMi[i] <- Mi(Doc_W[i], Doc_W[j]) ;
      Mi_Ave  <- AVE(CalcMi);
      if(Mi_Ave ≥ Doc_Mi)
        then(Entropy++);
    }
  }
  if(Compare_W(Doc_W, Template_Contents))
      then(Entropy++); //내용형 템플릿 증가
  if(Compare_W(Doc_W,
     Template_Advertisement))
      then(Entropy--); //홍보형 템플릿 증가
End
```

결속력 계산은 상호 정보량 계산 수식을 사용한다.

상호 정보량 수식은 단어와 단어 사이의 연관성을 정량적으로 나타내기 위해 사용된다[14]. 상호 정보량 수식을 사용하여 단어와 단어간의 결속력을 측정하여 유사계수 행렬을 구성한다.

유사도가 일정하게 클러스터링되면 그 안에서 키워드를 구한다. 또한 그 키워드에 연결된 웹 문서에서도 문서의 단어 수를 구하고, 현재의 홈페이지와 관련 유사도를 구하여 유사도가 높으면 같은 부류의 문서로 간주한다. 또한 템플릿을 이용하여 내용형과 홍보형으로 선택된 샘플 문서에 대하여 자주 출현한 단어에 대해 템플릿을 만들어 놓는다. 내용형 템플릿에 매칭되는 단어에 대해서는 내용형 가중치를 증가시키고, 홍보형 템플릿에 매칭되는 단어에 대해서는 홍보형 가중치를 증가시킨다.

내용형 문서의 선택은 가중치에 의해 계산된다. 내용형 문서 선택 방법에서 제시한 방법들에 의해서 내

용형에 가까우면 가중치를 증가하고, 가중치가 기준치 이상이면 내용형 문서로 선택한다.

본 논문에서의 키워드 추출방법은 식 (1)과 같이 설튼(Salton)이 제시한 키워드 추출 방법인 역문헌 빈도에 문서내 단어 빈도를 곱한 값을 가중치로 부여하는 식을 사용하였다[7].

$$IDF = \log_2 \frac{n}{DF} + 1$$
$$= TF(\log_2(n) - \log(DF) + 1) \quad \text{식 (1)}$$

여기서 용어 빈도(Term Frequency)와 역문헌 빈도(Inverted Document Frequency)는 단어 빈도가 높고, 문서 빈도가 낮을수록 큰 값을 가진다.

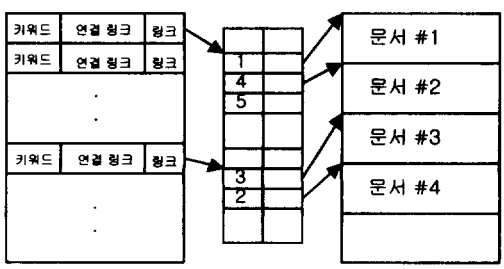
필터링된 내용형 문서에 대해서만 키워드를 추출하기 때문에 기존 문서의 키워드 색인 방법을 이용했을 경우보다 검색 효율이 좋다.

내용이 없는 문서들에 대해서는 웹 문서 특성인 태그 특성을 고려하여 키워드를 추출하며, 내용형 문서의 키워드 추출 방법, 제목에 대한 가중치, 글자의 크기, URL에 대한 가중치를 고려하여 키워드를 추출한다.

4.3 키워드 자료 구조

추출된 키워드를 저장하는 기법으로는 역화일(inverted file)을 이용하는 방법과 각 문서에 해당하는 키워드를 이진 열(bit stream)로 구성한 비트맵과 시그너쳐(signature) 파일을 이용하는 방법으로 나눌 수 있다. 이러한 저장 기법들은 빠른 속도와 효율적인 저장 공간을 이용해야 한다.

일반적인 역화일 구조는 (그림 8)과 같이 키워드를 포함한 문서로 연결하는 링크를 갖는 키워드의 색인으로 구성하며, 역화일 구조는 배열이나 트리, 트라이 등 다양한 구조를 가지고 있다.



(그림 8) 역화일 구조

대부분의 웹 정보 검색 시스템은 사용자가 웹 브라우져를 통해 검색이 이루어질 때 각각의 프로세스를 생성하는데 프로세스가 생성될 때마다 질의에 대한 검색을 위해 메모리로 역화일을 로드한다.

또한 많은 사용자가 검색을 수행할 때는 그만큼 역화일을 많이 로드해야 하기 때문에 검색 시스템의 성능이 저하된다. 그러므로 역화일의 크기가 클수록 검색 시스템의 검색 속도가 저하되므로 역화일의 크기는 작을수록 시스템의 성능이 좋아진다.

본 논문에서는 일반적인 역화일 구조에 비해 초기에 로드하는 파일의 크기가 작기 때문에 효율적으로 시스템 자원을 사용할 수 있는 트라이(trie) 구조를 사용한다[15]. 트라이 구조는 키워드의 첫 음절과 키워드에 대한 문서 식별자를 저장하며, 문서 식별자에 대한 각각의 문서 정보는 파일로 저장한다. 따라서 역화일내의 키워드를 검색할 때 전체 역화일을 로드하지 않아도 되고, 각각의 문서에 해당하는 파일만 로드하면 됨으로 검색 성능을 높일 수 있다.

5. 실험 및 평가

5.1 실험 데이터

실험 환경으로는 Windows NT 서버 4.0을 사용하였으며, DB로는 MS-SQL 서버 6.5를 사용하였다. 실험은 인하대학교 도메인내의 홈페이지 5,965개를 대상으로 수행하였다. 전체 URL 중 사용자가 자주 사용하는 40개의 키워드를 대상으로 실험하였으며, 평가는 정보검색 시스템의 검색 효율 척도 가운데 가장 많이 사용하고 있는 정확도와 재현율의 관계로 나타내었다.

연관 피드백 에이전트를 이용하여 사용자 선호 키워드 DB를 구축하는 과정은 다음과 같다. (그림 9)는 사용자 선호 키워드 '취업'에 대한 적합 문서 원문의 예이다.

```
<id> 07899
<title>취업 정보 서비스</title>
<abstract>
취업 정보 서비스 채용 속보 종합 정보 서류 전형 면접 시험 검사 자격증 안녕하십니까? 이용해 주셔서 감사합니다. 최고의 서비스를 위해 노력하고 있습니다.
노동부 부직 온라인 채용
</abstract>
```

(그림 9) 적합 문서 원문

(그림 10)은 (그림 9)와 같은 원문을 형태소 분석기를 이용하여 형태소를 분석한 결과이다.

```
[취업 ((취업) (N))]
[정보 ((정보) (N))]
[서비스 ((서비스) (N))]
[채용 ((채용) (N))]
        .
        .
        .
[노동부 ((노 동부) (N N)) ((노동 부) (N N))]
[부직 ((부직) (N))]
[온라인 ((온라인) (N))]
[채용 ((채용) (N))]
```

(그림 10) 원문의 형태소 분석 결과

적합 문서 원문을 형태소 분석한 결과에서 문서 1, 2와 같이 명사만을 추출하여 식 (1)의 역문헌 빈도를 이용하여 문서 1, 2의 키워드를 각각 구한다.

문서 1

채용 취업 광장 구인구직 등록하기 구인란 보기 구직 보기 고급인력 취업 속보 취업뉴스 기업탐방 기사

문서 2

취업 정보 관련 사이트 전국 대학교수 연구원초빙 노동부 홈페이지 리크루트 사이버 채용 박람회 취업 뱅크 신바람 일터 벼룩시장 구인 구직 인턴사원

적합한 문서들에서 추출된 키워드들과 문서와 키워드 사이의 가중치에 의해 순위 부여된 키워드들은 <표 2>와 같다.

〈표 2〉 적합 문서 키워드

적합 문서	키 워 드
문서 1	채용(1), 구인(2), 구직(3)
문서 2	구직(1), 채용(2), 인력(3), 구인(3)

적합 문서에서 추출된 키워드들 중에서 순위 부여 값이 높은 키워드를 연관 키워드로 추출한다.

<표 2>에서 순위 부여된 키워드 중에서 순위가 가장 높은 순서로 연관 키워드를 추출하면 <표 3>과 같다.

〈표 3〉 연관 키워드

선호키워드	연관키워드 1	연관키워드 2
취업	채용	구인 구직

연관 키워드를 추출하면 '취업'에 대한 선호 키워드에 대하여 구해진 연관 키워드를 역화일로 구축하고, 사용자가 추후에 '취업'을 검색할 때 연관 키워드 '채용', '구인, 구직'을 사용자 질의어 '취업'에 추가하여 검색한다.

5.2 분석 및 평가

연관 피드백에 필요한 주요 데이터는 각 키워드에 검색된 주요 적합 문서와 그 문서의 키워드이다. 사용자 선호에 의해 키워드가 순위 부여되면 그 키워드에 대한 적합 문서를 선별한다. 그리고 적합 문서의 키워드를 이용하여 연관 키워드를 구하며 연관 키워드는 <표 4>와 같은 구조의 데이터 테이블에 저장된다.

〈표 4〉 연관 키워드 데이터 테이블

선호 키워드	키워드1	키워드2	키워드3
수강신청	과 목	신청서	안 내
영 화	시네마	평 론	비디오
학사일정	일정표	행 사	학 사
학생회	선 거	활 동	
점심메뉴	식 당	식 단	
성 적	학 점	시험점수	평 가
검색엔진	정보검색	심마니	
동아리	써 클	대학문화	모 임
취 업	채 용	구인 구직	직 업
입학안내	신입생	전 형	

<표 4>의 연관 키워드는 데이터 베이스에 저장된다. 그리고 사용자의 새로운 선호 키워드에 대한 연관 키워드를 구하여 지속적으로 갱신한다.

본 논문에서의 성능 평가를 위하여 식 (3)과 같은 정확도와 재현율을 사용하였다[7,11,13,16].

$$\text{정확도} = \frac{\text{검색된 적합문서수}}{\text{검색된 문서 총수}}$$

$$\text{재현율} = \frac{\text{검색된 적합문서수}}{\text{적합문서 총수}} \qquad \text{식 (3)}$$

<표 5>는 사용자의 키워드만 가지고 검색을 수행한

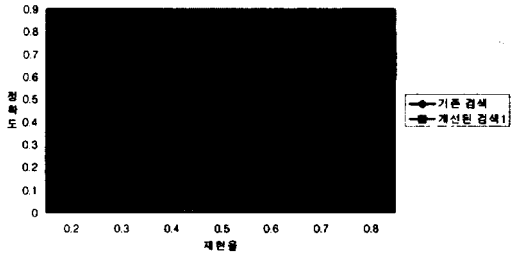
결과에 대한 정확도와 재현율과 본 논문에서 설계한 연관 피드백을 이용하여 한 개의 선호 키워드를 추가하여 검색을 수행하였을 때의 정확도와 재현율의 관계를 나타낸다.

〈표 5〉 검색효율 비교표

키워드	기존 검색		개선된 검색	
	재현율	정확도	재현율	정확도
수강신청	0.17	0.42	0.06	0.85
영 화	0.2	0.4	0.2	0.4
학사일정	0.7	0.18	0.3	0.3
학생회	0.48	0.3	0.16	0.63
점심메뉴	0.25	0.33	0.25	0.33
성 적	0.12	0.83	0.03	0.8
검색엔진	0.24	0.35	0.24	0.35
동아리	0.26	0.36	0.03	0.89
취 업	0.13	0.67	0.1	0.75
입학안내	0.5	0.2	0.5	0.2

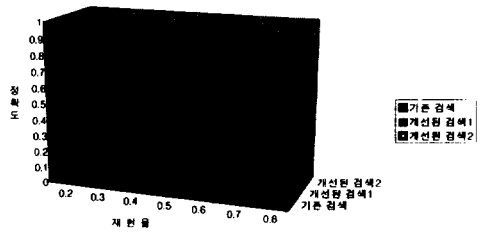
사용자의 키워드만 가지고 검색을 수행한 결과 60.4%의 정확도를 보였으나, 연관 피드백을 이용하여 키워드를 추가하여 검색을 수행하였을 때는 66.6%로서 정확도가 향상되었음을 알 수 있다.

기존 검색 시스템[15]과 연관 피드백 에이전트를 추가한 검색 시스템 사이의 정확도와 재현율 그래프는 (그림 11)과 같으며, 이를 통하여 재현율이 낮을때 정확도가 좀더 높게 나타남을 알 수 있다.



(그림 11) 정확도 재현율 곡선

연관 키워드를 세 개 이상 추가하였을 때의 정확도와 재현율은 (그림 12)의 [개선된 검색 2]와 같이 기존 검색보다 효율은 좋았으나, 재현율이 매우 감소하여 적합 문서가 검색되지 않음을 알 수 있는데, 이는 교내의 웹 문서만으로 실험한 한정된 문서 데이터에 기인한 것으로 추측된다.



(그림 12) 연관 키워드가 세 개 이상의 곡선

6. 결 론

본 논문에서는 기존의 검색 시스템에 연관 피드백 에이전트 시스템을 추가하여 사용자의 키워드에 적합 문서에서 추출한 키워드를 결합하여 검색 키워드로 사용함으로써 검색 결과의 정확도를 향상시켰다. 그리고 내용형 문서와 그밖에 내용형이 아닌 홍보형이나, 다른 형태의 문서를 필터링하여 선택된 특정 문서들만을 가지고 그 문서들에 대한 키워드 추출 방법을 적용함으로써 문서의 키워드 추출에 있어 정확도를 높일 수 있었으며, 내용형이 아닌 문서를 전부 저장할 필요가 없기 때문에 저장공간의 효율을 향상시킬 수 있었다. 또한 검색 시스템을 사용하는 비숙련자들에게 연관 키워드를 추가 제공하여 이용케 함으로써 숙련자가 검색하는 것과 유사한 검색 결과를 기대할 수 있다.

향후, 세 개 이상의 선호 키워드를 추가하여 검색을 수행하고, 비적합 문서에서 추출한 키워드에 대해서도 불리언 연산을 사용하여 검색하는 연구를 수행할 예정이다. 그리고 사용자나 집단의 정보 선호도를 기반으로 한 데이터 베이스를 구축하고, 개개인의 프로파일 데이터 베이스를 추가함으로써 검색의 정확도를 향상시킬 연구도 수행할 예정이다.

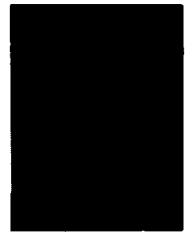
참 고 문 헌

[1] John Wiley & Song, *Agent Sourcebook*, Wiely Computer Publishing, 1997.

[2] 박세진, 강상배, 권혁철, "Relevance Feedback을 이용한 정보검색시스템의 검색 효율 향상", HCI 학술대회 발표 논문집, pp.3-8, 1997.

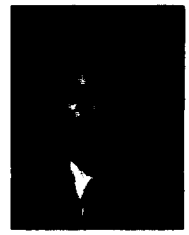
[3] 백혜정, 박영택, "웹 에이전트를 이용한 사용자 관심도 학습", 한국 정보과학회 논문지, Vol.24, No.

2, pp.89-92, 1997.
[4] Dillon, D. and Desper, J., "The Use of Automatic Relevance Feedback in Boolean Retrieval Systems," JD 36(3), pp.197-208.
[5] 정영미, *정보검색론*, 정음사, 1992.
[6] 박현주, 최재덕, 강상배, 박승, 박용욱, 권혁철, "인터넷 홈페이지 검색시스템 구현과 검색효율 향상", 제 9회 한글 및 한국어정보처리 학술발표 논문집, pp.63-67, 1997.
[7] William B. Frakes, Ricardo Baeza-Yates, *Information Retrieval*, Prentice-Hall, 1992.
[8] 최중민, "에이전트 개요와 연구방향", 정보과학회지, 제 15권, 제 3호, pp.17-28, 1997.
[9] Voorhees, E. and Buckley, C. and Salton G., "Boolean Query Formulation with Relevance Feedback," TR 83-539, 1983.
[10] 신봉기, "웹 에이전트", 정보과학회지 제 15권 제 3호, pp.61-68, 1997.
[11] 조영재, 이창훈, 박태순, "WWW 상의 자료검색을 위한 효과적인 에이전트 검색 알고리듬 구현", 한국정보과학회 학술발표 논문지, Vol.24, No.2, pp.77-80, 1997.
[12] Nicholas J. Belkin and W. Bruce Croft., "Information Filtering and Information Retrieval," Communications of the ACM, Vol.35, No.12, pp.29-38, December, 1992.
[13] 김영택, *자연언어 처리*, 교학사, 1994.
[14] 전미선, 박세영, "상호 정보를 이용한 어의 모호성 해소에 관한 연구", 제6회 한글 및 한국어정보처리 학술발표 논문집, pp.369-373, 1994.
[15] 오종인, "Client/Server에 기반한 웹 정보검색 시스템의 정확도 향상", 인하대학교 공과대학 전자계산공학과 석사학위 논문, 1998.
[16] Gerard Salton & Micheal J. McGill, *Introduction to Modern Information Retrieval*, 1983.



백 준 호

e-mail : paek100@nlsun.inha.ac.kr
1997년 건양대학교 전자계산학과 졸업(학사)
1999년 인하대학교 전자계산공학과 졸업(공학석사)
1997년~현재 인하대학교 대학원 전자계산공학과 재학
관심분야 : 자연어 처리, 정보검색



최 준 혁

e-mail : jhchoi@kimpo.ac.kr
1990년 경기대학교 전자계산학과 졸업(이학사)
1995년 인하대학교 대학원 전자계산공학과 졸업(공학석사)
1995년~현재 인하대학교 대학원 전자계산공학과 박사과정
1997년~현재 김포대학 컴퓨터계열 전임강사
1998년~현재 김포대학 전자계산소장
관심분야 : 자연언어처리, 정보검색, 신경망

이 정 현

e-mail : jhlee@dragon.inha.ac.kr
1977년 인하대학교 전자공학과 졸업(학사)
1980년 인하대학교 전자공학과 졸업(공학석사)
1988년 인하대학교 전자공학과 졸업(공학박사)
1979년~1981년 한국전자기술연구소 시스템 연구원
1984년~1989년 경기대학교 교수
1989년~현재 인하대학교 전자계산공학과 교수
관심분야 : 자연언어처리, 인간과 컴퓨터의 상호작용, 정보검색, 음성인식, 고성능 컴퓨터 구조