정보검색에서의 어의 중의성 해소를 위한 자동 키워드망의 이용

김 정 세[†]·장 덕 성[††]

요 약

문서 검색 시스템에서 자동 색인은 필수적이다. 그러나 자동 색인만으로는 최적합 문서들을 상위에 위치시키기 불가능하다. 뿐만 아니라 동음이의어를 갖는 부적합한 문서들이 상위에 위치되는 것을 막을 길이 없다. 본 논문에서는 이런 문제를 해소하고 검색 효과를 높이기 위해 2차 검색에 자동 키워드망을 이용하는 두 단계 검색시스템을 연구하였다. 1차 검색은 자동색인으로 만들어진 역색인 파일을 이용하며, 2차 검색은 단어 연관성을 기초로 만든 자동 키워드망을 이용한다. 2차 검색을 위한 문서 순위 재조정 식들을 여러 개 만들어 비교하였으며, 이 식들이 동음이의어의 어의 중의성 해소에 얼마나 효과가 있는지 성능을 평가하였다.

Resolving the Ambiguities in Word Sense by using Automatic Keyword Network in Information Retrieval

Jung-Sae Kim[†] · Duk-Sung Jang[††]

ABSTRACT

The automatic indexing is a compulsory part for the text retrieval system. However it is impossible to rank the appropriate texts at top. Furthermore, it is more difficult to prevent to rank the inappropriate texts having homonyms at top by only the automatic indexing. In this paper, we proposed the two-level retrieval system to enhance the retrieval efficiency, in which Automatic Keyword Network (AKN) is used at the second-level process. The first-level search is carried out with an inverted index file generated by the automatic indexing. On the other hand the second-level search exploits AKN based on the degree of association between terms. We have developed several formulas for rearranging the rank of texts at second-level search, and evaluated the performance of the effects of them on resolving the word sense ambiguities.

1. 서 론

오늘날 인간의 힘만으로 관리가 불가능 할 정도로 수많은 정보가 쏟아져 나오고 있다. 이러한 정보들을 컴퓨터에 저장하고 관리하여 사용자에게 제공하는 정보 검색 시스템(Information Retrieval System)이 널리 연구되고 있다.

일반적으로 방대한 문서의 집합에서 정보 요구자가 원하는 문서를 추출하기란 쉽지 않다[3, 5]. 정보 검색에서 질의 검색어들이 문서에 대하여 어느 정도의 중요도를 가지고 존재하느냐를 기준으로 문서를 순서화한다. 문서의 순서화는 단순히 사용자 질의를 만족하는 문서들의 집합을 검색하는 것이 아니라, 질의를 만족하는 정도에 따라 검색된 문서들에 순위를 부여하여 사용자들이 필요한 정보를 얻는데 소모되는 시간을 최

※ 본 논문은 1998년도 계명대학교 비사연구기금으로 이루어졌음.
† 정 회 원 : 한국전자통신연구원 음성언어팀 연구원
†† 정 회 원 : 계명대학교 컴퓨터전자공학부 교수
논문접수 : 2000년 7월 25일, 심사완료 : 2000년 11월 28일

소화하는 것이다[2]. 그러나 실제 순서화된 문서들을 보면 질의한 내용과는 다른 문맥의 문서들이 상위로 순서화되는 경우를 볼 수 있다. 이는 비록 질의 용어들이 해당 문서에서 보다 중요도를 갖지만 문서에서 다루고 있는 문맥과는 반드시 일치하지는 않기 때문이다.

문서순위 결정방법과 관련하여 적합성 피드백에 의해 질문을 수정하는 방법, 시소러스를 이용하여 색인어간의 개념적 밀접성을 계산하는 방법, 상호정보를 이용하여 시소러스로서는 부족한 색인어간의 관련도를 구하는 방법, 형용사나 부사와 같은 비주제어를 통해 성능을 개선하는 방법 등이 연구되고 있다.

적합성 피드백이란 검색된 문서의 내용을 보고 사용자가 적합문서와 부적합문서를 구분하고, 이들 문서에 나타난 단어를 중심으로 적합성 가중치 공식을 사용하여 질의를 수정하여 다시 질의하여 문서를 검색하는 것을 말한다[6]. 자동 질문 수정이 가능한 대표적 시스템으로 SMART 시스템이 있는데, 실험 결과 상당히 높은 검색효율을 가져오는 것으로 나타나 있다. 한번의 적합성 피드백으로 정확률을 10~20% 향상시킬 수 있으나, 적합성 판정의 애매함 때문에 잘못된 질의 수정으로 인해 원래의 의도를 벗어날 수 있다는 단점이 있다.

시소러스를 이용한 문서 순위 결정 방법은 문서에 포함된 색인어와 질의에 포함된 색인어간의 개념적 밀접성을 계산하여 문서들의 순위를 결정하는 방법이다. 적합성 알고리즘(Relevance), 거리 알고리즘(R-Distance, K-Distance)은 기본 거리 함수(Primitive Distance Function)를 기반으로 하여 질의와 문서 사이의 개념적 근접성 또는 개념적 거리를 계산한다. 기본 거리 함수 distance (t_i, t_j)는 시소러스에 포함되어 있는 임의의 두 색인어들 t_i, t_j사이의 개념적 거리를 나타낸다. 논문 [9, 11]은 이 방법들을 확장하여 모델인 KB-FSM(Knowledge Based Fuzzy Set Model)과 KB-EBM(Knowledge Based Extended Boolean Model) 모델을 제안했다.

재현율을 향상시키기 위해 질의 확장을 하는데 있어, 시소러스와 상호 정보를 동시에 이용하는 방법이 있다[1, 4, 8]. 시소러스는 용어간의 관계만이 정의되어 있을 뿐, 용어간의 관련도는 나타나지 않는다는 단점을 보완하기 위해 상호 정보를 이용한다. 시소러스에 나타나지는 않지만 상호 정보 값에 의해 상당한 관련도를 나타내는 용어들에 대해 새로운 개념관계로 시소러스에 등록하여 실제 상황에 유동적으로 시소러스를 구성한다. 질의 확장시 질의에 나타나는 용어와 관련된 용어들을 어느 정도 까지 포함할 것인가에 대한 척도도 상호 정보에 의해 제공된 용어간의 관련도에 의해 조절 가능하다.

키워드가 아닌 단어들 즉, 형용사나 부사, 수사 같을 것들을 비주제어라고 하고, 이들 비주제어를 통해 검색 효율을 개선하는 방법이 있다[10]. 키워드로 색인된 데이타로 1차 검색을 하고, 질의 내의 비주제어와 1차로 검색된 문서를 본문 검색하여 순서를 재조정한다. 그러나 정확한 의존관계에 기초하지 않아 비주제어와 키워드간의 연관관계가 잘못 생성될 수 있다. 이것을 단어 거리에 따라 가중치를 조정하여 줌으로서 해결한다. 이것은 질의 내에 비주제어가 있을 경우에 한정하여 사용할 수 있다.

본 논문에서는 자동 키워드망(AKN : Automatic Keyword Network)을 구성하여 문서순위를 재조정하는 방법을 제안하였다. 문서순위 결정을 위해 2단계 과정을 거친다. 1차로 검색을 위한 색인, 2차로 문서 순위 재조정을 위한 자동 키워드망을 구성한다. 1차로 검색된 문서들에 대하여, 질의어 내의 키워드와 1차로 검색된 문서내의 키워드들의 관계를, 자동 키워드망 내의 값들에 식을 적용하여 2차로 의미적인 관련성 정도에 따라 문서순위를 재조정하는 방법이다. 자동 키워드망을 이용하면 2차 문서순위의 재조정으로 정확성 향상은 물론 1차 검색에서의 어의 중의성 해석도 가능함을 보이고자 한다.

본 논문에서 사용된 실험 환경은 백과사전 개발을 위한 옥서[7]를 사용하여, (주)계몽사에서 발간한 크라운판 백과사전 여섯권(약 2,500쪽)을 전자도서화한 멀티미디어 CD-ROM 타이틀이다. 이 전자 백과사전은 약 23,000 표제어를 가지며 자연어 검색 등의 기능을 제공한다. 옥서를 이용하여 전 백과사전에 대해 AKN을 구성하고, 표제어뿐만 아니라 설명문에 대한 내용 검색이 가능하도록 하였다. 여기서 표제어란 사전의 entry로 백과사전에서 '가나다' 순으로 나열된 단어이다.

2장에서는 1차 검색을 위한 자연어 색인의 구축에 대하여 설명하고, 3장에서는 자동 키워드망 구축 방법을 설명한다. 4장에서는 문서내의 키워드 빈도를 계산하여 키워드 정보 파일을 구성하는 방법을 논한다. 5장은 검색 시스템으로서 1차 자연어 검색 및 2차 문서 순위의 결정 방법에 대해 다룬다. 6장에서는 실험 및 평가 결과를 분석하고, 마지막 7장에서는 결론과 추후 연구해야 할 분야에 대하여 기술하였다.

2. 1차 검색을 위한 자연어 색인

2.1 자동색인

색인이란 저장된 문서를 미리 분석하여 주요 단어 또는 어구를 추출한 후, 찾기 편리한 형태로 저장하는 것을 말하며, 검색시 문서 내용 전체를 검색하지 않고 주요 단어 즉 키워드만을 검색함으로써 빠른 시간에 사용자의 요구를 만족시킬 수 있게 한다. 그러나 방대한 정보에 대해 일일이 사람 손으로 색인어를 추출해야 한다면, 대단히 비경제적이고 색인어 선택시 일관성을 유지하기도 힘들기 때문에, 자동 색인하는 것이 일반적이다. 자동 색인 기법은 통계적 기법과 형태소 분석, 구문분석, 의미분석을 하는 언어학적인 기법으로 나눌 수 있다.

키워드의 빈도를 계산하는 통계적 방법은 한국어에 적용하기에는 어느 정도 한계를 지니고 있고, 구문 구조를 파악하는 구문분석이나, 문장을 완전히 이해하기 위한 의미분석은 구현하기 어렵다. 따라서 본 논문에서는 각 어절에서 형태소를 분리한 후, 조사, 보조사, 접미어 등을 분리하고, 키워드로 사용 가능한 명사들을 모두 추출한 후, 명사들을 색인어로 취하는 형태소 분석 방법을 택하였다.

색인 파일의 형식은 요약 파일 형식과 역색인 파일 형식이 있다. 본 논문에서는 부가적인 저장 공간을 많이 필요로 하지만 검색 성능이 좋은 역색인 파일 형식을 사용하였다. 역색인 파일의 색인 레코드는 기본적으로 검색어와 이 검색어를 포함하고 있는 문서 레코드의 번호들로 구성된다. 역색인 파일의 형식은 다음과 같다.

[역색인 파일 형식] : 키워드^문서id^가중치^문서id^가중치^....

2.2 가중치 설정

역색인 파일 구성 시 중요한 요소 중 하나는 가중치 설정 방법이다. 일반적으로 키워드와 문서간의 유사도에 따라 그 가중치를 설정한다. 본 논문에서는 일반적으로 가장 많이 사용하는 Salton이 제안한 문서 빈도에 의해 키워드의 가중치를 계산한다[5]. 다음은 Salton의 가중치 계산식을 개선한 것이다.

$$Weight_{ij} = (Freq_{ij}*k)*(\log(n)-\log(Docfreq_j)+1)$$

$Weight_{ij}$는 문서 i에서의 키워드 j의 가중치이다. $Freq_{ij}$는 문서 i에서의 키워드 j의 빈도수이고, n은 문서의 수이다. $Docfreq_j$는 키워드 j가 출현한 문서의 수이다. 가중치는 기본적으로 빈도에 관계가 있다. 여기서 k를 낮은 값으로 곱해줌으로 단순 빈도를 약간 인정하는 값으로 사용한다. 그 뒤에 곱해지는 값은 전통적으로 정규화된 역문서 빈도의 값이다. 그리고 사전의 엔트리의 가중치를 다른 키워드들에의 가중치보다 높게 부여하여 키워드 색인을 한다[7].

1차 검색을 위한 자연어 색인 구축 예를 들면 (그림 1)과 같다. 우선 각 문서에 자동으로 키워드 마크를 한다. 다음 그 문서에서 키워드를 추출해 내고, 키워드를 가,나,다순으로 소트한다. 가중치 계산을 한 후 역색인 화일로 구성하는 것이다.

step.1 : 키워드 마크

! 105인 사건
1912년 \일제\가 \데라우치\\총독\\암살\\음모\를 \구실\로 \신민회\\회원\을 \체포\하여 \고문\한 \사건\. 1910년에 \평북\ \선천\에서 \안명근\에 의한 \총독\ \암살\ \미수\\사건\이 일어나자, \일제\는 이를 \계기\로 \애국\\지사\들을 \체포\할 것을 \계획\하고, \신민회\가 \총독\\암살\을 \준비\하고 있다는 \구실\을 붙여 그 \회원\ 등 600여 \명\을 \검거\하였다. \일제\는 이들을 \고문\한 \끝\에 그 \대표자\ 105 \명\을 \재판\에 붙여 \투옥\하였는데, 이 \사건\으로 \신민회\는 많은 \타격\을 받아 \자연\히 \해체\되고 말았다.

step.2 : 문서내의 키워드 추출

105인 사건	1	15(제목가중치)
일제	1	3
데라우치	1	1
총독	1	3
암살	1	3
음모	1	1

step.3 : 소트

가스가열기구	258	15	가스관	20676	3
가스가열기구	263	1	가스관이음	20676	1
가스가열기구	264	1	가스광업	2083	1
가스검지기	259	15	가스괴저	1491	1
가스계량기	260	15	가스교환	18529	1
가스공급	133281	1	가스교환	22209	1

(그림 1) 자연어 색인의 구축 예(계속 됨)

step.4 : 역색인 파일로 구성

가스가열기구	3	264	48734	263	48734	258	250000	
가스검지기	1	259	250000					
가스계량기	1	260	250000					
가스공급	1	133281	53560					
가스관	2	20676	50496	20672	50496			
가스관이음	1	20676	53506					
가스광업	1	2083	53506					
가스괴저	1	1491	53506					
가스교환	4	3105	47485	22211	47485	22209	47485	8529 47485

step.5 : 역색적인 파일의 인덱스 구성

가스가열기구	59737
가스검지기	59748
가스계량기	59809
가스공급	59834
가스관	59858
가스관이음	59892
가스광업	59918
가스괴저	59941
가스교환	59964
가스기관	60023

(그림 1) 자연어 색인의 구축 예

2.2.1 자동 키워드 마크

간단한 형태소 분석을 통하여 키워드를 마크한다. 일반적인 형태소 분석은 문장에서 어절 단위로 추출한 후 그들의 기본형과 특성을 추출하는 것이다. 그러나 정보 검색 시스템에서는 키워드에 해당하는 명사에 대해서만 추출한다. 추출방법은 조사 최장일치법과 조사를 떼어낸 나머지가 명사사전에 있는지를 확인한다. 명사가 있으면 마크를 하고 없으면 조사를 줄여서 즉, 명사를 확장해서 다시 명사사전과 검사한다.

2.2.2 문서내의 키워드 추출과 소트

- 각 문서에서 마크된 키워드를 문서 번호와 빈도로 나타낸다. 여기서 제목에 대해서는 15의 빈도를 준다. 즉 제목 가중치를 준다.
- 각 문서에서 추출된 키워드들을 가,나,다 순으로 소트한다.

2.2.3 역색인 화일의 구축

먼저 가중치 계산을 한 후 역색인 화일로 구축한다.

2.2.4 역색인 화일의 인덱스 구축

역색인 화일의 인덱스는 키워드와 그 키워드의 역색인 화일에서의 주소로 구축한다.

3. 2차 검색을 위한 자동 키워드망

3.1 자동 키워드망의 정의와 구축

자동 키워드망이란 문서의 집합에서 용어들간의 관련도를 구하는 방법으로 백과사전 키워드 관계에 대한 구조에 따라 임의로 가중치를 주어 형성한 것을 말한다.

- 키워드의 연관성을 측정하기 위하여 백과 사전을 입력 자료로 선정하였으며, 다음 특수기호들을 사용하여 키워드들의 문장에서의 역할을 구분하였다. 하나의 사전의 엔트리와 그 내용들을 하나의 문서라 한다.

  <id> 사전 엔트리의 고유번호로 문서 번호라 한다.
  ! 사전의 엔트리 (백과사전의 표제어로 약 23,000여개)
  # 엔트리가 2개 이상(동음이의어 : 배[1] ,배[2])
  @ 해설옮김(동의어 : 가극, 오페라)
  $ 참조어(관련어 : 가계, 가정경제)
  \ 의미있는 키워드로 앞뒤에 '\' 기호를 붙였다.

- 연관성은 키워드를 중심으로 이루어 지는데 백과사전의 경우 표제어와 해설옮김은 동의어의 의미임으로 가중치 값을 100으로 주어 연관을 시키고 또, 백과사전의 경우 표제어에 대한 첫 문장의 경우 설명을 대표하는 문장이므로 가중치 값을 30으로 주어 연관을 시킨다. 그 외 표제어와 나머지 문장의 설명 단어에 대한 연관과 표제어의 참조어와 설명 단어사이의 연관을 시킨다. 가중치 값에 대한 연관 내용은 <표 1>과 같다.

〈표 1〉 키워드들 간의 가중치 값

문장내에서의 관계	가중치값
사전의 엔트리와 해설 옮김	100
사전의 엔트리와 참조어	50
사전의 엔트리와 첫 문장내의 키워드들	30
사전의 엔트리와 나머지 문장내의 설명 키워드들	10
참조어와 설명 키워드들	5
Window size 내의 이웃한 키워드들	10

추출된 키워드 쌍은 어떤 키워드 A와 B가 서로 연관이 있다면 B도 A와 같은 표1-1의 가중치를 가지는 대칭성을 인정하여 대칭파일을 만든다. 키워드 A와 B가 다른 문서에서로 나타남으로 같은 키워드 쌍이 나오면 가중치를 합하여 중복성을 제거한다. 위의 가중

치 값에다 여러 문서에서 나타난 키워드 쌍들에 대해서 중복성을 제거하면 AKN값이 된다. 검색 대상 문서들에서 나타나는 키워드가 n개 존재한다면 N*N 행렬이 가능하다. 그러나 대부분의 행렬 값이 0이 되는 Sparse Matrix가 되므로 마지막 결과는 각 키워드에 대한 타 키워드들의 리스트 형식을 갖게 한다.

본 논문에서 윈도우 사이즈를 5로 채택한다. 이유는 키워드간의 의미를 한정하는 데에 적당하기 때문이다 [1]. 윈도우 사이즈란 키워드 x와 y가 몇 개의 키워드 거리 이내에 있을 때 서로 영향을 미치는가 하는 문제인데, 실제 문장에서는 x와 y는 거리보다는 문장의 구조에 영향을 더 많이 받는다. 그러나 파싱해야 하는 어려움이 따르기 때문에 키워드간의 거리로 처리한다. 한 키워드가 가장 영향을 많이 받는 키워드는 바로 앞뒤의 키워드이지만, 그러나 두 번째 혹은 세 번째 키워드도 영향을 미치므로 그것을 배제 할 수 없다. 또한 너무 멀리 떨어진 키워드간의 상관관계를 따지는 것도 무의미하기 때문에 윈도우 사이즈 5~10 정도의 키워드들이 가장 밀접한 관계를 가진다고 보여진다.

3.2 구축 예

자동 키워드망의 구축을 예로 들어 설명하면 (그림 2)와 같다.

3.2.1 자동 키워드쌍 추출

1. 제목추출
2. @ (해설옮김)나 $(참조어) 추출
   - 2.1 @(해설옮김)이면 제목과 그 단어의 값이 100을 갖는 쌍을 만든다. 역 쌍의 구성
   - 2.2 $(참조어) 이면 제목과 참조어의 값이 50을 갖는 쌍을 만든다. 역 쌍의 구성
3. 모든 마크된 단어에 대해
   - 3.1 첫 문장내의 마크된 단어에 대해 제목과 30, $(참조어)와 5의 값을 갖는 쌍을 만든다. 역쌍 구성
   - 3.2 나머지 문장에 대한 마크된 단어에 대해 제목과 10, $(참조어)와 5의 값을 갖는 쌍을 만든다. 역쌍 구성
4. 한 문장의 끝까지 window size내의 단어쌍에 대해 10을갖는 쌍을 만든다. 역쌍 구성.

3.2.2 추출된 키워드쌍들을 소트한다.

! \가가라강\

\인도\ \갠지스강\의 큰 \지류\의 하나. \길이\는 약 960km이다. \티베트\ \고원\ 마나사로와르 \호수\ \남쪽\에서 \시작\되어 \네팔\ \서부\를 \남쪽\으로 흘러, 힌두스탄 \평원\의 \중앙부\에 있는 \도시인\ \파트\나 \근방\에 \이르러\ \갠지스강\과 합쳐진다. 힌두스탄 \평원\의 \서부\ \지역\을 \차지\하는 \우타르프라데시주\의 \상업상\ \중요\한 \내륙\ \수로\가 되며, \유역\의 \물대기\에 큰 \몫\을 한다.

키워드 쌍의 추출

가가라강\ 인도\ 30	인도\ 가가라강\ 30
가가라강\ 갠지스강\ 30	갠지스강\ 가가라강\ 30
가가라강\ 지류\ 30	지류\ 가가라강\ 30
가가라강\ 길이\ 20	길이\ 가가라강\ 30
가가라강\ 티베트\ 20	티베트\ 가가라강\ 30
인도\ 갠지스강\ 10	갠지스강\ 인도\ 10
인도\ 지류\ 10	지류\ 인도\ 10
갠지스강\ 지류\ 10	지류\ 갠지스강\ 10

소트

가가라강\ 갠지스강\ 20
가가라강\ 갠지스강\ 30
가가라강\ 고원\ 20
가가라강\ 근방\ 20
가가라강\ 길이\ 20
가가라강\ 남쪽\ 20
가가라강\ 내륙\ 20
가가라강\ 네팔\ 20

자동 키워드망의 구축

가가라강 27 갠지스강50 남쪽40 서부40 평원40 지류30
인도30 네팔20 도시인20 몫20 물대기20 상업상20 근방20
수로20 시작20 우타르프라데시주20 유역20 이르러20
길이20 중앙부20 중요20 고원20 지역20 차지20 티베트20
파트20 내륙20 호수20

갠지스강 87 유역170 인도130 강90 북부60 마가다왕국50
가가라강50 지방40 …….

(그림 2) 자동 키워드망의 구축 예

3.2.3 자동 키워드망구축

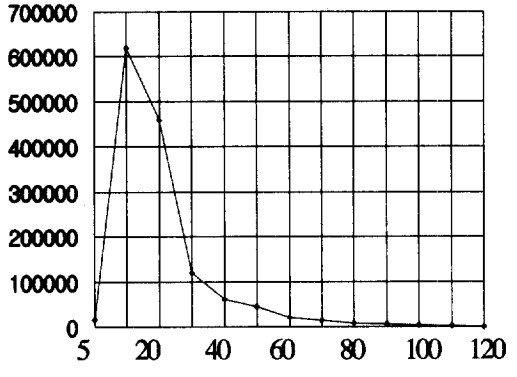
검색 대상 문서들에서 나타나는 키워드가 n개 존재한다면 N*N행렬이 가능하다. 그러나 대부분의 행렬값이 0이 되는 Sparse Matrix가 되므로 마지막 결과는 각 키워드에 대한 타 키워드들의 리스트 형식을 갖게 한다. 같은 쌍이 나오면 가중치를 합하여 중복성을 제거한다.

3.3 결과 및 분포도

백과사전에 대한 자동 키워드망의 크기는 약 20M 바이트이며, 가장 많이 관련된 키워드는 '위'로 관련된 키워드의 갯수는 9285개이다. 가장 높은 자동 키워드망 값을 가진 키워드 쌍은 "우리, 나라"로 31760의 값을 가진다.

다음은 키워드 쌍들의 값에 대한 분포도이다. x축은 자동 키워드망의 값이고, y축은 자동 키워드망의 값에 대한 키워드 쌍의 갯수이다.

(그림 3)에서 보면 키워드 쌍들의 값이 평균 20도 안 된다. 이것은 거의 대부분의 값이 윈도사이즈 내의 값 즉 10으로 구성되기 때문이다.



(그림 3) 자동 키워드망에서 키워드 쌍둘의 값에 의한 분포도

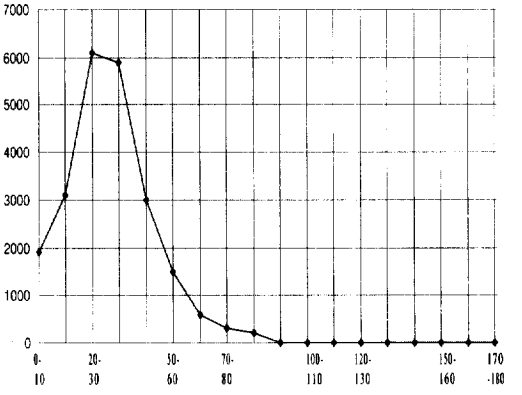
4. 문서 내의 키워드 정보의 구성

자동 키워드망 이외에 백과사전 내의 문서에서 키워드들을 검색하기 위해 표제어별로 문서번호를 부여하고 하나의 문서내의 키워드의 개수 그리고 실제의 각 키워드의 문서내의 빈도를 알 수 있는 문서의 키워드 정보 파일을 구성한다. 다음은 키워드 정보 파일의 형식이다.

[키워드 정보 파일의 형식]

문서 번호^전체 키워드 수^중복을 제거한 키워드 수^사전 엔트리^빈도^문서내의 키워드1^빈도^,문서내의 키워드2^빈도...

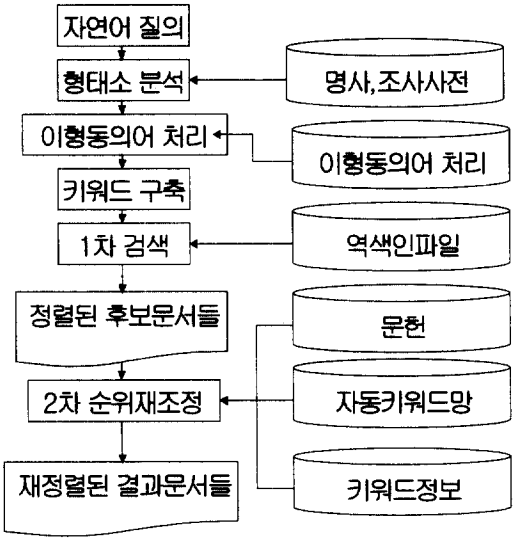
하나의 문서에서 마크된 키워드를 뽑아서 빈도순으로 소트(sort)한 후 키워드 정보 일의 형식으로 구성한다. 백과사전의 경우 약 4M 바이트 크기가 된다. 가장 크기가 큰 것은 "조선"으로 953개의 키워드를 가진다. (그림 4)는 문서내의 키워드 수에 대한 분포도이다. x축은 문서내의 키워드 수이고, y축은 키워드 수에 따른 문서의 수이다.



(그림 4) 문서내의 키워드 수에 대한 분포도

5. 검색 시스템

전체 시스템 구성을 간략히 나타내면 (그림 5)와 같다. 질의어에 대한 형태소 분석, 이형동의어 등의 처리를 하여 키워드를 추출한다. 검색된 키워드로 1차 검색 한 후, 2차로 문서 순위 결정식들을 적용하여 문서들의 순위를 재조정한다.



(그림 5) 검색 시스템 구성도

5.1 1차 자연어 검색

1차 검색은 역색인 파일 검색 기법을 사용한다. 자연어 질의에서 추출한 키워드가 나타나는 문서들을 역색인 파일을 검색하여 찾아서 가중치를 계산한다. 찾아낸 후보 문서들을 가중치순으로 특정 개수만큼 추출한다.

5.1.1 형태소 분석

자연어 색인에서의 키워드 마크와 동일한 과정으로 질의 내에서 키워드를 추출해 낸다.

5.1.2 이형 동의어 처리

이형 동의어란 하나의 뜻을 나타내는 다른 형태의 키워드들이 존재하는 말로 표준 형태로 바꾸어 준 후 검색하여야 한다.

예) 아메리카(미국), 크레디트카드(신용카드) 등

5.2 2차 문서 순위의 결정 방법

2차 문서 순위 재조정은 자동 키워드망을 이용하여 질의어 내의 키워드들과 1차 검색에서 추출된 문서내의 키워드들간의 자동 키워드망 값을 구하고 여러 식들을 적용하여 검색의 효율을 높이는데 그 목적이 있다. 다음은 식이 나오게 된 이유이다.

1) 자동 키워드망의 구성시 높은 빈도의 키워드는 상대적으로 많은 키워드들과 쌍을 이룰것이고 그 값 또한 커질 것이다. 단지 연관된 키워드의 수가 많다는 이유만으로도 문서 순위의 재조정시 높은 빈도의 키워드가 많은 문서들이 상위로 올라오게 된다(식 (1)).
2) 문서내의 키워드가 많은 즉, 상대적으로 큰 문서에서는 키워드 쌍들이 많아서 당연히 높은 순위를 차지하게 된다.
3) 1, 2를 해결하기 위해 자동 키워드망의 정규화와 문서의 정규화를 한다.
4) 절대값을 이용한 정규화로 식 (2)와 식 (3)이 도출되었고 식 (2)는 자동 키워드망 값에 키워드의 빈도로 나눈 식이고 식 (3)은 문서의 키워드 수로 나눈 식이다.
5) 상대값을 이용한 정규화로 식 (4)와 식 (5)가 도출되었고 식 (4)는 자동 키워드망에서 키워드의 빈도가 가장 큰 값과 해당 키워드의 빈도로 나눈 값을 자동 키워드 망값에 더한 식이고, 식 (5)는 가장 많은 키워드를 갖는 문서의 총 키워드 수에 해당 문서의 키워드 수로 나눈 값을 자동 키워드 망값에 더한 식이다.
6) 그리고 식 (4)와 식 (5)에다 로그(log)를 취한 정규화식이 식 (7)과 식 (8)이다.
7) 자동 키워드망의 정규화와 문서의 정규화를 합하는 식이 식 (6), 식 (9)이다.

다음은 식에서 쓰이는 변수에 대한 설명이다.

k는 질의어 내의 키워드 수이고,
l은 문서 내의 키워드 수이다.
$AKNV_{ij}$는 질의어 내의 키워드 i와 문서의 내의 키워드 j의 AKN의 값이다.
AKV_i는 질의어 내의 키워드 i의 AKN에서 발생 빈도이다.
DK_j는 문서 번호 j가 가지고 있는 키워드 수이다.
$MAXAKV$는 AKN에서의 최대 발생 빈도이다.
$MAXDK$는 문서들 중 최대 키워드 빈도이다.

AKV_i의 예는 (그림 2)의 경우 '갠지스강'은 총 50개의 키워드로 구성되어 있어 AKV(갠지스강)은 50이 된다.

DK_j의 예로서 DK(조선)은 '조선'이라는 엔트리의 키워드 수인 953이 된다.

$MAXAKV$의 예는 백과사전에서 '우리, 나라'로 31,760이다.

$MAXDK$의 예는 백과사전에서 '조선'이라는 엔트리의 키워드 수인 953이 된다.

식 (1) 질의어의 키워드와 문서내의 키워드가 자동 키워드망에 형성되어 있으면 이들을 모두 합한 값을 질의와 문서 사이의 유사도 값으로 사용하는 것이다. 이 유사도 값에 따라 순서를 재정렬한다.

$$SIM(Q, D) = \sum_{i=1}^{k} \sum_{j=1}^{l} AKNV_{i,j}$$

식 (2) 질의어 키워드의 최대 발생 빈도 값으로 자동 키워드망의 값을 나눈 값을 모두 합한다. 키워드의 절대 값에 의한 정규화(normalization)의 의미를 갖는다.

$$SIM(Q, D) = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{AKNV_{i,j}}{AKV_i}$$

식 (3) 문서의 키워드 수로 자동 키워드망의 값을 나눈값을 모두 합한다. 이의 의미는 문서의 절대 크기 값에 의한 정규화이다.

$$SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{AKNV_{i,j}}{DK_j}$$

식 (4) 자동 키워드망의 값과 해당 질의어 키워드에 대한 상대 키워드 값에 의한 정규화값(즉, 키워드의 최대수를 해당 질의어 키워드로 나눈 값)을 곱한것을 모두 합한 것이다.

$$SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l}AKNV_{i,j}\times\left(\frac{MAXAKV}{AKV_i}\right)$$

식 (5) 자동 키워드망의 값과 해당 문서의 상대 정규화 값(즉, 키워드를 최대로 갖는 문서의 키워드 수를 해당 문서의 키워드의 수로 나눈 값)을 곱한 값을 모두 합한다.

$$SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l}AKNV_{i,j}\times\left(\frac{MAXDK}{DK_j}\right)$$

식 (6), 식 (4)와 식 (5)을 모두 적용하면 다음과 같은 식을 유도할 수 있다.

$$SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l}AKNV_{i,j}\times\left(\frac{MAXAKV}{AKV_i}\right)\times\left(\frac{MAXAKV}{DK_j}\right)$$

7번 이하 식들은 각 값의 로그(log)를 취한 값(즉, 일정한 구간 값으로 취한 값)을 적절히 혼합하여 질의어와 문서 간의 유사도를 계산한다. 여기에서 C를 더해주는 것은 곱해지는 값이 0이 되는 것을 방지하기 위한 값으로, 본 실험에서는 C값을 0.5로 두어서 사용한다.

식 (7) 식 (4)에 로그를 취한다.

$$SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l}AKNV_{i,j}\times(\log_2 MAXAKV-\log_2 AKV_i+C)$$

식 (8), 식 (5)에 로그를 취한다.

$$SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l}AKNV_{i,j}\times(\log_2 MAXDK-\log_2 DK_j+C)$$

식 (9), 식 (6)에 로그를 취한다.

$$SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l}AKNV_{i,j}\times(\log_2 MAXAKV-\log_2 AKV_i+C)\times(\log_2 MAXDK-\log_2 DK_j+C)$$

6. 실험 및 평가

본 논문에서 계몽사 학생 대백과 사전에서의 텍스트 자료 10M정도 크기의 데이타에 약 23,000여개의 엔트리로 구성된 자료를 가지고 약 10만여개의 키워드를 추출하여 자동으로 1차 자연어 검색을 위한 색인파일과 2차 문서 순위 조정을 위한 자동 키워드망을 구축하고, 아래와 같은 평가방법을 적용하여 평가하였다.

6.1 정확률에 대한 평가

- 질의의 키워드와 문서내의 키워드가 같을 경우에 어떤 값을 취해야 좋은지를 판정하는 기준으로 2가지 망의 분포의 최소(MIN)값은 0, 최대(MAX)값은 100, 중간값(MID)은 20을 주어 검색한다.

질의 갯수 46개의 적합성 정보를 이용한다. 질의어는 9명의 일반인들이 백과사전에 있는 내용에 대해 알고자 하는 사항을 5개씩 질의하게 하였다. 적합성 정보는 4명의 전문자로부터 자연어 질의에 적합한 백과사전 표제어 함목 중 2명 이상이 일치한 백과사전 항목 정보를 사용한다. 평가 방법으로는 정확률과 재현률의 식을 이용하였다.

정확률 = 검색된 적합 문서수 / 검색된 문서총수
재현률 = 검색된 적합 문서수 / 적합 문서총수

다음은 46개 자연어 질의 중 어의중의성을 가진 6개 질의이다

1. 우리나라의 꽃(국화)이름은
2. 지구의 자전과 공전에 대하여 설명해 주십시오.
3. 팔만대장경의 자수가 50만쯤 입니까?
4. 아침에 풀잎에 맺힌 물방울은 이슬인가, 아니면 서리인가?
5. 당나귀와 말의 차이점은?
6. 눈이 나쁜사람은 왜 안경을 써야 하는지 알고 싶다.

다음은 위의 6개 질의에 대한 적합성 정보이다.

1. 국화1, 무궁화

〈표 2〉 2차 검색 결과와 1차 검색과의 비교표(1차검색 = 0.0%)

재현율시점	0.2	0.25	0.3	0.4	0.45	0.5	0.55	0.6	0.65	평균 증가치
1차검색	0.71	0.64	0.58	0.49	0.47	0.43	0.39	0.38	0.35	0.0(%)
02-MIN	0.82	0.77	0.71	0.62	0.59	0.55	0.49	0.43	0.39	10.8
04-MIN	0.82	0.76	0.70	0.62	0.59	0.55	0.49	0.44	0.39	10.5
09-MIN	0.71	0.64	0.60	0.56	0.51	0.49	0.45	0.41	0.36	3.5
02-MID	0.82	0.76	0.71	0.62	0.59	0.56	0.49	0.43	0.39	10.8
04-MID	0.82	0.76	0.71	0.62	0.59	0.56	0.49	0.44	0.39	10.9
09-MID	0.71	0.6	0.60	0.56	0.51	0.49	0.45	0.41	0.36	3.5
02-MAX	0.79	0.75	0.71	0.60	0.58	0.56	0.49	0.44	0.40	10.2
04-MAX	0.79	0.75	0.71	0.60	0.58	0.56	0.49	0.43	0.40	10.1
09-MAX	0.72	0.66	0.61	0.55	0.52	0.49	0.45	0.42	0.36	4.1

2. 공전주기, 자전2, 공정3, 지구1, 광행차, 일주 운동, 푸코진자의 실험
3. 고려대장경, 대장경, 팔만대장경, 해인사 대장경판
4. 이슬, 서리1, 이슬점
5. 노새, 말1, 당나귀, 나귀, 간생
6. 굴절이상, 난시, 노안, 눈1, 안과, 원시, 근시, 안경, 보안경

질의내의 키워드가 사전의 엔트리와 동일할 경우에는 1차 검색에서 사전의 엔트리 가중치를 가지고 들어가서 상위로 올라오게 된다. 2차 문서 순위 재조정에서는 망 자체가 구조 정보 즉 엔트리와 문서내의 키워드간의 값을 갖고 있어서 질의내의 키워드가 엔트리이면 많은 향상이 있게 된다.

<표 2>는 자동 키워드망을 이용한 2차 검색 결과중 식 (2), 식 (4), 식 (9)에 대해서 질의의 키워드와 문서내의 키워드가 같을 경우에 최소(MIN)값은 0, 최대(MAX)값은 100, 중간값(MID)은 20을 주어 판정한 결과표이다. 여기서 증가치란 각각의 재현율 포인터에서 정확률 차이값을 더하고, 그 값을 재현율 포인터수로 나눈 평균 증가치이다.

- 식 (4)에서 1차 검색 보다 최고 약 10.9%의 향상을 보였다. 이것은 키워드망의 값에 의해 가중치를 조절함으로써 순위 재조정이 보다 의미적으로 가까운 문서가 상위로 올라온다는 것을 보여준다. 즉 내용에 민감(CONTEXT SENSITIVE)하게 작용 한다는 것이다. 1차 검색 결과 문서 내에서의 키워드 수로 정규화하기 보다는 자동 키워드망의 발생 빈도 수에 의해 정규화하는 것이 더 좋은 결과를 보임을 알 수 있고 이것은 곧 2차 검색시 자동 키워드망에 의존한다는 것을 알 수 있다.
- 질의의 키워드와 문서내의 키워드가 같을 경우에 취하는 최소(MIN), 최대(MAX), 중간값(MID)에 대해서는 크게 신경을 쓰지 않아도 된다. 단지 최대값을 주었을 경우는 한번 더 엔트리 가중치를 부가하는 것과 같은 효과로 최대값을 부여한 가중치로 인한 잘못된 효과(side effect)를 가져올 수 있다. 왜냐하면 동음이의어들에 대한 엔트리들이 상위를 차지하기 때문이다.
- 자동 키워드망은 키워드 쌍들에 대해서는 어느 정도 속성이 부여되어 만들어져 있어서 정규화를 하면 할수록 속성을 상대적으로 잃어버리게 되어 식 (9)에서는 많이 향상되지는 않는다.

6.2 어의 중의성 해소측면의 평가결과

1차 자연어 검색의 문제점중의 하나는 엔트리 가중치 조절로 동음이의어인 "가장, 국화, 자전, 공전, 다리, 자수" 등의 엔트리들이 상위를 차지하는 것이다[7]. 어의 중의성 해소측면에서는 대체적으로 모든 식이 어의 중의성을 해소한다. 즉 1차 검색에서의 어의 중의성은 2차 검색에서 문서 순위가 재조정됨으로써 가려지게 된다. 다음은 각 식별 어의중의성 해소의 비교를 보여준다.

〈표 3〉 자동 키워드망을 이용한 식별 어의 중의성해소 비교표(1차검색 = 0/10)

식	1	2	3	4	5	6	7	8	9
MAX	7	9	3	9	3	5	8	7	8
MID	7	9	3	9	3	5	8	7	8
MIN	7	9	3	9	3	5	8	7	8

- 적합성 정보를 가진 질의 46개 중에 동음이의어를 가진 질의어 10개에 대해 실험한 결과, 식 (2), 식

(4)번에서 9개의 어의중의성이 해소됨을 보였다.

- 질의내의 키워드와 문서내의 키워드가 같을 경우의 MAX, MIN, MID값을 취하는데 이것은 어의 중의성 해소와는 관계없음을 알았다.

다음의 예는 1차 검색에서는 자전과 공전이 동음이의어로 인한 문제를 가지고 있으나, 자동 키워드망을 이용한 식 (4)에서의 문서순위 재조정에서 어의중의성을 해결하고 있음을 보여준다.

예) 질문 : 지구의 자전과 공전에대해 설명해 주십시오.

1) 공전[3] (1차 검색 후 2위 -> 2차 검색 후 1위)
   => 지구의 공전
2) 자전[2] (1차 검색 후 1위 -> 2차 검색 후 2위)
   => 지구의 자전
3) 중력 (1차 검색 후 8위 -> 2차 검색 후 3위)
4) 일주운동(1차 검색 후 10위 -> 2차 검색 후 4위)
5) 푸코 진자의 실험
   (1차 검색 후 7위 -> 2차 검색 후 5위)
6) 자전[1] (1차 검색 후 6위 -> 2차 검색 후 6위)
   => 옥편이라고도 한다.
7) 공전[2] (1차 검색 후 3위 -> 2차 검색 후 7위)
   => 관청이 가지고 있는 토지
8) 공전[1] (1차 검색 후 4위 -> 2차 검색 후 8위)
   => 조선시대 육전의 하나.
9) 지구의 (1차 검색 후 5위 -> 2차 검색 후 9위)

7. 결 론

본 논문에서는 1차로 역색인 파일을 기반으로 한 문서 순서화, 2차로 자동 키워드망을 이용한 문서 순위의 재조정을 시도함으로써, 정확성 향상은 물론 역색인 파일만으로는 해결할 수 없는 어의 중의성을 해소하였다.

질의어의 키워드 수, 문서내의 키워드 수, *AKNV, AKV, DK, MAXAKV, MAXDK* 등을 이용하여 여러 식을 만든 다음, 어느 식이 가장 검색 효율이 높은지 실험해 보았다. 백과사전으로 실험한 결과, 4번 식이 MID 값에서 약 10.9%의 향상을 보이는 양호한 결과를 얻었다. 이것은 키워드망의 값에 의해 가중치를 조절하면, 보다 의미적으로 가까운 문서가 순위 재조정에서 상위로 올라온다는 것을 보여준다. 다시 말하면 1차 검색 결과 문서 내에서의 키워드 수로 정규화하기보다는 자동 키워드망의 발생 빈도 수에 의해 정규화하는 것이 더 좋은 결과를 보임을 알 수 있다.

어의 중의성 해소측면에서는 대체적으로 모든 식이 어의 중의성을 해소함을 알 수있다. 즉 1차 검색에서의 어의중의성은 2차 검색에서 문서 순위가 재조정됨으로써 가려지게 된다. 어의중의성 해소에 있어서는 적합성 정보를 가진 질의 46개 중에 동음이의어를 가진 질의어 10개에 대해 실험한 결과, 식 (2), 식 (4)에서 9개의 어의 중의성을 해소할 수 있었다.

본 논문에서 제안한 자동 키워드망을 이용한 2차 문서 순위 결정 방법은 일반 순서화 된 문서의 순위 재조정이나 질의 확장, 적합성 피드백 다음에, 최종적으로 문서 순위를 재조정 할 수 있는 방법이다. 왜냐하면 재현율도 중요하기 때문에, 재현율을 유지하면서 정확성을 높일 수 있기 때문이다.

앞으로의 연구는 구조정보 없는 일반적인 상호정보에 대한 연구와, 백과사전이 아닌 일반 문서들에 대한 적용, 그리고 문서들의 삽입이나 변경이 빈번한 시스템에 대한 동적 색인 관리 및 상호정보망, 키워드망 관리에 대한 연구가 필요할 것이다.

참 고 문 헌

[1] K. W. Church and P. Hanks, Word Association Norms, Mutual Information, and Lexicography, Computational Linguistucs, Vol.16, No.1, pp. 22-29, 1990.

[2] D. Harman, Ranking Algorithms, in Information Retrieval : Data Structure and Algorithms, W.B. Frakes and R. Baeza-Yates , Prentice-Hall, Englewood Cliffs, NJ, pp.363-392, 1992.

[3] D. Harman and G. Candela, Retrieving Records from a Gigabyte of Text on a Minicomputer using Statistical Ranking, Journal of the American Society for Information Science, Vol.41, No.8, pp.581-589, 1990.

[4] D. M. Magerman and M. P. Marcus, Parsing a Natural Language Using Mutual Information Statistics, National Conference on Artificial Intelligence (AAAI-90), pp.984-989, 1990.

[5] G. Salton, Automatic Text Processing : The Trans-

formation, Analysis, and Retrieval of Information by Computer, Addition-Wesley Publishing Company, 1989.

[6] G. Salton and C. Buukley, "Improving Retrieval Performance by Relevance Feedback," Joural of the American Society for Information Science, Vol.41, No.4, pp288-297, 1990.

[7] 강현규, 옥서의 자연어 검색 성능 분석 및 개선, 한국정보처리학회 춘계 학술발표논문집, 제22권 제1호, pp.56-59, 1995.

[8] 강현규, 박세영, 최기선, 자연언어 정보 검색에서 상호정보를 이용한 2단계 문서 순위 결정방법, 한국정보과학회 논문지. 제23권 제8호, pp.852-861, 1996.

[9] 김대진, 정상철, 신동욱, "시소러스를 기반으로 하는 문서순위 결정 방법에 관한 연구", 한국정보과학회 봄 학술발표논문집 제21권 제1호, pp.177-180, 1994.

[10] 이승률, 강현규, 박세영, 이상조, "자연어 질의 정보검색 시스템의 비주제어 탐색방법을 통한 성능 개선", 제6회 한글 및 한국어 정보처리 학술발표논문집, pp.374-377, 1994.

[11] 이준호, 시소러스의 연관성 정보를 이용한 문서의 순위 결정 방법, 한국정보처리학회지, 제10권 제2호, 1993.

장 덕 성

e-mail : dsjang@kmu.ac.kr

1979년 경북대학교 컴퓨터공학과 졸업(학사)

1981년 서울대학교 전산과학과 (이학석사)

1988년 서울대학교 컴퓨터공학과 (공학박사)

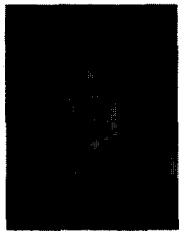
1982년~1985년 동아대학교 전산공학과 조교수

1985년~현재 계명대학교 컴퓨터공학과 교수

1992년~1993년 University of Colorado 방문연구교수

1998년~현재 계명대학교 기획정보처 전산원장

관심분야 : 컴파일러, 시각프로그래밍, 자연어처리, 정보검색, 에이전트 등



김 정 세

e-mail : jungskim@etri.re.kr

1994년 계명대학교 컴퓨터공학과 졸업(학사)

1996년 계명대학교 컴퓨터공학과 (공학석사)

1996년~2000년 한국정보시스템 근무

2000년~현재 한국전자통신연구원 음성언어팀 근무

관심분야 : 자연어처리, 정보검색, 음성이해 등