자동 키워드망과 2단계 문서 순위 결정에 의한 자연어 정보검색 모델

강현규*, 박세영*, 최기선**

*ETRI 자연어처리 연구실, **한국과학기술원 전산학과

A Natural Language Information Retrieval Model using Automatic Keyword Network and Two-level Document Ranking

Hyun-Kyu Kang*, Se-Young Park*, Key-Sun Choi**

*Natural Language Processing Section/ETRI, **Dept. of Computer Science/KAIST,

요 약

본 논문은 정보검색에서 사용자에게 순서화된 문서를 제시하기 이전에 1차로 검색된 문서들에 대하여 자동 키워드망과 2단계로 문서 순위 결정하는 모델에 대하여 논하였다. 자연어 검색을 위한 색인은 자동으로 구축된 키워드 색인으로 1차로 자연어 검색을 하고, 2차로 자동 키워드망을 이용한 순위 재조정을 통해 검색효율의 향상에 관해 검색 효율을 평가하여 1차 검색 결과보다 최대 10.9%의 검색효율 향상을 보였다. 또한 문서 순위 조정 방법에 있어서 여러 가지 공식을 비교 분석하였으며 내용 검색을 반영하는 공식을 찾았다. 본 논문에서 제시한 2단계 순위 결정 방법은 텍스트를 기반으로하는 정보 검색의 분야에 적용되어 검색효율을 높일 수 있는 한가지 방법이 될 수 있을 것이다.

1. 서론

현대 정보화 사회에서 정보는 인간의 관리가 불가능 할 정도로 많이 쏟아져 나오고 있다. 따라서 이러한 수많은 정보들을 컴퓨터에 저장하고 이를 관리하여 필요에 따라 사용자에게 서비스하는 정보 검색 시스템(Information Retrieval System)이 널리 이용되고 있다. 정보 검색 이란 정보 탐색자가 원하는 요구에 가장 적합한 정보를 검색하는 것이다. 그러나 일반적으로 방대한 문서의 집합에서 정보 요구자가 원하는 문서를 추출하기란 쉽지 않다[1,2]. 또한 정보검색 시스템의 중요한 목적중의 하나는 단순히 사용자 질의를 만족하는 문서들의 집합을 검색하는 것이 아니라, 질의를 만족하는 정도에 따라 검색된 문서들에 순위를 부여함으로써 사용자들이 필요한 정보를 얻는데 소모되는 시간을 최소화하는 것이다[3].

정보검색에서 일반적으로는 질의의 키워드들이 문서에 대하여 어느 정도의 중요도를 가지고 존재하느냐를 기준으로 문서를 순서화 한다. 그러나 실제 순서화된 문서들을 보면 질의한 내용과는 다른 문맥의 문서들이 상위로 순서화 되는 경우를 볼 수 있다. 이는 비록 질의 키워드들이 해당 문서에서 보다 중요도를 갖지만 문서에서 다루고 있는 문맥과 반드시 일치하지 않기 때문이다.

또한 자연어 질의시에 그 속에 숨겨진 의미를 찾아내어 해당 문서을 검색하기는 힘들다. 따라서 정보 검색의 효율을 높이기 위하여 지식에 바탕을 둔 지적 정보 검색 시스템들이 등장하고 있다. 이러한 시스템들은 그 지식구조로서 실세계의 개념과 그 관계를 표현하는 지식베이스를 이용하고 있다. 이런 지식베이스의 종류로는 의미망, 시소러스, 상호 정보등이 있으며, 이를 이용하여 정보 검색의 성능을 향상시키려는 노력이 계속되고 있다[4,5].

상호 정보망이란 사용할 문서의 집합에서 용어들간의 관련도를 구하는 방법으로, 이것은 코퍼스(Corpus)에 의존하는 실제적인 관련도 이다[6,7].

본 논문은 최종적으로 사용자에게 순서화된 문서를 제시하기 이전에 1차로 검색된 문서들에 대하여 질의의 키워드와 1차검색된 문서내의 키워드들의 관계를 자동으로 구축된 상호정보를 이용하여 다시한번 2차로 보다 의미적인 관련성 정도에 따라 재 순서화 하여 검색 결과를 제시하는것을 제안하고자 한다.

이를 위하여 먼저 자동으로 상호정보망을 구축한다. 1단계로 일반적인 통계 기법인 벡터공간 모델을 이용하여 TF*IDF에 따라 후보 문서 순위를 검색한다[1]. 2단계로 상호정보값을 이용하는 여러가지 공식을 적용하여 최종적으로 1차의 문서 순위를 재 조정 함으로써 정보 검색의 검색효율을 높인다.

제 2장에서는 키워드 색인, 자연어 검색 그리고 기존 문서 순위 결정의 문제점에 대하여 언급한다. 제 3장에서는 2단계 문서 순위 모델과 자동 키워드망 구축 방법 및 2차 문서순위의

조정을 위한 여러가지 공식들을 설명하며, 제 4장에서는 실험 및 평가 결과를 분석한다. 마지막으로 제 5장에서 결론을 맺는다.

2. 기존 색인, 검색 및 문서 순위 결정의 문제점

2.1 키워드 색인

자연어 검색을 위한 키워드 색인은 먼저 키워드를 자동으로 추출한다. 추출된 키워드들에 대하여 문서 내용과 관련 중요도를 갖는 역 색인 화일을 생성한다. 다음은 빈도에 의해 키워드의 중요도를 계산하는 식이다[8].

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

$Weight_{ij}$는 문서 i에서의 키워드 j의 중요도이다. $Freq_{ij}$는 문서 i에서의 키워드 j의 빈도수이고, n은 키워드 j를 갖는 문서의 수이다. 중요도는 기본적으로 빈도에 관계가 있다. 여기서 k를 낮은 값으로 곱해줌으로 단순 빈도만을 약간 인정하는 값으로 사용한다. 또한 그 뒤에 곱해지는 값은 전통적으로 정규화(nomalize)된 역문헌 빈도(IDF: Inverse Document Frequency)의 값이다. 그리고, 제목의 가중치를 다른 키워드들에의 가중치보다 높게 부여하여 키워드 색인을 한다[8].

2.2 자연어 검색

자연어 검색은 역 색인 화일 검색 기법을 사용하여 수행한다. 1차 자연어 검색에서는 IDF(Inverse Document Frequency) 조절 및 제목가중치 부가 중심으로 문서를 순서화를 한다[8].

자연어 질의에서 나타난 키워드 리스트에 있는 키워드가 나타나는 문서 들을 역색인 화일을 검색하여 찾아내어 가중치를 계산한다. 찾아낸 후보 문서들을 가중치순으로 특정 갯수만큼 추출한다.

2.3 기존 순위 결정의 문제점

i개의 유일한(unique) 키워드를 갖는 데이타 집합(textual set)이 주어졌다고 가정하자. 그러면 문서는 벡터 $(t_1, t_2, t_3, \ldots, t_n)$에 의해 표현될 수있다. 여기서 t_i는 만일 키워드 i가 존재하면 가중치값을 갖고 만일 문서에 키워드 i가 존재하지 않으면 0의 값을 갖는다. 같은방법으로 질의도 표현될 수 있다. 다음은 11개의 유일한 키워드를 갖는 데이타 집합의 표현을 보여준다. 제일 윗부분은 이 데이타 집합의 11개 키워드를 보여준다. 두번째 부분은 자연어 질의를 보여주고 질의의 단어가 있으면 그 벡터위치에 가중치값을, 없으면 0를 개념적 벡터에 변환되어 표현되어 있다. 세번째 부분은 유사하게 데이타 집합의 세가지 문서에 대한 개념적 표현을 보여주고 있다. 질의와 가장 잘 부합하는 문서를 결정하기 위하여 질의 벡터와 문서 벡터의 유사도값이 네번째 부분에 만들어져 있고 결과로 문서가 순서화되어 나타나 있다.

예에서 가중치값은 일반적인 TF*IDF의 값에 따라 할당되고 질의와 문서사이의 유사도 값은 일반적인 cosine measure에 의해 할당되어 있다[1].

개념적 용어 가중치

감인자 구리활자 이천 자격루 자오선 장영실 천문시계 천문의 측우기 판본 해시계

질의	장영실
벡터	(0 0 0 0 0 1 0 0 0 0 0)
문서 1(감인자)	감인자 구리활자 이천 장영실 판본
벡터	(0.3 0.3 0.1 0 0 0.1 0 0 0 0.3 0)
문서 2(혼천의)	혼천의 천문시계 자오선 이천 장영실
벡터	(0 0 0.1 0 0 0.4 0.1 0.4 0 0 0)
문서 3(장영실)	장영실 이천 천문의 자격루 해시계 측우기
벡터	(0 0 0.1 0.3 0 0.1 0 0.3 0.3 0 0.4)

유사도 검색

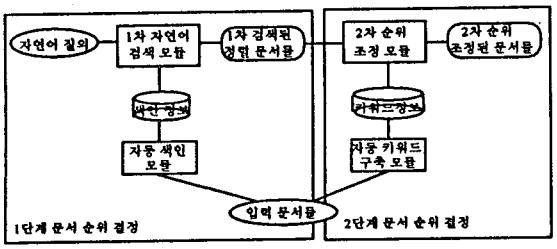
질의	(0 0 0 0 0 1 0 0 0 0 0)
문서1	(0.3 0.3 0.1 0 0 0.1 0 0 0 0.3 0)
	(0 0 0 0 0 0.1 0 0 0 0 0) = 0.1
질의	(0 0 0 0 0 1 0 0 0 0 0)
문서2	(0.3 0.3 0.1 0 0 0.1 0 0 0 0.3 0)
	(0 0 0 0 0 0.1 0 0 0 0 0) = 0.1
질의	(0 0 0 0 0 1 0 0 0 0 0)
문서3	(0.3 0.3 0.1 0 0 0.1 0 0 0 0.3 0)
	(0 0 0 0 0 0.1 0 0 0 0 0) = 0.1

위의 예에서 문서1 (감인자), 문서2 (혼천의), 문서3 (장영실) 은 질의 "장영실에 대하여" 와의 유사도가 동등하게 0.1의 값을 가지므로 임의의 순서에 의해 순서화 될 수 밖에 없다. 따라서 질의 "장영실에 대하여" 는 문맥적으로 문서3 (장영실) 을 최상위의 순위로 두어야 함에도 불구하고 임의의 순서인 최하위에 머무를 수 밖에 없다. 그러므로 보다 질의와 문맥적으로 가까운 문서3을 상위에 둘 수 있는 2단계 문서 순위 결정 모델을 제안하여 자연어 정보검색의 검색효율을 향상 시키고자 한다.

3. 2단계 문서 순위 결정

3.1 2단계 문서 순위 결정 모델

2단계 문서 순위 결정 모델은 그림 1과 같다. 1단계로서 자동 키워드 색인을 거친 색인 정보에 의해 1차 자연어 검색을 하며 1차적으로 검색된 순위 결정된 후보 문서를 결정한다. 2단계로서 1단계의 검색된 정렬 문서들 가운데 상위의 문서들에 대하여 이미 자동으로 구축된 키워드망을 이용하여 2차의 순위를 재 조정하여 최종적으로 문서의 순위를 결정한다.



<그림 1> 2단계 문서 순위 결정 모델

3.2 자동 키워드망의 구축

자동 키워드망(AKN : Automatic Keyword Network)이란 사용할 문서의 집합에서 키워드들간의 관련도를 구하는 방법으로 백과사전 단어 관계에 대하여 구조에 따라 임의로 부가 가중치를 주어 키워드 망을 형성시킨다.

- 단어의 연관성을 측정하기위하여 백과 사전을 입력 자료로 선정하여 아래와 같은 특수기호들에 의미를 부여하였다.

  ! 표제어
  # 표제어 2개이상
  @ 해설옮김
  $ 참조어
  \ 의미있는 최소 단어로 자동 키워드 마크 앞뒤에 '\' 기호를 붙였다.

- 연관성은 의미있는 최소의 단어를 중심으로 이루어졌다. 추출된 단어쌍은 어떤 단어 A와 B가 서로 연관이 있다면 B도 A와 같은 가중치를 가지는 대칭성을 인정하여 대칭화일을 만든다. 중복 단어쌍이 발견되면 가중치를 모두 합하여 중복성을 없앤다.

표제어와 해설 옮김	100
표제어와 참조어	50
표제어와 한 문장내의 단어들	30
표제어와 나머지문장내의 설명 단어들	20
참조어와 표제어 설명 단어들	5
window size 5내의 이웃한 단어들	10

실제 문장에서는 단어와 단어는 그 단어간의 거리보다는 문장의 구조에 의해서 영향을 받는다. 그러나 문장의 구조를 알기위해서는 파싱을 해야하는 문제가 발생함으로 단어간의 거리에 제한을 두는 것이 타당하다.
여기서 window size란 blank를 기준으로 한 것으로 중간점과 '-'도 여기에 포함시켰다. 괄호 사이의 내용은 주로 앞의 단어를 해석하는 내용이므로 AKN에 넣지 않았다.
window size가 5인 이유는 단어간의 의미를 한정하는데 에 적당하기 때문이다[6].
AKN에서의 최대 키워드 수와 최대 값이 <표 1>에 나타나 있다.

<표 1> AKN의 최대 공기 키워드수와 최대 값

	AKN
키워드 수	약 57,000
MAX값	31760 (나라, 우리)
MAX 공기 키워드수	9285(위)

3.3 문서내의 키워드 정보의 구성

AKN망 이외에 이를 효율적으로 잘 이용하고 2단계 문서 순위 결정을 할 수 있도록하기 위한 문서 번호 및 문서내의 키워드의 갯수 그리고 실제의 각 키워드의 문서내의 빈도를 알 수 있는 문서의 키워드 정보 화일을 구성한다.

3.4 2차 문서 순위 조정

2차 문서 순위 조정은 AKN을 이용하여 질의 내의 키워드들과 문서내의 키워드들간의 상호 정보값을 구하고, 그 값과 문서내의 키워드 정보를 이용하여 다음의 공식으로 문서순위를 재조정한다.

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

$$2.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} MINV_{i,j} / MIV_i$$

$$3.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} MINV_{i,j} / DK_j$$

$$4.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} MINV_{i,j} * (MAXMIV / MIV_i)$$

$$5.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} MINV_{i,j} * (MAXDK / DK_j)$$

$$6.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} MINV_{i,j} * (MAXMIV / MIV_i) * (MAXDK / DK_j)$$

$$7.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} MINV_{i,j} * (\log MAXMIV - \log MIV_i + 0.5)$$

$$8.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} MINV_{i,j} * (\log MAXDK - \log DK_j + 0.5)$$

$$9.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} (\log MINV_{i,j} + 0.5) * (\log MAXMIV - \log MIV_i + 0.5)$$

$$10.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} (\log MINV_{i,j} + 0.5) * (\log MAXDK - \log DK_j + 0.5)$$

$$11.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} MINV_{i,j} * (\log MAXMIV - \log MIV_i + 0.5) * (\log MAXDK - \log DK_j + 0.5)$$

$$12.\ SIM(Q,D)=\sum_{i=1}^{k}\sum_{j=1}^{l} (\log MINV_{i,j} + 0.5) * (\log MAXMIV - \log MIV_i + 0.5) * (\log MAXDK - \log DK_j + 0.5)$$

여기에서
k는 질의의 키워드 수이고,
l은 문서의 키워드 수이다.
$MINVij$는 질의의 키워드 i와 문서의 키워드 j의 AKN의 값이다.
$MIVi$는 질의의 키워드 i 의 AKN에서 공기(co-occurence) 빈도이다.
DKj는 문서 j의 키워드 수이다.
$MAXMIV$는 AKN에서의 최대 공기 빈도이다.

*MAXDK*는 문서들 중 최대의 키워드 빈도이다.

- 1번 공식의 의미는 질의의 키워드와 문서내의 키워드가 AKN에 형성이되어 있으면 이들을 모두 합한 값을 질의와 문서 사이의 유사도 값으로 사용하는 것이다. 그래서 이 유사도 값에 따라 순서를 재 정렬한다.
- 2번 공식의 의미는 질의 키워드의 절대 공기(co-occurence)값으로 AKN값을 나눈 값을 모두 합한다. 키워드의 절대 값에 의한 정규화(normalization)의 의미를 갖는다.
- 3번 공식은 문서의 키워드 수로 AKN값을 나눈값을 모두 합한다. 이의 의미는 문서의 절대 크기 값에 의한 정규화이다.
- 4번 공식은 AKN값과 해당 질의 키워드에 대한 상대 키워드 정규화 값 (즉, 키워드의 최대수를 해당 질의 키워드로 나눈 값)을 곱한것을 모두 합한 것이다.
- 5번 공식은 AKN값과 해당 문서의 상대 정규화 값 (즉, 키워드를 최대로 갖는 문서의 키워드 수를 해당 문서의 키워드의 수로 나눈 값) 을 곱한 값 을 모두 합한다.
- 6번 공식은 4번 공식과 5번 공식을 모두 적용한 것이다.
- 7번이하 공식들은 각 값의 log를 취한 값 (즉, 일정한 구간 값으로 취한 값)을 적절히 혼합하여 질의와 문서 간의 유사도를 계산한다. 여기에서 0.5를 더해 주는 것은 곱해지는 값이 0이되는 것을 방지하기 위한 값이다.

4. 실험 및 평가

4.1 실험 데이타

한국 전자통신 연구소 자연어 처리 연구실에서 개발한 멀티미디어 한국어 전자 백과 사전검색 시스템인 옥서를 기반으로 하였다. 옥서는 계몽사 학생 대백과 사전을 근간으로 하여 10M byte정도의 데이타, 약 23,000개의 표제어에 10만여개의 키워드를 갖추고 있는 사전 검색 시스템이다.

본 논문에서는 자동으로 키워드를 추출하여 1차 자연어 검색을 위한 색인화일과 2차문서 순위 조정을 위한 자동 키워드망 (AKN) 을 구축하였다.

그리고 질의 키워드와 문서내의 키워드가 같을 경우에 어떤 값을 취해야 좋은지를 판정하는 기준으로 AKN 값 분포의 최소 (MIN) , 최대 (MAX) , 중간값 (MID)을 다음 표 2와 같이 사용한다.

<표 2> 질의내의 키워드와 문서내의 키워드가 같을 경우의 MAX,MIN,MID값

	MAX	MIN	MID
AKN 분포	100	0	20

4.2 평가 방법

- 질의 갯수 46개와 적합성 정보를 이용한다.

질의 : 9명의 일반인 들이 백과사전에 있는 내용에 대하여 알고자하는 사항을 자연어로 5개씩 질의한 45개의 자연어 질의를 사용한다.

적합성 정보 : 4명의 전문가로부터 자연어 질의에 적합한 백과사전 표제어 항목 중 2명 이상이 일치한 백과사전 항목 정보

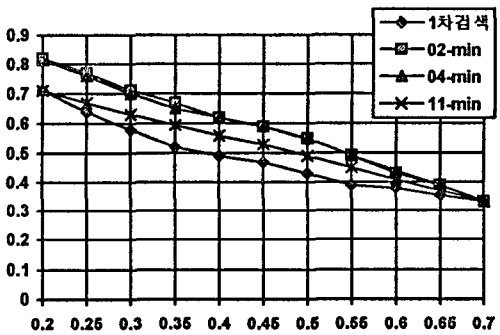
- 기존의 정확률과 재현률의 공식을 이용한다[1].

정확률(Precision) = 검색된 적합 문서수 / 검색된 문서 총수
재현률(Recall) = 검색된 적합 문서수 / 적합문서 총수

45개의 자연어 질의에 대하여 1차 검색을 한 후, 자연어 질의의 키워드와 검색된 문서의 키워드간의 상호 정보값을 식들을 이용하여 계산하여 문서 순위를 재조정하였다. 그리고 재현율과 정확률을 계산하였다.

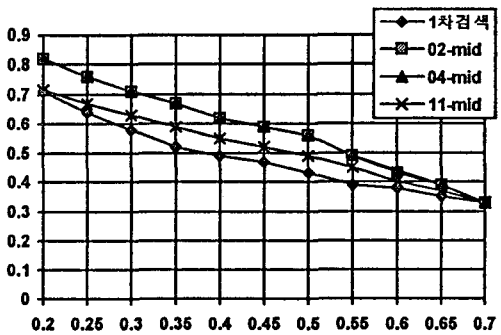
4.3 평가 결과

다음 그림 2는 여러가지 검색 효율 중 MIN 값을 중심으로 하고 그중 검색 효율이 좋은 3가지 공식의 결과이다.



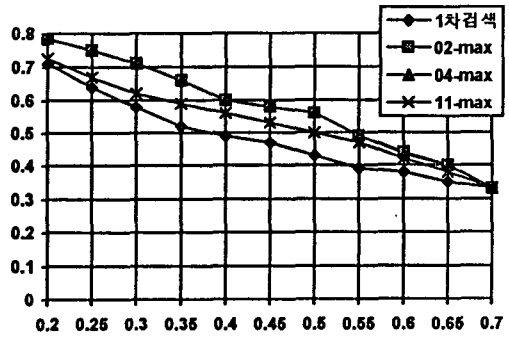
<그림 2>1차검색 및 2차조정 min의 재현율(x 축)과 정확률(y 축)

여기에서 기본 곡선은 1차 자연어 검색의 효율이며 나머지 곡선은 각각 2,4,11번 공식의 MIN 적용시 검색 효율이다. 그래프 상에서 2,4번 공식이 우세한 것으로 나타나 있다. 그림 3은 MID 값을 중심으로 한 결과 이다.



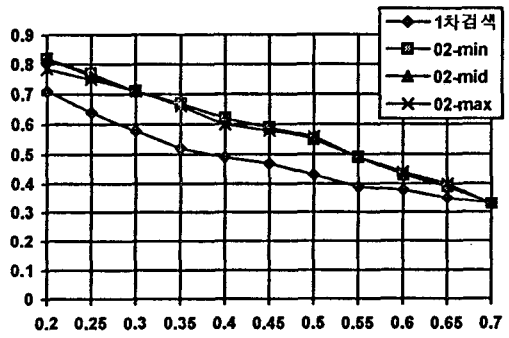
<그림 3>1차검색 및 2차조정 mid의 재현율(x 축)과 정확률(y 축)

마찬가지로 2,4번 공식이 우세한 것으로 나타나 있다. 그림 4는 MAX 값을 중심으로 한 결과이다.



<그림 4>1차검색 및 2차조정 max의 재현율(x 축)과 정확률(y 축)

MIN 이나 MID 보다는 약간 떨어지지만 2,4번 공식이 우세하다. 그림 5는 1차 검색에 대한 2차 순위 조정의 2번 공식을 중심으로 본 MIN, MID, MAX의 결과이다.



<그림 5> 1차검색 및 2차 조정 2번 공식의 min, mid, max의 재현율 (x 축)과 정확률 (y 축)

공식에 따른 MIN, MID, MAX는 거의 유사하게 나타난다. 마찬가지로 4번, 11번에서도 MIN, MID, MAX는 거의 유사하게 나타난다. 다음 표 3은 전체적인 검색 효율 향상표 이다. 각각의 재현율 포인트에서 정확률 차이값을 더하고 그값을 재현율 포인트 수로 나눈 평균 증가치이다.

<표 3> 백과사전 AKN과 1차 검색과의 비교표 (1차검색=0.0%)

	MIN	MID	MAX	평균
02	+10.80%	+10.80%	+10.15%	+10.58%
04	+10.45%	+10.90%	+10.10%	+10.48%
11	+3.50%	+3.50%	+4.10%	+3.70%
평균	+8.25%	+8.40%	+8.12%	

- 1차검색 결과보다 최고 약 10.9%의 향상을 보였다. 이것은 AKN의 값에 의해 가중치를 조절함으로서 순위재조정이 보다 의미적으로 가까운 문서가 상위로 올라온다는 것이다. 즉 문맥에 민감하게 (context sensitive) 작용한다는 실험 증거이다.
- 2번 공식과 4번 공식이 좋고, MIN, MID, MAX는 비슷한 결과를 보인다.

5. 결 론

본 논문에서는 AKN을 이용한 2단계 문서 순위 결정 방법에 의한 자연어 검색 모델을 제안했다. 지금까지 자동으로 구축된 키워드 색인에서 1차로 자연어 검색을하고 2차로 AKN을 통한 정확성 향상에 관해 검색 효율을 평가하여 1차검색 결과보다 최대 10.9%의 향상을 보였다. 공식에서는 AKN에서 4번 공식의 MID값이 더 낫다는 것을 알수 있다. 그러나 2번이나 4번 공식의 MIN 이나 MID도 유사한 검색 효율 향상을 이룬다고 할 수 있다.

본 논문에서 제안한 2단계 문서 순위 결정 방법은 일반 순서화된 문서의 순위 재조정이나 질의확장, relevance feedback 등의 후에 최종적으로 순서를 재조정 할 수 있는 하나의 방법이다. 왜냐하면 재현율도 중요함으로, 재현율을 유지하면서 정확성을 높일 수 있는 한가지 방법이기 때문이다.

앞으로 임의 부가 가중치가 아닌 상호 정보망을 이용 하여 실험 할 예정이다.

참고 문헌

[1] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley Publishing Company, 1989.

[2] 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.

[3] T. Noreault, M. Koll, and J. J. McGill, "Automatic Ranked Output from Boolean Searches in SIRE," *Journal of the American Society for Information Science*, Vol. 28, No. 6, pp. 333-339, 1977.

[4] J. H. Lee, M. H. Kim, and Y. J. Lee, "Ranking Documents in Thesaurus-based Boolean Retrieval Systems," *Information Processing and Management*, Vol. 30. No. 1, pp. 79-91, 1994.

[5] Y.W. Kim, J. H. Kim, "A Model of Knowledge Based Information Retrieval With Hierarchical Concept Graph," *Journal of Documentation*, Vol. 46. No. 2, pp. 113-136, 1990.

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

[7] D. Biber, "Co-occurrence Patterns among Collocations : A Tool for Corpus-Based Lexical Knowledge Acquisition," *Computational Linguistics*, Vol. 19, No. 3, pp. 531-538, 1993.

[8] 강현규, "옥서에서의 자연어 검색 성능분석 및 개선", 한국정보처리학회 95춘계 학술발표 논문집 ,pp56-59,1995.

[9] 김명철, 이윤재, 최기선, 김길창, "시소러스 작성을 위한 개념 획득 도구", 한글 및 한국어 정보처리, pp39-50, 1992.