인접한 단어와 키워드 주제어 정보에 기반한 유사 문헌 검색 시스템 개발

Development of Similar Bibliographic Retrieval System based on Neighboring Words and Keyword Topic Information

김 광 영(Kwang-Young Kim)*
곽 승 진(Seung-Jin Kwak)**

〈 목 차 〉

초 록

유사 문헌 검색 시스템은 추출된 색인어 중에서 어떤 것을 선택하는가에 따라 검색 결과에 많은 차이점이 발생한다. 본 연구에서는 추출된 후보 색인어의 선정의 오류를 최소한으로 하는 방법을 제공한다. 본 연구에서는 유사 문헌에서 추출된 후보 색인어들을 이용하여 인접한 단어들의 정보와 추출된 키워드 주제어 정보를 이용하였다. 그리고 관련 저자들 정보와 검색 결과의 재순위화 방법을 이용하여 보다 정확도가 높은 유사 문헌 검색 시스템을 개발하였다. 본 논문에서는 과학기술 학회마을 데이터베이스를 이용하여 실험하였다. 실험과 사용자 평가를 통해서 유사 문헌 검색 시스템의 성능을 입증하였다.

키워드: 정보 검색, 정보 검색 시스템, 주제어, 유사 문서, 가중치

ABSTRACT

The similar bibliographic retrieval system follows whether it selects a thing of the extracted index term and or not the difference in which the similar document retrieval system There be many in the search result is generated. In this research, the method minimally making the error of the selection of the extracted candidate index term is provided In this research, the word information in which it is adjacent by using candidate index terms extracted from the similar literature and the keyword topic information were used. And by using the related author information and the reranking method of the search result, the similar bibliographic system in which an accuracy is high was developed. In this paper, we conducted experiments for similar bibliographic retrieval system on a collection of Korean journal articles of science and technology arena. The performance of similar bibliographic retrieval system was proved through an experiment and user evaluation.

Keywords: Information Retrieval, Information Retrieval System, Topic Information, Similar Document, Weight

* 한국과학기술정보연구원 정보기술연구실 선임연구원(kykim@kisti.re.kr)(제1저자)
** 충남대학교 사회과학대학 문헌정보학과 조교수(sjkwak@cnu.ac.kr)(교신저자)
• 접수일: 2009년 8월 21일 • 최초심사일: 2009년 8월 25일 • 최종심사일: 2009년 9월 21일

I. 서 론

인터넷의 발달과 PC보급의 확대로 웹문서, 전자책, 전자저널 등의 디지털 정보가 급속도로 증가하고 있다. 정보의 홍수 속에서 사용자가 원하는 정보를 빠르고 정확하게 찾기는 점점 더 쉽지 않다. 단순히 검색엔진의 사용만으로 정보문제를 근본적으로 해결 할 수 없기 때문에 웹 포털 등 검색 서비스에서도 다양한 서비스 기능을 개발하고 있다.

대부분의 검색 시스템은 정확한 검색 기능뿐만 아니라 유사 문서 검색 기능을 제공하고 있다. 이러한 문서간 유사도 측정기술의 응용 분야로는 정보검색, 문서동일성 여부 검증, 문서 분류 및 클러스터링(clustering) 등이 있다. 유사 문서 검색 시스템은 여러 문서의 유사도를 검사해 원본과 일치하는지 여부를 판단하여 동일한 리포트 조사 등에 사용되고, 논문 서비스 시스템에서는 표절이 의심되는 논문 등을 가려내는데 사용된다. 최근에는 논문 표절과 저작권 침해 등의 문제가 사회적 이슈로 떠오르면서 이를 판별하는 시스템에 관심이 많아지고 있다.

유사 문서를 판별하기 위해서는 가장 중요한 것은 용어들을 선정하는 방법이고 시스템마다 다양하다. 대부분의 시스템은 유사한 문서를 검색하기 위해서 후보 색인어들을 추출하지만 후보 색인어로 추출된 용어들은 단어의 연결성 정보를 잃어버리게 된다. 이것은 정보 검색 시스템이 색인을 위해서 문서의 단어들을 색인화 시키는 과정에서 발생한다. 따라서 추출된 후보 색인어의 단어 연결성 유지가 무엇보다 중요하다.

유사 문서 검색 시스템은 사용되는 분야에 따라 다양한 모델이 적용될 수 있다. 현재 대부분의 정보 검색 시스템은 문서에 나타난 모든 용어를 추출해서 문서간의 유사도 판별에 사용하고 있지만 실질적으로 유사도 판별에 중요한 작용을 하는 용어는 빈도가 높거나 혹은 기타 여러 가지 용어 가중치 기법을 이용해서 높은 가중치를 가지는 용어만이 문서 간의 유사도에 영향을 주는 편이다.[1)]

현재 사용되고 있는 대부분의 유사 문서 검색 시스템은 문서에서 출현하는 빈도만을 이용하여 유사도를 측정하여 서비스를 하고 있다. 따라서 이용자에게 보다 정확한 유사 문서를 제공하기 위해서는 다양한 방법과 모델들을 개발하고, 운영되고 있는 시스템에 적용하여 실험을 통하여 평가할 필요가 있다.

본 연구의 목적은 유사 문서 검색 시스템의 성능 향상을 위한 새로운 모델을 제안하고, 현재 운영 중에 있는 KISTI의 과학기술 학회마을[2)] 논문 정보서비스에 적용하여 정확도가 높고 유사한 논문을 찾아 주는 정보 검색 시스템을 개발하고 평가하는 것이다.

본 연구에서는 유사 문헌을 판별하기 위하여 중요한 용어를 선정하는 방법으로 다음과 같이 세

1) 장성호, 강승식, "용어 선별 기법에 의한 유사 문서 판별 시스템," 한국정보과학회학술지, 제30권, 제1호(2003), pp.534-536.

2) KISTI 과학기술학회마을 홈페이지, 〈http://society.kisti.re.kr/〉 [cited 2009. 08. 19].

가지 모델을 개발하고 실험을 통하여 평가하였다. 첫째, 연결성 정보를 위해서 인접한 단어들이 후보 색인어로 선정 될 수 있도록 시스템을 설계하였다. 둘째, 주제어 정보를 이용하여 후보 색인어를 주제 정보와 가까운 용어들이 선정 될 수 있도록 하였다. 셋째, 일반적으로 논문서지 DB에 기술되어 있는 공동 저자들이 관련된 유사한 연구를 할 경우가 매우 높다는 점을 이용하여 유사 문서 검색 결과에서 관련 저자들이 나타나는 문서들을 상위로 랭킹 시킬 수 있는 모델을 개발하였다.

이런 정보들을 이용하여 보다 성능이 향상된 유사 문헌 검색 시스템을 개발하고, 과학기술학회마을 학술DB를 이용하여 실험하고 평가자들이 직접 평가하도록 하였다. 평가 방법은 평가자들이 유사 문헌을 검색하여 상위 10위까지의 유사 문헌들에 대해서 적합한 문서가 나타나는지를 역순위 평균(Mean Reciprocal Rank) 평가 방법으로 평가자가 직접 평가를 하였다.

Ⅱ. 관련연구

장성호와 강승식은 추출된 용어만으로 문서 간의 유사도 비교를 하는 경우에는 문서 간에 공통 용어의 수와 각 문서의 총 용어 수를 이용하였다. 공통 용어의 수와 각 문서에서 추출된 용어의 수의 관계로 임의의 두 문서가 서로 유사한지를 판단하기 위해서 문서 간의 관계를 보는 방식으로 구현하였다.[3)]

문서 간의 유사도 측정을 위한 색인추출 방법 중 하나는 개별 단어로 구성된 색인을 생성하는 방법이다. 단어 색인은 단일단어로 구성되어 적은 양의 데이터에도 많은 색인을 추출할 수 있다는 장점이 있는 반면에 문맥정보를 포함할 수 없다는 단점이 있다고 박수용 등은 주장하였다.[4)]

그러나 유사 문서 검색 또는 일반 정보 검색 시스템에서 단어와 문서간의 가중치를 계산하는 가장 일반 적인 방법은 Cosine 유사도를 측정하는 방법이다. 또한 위와 같이 단어 색인 방법의 특징으로 두 문서간의 일치하는 색인어의 수, 상호정보량, 평균조건확률, 색인어 출현 빈도 등의 특징을 이용하기도 한다.

구색인 방법[5)6)]은 공기 정보 등을 추출하기 위해서 단어쌍을 색인으로 추출한다. 공기(Collocation 또는 Co-occurrence)란 두 단어가 동일 문서, 동일 문단, 동일 문장 또는 일정한 크기의 단어창 안에서 같이 발생하는 현상을 말하며 공기 빈도수가 클수록 두 단어가 밀접한 관련이 있다고 간주

3) 장성호, 강승식, 전게논문, pp.534-536.
4) 박수용 등, "유사도 측정 기법을 이용한 효율적인 요구 분석 지원 시스템의 구현," 한국정보과학회논문지, 제27권, 제1호(2000), pp.13-23.
5) Baker, McCallum, "Distributional Clustering of Words for Text Classification," *Proceedings of SIGIR* (1998), pp.96-103.
6) 정준호, 김미진, "문서 요약 시스템을 위한 수사 구조 트리 생성," 한국정보과학회논문지, 제26권, 제2호(1999), pp.175-177.

한다.[7] 즉, 임의의 두 단어가 일정영역에서 동시에 출현하는 빈도가 높을수록 두 단어의 관련이 깊다는 것이다.

따라서 단어쌍을 기반으로 계산되는 유사도는 두 문서의 색인 파일에 공통으로 존재하는 단어쌍의 가중치 값의 곱들을 합산한 것으로 이 값이 클수록 두 문서 간의 유사도가 높아진다. 단어쌍을 색인어로 하는 구색인 방법은 문맥의 정보를 어느 정도 반영할 수 있기에 단어색인 방법에 비해 문서의 내용을 더 잘 내포할 수 있다. 하지만 색인어가 되는 단어쌍의 추출에 드는 시간적인 비용이 크다.[8]

문서 분류 분야에서도 정확한 주제어 정보를 추출하는 것이 중요한 문제점으로 부각되고 있다. 안희국과 노희영은 주제어 추출을 위해서 한국어의 기초 문형 구조를 분석하여 문장의 역할(주어) 정보를 추출하는 방식을 사용하였다. 이와 같은 방식을 사용하여 더욱 세분화하여 주제어를 추출하는 것이 가능함을 확인 할 수 있다.[9]

이러한 유사 문서 검색은 문서 클러스터링 및 분류 분야에서 문서 간의 관련성을 정량적으로 측정하기 위해 문서간의 유사도를 계산해야하는데 사용되며, 학습 기반을 이용하는 방법 등 다양한 방법과 모델을 이용하여 사용되고 있다.

Ⅲ. 유사 문헌 검색 시스템

본 연구에서 제안한 유사 문헌 검색 시스템은 구색인 방법과 비슷하지만 빠른 속도로 처리하기 위해서 추출된 색인어들에 대해서만 처리를 한다. 즉, 구색인 방법에서 인접한 단어 사이의 공기 정보를 추출하기 위해서 슬라이딩 윈도 기법을 사용하여 맨 앞의 내용어와 다음 내용어들 간의 쌍을 추출하는 방법[10]과는 다르게 색인어로 추출된 명사들 중에서 서로 인접한 단어에 대해서 가중치를 부여 받는 방식을 사용하였다. 그리고 마지막에는 키워드 주제어 정보에 맞는 색인어들에 대해서 가중치를 추가로 더 받게 된다. 위와 같은 방식으로 최대한 정확한 후보 색인어를 선정하여 유사한 문헌을 검색을 할 수 있는 시스템을 구현하였다.

〈그림 1〉에서와 같이 질의어를 이용한 문서를 검색한 후 유사 문헌 검색을 선택할 때 선택된 문서에서 문서의 색인어를 추출한다. 추출된 단어들에 대해서는 인접한 단어의 가중치와 주제어 검색을 통한 주요 핵심 단어들에 대해서 2, 3차 가중치를 다시 계산하게 된다. 여기서 선정된 중심

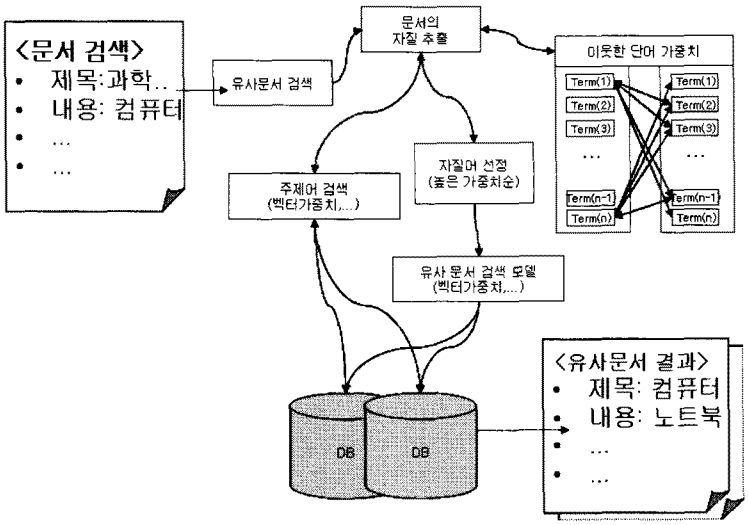
7) Hajime, Takeo, Manbu, "Text Segmentation with Multiple Surface Linguistic Cues," *Proceedings of the COLING-ACL*(1998), pp.881-885.

8) 김혜숙, "단어/단어쌍 특징과 신경망을 이용한 두 문서간의 유사도 측정," 한국정보과학회논문지, 제31권, 제12호(2004), pp.1660-1671.

9) 안희국, 노희영, "문서 분류를 위한 문장 응집도와 주어 주도의 주제어 추출," 한국정보과학회, 한국컴퓨터종합학술대회 논문집, 제32권, 제1(B)호(2005), pp.463-465.

10) 김혜숙, 전게논문, pp.1660-1671.

색인어를 이용하여 유사한 문헌 검색을 수행하게 된다. 그리고 최종 결과에서 다시 관련 저자들의 정보를 이용하여 검색된 문서 중에서 관련된 저자들이 나타날 경우에는 그 문서가 더 높은 가중치를 받게 되어서 상위 문서로 랭킹 시키는 방식으로 시스템을 구성하였다.

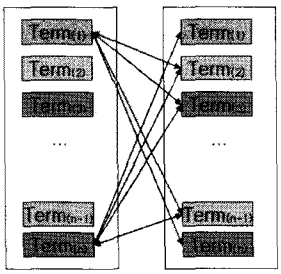


〈그림 1〉 시스템 구조도

1. 인접한 단어들을 이용한 후보 색인어 추출

유사 문헌 검색을 위해서 색인기에서 추출된 색인어들은 정확한 후보 색인어로 선정되기 위해서 인접한 단어들을 정보를 이용한다. 추출된 단어들 간에 서로 가까운 거리에 있는 단어들끼리는 상호 가중치를 더 받게 되어서 유사 문헌의 후보 색인어의 선정되도록 한다.

구색인 방법으로 단어의 쌍을 추출하는 것이 아니라 추출된 일반 명사들 중에서 서로 가까운 단어들에 대해서 한 번 더 가중치를 산정하게 된다.



〈그림 2〉 인접한 단어 간의 상호 가중치 계산

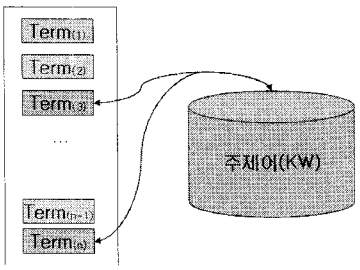
〈그림 2〉는 후보 색인어들 간의 위치 정보 색인 데이터베이스[11]를 이용하여 거리를 계산한다. 〈그림 2〉와 같이 모든 후보 색인어(Term)들 간에 거리를 계산하는 것을 나타내고 있으며 N×N 번을 수행하게 된다. 실제로 위치 정보를 이용하여 거리가 가장 짧은 것을 선택한다. 그리고 복합명사에 대해서는 두 단어가 분리될 경우에는 대해서는 제거하였다. 예를 들면 "정보검색"이라는 단어서 "정보"와 "검색"이 나타날 경우에는 제외시킨다. 이와 같은 복합명사 단어는 한단어로 인식을 한다.

위와 같은 방식을 사용함으로써 서로 가까운 색인어들은 후보 색인어로 선정될 수 있도록 시스템을 설계하였다.

예를 들면 "... 비만 고양이 ..."에서 "비만"과 "고양이"는 아주 인접한 단어임으로 서로 가중치를 받게 되어서 〈식 3〉에서 와 같이 distWeight의 가중치를 추가하여 유사 문헌 검색에서 후보 색인어로 선정이 될 확률이 다른 후보 색인어 보다 높아지게 된다. 이와 같이 단어 간의 연결성 정보를 이용하였다.

2. 키워드 주제 정보를 이용한 후보 색인어 선정

후보 색인어들 중에서 논문 저자들의 키워드 주제어 정보가 반영이 될 수 있도록 저자들의 키워드 주제어들을 추출하였다. 추출된 키워드 주제어 정보를 이용하여 유사 문헌 검색을 위해서 추출된 색인어들이 나타날 경우에 가중치를 추가한다.



〈그림 3〉 주제어와 추출어 간의 가중치 계산

〈그림 3〉은 저자들의 키워드 정보가 나타나는 후보 색인어(Term)들에 대해서 주제어 정보를 찾아서 가중치를 추가하는 것을 나타내고 있다. 주제어 정보는 논문에서 키워드 정보를 중심으로 추출되며 사용자가 선정한 문서에 대해서 검색을 하여 상위 30개의 문서에 대해서 키워드 정보를

11) 위치 정보 색인 데이터베이스 : 한 단락을 중심으로 각 단어에 번호를 부여하여 단어의 위치 번호(1, 2, 3, ...)를 색인 DB에 저장 하여 관리 함.

추출한다. 상위 30개의 문서를 선정한 이유는 검색 결과가 최소 상위 30%까지는 정확하다고 가정을 하였다. 물론 검색 결과 전체를 사용해도 가능하나 많은 문서를 선정할 때에는 시스템의 속도에 많은 영향을 준다. 본 논문에서는 실험적으로 상위 30개로 한정을 하여 실험을 하였다.

저자들의 키워드 정보의 가중치는 첫 번째, 두 번째, 세 번째 순으로 가중치를 계산한다.

〈식 1〉 $$T(w_t) = \sum_{n,m,l} (n\alpha + m\beta + l\gamma) w_t \qquad (\text{단, } \alpha > \beta > \gamma, \alpha+\beta+\gamma = 1.0)$$

〈식 1〉에서와 같이 키워드 중에서 첫 번째 것이 가장 높은 가중치를 가진다. n, m, l은 가중치 Wt에서 전체 첫 번째, 두 번째, 세 번째 키워드들에 대해서 일치한 전체 α, β, γ 값의 합한 값을 의미한다. 본 논문에서는 α, β, γ 값은 상수 값으로 조정 될 수가 있다. 본 논문에서는 α=0.6, β=0.3, γ=0.1로 실험적인 값을 사용하였다. 이것은 몇 번째 키워드에 대해서 가중치를 더 많이 줄 것인가를 결정할 수 있다. 본 연구에서는 저자들이 첫 번째 키워드가 가장 핵심 주제 정보를 그 다음 순위로 저자들이 키워드를 작성할 것이라고 가정을 하였다. 위와 같은 가중치로 추출된 주제어 정보를 이용하여 유사 문헌에서 추출된 색인어들에 대해서 주제어와 같은 단어들에 대해서는 후보 색인어로 더 가중치를 추가하게 된다. 가중치는 유사 문헌에서 추출된 전체 가중치의 평균값을 더 한다.

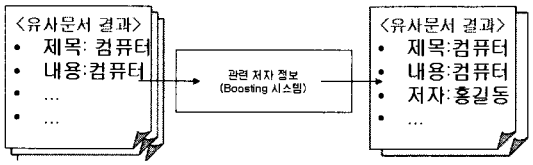
본 연구에서는 위와 방식을 사용함으로 추출된 후보 색인어 중에서 주제어 정보와 가장 가까운 단어들을 먼저 선정할 수 있도록 하였다. 주제어 정보는 유사 문헌 검색에서 가장 중요한 핵심 정보이고 정확한 유사 문헌을 검색할 수 있는 중요한 변수이다.

예를 들면 논문의 주제어가 "정보 검색 시스템"이라는 주제어가 들어 있으면 후보 색인어 중에서 "정보", "검색" 및 "시스템"의 색인어들은 유사 문헌 검색 시스템의 후보 색인어로 선정될 확률이 높아진다.

3. 관련 저자 정보를 이용한 가중치

본 연구에서는 유사한 논문을 찾기 위해서 관련 저자 정보[12]들을 이용하는 방식을 추가하여 사용한다. 보통 논문에서는 관련된 저자들이 비슷한 주제 정보를 이용하여 논문을 작성하는 성향들이 있다. 본 논문에서는 이러한 점을 이용하여 서지 DB에서 관련 저자들 정보를 이용하는 시스템을 구성하였다.

12) 관련 저자 정보 : 논문의 저자들의 정보(첫 번째 저자, 두 번째 저자, 교신 저자 정보).



〈그림 4〉 관련 저자 정보 Boosting 시스템

본 논문은 논문의 서지 DB에 있는 관련 저자 정보 중에서 첫 번째와 마지막 저자(교신 저자)를 중심으로 관련된 유사 논문을 찾는 방식을 사용하였다. 〈그림 4〉는 유사한 문서들의 검색 결과에서 나온 문서들 중에서 관련된 저자들의 정보를 이용하여 상위 문서로 랭킹 시키는 시스템이다. 색인어 추출에서 선정된 후보 색인어들과 관련된 저자들의 정보를 이용한 보다 더 정확도가 높은 시스템을 구성하기 위한 것이다.

본 연구에서는 앞의 방식에서 나온 최종 결과 문서들 중에서 관련된 저자가 있을 경우에 그 문서를 재순위화를 시킴으로써 정화도 높은 유사 문서를 제공할 수 있도록 시스템을 구성하였다. 하지만 앞 단계에서 후보 색인어 추출하여 검색한 결과 관련된 저자들이 없을 경우에는 적용되지 않는다.

Ⅳ. 실험 및 결과

본 연구의 실험을 위해서 사용한 DB는 KISTI가 운영하는 과학기술 학회마을에서 서비스 중인 40만 건의 국내 논문의 서지 DB를 사용하였다. 논문은 제목, 초록, 키워드 정보가 포함 된 것들이다. 검색 시스템의 서버 사양은 리눅스 Redhat 4.1.2, 메모리 12G, 2CPU 인텔 Xeon 1.6GHz를 사용하였다. 과학기술학회마을 DB의 섹션 구성은 〈표 1〉과 같다.

〈표 1〉 과학기술학회마을 섹션 구성

섹션이름	설명
TIK	한글 제목
TIE	영어 제목
ABK	한글 초록
ABE	영문 초록
KW	키워드 정보
AUK	한글 논문 저자명
…	…

본 논문에서는 〈표 1〉과 같이 한글 제목, 영문 제목, 초록, 영문 초록, 키워드, 저자명 등을 각각

하나의 섹션으로 나누고 이것들이 전체 하나의 문서를 이루는 구조로 적재하고 색인하여 사용하였다.

본 연구에서는 평가 방법으로는 정답 후보 집합을 만들어서 정확도를 평가해야 하나, 직접 평가자들이 검색하여 적합하다고 생각하는 문서가 상위에 얼마나 분포하는지 여부를 조사하기 위해서 역순위평균(MRR) 방식을 이용하였다.

역순위 평가 방법은 일반적으로 질의/응답 시스템의 평가를 위해서 사용되는 방법으로 정답 문서가 1번째의 순위에 나타나면 1 점, 2번째에 나타나면 1/2 점, N번째의 순위에 나타났으면 1/N으로 점수를 부여하는 방식이다.

실험을 결과를 평가하기 위해서 관련 분야 3명의 평가자가 직접 결과를 보고 평가를 하는 방식을 사용하였다. 평가자들은 실험의 결과에 적합한지 문서인지 여부를 평가하여 제출하는 방식을 사용하였다. 보다 정확한 평가를 위해서 많은 사용자들이 평가에 참하는 것이 바람직하나 실험적인 결과를 고찰하기위해서 3명만이 평가에 참가를 하였다.

1. 인접한 단어들을 이용한 후보 색인어 추출

가. 실험 방법

인접한 단어들을 이용한 실험 방법으로는 (1) 한 문서를 〈표 1〉과 같이 각각의 섹션으로 보고 각 섹션별로 TF(Term Frequency)를 선정된 후보 색인어를 이용하여 유사 문서 검색을 수행 (2) 한 문서에서 발생하는 TF를 이용하여 각 섹션별로 검색을 수행 (3) 한 문서를 각각의 섹션으로 보고 각 섹션별로 TF를 선정된 후보 색인어 범위를 넓게 잡고 선정된 후보 색인어 간의 인접한 단어의 정보를 이용하여 가중치를 산정 후에 선정된 후보 색인어를 이용하여 유사 문서를 검색을 수행하였다. 실험 (4)에서는 한 문서를 전체로 보고 TF 정보를 추출하고, 추출된 단어들에 대해서 DF(Document Frequency) 정보를 산출하고 〈식 1〉을 기본으로 후보 색인어에 대한 가중치를 산출한다. 그리고 실험(3)의 인접한 단어 간의 유사도를 측정하여 평균 가중치를 추가하는 방식으로 실험을 수행하였다.

실험(1)에서는 문서의 기본적인 TF 정보를 추출한다. 추출된 단어들에 대해서 DF정보를 DB에서 구하고 문서에서 추출된 색인어에 대해서 가중치를 〈식 2〉에서 계산한다.

〈식 2〉 $$w(d_i k_j) = tf_{ij} \log(N/df_j)$$

후보 색인어는 추출된 단어 중에서 최대 70개로 선정하고 최종 후보 색인어는 15로 제한을 하여 유사 문헌 검색을 수행한다. 모든 용어를 추출한 경우보다 40~60개의 중요 용어만 추출한 경우에 비슷하거나 더 좋은 성능을 나타낸다는 것을 알 수 있으며, 이것으로 문서내의 모든 용어 추출이

반드시 필요한 것이 아니라는 것을 알 수 있다.[13]

후보 색인어를 너무 많이 선정하면 유사 문헌의 가중치 값이 낮아질 수 있으므로 정확한 문서를 찾기에 어려움이 있다. 그렇다고 후보 색인어를 너무 작게 선정하며 유사 문서의 가중치 값이 높아질 수 있지만 실제 정확한 유사 문헌으로 찾을 수 없게 된다. 본 실험에서는 최종 후보 색인어 선정을 실험을 통해서 15정도로 고정을 하고 실험을 수행하였다.

실험(2)에서는 한 문서를 전체로 보고 TF 정보를 추출하고, 추출된 단어들에 대해서 DF 정보를 산출하고 식(2)을 기본으로 후보 색인어에 대한 가중치를 산출한다.

실험(3)에서는 실험(1)과 같이 수행 후에 후보 색인어 70개를 이용하여 상호간의 인접한 단어의 정보를 이용하여 아래 〈식 3〉과 같이 추가적으로 가중치를 더 산정한다.

〈식 3〉 $$w(d_{i,k_j}) = tfidf_{ij} + distWeight_j$$

〈식 3〉의 distWeight값은 두 단어가 가장 가까운 단어들에 대해서(거리 값이 최소인 단어들끼리) 가중치에 대한 평균값을 서로 추가하였다.

나. 실험 결과

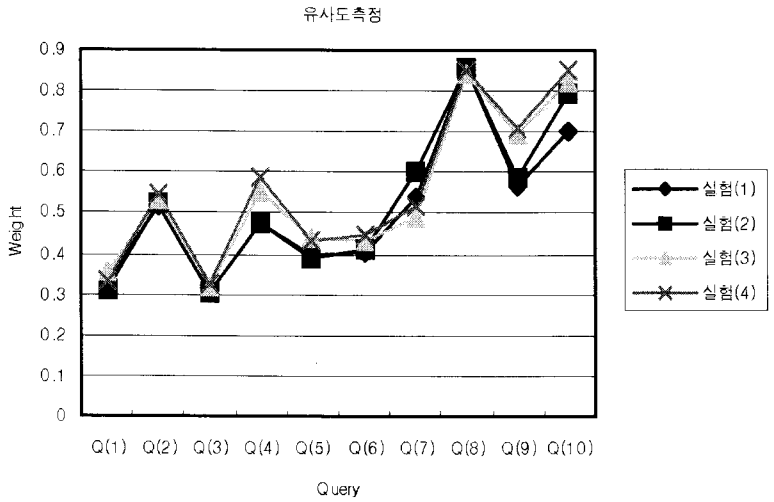
본 실험에서는 검색한 결과에 대해서 벡터 모델을 이용하여 검색한 문서의 가중치 중에서 최댓값을 1.0으로 100%로 볼 때, 유사한 정도가 30%(0.3)이상인 문서에 대해서 어느 정도 적합하다고 보았다. 즉 유사 문헌 검색 결과에 대해서 유사한 정도인 임계값이 0.3이상인 문서에 대해서 아래의 실험 결과를 산출하였다.

〈표 2〉는 10개의 질의어로 검색한 문서에 대해서 특정 문서를 선택하여 유사 문서를 검색을 수행한다. 그리고 그 결과 값 중에서 가장 높은 유사도 값을 계산하였다. 즉 후보 색인어의 선정에 따라 유사도 값이 다르게 나타 날 것이다. 본 논문에서는 〈표 2〉와 같이 자릿수 통일을 위해서 소수점 3자리 이하는 절삭하였다.

〈표 2〉 실험별 유사도 값 측정

구 분	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
실험(1)	0.335	0.518	0.305	0.477	0.394	0.402	0.536	0.855	0.566	0.701
실험(2)	0.310	0.522	0.305	0.477	0.391	0.413	0.6	0.855	0.587	0.794
실험(3)	0.355	0.528	0.319	0.552	0.438	0.429	0.491	0.846	0.697	0.819
실험(4)	0.336	0.546	0.323	0.591	0.430	0.448	0.514	0.847	0.711	0.849

13) 장성호, 강승식, 전게논문, pp.534-536.



〈그림 5〉 실험별 유사도 값 측정

〈그림 5〉의 그래프는 유사도 값이 30%이상인 것 중에서 가장 높은 유사도 값을 산출 한 것이다. 위의 그래프에서 나타난 것과 같이 전체적으로 실험(4)에 대해서 유사도 측정값이 높은 것으로 나타나고 있다.

〈표 3〉은 10개의 질의어로 검색한 문서에 대해서 특정 문서를 선택하여 유사 문헌 검색을 수행한다. 그리고 그 결과 문서 건수 나타내고 있다.

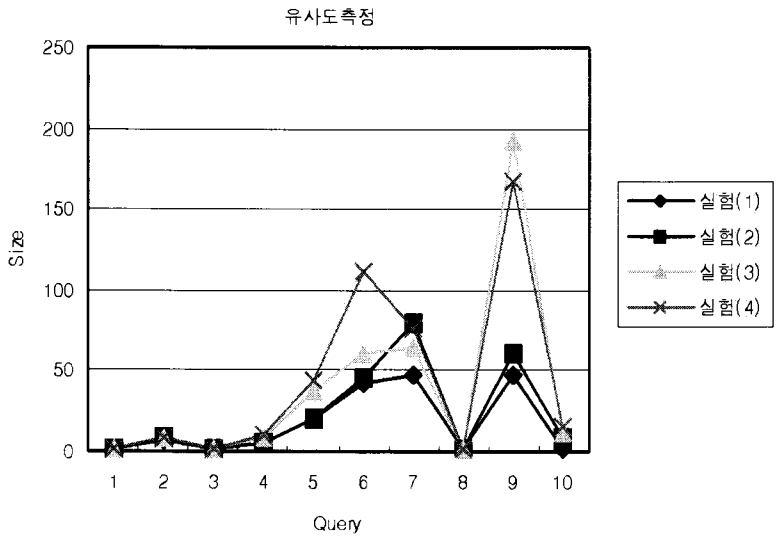
〈표 3〉 실험별 유사 문서 건수

구 분	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
실험(1)	1	7	1	5	20	42	48	1	47	2
실험(2)	1	8	1	5	20	46	80	1	60	8
실험(3)	1	8	1	8	37	61	65	2	193	11
실험(4)	2	9	1	10	44	112	76	2	167	16

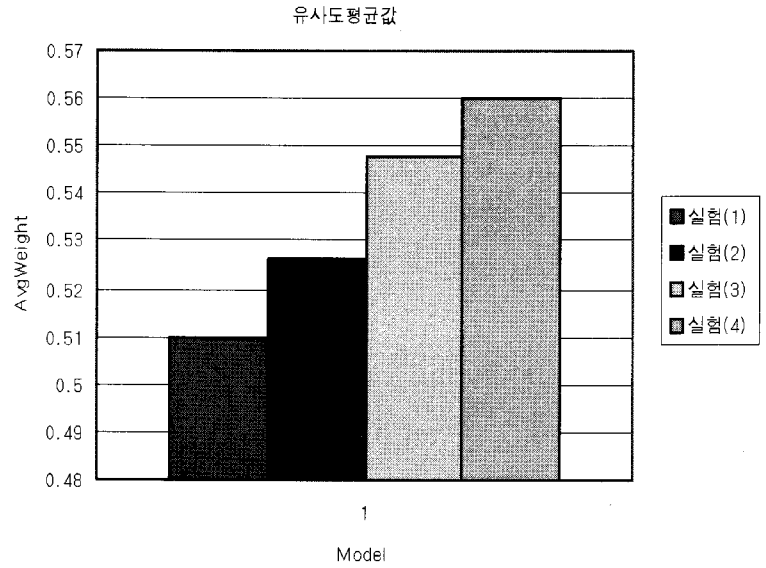
〈그림 6〉은 유사도 값이 30%이상인 것에 대해서 유사 문헌 건수를 나타내고 있다. 후보 색인어의 선정에 따라 문서 건수가 달라질 수 있다. 빈도가 높은 후보 색인어가 선정될 경우에는 이것이 주제어와 관련이 있든 없던 많은 문서들이 나타날 것이다. 위 실험은 후보 색인어의 선정이 기존 것과 다르게 변화되고 있다는 것을 의미하는 것이다.

〈그림 7〉은 전체 유사도 값에 대해서 평균값을 산출한 것이다. 〈그림 7〉그래프에서 나타난 결과와 같이 평가자들이 평가한 결과 값으로 볼 때 실험(4)에 대해서 높은 성능 값을 나타내고 있음을

볼 수가 있다. 즉, 문서를 전체로 보고 TF 정보를 추출하고 추출된 단어들에 대해서 DF 정보를 산출하여 근접한 후보 색인어들에 대해서 가중치를 다시 계산하여 나온 결과가 정확도가 가장 높다는 것을 나타내고 있다.



〈그림 6〉 실험별 유사도 별 건수



〈그림 7〉 전체 유사도 값의 평균값

2. 주제어 정보를 이용한 후보 색인어 선정

가. 실험 방법

앞의 실험의 인접한 단어들을 이용한 후보 색인어 선정하는 방법이 더 좋은 성능을 보여준다는 것을 알 수 있었다. 이번 실험은 앞의 실험에서 가장 성능이 좋은 방법을 이용하고, 키워드 주제어 정보를 이용한 후보 색인어 선정 방법에서 후보 색인어에 대해서 가중치를 추가 하였다. 즉 후보 색인어들 중에서 키워드 주제어 정보와 일치하는 용어들을 선별하는 작업이다. 후보 용어들 중에서 키워드 주제어와 일치하는 용어들에 대해서 다시 가중치를 추가하는 방식을 사용하였다.

〈식 4〉 $$w(d_{i,k_j}) = tfidf_{ij} + distWeight_j + \alpha$$

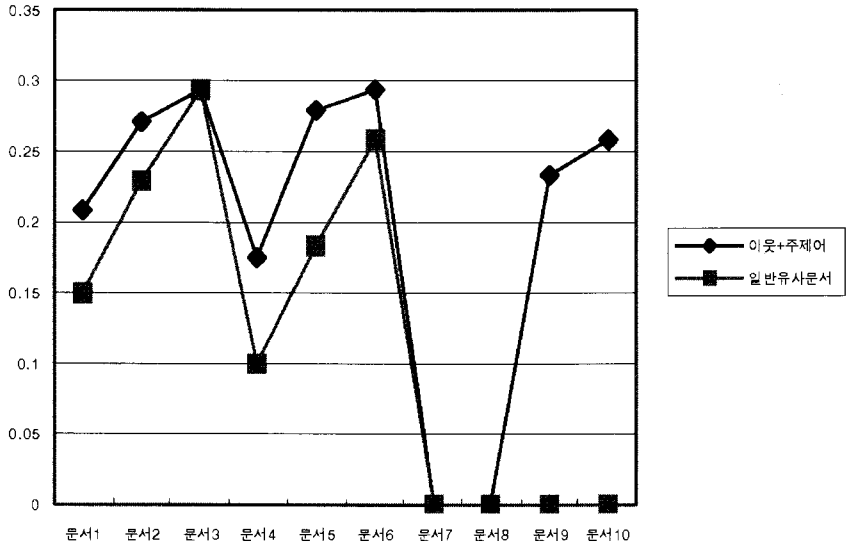
〈식 4〉와 같이 키워드 주제어 정보와 일치하는 후보 색인어들에 대해서는 0~1.0 사이 값인 알파 값을 추가하게 된다. 본 실험에서는 전체 가중치의 평균값을 할당하였다. 즉, 키워드 주제어 정보의 단어가 있을 경우에는 주제어 정보의 단어가 후보 색인어로 선정될 수 있도록 전체 후보 단어들의 평균값을 할당하였다. 평균값, 로그 값, 루트 값 중에서 실험을 통해서 볼 때 평균값을 택한 경우가 가장 유사도가 높게 반영됨으로 그 값을 선택하였다.

〈식 5〉 $$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}$$

또한 위의 실험을 위해서 평가자가 직접 유사한 문서들에 대해서 검색을 하여 평가를 하였다. 즉, 평가자가 직접 검색한 유사 문헌에 대해서 맞는다고 생각하는 것을 〈식 5〉과 같이 역순위평균(Mean Reciprocal Rank) 방식을 사용하여 사용자가 평가를 하였다. 평가자가 임으로 10개의 문서를 선택하여 10개의 문서에 대해서 유사 문헌 검색을 수행한다. 수행한 결과에 대해서 상위 1~10위 안에 유사한지 문서가 있는지 여부를 MRR 방식으로 평가하여 점수를 계산하였다.

나. 실험 결과

추출된 후보 색인어들 중에서 주제어 정보를 이용하여 가중치를 추가하는 실험을 수행하였다. 실험(4)에 대해서 10개의 문서를 선정하여 MRR 평균값을 3명의 평가자들이 각각 10개의 문서를 선정하여 사용자가 유사 문서를 검색하고 검색된 결과 상위 1~10개의 문서에 대해서 MRR 평균값을 계산하였다. 첫 번째 평가자가 평가 결과는 〈그림 8〉과 같다.



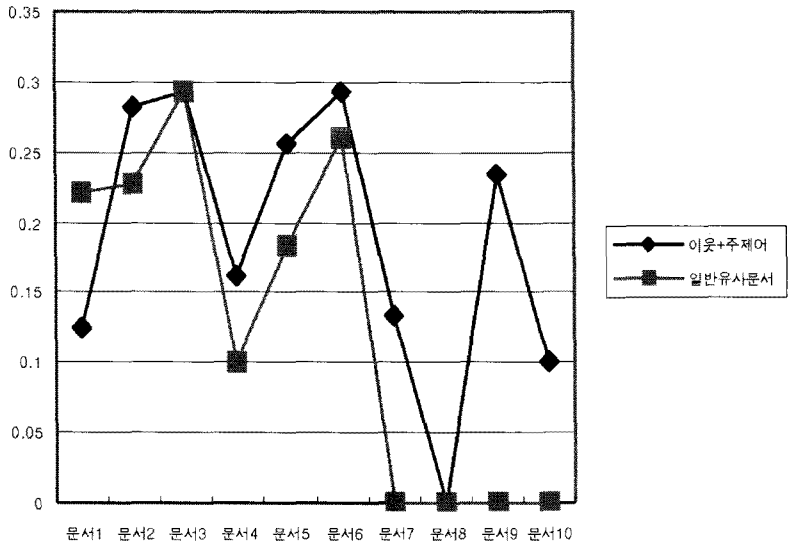
〈그림 8〉 상위 10개의 문서에 대해서 MRR 평균값

〈표 4〉 MRR 평균값

10위 MRR	이웃단어+주제어	일반유사문서
문서1	0.208	0.150
문서2	0.271	0.228
문서3	0.292	0.292
문서4	0.175	0.100
문서5	0.278	0.183
문서6	0.292	0.259
문서7	0	0
문서8	0	0
문서9	0.232	0
문서10	0.259	0

본 논문에서는 상위 1~10위에 정답 문서가 있는 지를 평가하는 방법으로 문서1은 일반 문서에 비해서 13%가 더 나타날 확률이 높다. 그러므로 〈그림 8〉과 〈표 4〉와 같이 전체적으로 평균 MRR 값이 0.07% 정도인 향상되는 것을 보인다. 첫 번째 평가자 평가의 결과에서 나타난 것처럼 주제어에 관련된 후보 색인어를 선정함으로써 유사한 문서들이 상위에 랭킹 되는 것을 알 수 있다. 즉 일반 유사 문헌 방식보다 평균 MRR값들이 전체적으로 높게 평가된다는 것을 알 수가 있다.

두 번째 평가자의 결과는 〈그림 9〉 및 〈표 5〉와 같다.

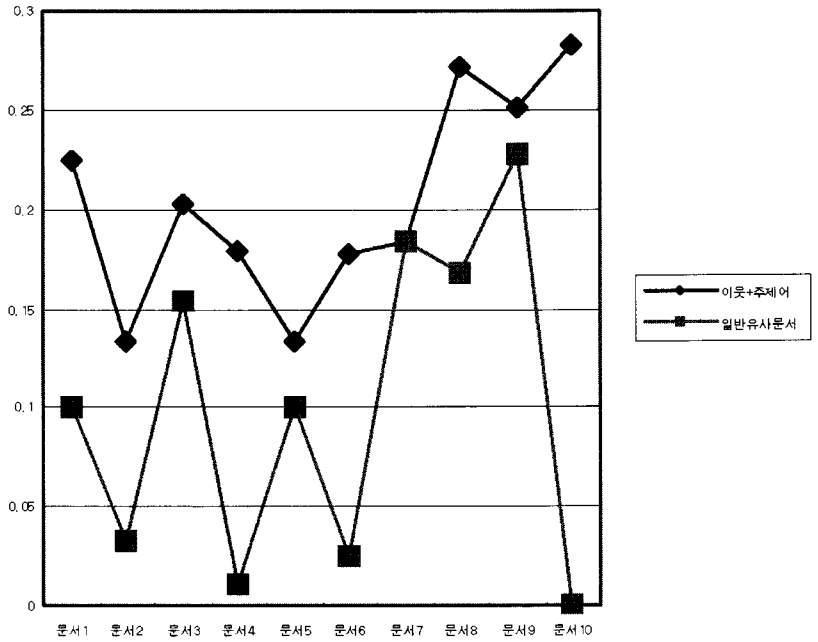


〈그림 9〉 상위 10개의 문서에 대해서 MRR 평균값

〈표 5〉 MRR 평균값

10위 MRR	이웃단어+주제어	일반유사문서
문서1	0.225	0.100
문서2	0.133	0.033
문서3	0.203	0.153
문서4	0.179	0.011
문서5	0.133	0.100
문서6	0.176	0.025
문서7	0.183	0.183
문서8	0.271	0.168
문서9	0.251	0.228
문서10	0.282	0

두 번째 평가자의 평가에서는 상위 10개의 문서에 대해서 평균 MRR값이 0.06으로 향상 된 것을 볼 수 있었다. 위의 두 번째 평가자의 평가의 결과에서 나타난 것처럼 주제어에 관련된 후보 색인어를 선정함으로써 유사한 문서들이 상위에 랭킹 되는 것을 알 수 있다. 즉 일반 유사 문서 방식보다 평균 MRR값들이 전체적으로 높게 평가된다는 것을 알 수가 있다. 세 번째 평가자의 결과는 〈그림 10〉 및 〈표 6〉과 같다.



〈그림 10〉 상위 10개의 문서에 대해서 MRR 평균값

〈표 6〉 MRR 평균값

10위 MRR	이웃단어	일반유사문서
문서1	0.225	0.100
문서2	0.133	0.033
문서3	0.203	0.153
문서4	0.179	0.011
문서5	0.133	0.100
문서6	0.176	0.025
문서7	0.183	0.183
문서8	0.271	0.168
문서9	0.251	0.228
문서10	0.282	0

세 번째 평가자의 평가에서는 상위 10개의 문서에 대해서 평균 MRR값이 0.07로 향상 된 것을 볼 수 있었다. 3명의 평가자들의 평가에서 유사 문헌 검색을 위해서 추출된 색인어 중에서 주제어 정보에 관련된 단어를 선정하는 것이 좋은 결과를 제공하고 있다는 것을 나타내고 있다.

유사 문헌 검색을 위해서 추출된 많은 색인어 중에서 어느 것을 선정하느냐에 따라서 검색 결과가 많이 상이할 수 있다. 하지만 키워드 주제어 정보를 이용하면 그 문서의 주요한 키워드를 후보 색인

어로 선정하여 유사 문헌 검색에 반영이 되어 정확도의 유사 문헌 검색 결과를 제공할 수 있었다.

3. 관련 저자 정보를 이용한 가중치

가. 실험 방법

유사한 문헌을 찾기 위해서 관련된 저자 정보를 추가하여 보다 성능을 향상시키는 방법을 이용하였다. 같은 사람들이 관련된 유사한 논문을 작성하는 경우가 많다.

본 연구에서는 이러한 유사한 사용자들의 정보를 이용하여 실험하였다. 검색한 결과에서 관련 저자 정보를 추가함으로써 보다 높은 성능의 유사 문헌을 검색하도록 하였다. 가중치를 계산하는 방식은 기존의 유사 문헌 검색에 가중치 값과 관련 저자 정보의 가중치를 합한 값을 2로 나누어 전체 1.0의 가중치를 가지도록 처리 하였다.

〈식 6〉과 같이 W(d)는 유사 문서의 가중치이며 알파 값은 관련 사용자 정보들이 나타나는 문서들의 값이다.

〈식 6〉 $$Weight_{total}(d) = \frac{(W(d)+\alpha)}{2} \quad 0<\alpha\leq 1$$

〈식 6〉과 같이 알파 값은 0에서 1.0사이의 값을 가진다. 최댓값은 1.0을 받게 된다. 본 연구에서 저자별로 가중치를 다르게 계산을 하였다. 첫 번째와 마지막(교신) 저자의 가중치를 가장 높게 계산을 하고 중간 저자들의 가중치는 첫 번째와 마지막 저자의 가중치보다 낮다.

만약 유사도 가중치 값이 1.0이고 관련 저자 정보 가중치 값이 1.0이 되면 최종 가중치 값은 1.0이 되는 방식을 사용하여 관련 저자 정보를 이용하여 가중치를 더하는 방식을 취하였다. 즉, 검색한 결과에서 관련된 저자들이 나타날 때는 그 문서를 상위로 재순위화 시키기 위한 알고리즘이다. 같은 저자와 공동 저자들이 비슷한 주제 분야를 연구할 가능성이 높기 때문이다. 그러므로 위와 같은 가중치를 이용하여 유사 문헌으로 검색된 결과들에 대해서 다시 한 번 재순위화를 시킴으로써 정확도가 높은 유사 문헌을 제공할 수 있다.

또한 위의 실험을 위해서 평가자가 직접 유사한 문서들에 대해서 검색하고 평가를 하였다. 즉, 평가자가 직접 검색한 유사 문서에 대해서 맞는다고 생각하는 것을 〈식 5〉와 같이 역순위평균(Mean Reciprocal Rank) 방식을 사용하여 평가를 하였다.

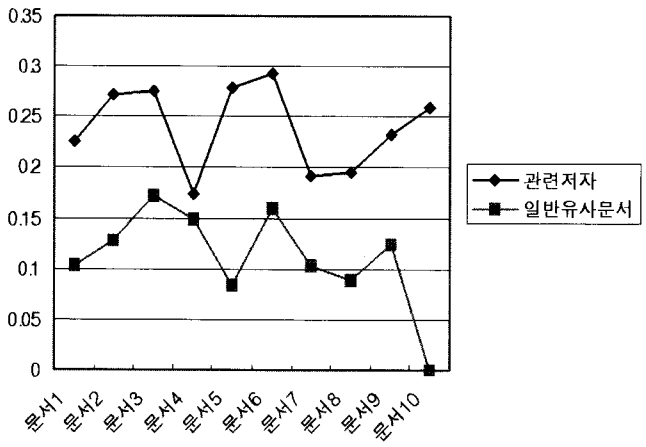
나. 실험 결과

실험을 위해서 10개의 문서를 선정하여 MRR 평균값을 3명의 평가자들이가 직접 평가를 하였

다. 유사 문서 결과 중에서 상위 10개의 문서를 선정하여 사용자가 유사 문서만을 이용한 검색과 관련 저자 정보를 동시에 이용한 유사 문서 검색된 결과 상위 1~10개의 문서에 대해서 MRR 평균값을 계산하였다. 첫 번째 평가자의 평가 실험 결과는 〈그림 11〉과 같다.

〈표 7〉 관련 저자 MRR 평균값

10위 MRR	관련저자	일반유사문서
문서1	0.225	0.103
문서2	0.271	0.128
문서3	0.276	0.172
문서4	0.175	0.150
문서5	0.278	0.0833
문서6	0.292	0.159
문서7	0.191	0.103
문서8	0.195	0.089
문서9	0.232	0.125
문서10	0.259	0



〈그림 11〉 상위 10개의 문서에 대해서 MRR 평균값

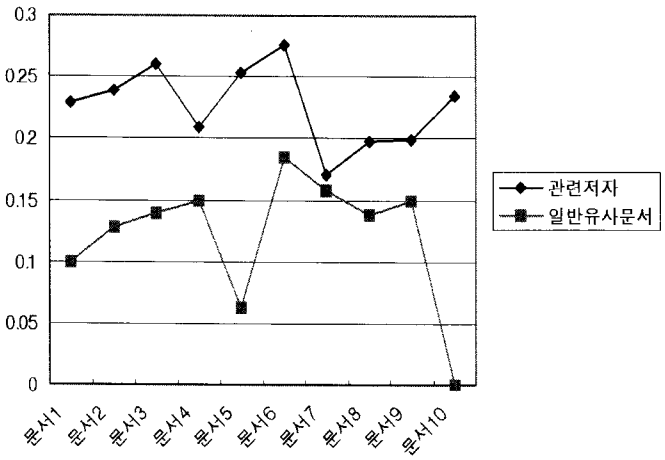
첫 번째 평가자의 평가에서는 상위 1~10개의 문서에 대해서 평균 MRR값이 0.116으로 향상 된 것을 볼 수 있었다. 〈그림 11〉과 같이 전체적으로 관련된 저자들 문서를 상위에 분포시킴으로써 사용자들은 관련된 비슷한 연구 주제를 쉽게 찾을 수가 있다. 첫 번째 평가자의 평가 결과에서 나타난 것처럼 관련 저자들의 정보를 사용함으로써 유사한 문서들이 상위에 랭킹 되는 것을 알 수 있다. 즉 일반 유사 문서 방식보다 평균 MRR값들이 전체적으로 높게 평가된다는 것을 알 수가 있다.

두 번째 평가자의 평가에서도 10개의 문서를 선정하여 MRR 평균값을 평가자가 평가를 하였다.

10개의 문서를 선정하여 사용자가 유사 문서를 검색하고 검색된 결과 상위 1~10개의 문서에 대해서 MRR 평균값을 계산하였다. 실험 결과는 〈그림 12〉와 같다.

〈표 8〉 관련 저자 MRR 평균값

10위 MRR	관련저자	일반유사문서
문서1	0.228	0.100
문서2	0.238	0.128
문서3	0.259	0.139
문서4	0.209	0.150
문서5	0.253	0.062
문서6	0.276	0.184
문서7	0.170	0.158
문서8	0.197	0.137
문서9	0.199	0.150
문서10	0.234	0



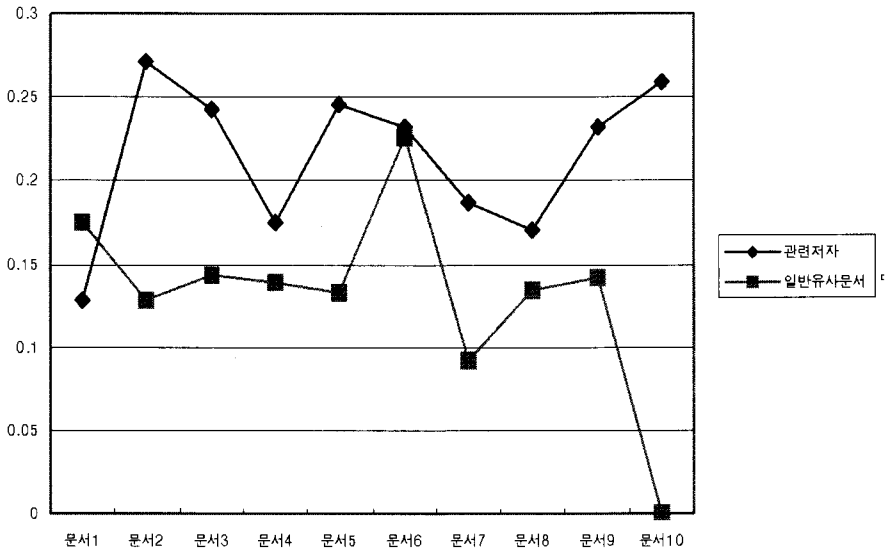
〈그림 12〉 상위 10개의 문서에 대해서 MRR 평균값

두 번째 평가자의 평가에서는 상위 10개의 문서에 대해서 평균 MRR값이 0.082로 향상 된 것을 볼 수 있었다. 또한 두 번째 평가자의 평가 결과에서 나타난 것처럼 관련 저자들의 정보를 사용함으로써 유사한 문서들이 상위에 랭킹 되는 것을 알 수 있다. 즉 일반 유사 문서 방식보다 평균 MRR 값들이 전체적으로 높게 평가된다는 것을 알 수가 있다.

마지막으로 세 번째 평가자의 평가에서도 10개의 문서를 선정하여 MRR 평균값을 평가자가 평가를 하였다. 10개의 문서를 선정하여 사용자가 유사 문서를 검색하고 검색된 결과 상위 1~10개의 문서에 대해서 MRR 평균값을 계산하였다. 실험 결과는 〈그림 13〉과 같다.

〈표 9〉 관련 저자 MRR 3차 평균값

10위 MRR	관련저자	일반유사문서
문서1	0.128	0.175
문서2	0.271	0.128
문서3	0.242	0.142
문서4	0.175	0.139
문서5	0.245	0.133
문서6	0.232	0.225
문서7	0.186	0.091
문서8	0.170	0.134
문서9	0.232	0.141
문서10	0.259	0



〈그림 13〉 상위 10개의 문서에 대해서 MRR 3차 평균값

세 번째 평가자의 평가에서는 상위 10개의 문서에 대해서 평균 MRR값이 0.064로 향상 된 것을 볼 수 있었다. 마지막으로 세 번째 평가자의 평가에서 문서 10개를 선정하여 유사 문서를 수행한 결과 위의 그림과 같다. 3명의 평가자들의 평가에서 유사 문헌 검색을 수행한 결과 중에서 관련 저자 정보를 이용하여 관련된 저자들의 문서를 상위로 높여 주는 방식을 제공함으로써 보다 좋은 결과가 나타났다.

본 연구에서는 KISTI의 과학기술 학회마을에서 서비스 중인 논문들에 대해서 유사한 문헌 검색 시스템을 제공하기 위한 정보 검색 시스템이다. 논문들은 보통 관련된 저자들이 비슷한 논문들을 계속해서 작성한다는 점이다. 이 점을 이용하여 본 연구에서는 유사 문헌 검색 결과에 반영하여

정확도의 유사 문헌 검색 결과를 제공할 수 있었다.

V. 결론 및 제언

본 연구는 유사 문헌 검색 시스템의 성능 향상을 위한 새로운 모델을 제안하고, 현재 서비스되고 있는 과학기술 학회마을 논문 정보서비스에 적용하여 정확도가 높은 유사 문헌을 찾아 주는 정보 시스템을 개발하고 평가하였다. 즉 표절이 의심되는 논문을 찾는 것보다 사용자가 검색한 문서와 관련된 가장 유사한 주제를 가진 문서를 제공하는 방법을 실험하였다.

연구 결과를 정리하면 다음과 같다. 첫째, 유사 문헌을 판별하는 위해서 색인어로 추출된 단어들을 인접한 단어로 선정함으로써 보다 정확성을 높였다. 즉 문서에서 추출된 후보 색인어 중에서 단어 간의 연결성 정보를 유지하는 것이 더 정확한 유사 문헌을 검색할 수 있었다.

다음으로 키워드 주제어 정보를 이용하여 후보 색인어들이 주제 정보와 가까운 용어들이 선정될 수 있도록 함으로써 유사 문헌 검색 결과에 반영하여 정확도가 높은 유사 문헌 검색 결과를 제공할 수 있었다. 키워드 주제어 정보를 선정하는 방식을 이용함으로써 내용 기반의 유사한 문헌 정보를 제공할 수 있었다. 이 실험에서는 주제어 정보와 인접한 단어 간의 정보를 이용하여 보다 정확한 후보 색인어를 선정하도록 하였다. 그 결과 유사 문헌에서 평균 MRR값이 0.06~7정도 향상되는 것을 볼 수 있었다.

마지막으로 일반적으로 논문서지 DB에 기술되어 있는 공동 저자들이 관련된 유사한 연구를 할 경우가 매우 높다는 점을 이용하였다. 유사 문헌 검색 결과에서 관련 저자들이 나타나는 문서들을 상위로 랭킹 시킬 수 있는 모델을 제시함으로써 유사 문헌 검색 시스템에 보다 높은 성능의 유사 문헌 검색 시스템을 제공할 수 있었다. 실험 결과로는 평균 MRR값이 0.064~0.116으로 향상되는 것을 볼 수 있었다.

현재 대부분의 논문 검색 시스템에서는 원문 전체를 이용한 검색과 유사 문헌 검색 방법을 사용하고 있다. 보다 정확한 주제어를 추출하기 위하여 논문서지 DB에 기술된 정보를 이용하는 것보다 원문 전체를 이용하여 후보 색인어를 선정하는 방법과 주제어가 제목, 초록, 본문 등에 나타나는 정보에 따라 가중치를 다르게 하여 유사도를 측정하는 모델 등의 연구도 요구된다. 그리고 유사 문헌 후보 정답 집합을 구축하여 정확한 정확도를 평가할 필요가 있다.

〈참고문헌은 각주로 대신함〉