# 유사과제파악을 위한 검색 알고리즘의 개발에 관한 연구

## A Study on the Development of Search Algorithm for Identifying the Similar and Redundant Research

박동진*, 최기석**, 이명선**, 이상태***
공주대학교 산업시스템공학과 교수*, 한국과학기술정보연구원**, 한국표준과학연구원***

Dong-Jin Park(mispdj@kongju.ac.kr)*, Ki-Seok Choi(choi@kisti.re.kr)**,
Myung-Sun Lee(mslee@kisti.re.kr)**, Sang-Tae Lee(stlee@kriss.re.kr)***

**요약**

국가적으로 그리고 각 연구기관에서는 투자의 효율성을 기하기 위하여 연구사업 선정과정에서 데이터베이스로부터 중복과제 혹은 유사과제를 검색하는 과정을 거친다. 최근 부얼리언 기반의 키워드 매칭 검색알고리즘의 발전 및 이를 채택한 검색엔진의 개발로 인하여 검색의 정확도가 많이 향상되었지만, 사용자가 입력하는 제한된 수의 키워드들에 의한 검색은 유사과제 파악과 우선순위의 결정에 어려움이 있다. 본 연구에서는 제안된 과제의 문서를 분석하여 다수의 색인어들을 추출하고, 이들에게 가중치를 부여한 후, 기존의 문서들과 비교하여 유사과제를 찾아내는 문서단위의 검색 알고리즘을 제안한다. 구체적으로 벡터공간검색(Vector-Space Retrieval)모델의 한 종류인 TFIDF(Term Frequency Inverse document Frequency)를 기본 구조로 채택한다. 또한 개발되는 알고리즘에는 연구과제 제안문서의 구조에 적합한 속성별 가중치(feature weighting)를 반영하고 검색속도의 향상을 위하여 K-최근접 문서(KNN: K-Nearest Neighbors) 기법도 반영한 알고리즘을 제시한다. 실험을 위하여 실제 연구제안 문서와 구조가 동일한 기존의 보고서를 사용하였는데, KISTI에서 운영하는 과학기술정보포털서비스인 NDSL에서 이미 분류해 놓은 4분야의 1,000 개 연구 보고서 문서를 발췌하여 실험을 하였다.

■ 중심어 : | 유사과제 | 검색엔진 | TFIDF | K-최근접법 |

**Abstract**

To avoid the redundant investment on the project selection process, it is necessary to check whether the submitted research topics have been proposed or carried out at other institutions before. This is possible through the search engines adopted by the keyword matching algorithm which is based on boolean techniques in national-sized research results database. Even though the accuracy and speed of information retrieval have been improved, they still have fundamental limits caused by keyword matching. This paper examines implemented TFIDF-based algorithm, and shows an experiment in search engine to retrieve and give the order of priority for similar and redundant documents compared with research proposals. In addition to generic TFIDF algorithm, feature weighting and K-Nearest Neighbors classification methods are implemented in this algorithm. The documents are extracted from NDSL(National Digital Science Library) web directory service to test the algorithm.

■ keyword : | Similar Redundant Proposal | Search Engine | TFIDF | KNN |

접수번호 : #090909-001
접수일자 : 2009년 09월 09일
심사완료일 : 2009년 09월 16일
교신저자 : 이명선, e-mail : mslee@kisti.re.kr

# I. 서론

## 1.1 연구의 필요성

국가적으로 그리고 각 연구기관에서는 R&D 투자의 확대와 아울러 투자의 효율성을 기하기 위하여 연구사업 선정과정에서 중복과제 혹은 유사과제를 검토하는 과정을 거친다. 현재는 사업선정위원들이 각자의 전문지식을 기반으로 각 기관에서 보유하고 있는 데이터베이스를 활용하거나, 연구성과관리 검색엔진의 검색결과에 의존하여 유사(중복)과제를 파악하고 있다. 그러나 이러한 방식은 선정위원들이 첫째, 기관에 국한된 자료에 근거하거나, 둘째, 키워드 매칭 검색결과의 단점인 제한된 혹은 너무 광범위한 정보를 기반으로 판단해야 하므로 중복과제를 정확히 파악하기 어렵다.

따라서 국가가 지원하는 연구개발지원사업에 있어서 동일, 유사한 연구과제를 정부 각 부처에 중복 제안하여 지원받는 것을 사전에 방지하고 국가연구개발 투자의 효율성을 제고하기 위해서는 먼저 검색대상이 되는 연구성과관리를 위한 통합시스템이 필요하다. 또한 연구자의 입장에서도 연구주제와 관련하여 다른 연구자에 의해서 제안되었거나, 현재 혹은 과거에 수행된 유사한 연구과제를 사전에 파악하는 것도 연구를 기획하는데 있어서 매우 중요하다.

최근 부얼리언 기반의 키워드 매칭 검색알고리즘의 발전 및 이를 채택한 검색엔진의 개발로 인하여 검색의 정확도가 많이 향상되었지만, 사용자가 입력하는 제한된 수의 키워드들과 이들을 포함하는 문서들을 검색하는 것만으로 유사과제문서를 파악하는 것에 다음과 같은 문제점이 있다. 첫째, 입력하는 몇 개의 키워드가 제안하는 문서를 대표할 수가 없다. 둘째, 키워드의 정확한(exact) 매칭의 결과만을 제공하는 부얼리언 알고리즘에서는 너무 많은 수의 혹은 너무 적은 수의 결과가 나올 수 있다. 셋째, 제목, 저자, 요약문 등과 같이 연구과제 제안문서에 포함된 구조적인 특성을 검색시 반영할 수 없다. 예를 들면 한국과학기술정보연구원(KISTI)에서 운영중인 과학기술정보서비스[1][2]나 정보통신연구진흥원(IITA)의 중복지원방지시스템(Naris)[3]시스템에서는 키워드 매칭에 따른 검색을 통해서만 가능하며 이상과 같은 문제점들을 포함하고 있다.

## 1.2 연구의 목적 및 방법

본 논문은 국가과학기술종합정보서비스(NTIS)의 하위시스템 중의 하나인 유사과제 검색시스템에 필요한 알고리즘을 연구하는 데 있다. 연구성과물에 대한 단순 키워드매칭 기반의 검색을 지양하고, 연구과제 제안서 문서를 기반으로 기존의 연구 중인(혹은 완료된) 과제의 문서 및 연구성과물을 검색하여 유사한 것들을 찾아내는 알고리즘이다. 즉, 본 연구에서는 문서단위의 비교 검색을 함으로써 키워드 매칭 검색엔진의 근본적인 문제점을 해결한다.

본 연구에서는 문서의 색인어들에 대한 가중치를 부여하는 벡터공간검색(Vector-Space Retrieval)모델의 한 종류인 TFIDF(Term Frequency Inverse document Frequency)를 기본 구조로 채택한다. 또한 개발되는 알고리즘에는 연구과제 제안문서의 구조에 적합한 속성별 가중치(feature weighting)를 반영하고 검색속도의 향상을 위하여 K-최근접 문서(KNN: K-Nearest Neighbors) 기법을 반영한 알고리즘을 제시한다. 본 연구를 통해서 개발된 알고리즘은 유사과제를 검색하는 시스템에 적용될 수 있을 뿐 아니라 디렉토리 검색서비스를 위한 자동문서 분류시스템에도 적용가능하다. 본 연구는 서론에 이어 제2장에서는 이론적 배경을 기술하였고, 제3장에서는 알고리즘의 개발을, 그리고 4장에서는 실험 및 알고리즘의 평가를 다루고, 마지막으로 제5장에서는 결론 및 연구의 한계점을 기술한다.

# II. 이론적 배경

## 2.1 TFIDF

색인어에 가중치를 부여하는 정보검색 모델은 크게 부울리언 모델과 벡터모델이 있다. 본 연구에서는 문서의 색인어들의 가중치 부여를 위해서 벡터공간검색의 한 종류인 TFIDF를 채택한다[4]. TFIDF는 부분매칭을 가능하게 하고, 부울리언 방법과는 달리 비이진 가중치를 부여하고 이 가중치를 이용해서 유사도 점수를 계산

한다. TFIDF에서는 하나의 문서 i 가 N개의 색인어를 가지면 N차원의 벡터인 i = < Wi1, Wi2, Wi3, Wi4, WiN-1, WiN > 로 표현한다. Wik는 i 문서의 색인어 k의 가중치이며, 아래와 같은 식으로 계산할 수 있다.

$$W_{ik} = tf_{ik} \log(\frac{\text{number of document}}{n_k})$$

Wik = 문서 i안의 색인어 k의 가중치

nk = 색인어 k를 포함하는 전체 문서의 수

tfik = 문서 i 안에서 색인어 k 가 발생된 수 / 문서 i 안에서 가장 많이 발생된 색인어의 수, 즉 fik/maxfik 가 된다.

본 연구에서는 입력되는 질의 문서와 데이터베이스에 저장된 문서(검색 문서) 중에서 가장 유사한 문서를 찾아내는 것이다. 질의 문서와 검색 문서간의 유사도는 아래와 같이 각 문서의 색인어 가중치 벡터를 이용해서 계산한다[5].

$$sim(D_i, D_j) = \frac{\sum_{k=1}^{N} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{N} w^2_{ik} \sum_{k=1}^{N} w^2_{jk}}}$$

Di, Dj = 문서 i, j

N = 색인어의 수

Wik, Wjk = 문서 i, j 의 색인어 가중치 벡터

## 2.2 K-NN Feature Weighting TFIDF 방식

모든 검색대상 문서에 대하여 유사도를 계산하는 것은 지나치게 많은 시간이 소요된다. 따라서 유사도가 높을 것으로 예상되는 문서를 제한된 수만큼 발췌하여 유사도를 계산하고 가장 높은 유사도 값을 가진 검색문서를 제시하는 방법들이 개발되었다. K-NN TFIDF 방식이 그 중의 하나이다[6-8].

문서간의 유사도를 계산하는데 있어서 각 속성들은 서로 다른 중요도를 가진다[9]. 예를 들면 상식적으로 두문서간에 제목 속성에 소속된 색인어들이 서로 매칭되면 다른 속성들에 비해서 더 많은 가중치를 부여하는 것이 바람직하다. 다기준 의사결정문제를 해결하는 기법중의 하나인 계층분석과정(AHP: Analytical Hierarchy Process)방법을 이용하여 각 속성에 대한 중요도를 파악할 수 있다. AHP 방법을 통하여 각 속성들 간의 중요도를 파악하여 유사도 계산시 반영하는 것이 고려될 필요가 있다. 일반적인 제안서(계획서) 문서의 구조는 [표 1]과 같으며 각 항목은 문서의 속성(feature)으로 파악된다. 유사문서 검색시스템에서는 전산으로 입력된 항목 중 분별력이 강한 항목을 선택하여야 한다.

**표 1. 제안서 항목**

| 항목 \ 방식 | 전산 입력 | 문서 제출 |
|---|---|---|
| 사업구분 | ○ | ○ |
| 제목(한글과 영문) | ○ | ○ |
| 연구책임자 및 연구원 | ○ | ○ |
| 연구기관 | ○ | ○ |
| 연구기간 | ○ | ○ |
| 연구의 목적 | ○ | ○ |
| 연구요약 | ○ | ○ |
| 키워드 | ○ | ○ |
| 연구내용(상세내용) | | ○ |
| 예산계획 | ○ | ○ |

# III. 실험환경 및 알고리즘의 개발

## 3.1 실험환경

본 연구에서는 연구과제 제안서를 저장하고 있는 데이터베이스 확보의 어려움으로 인해서 제안서의 주요 구성항목과 일치하는 연구보고서 데이터베이스를 활용한다. 구체적으로 KISTI에서 운영하는과학기술정보포털서비스인 YesKiSTi(현재의 NDSL)의 디렉토리 검색서비스인 "표준주제검색" 서비스에서 문서를 발췌하였다. 본 연구에서는 아래와 [표 2]와 같이 문서를 발췌하였다. "표준주제검색"의 대분류인 기계금속분야(BA), 전기전자분야(BB), 정보통신분야(BI), 화학화공분야(BK), 생명분야, 환경건설분야, 이상 6개 분야 중 코드화한 4개의 각 분야에 250개의 문서, 총 1,000개의 문서를 발췌하였다. <첨부 1> 기계금속분야의 코드체계이다.

표 2. 실험 문서 분류구조

| 대분류 | 중분류 | 소분류 | 소속 문서수 | 대분류별 총문서수 |
|---|---|---|---|---|
| BA | 5 | 5 | 10 | 250 |
| BB | 5 | 5 | 10 | 250 |
| BI | 5 | 5 | 10 | 250 |
| BK | 5 | 5 | 10 | 250 |
| | | | | 1000 |

코드체계: 대분류(BA, BB, BI, BK),
중분류(01~05), 소분류(01~05),
문서(01~10)
문서의 샘플코드 BA010101, BI030210, BK050510

각 문서의 속성은 제목, 저자, 정보출처, 초록으로 한정하였다. 현재 소분류 별로 소속된 문서들은 전문가가 분류한 것이 아니라, KISTI 검색엔진을 이용한 몇 개의 핵심검색어 매칭에 의해서 순위 없이 나온 것들이다. 따라서 포함된 문서가 정확히 해당 분야에 소속된다고 볼 수 없으며, 분야별 적합도를 보여주고 있지 않다. 따라서 본 연구에서는 각 대분류 분야 전공별로 자문교수를 초빙하여 문서를 재분류 하였다. 자문교수단에 의한 문서분류는 첫째, 각 문서의 대·중·소분류의 소속을 확정하였다. 둘째, 각 소분류에 소속된 문서들 10개 중에서 첫 번째 문서(문서코드 xx-xx-xx-01에 해당되는 문서)를 기준으로 나머지 9개를 전문가의 판단으로 유사도가 높은 순위를 매겼다. 이는 추후 알고리즘의 정확도를 판단하는데 기준이 될 것이다.

소프트웨어 환경은 다음과 같다. 형태소분석을 거쳐 Excel 포맷으로 된 데이터를 통합하여 Access 테이블로 전환하고 이를 불용어를 처리한다. 불용어가 처리된 데이블을 학습테이블 및 질의테이블로 구분한 후 각 색인어에 대한 가중치(weight)를 계산한다. 실험용 S/W는 window2000 서버에서 Visual Basic으로 응용프로그램을 작성하고, 데이터 처리는 SQL 서버 DBMS를 이용한다.

## 3.2 알고리즘의 개발

알고리즘은 크게 3단계를 거친다. 첫째는 색인화를 하는 것이며, 둘째는 가중치를 구하고, 마지막으로 유사도를 계산하는 것이다.

### 3.2.1 자동 색인화(Automatic Indexing)

본 단계에서는 한글 형태소 분석 소프트웨어 모듈을 사용하여 단순히 조사와 부사, 접속사, 그리고 의미를 갖지 않는 기호를 제거한다. 유사문서 검색시스템은 먼저 신규 제안서 문서에 포함된 단어를 추출하기 위해서 불용어를 제거하는 형태소 분석이 먼저 이루어진다. 한글 형태소 분석 소프트웨어인 HAM: Hangul analysis Module, KLT Version 2.1b 등을 이용하여 형태소 분석을 할 수 있다.

기본적인 한글 불용어 사전외에 연구제안서 도메인 분야에서 필요한 불용어 리스트를 구성하여야 한다. 형태소분석이 끝난 색인어(학습용 테이블) 46,556 단어 중 상위 15%에 해당되는 고빈도 색인어를 불용어 처리 기준으로 설정하였다. 예를 들면 "연구개발, 내용, 증가, 확립, 변화, 요구, 제시, 시스템 ..." 등과 같은 단어들은 R&D 도메인에서 너무나 많이 사용되는 단어이므로 식별력이 없기 때문에 이들을 불용어로 처리한다는 것이다. 그러나 그 중에서 식별 색인어로서 가치가 있다고 판단되는 색인어 들은 불용어로 처리하지 않았다. 결국 14.3%를 불용어로 처리하고 총 40,156 색인어를 최종적으로 선택하였다. <첨부 2>는 연구과제 제안 도메인의 불용어 리스트이다.

### 3.2.2 가중치 부여(Calculating Weight)

형태소분석과 불용어를 처리한 후 실험을 위한 최종 테이블로 학습용 테이블과 질의용 테이블을 준비한다. 학습용 테이블은 검색대상이 되는 문서들의 색인어를 포함하는 테이블이며, 질의용 테이블은 테스트를 하기 위한 문서들의 색인어를 포함하는 테이블이다. 실험에 사용된 데이블의 구조는 다음 [표 3]과 같으며, 기타 최종적으로 가중치(weight)를 계산하기 위하여 다수의 임시테이블이 존재한다.

표 3. 실험테이블의 구조

| 필드이름 | 설명 | 학습용 테이블 | 질의용 테이블 |
|---|---|---|---|
| serial_number | 일련번호 (PK) | ○ | ○ |
| container | 문서번호 | ○ | ○ |
| word | 색인어 | ○ | ○ |
| class | 세부소속번호 | ○ | ○ |

| | | | |
|---|---|---|---|
| tf | 단어반복회수 | ○ | ○ |
| max_tf | 단어최대반복회수 | ○ | ○ |
| normalized_tf | 정규화단어회수(tf/max_tf) | ○ | ○ |
| in_docs | 단어포함문서수 | ○ | |
| total_number_of_documents | 전체문서수 | ○ | |
| df | 출현비중(in_docs/전체문서수) | ○ | |
| idf | 역출현비중 (log(전체문서수/in_docs)) | ○ | ○ |
| weight | 가중치 | ○ | ○ |

필드중 class 필드의 값은 1-4의 값을 가질 수 있는데, 각각 1(제목), 2(저자), 3(소속단체), 4(요약) 이다. 이는 색인어가 문서의 구조상 어디에 소속되느냐에 따라서 값이 결정된다. 이것은 색인어가 어디에 소속되어 있느냐에 따라서 비중을 다르게 하고자 하는 Featuring Weighting TFIDF를 적용하기 위하여 구분해 놓은 것이다. tf 필드값부터 최종적으로 weight 필드값을 구하기 위해서 다수의 SQL 질의문장들이 포함된다.

### 3.2.3 유사도 계산(Calculating Similarity)

여기에서는 질의문서와 저장된 학습문서와의 유사도를 계산하여 유사도가 가장 높은 문서를 보여준다. 알고리즘에서는 먼저 몇 개의 질의 문서를 검색 할 것인가(알고리즘에서 검색할 테스트 문서의 개수)를 결정한다.

다음으로 k-nn 휴리스틱을 적용하는 방법으로 각 질의 문서의 색인어들 중에서 가중치 값이 상위인 k개를 결정한다. 예를 들면 k = 15 라면 질의 문서 중에서 개별 가중치가 높은 15개의 색인어만 선택하여 이를 기준으로 계속 진행한다는 것이다. 다음으로 선택된 k개의 색인어들 중에서 하나라도 같은 단어를 포함하고 있는 학습 문서를 전체 데이터베이스로부터 발췌한다. 발췌된 검색 문서들 중에서 질의 문서에 있던 n 개의 색인어와 같은 것들에 대해서만 다시 모든 색인어들의 가중치 합을 구한다. 즉 발췌된 모든 검색 문서에 대하여 단순 가중치 합을 구하는 것이다. 가중치 합을 기준으로 다시 상위 n 개에 해당하는 검색문서들을 발췌한다. 여기서 발췌된 n 개의 문서와 원래 질의 문서와의 유사도를 계산해서 가장 높은 유사도 값을 가진 문서가 추천된다. 이는 많은 질의문서와 학습문서에 다수의 색인어가 포함되어 있을 때 검색의 질은 유지하며 검색 시간을 급격히 줄이는 효과가 있다. 마지막으로 질의 문서와 발췌된 문서들 각각과의 유사도를 계산해서 높은 유사도를 가진 문서 순서대로 시스템에서 추천한다. 이상의 내용의 알고리즘은 [그림 1]과 같다.

```
For i = 1 to 검색할 질의 문서의 개수
    Get 질의 문서의 색인어, 가중치 from tblFilteredTestWeight_New
    K-nn을 위한 k값을 입력받음
    Call GetKNN(질의문서 레코드)
    For j = 1 to lstScore 리스트의 숫자
      if 테스트 문서의 각 색인어와 비교하여 일치하는 경우
        TFIDF방법에 의한 유사도를 계산하고 누적시킴
    Next j
    비교된 모든 학습문서들간의 유사도를 내림차순으로 정렬
    상위 k개 문서 만큼만 결과로 보여줌
  Next i
GetKNN(질의문서 레코드)
{
 질의문서에 소속된 색인어를 Array에 저장
 학습문서 테이블에서 질의문서에 소속된 색인어와 일치하는 모든
    단어 및 가중치를 발췌
 Array에 발췌된 문서와 단어 그리고 Weight를 저장
 저장된 문서중에서 가중치를 합산(WeightSum)하여 상위 n 만큼의
    문서를 lstScore 리스트박스에 재배열
}
```

그림 1. 유사도계산 알고리즘

## IV. 실험 및 알고리즘의 평가

[그림 2]는 실험을 위한 인터페이스 화면이다. 각 항목에 대한 설명은 다음과 같다. Menu(M)에는 테스트 시작, 결과저장 등의 하위메뉴가 있다. 아래 리스트 박스에서는 테스트 문서를 선택할 수 있다. 100개의 테스트 문서를 한꺼번에 테스트를 할 수도 있으며 10개의 문서를 한단위로 실험을 실시 할 수도 있다.

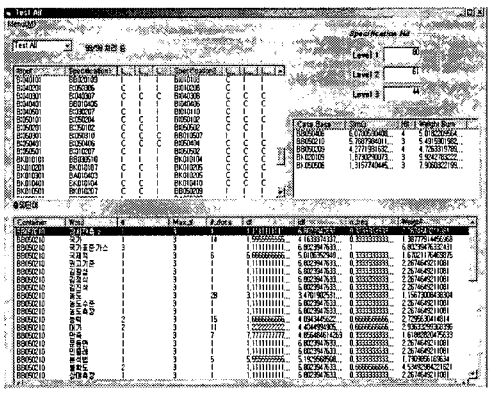


그림 2. 실험화면

리스트박스에서 Input 필드에서는 입력된 질의문서의 코드를 보여준다. Specification1은 각 질의문서에 대하여 유사도 계산 후 제일 높은 점수를 보인 학습문서를 보여주고, Specification2는 차점을 갖는 학습문서의 코드이다. L1, L2, L3는 순서에 따라서 Level1(대분류), Level2(중분류), Level3(소분류)를 말하는 것으로 질의문서와 발췌된 학습문서가 각 Level 별로 일치가 되었는가 혹은 아닌가를 보여준다. C는 일치되었음을 그리고 I는 일치하지 않았음을 보여준다. [그림 4.1]은 100개의 질의문서(Test All 선택)를 검색한 화면이다. Specification Hit에서 보면 100개의 질의문서 중에서 유사한 문서를 찾는데 있어서 대분류는 80개, 중분류까지는 61개, 소분류까지는 44개 맞추었다는 것이다. 그리고 좌하단의 리스트박스는 현재 비교되는 학습문서의 단어들을 보여주고, 우측 중간에 있는 리스트박스는 마지막으로 테스트하고 있는 문서와 비교한 상위 5개의 학습문서를 보여준다. 본 실험을 통하여 최종적으로 100개의 테스트문서를 10개씩 그룹으로 알고리즘을 평가해 본 결과는 [표 4]와 같다.

이상의 결과는 다음과 같은 의미를 갖는다. 첫째, 기존의 키워드 매칭에 비해서 다수의 색인어를 포함하는 문서 대 문서의 비교이므로 정확도는 당연히 뛰어날 수 밖에 없다. 그러나 색인어를 정확하게 유지하는 것이 매우 중요하다. 즉 불용어사전을 정확하게 유지 및 운영하는 전략이 필수적이다. 현재 적중률이 다소 떨어진 것은 불용어처리가 미흡하기 때문이다. 둘째, 기존의 키워드 매칭에 의한 검색 결과는 유사도의 개념이 없고 검색시스템의 기준에 따른 나열이다. 그러나 본 알고리즘의 결과는 유사도 계산결과에 따라 자동으로 배열된다. 이는 전문가의 판단에 따른 순위 배열과 매우 유사한 결과를 보인다. 셋째, 더 많은 케이스가 학습문서에 저장될수록 각 단어에 대한 가중치가 더욱 정교해지므로 시간이 갈수록 정확도가 증가되는 것이 본 알고리즘의 중요한 특성이다. 넷째, 본 실험은 제안서와 같은 구조를 같는 연구보고서를 이용하였으며, 실험에 사용된 문서의 크기는 매우 작다. 그래서 각 문서의 색인어 수가 매우 작기 때문에 매칭되는 단어의 수가 작을 수 밖에 없다. 그러나 정식의 제안서처럼 문서의 크기가 크면 자동적으로 정확도는 올라갈 것이다. 예를 들면 현재의 경우도 색인어 매칭의 숫자가 7개 이상인 경우에는 상위 5개의 문서속에 유사한 문서를 찾아낼 확률이 95% 이상인 것을 보면 알 수 있다. [표 4-1]은 실험결과를 정리한 것이다.

**표 4. 실험결과**

| | L1 | L2 | L3 |
|---|---|---|---|
| Test 1 | 5 | 4 | 3 |
| Test 2 | 8 | 7 | 4 |
| Test 3 | 8 | 5 | 4 |
| Test 4 | 9 | 6 | 4 |
| Test 5 | 9 | 7 | 5 |
| Test 6 | 10 | 8 | 5 |
| Test 7 | 8 | 4 | 4 |
| Test 8 | 9 | 8 | 6 |
| Test 9 | 7 | 5 | 4 |
| Test 10 | 7 | 7 | 5 |
| 합계 | 80 | 61 | 44 |

## V. 결론 및 한계

본 연구를 통하여 개발된 알고리즘은 중복과제를 파악함에 있어서 기존의 단순 키워드 매칭에 따른 문제점을 해결하였다. 즉 문서 대 문서로 비교함으로써 정확도를 향상시켰으며 우선순위별로 배열이 가능하다. k-nn 기법을 도입하여 검색시간을 현저하게 줄이고, Feature weighing 기법을 도입함으로서 전문가의 판단을 객관적이고 일관성 있게 반영하였다. 본 알고리즘은 유사과제 판단을 위한 것 뿐 아니라 다양한 분야에 적용할 수 있음을 확인 하였다. 예를 들면 디렉토리서비스를 위한 자동 문서 분류의 알고리즘으로 사용가능하다. 본 알고리즘은 비교적 단순한데 반하여 성능은 매우 뛰어남을 알 수 있다. 따라서 장차 국가가학기술정보시스템(NTIS)의 하위시스템 중의 하나인 유사과제 검색시스템에 필요한 핵심 알고리즘으로 적용될 수 있다.

본 연구는 중복과제 파악을 위한 시스템의 초기단계의 연구에 이므로 아직 많은 것들에 대한 고려가 부족

하다. 특히 불용어사전, 동의어, 유사어 등에 대한 처리가 결정적으로 부족함 부분이 있다. 실험에서는 제안서와 같은 구조를 갖는 연구보고서 문서를 사용하였다. 이는 문서의 크기가 실제 제안서에 비교해서 매우 작으며, 또한 테스트 문서를 100개 그리고 기존의 문서를 900개로 제한하여 실험하였으므로 결과에 대한 타당성에 대한 의문이 있을 수 있다. 따라서 본 실험을 통한 성공률은 실제 제안서를 대상으로 하는 결과와는 차이가 있을 수 있다. 개발된 알고리즘에 대한 복잡도 및 소요시간 등에 대한 객관적인 분석이 이루어지지 않았으므로 실제 적용을 위해서는 이를 위한 체계적인 분석이 필요하다. 알고리즘의 실험도 파라메타 값을 다양하게 하고, k-nn 및 feature weighting 도 알고리즘에 적절하게 추가적으로 반영하여야 하는 데 이에 대한 고려가 부족하였다. 따라서 추후에는 알고리즘의 정교화를 위한 다양한 시도가 있어야 하며 불용어 처리 방법을 개선하고, 시소러스와 같은 용어의 의미 및 관계를 보여주는 것들에 대한 보완이 이루어져야 한다.

## 참 고 문 헌


[1] 과학기술정보통합서비스, http://www.ndsl.kr

[2] 국가과학기술종합정보서비스, http://www.ntis.go.kr

[3] 중복지원방지시스템, https://www.naris.re.kr/naris

[4] Goffinet L and Noirhomme-Fraiture M (1995) Automatic hypertext link generation based on similarity measures between documents. Research Paper, RP-96-034, Institut d'Informatique, FUNDP. Available at http://www.fundp.ac.be/~lgoffine/Hypertext/semantic_links.html (visited November, 2002).

[5] 최준영, 배환국, 김기태, "하이퍼링크 정보를 이용한 웹문서의 핵심어 추출 및 개념구성," 98 ES 및 MIS 춘계학회 자료집, 1998.

[6] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, In Proc. of the European Conference on Machine Learning, Springer, 1998.

[7] Y. Yang and X Liu, A reexamination of text categorization methods, In SIGIR-99, 1999.

[8] 이종운 "사례기반추론을 이용한 한글 문서분류 시스템의 성능 향상에 관한 연구", 아주대학교 대학원 경영정보학과 석사학위논문, 2001.

[9] F. Debole and F. Sebastiani, Supervised tern weighting for automated text categorization, In Proc. of SAC-03, 18th ACM Symposium of Applied Computing, pp.784-788, 2003.

〈첨부 1〉 기계금속분야 코드체계

| | | | |
|---|---|---|---|
| 기계금속분야 | 공기조화 냉동 (BA01) | 1. 가스 보일러 (BA0101) | 10문서 |
| | | 2. 극저온 냉동기 (BA0102) | 10문서 |
| | | 3. 원심(터보)압축기 (BA0103) | 10문서 |
| | | 4. 유동층 연소장치 (BA0104) | 10문서 |
| | | 5. 유속 측정 (BA0105) | 10문서 |
| | 공장자동화 (BA02) | 1. 공장자동화[FA] (BA0201) | 10문서 |
| | | 2. 무인운반 시스템 (BA0202) | 10문서 |
| | | 3. 물류 시스템 (BA0203) | 10문서 |
| | | 4. 바코드 기술 (BA0204) | 10문서 |
| | | 5. 생산관리 및 품질관리 (BA0205) | 10문서 |
| | 금속가공 (BA03) | 1. 가압 주조법 (BA0301) | 10문서 |
| | | 2. 강섬유(Steel Fiber)제조(BA0302) | 10문서 |
| | | 3. 강주물 (BA0303) | 10문서 |
| | | 4. 금속 초극세사 (BA0304) | 10문서 |
| | | 5. 금형주조 (BA0305) | 10문서 |
| | 기계가공 (BA04) | 1. CNC공작기계 (BA0401) | 10문서 |
| | | 2. 구멍가공기술 (BA0402) | 10문서 |
| | | 3. 레이저 가공기술 (BA0403) | 10문서 |
| | | 4. 레이저 절단기술 (BA0404) | 10문서 |
| | | 5. 모서리 절삭기 (BA0405) | 10문서 |
| | 열처리 및 표면처리 (BA05) | 1. 가공열처리 (BA0501) | 10문서 |
| | | 2. 경사 기능 재료 (BA0502) | 10문서 |
| | | 3. 다이아몬드(DLC)코팅 (BA0503) | 10문서 |
| | | 4. 이온 플레이팅 (BA0504) | 10문서 |
| | | 5. 플라즈마 코팅 (BA0505) | 10문서 |

〈첨부 2〉 불용어 리스트

연구개발, 내용, 증가, 확립, 변화, 요구, 제시, 지원, 효율, 운영, System, 관한, 기능, 자료, 방안, 영향, 확보, 응용, 효과, 개선, 관리, 활용, 구현, 목표, 처리, 분야, 기술개발, 결과, 국내, 가능, 기존, 제공, 적용, 향상, 목적, 방법, 조사, 평가, 구축, 서비스, 성능, 기반, 수행, 사용, 특성, 제작, 분석, 설계, 이용, 시스템, 기술, 연구, 개발, 최근, 한국, 보고서, 연구과제, 특징, 장점, 확인, 문제, 가능성

## 저 자 소 개


**박 동 진(Dong-Jin Park)** 정회원

- 1983년 2월 : 아주대학교 산업공학과(공학사)
- 1988년 2월 : 한국외국대학교 경영정보학과(경영학석사)
- 1994년 8월 : 아주대학교 경영대학 경영정보학전공(경영학박사)
- 1995년 3월 : 남서울대학교 경영학과 조교수
- 1998년 3월 : 공주대학교 산업시스템공학과 교수

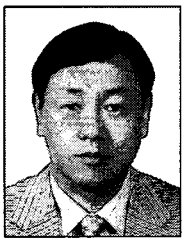
<관심분야> : 메타데이터, 시맨틱웹, 제조정보시스템(ERP, MES)

**최 기 석(Ki-Seok Choi)** 정회원

- 1988년 2월 : 서울대학교 계산통계학과 (공학사)
- 1997년 2월 : KAIST정보및통신공학과 (공학석사)
- 1988년 1월 ~ 현재 : KISTI 책임연구원

<관심분야> : 데이터베이스, 클러스터링

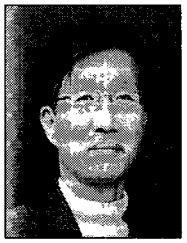
**이 명 선(Myung-Sun Lee)** 정회원



- 1983년 2월 : 아주대학교 전자공학과 졸업(공학사)
- 1996년 8월 : 한남대학교 컴퓨터공학과 졸업(공학석사)
- 2005년 2월 : 한남대학교 컴퓨터공학과 졸업(공학박사)

<관심분야> : 정보보안, 정보통신,정보시스템

**이 상 태(Sang-Tae Lee)** 정회원



- 1977년 2월 : 아주대학교 전자공학과 졸업(공학사)
- 1992년 8월 전북대학교 전자 및 컴퓨터공학과 졸업(공학석사)
- 1998년 2월 전북대학교 전자 및 통신공학 졸업(공학박사)
- 1985년 12월 ~ 현재 : 한국표준과학연구원 전산정보팀장
- 2009년 1월 ~ 현재 : 한국감성과학회 회장

<관심분야> : 정보통신, IT융합, 감성공학