Journal of the Korea Academia-Industrial cooperation Society
Vol. 12, No. 8 pp. 3677-3688, 2011

http://dx.doi.org/10.5762/KAIS.2011.12.8.3677

텍스트 마이닝을 이용한 특허정보검색 개발에 관한 연구

고광수[1], 정원교[1], 신영근[1], 박상성[1], 장동식[1*]
[1]고려대학교 산업경영공학부

A Study on Development of Patent Information Retrieval Using Textmining

Gwang-su Go[1], Won-Kyo Jung[1], Young-Geun Shin[1], Sang-Sung Park[1] and Dong-Sik Jang[1*]
[1]**School of Industrial Management Engineering, Korea University**

요 약 특허정보검색의 목적은 다양한 목적성을 지니고 있다. 일반적으로 특허정보검색은 제한된 키워드들에 의한 검색으로 이루어지며, 선행 특허권과 유사특허를 파악하기 위하여 반복적인 검색과 검토의 노력이 필요하다. 본 논문에서는 특허문서의 전체 텍스트를 분석하여 특징치를 찾아내는 내용기반 검색방법을 제안하고 검색결과를 질의문서와 유사한 문서 순으로 우선 배치하여 검색에 효율을 높일 수 있는 방법을 제안한다. 즉, 제안된 알고리즘은 텍스트 분석과정을 통해 각 문서별로 특징치가 부여되고 문서 간 특징치 비교를 통해 유사문서를 찾고 문서를 랭킹하여 유사정보를 제공한다. 텍스트 분석과정은 Stop-word과정, 핵심단어 추출과정, 핵심단어 가중치 산출 과정으로 이루어진다. 실험결과에서는 정확도 측정을 실시하여 일반검색엔진과 본 논문에서 제안한 알고리즘의 검색 정확도를 비교하였다. 본 논문은 검색결과를 질의한 문서와 유사한 문서 순으로 랭킹하기 때문에 검색이용자가 검색결과 검토과정에서 유사한 문서를 먼저 검토할 수 있도록 하여 검토시간을 줄이고 검색의 효율을 높일 수 있다. 또한 특허문서 전체 텍스트를 입력받아 사용하기 때문에 특허검색에 익숙하지 않는 이용자도 검색을 쉽고 빠르게 이용할 수 있다. 그리고 내용 기반 검색이 이루어지기 때문에 키워드 및 검색 식을 이용하는 방법보다 검색범위를 넓힐 수 있어서 검색에 누락되는 데이터를 줄일 수 있는 효과를 가진다.

Abstract The patent information retrieval system can serve a variety of purposes. In general, the patent information is retrieved using limited key words. To identify earlier technology and priority rights repeated effort is needed. This study proposes a method of content-based retrieval using text mining. Using the proposed algorithm, each of the documents is invested with characteristic value. The characteristic values are used to compare similarities between query documents and database documents. Text analysis is composed of 3 steps: stop-word, keyword analysis and weighted value calculation. In the test results, the general retrieval and the proposed algorithm were compared by using accuracy measurements. As the study arranges the result documents as similarities of the query documents, the surfer can improve the efficiency by reviewing the similar documents first. Also because of being able to input the full-text of patent documents, the users unacquainted with surfing can use it easily and quickly. It can reduce the amount of displayed missing data through the use of content based retrieval instead of keyword based retrieval for extending the scope of the search.

Key Words : Text Mining, TF-IDF, Precision, Stop-word, Patent Information Retrieval

---

본 논문은 2011년도 두뇌한국 21사업에 의하여 지원되었음.
본 논문은 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.
(한국연구재단-R1A4A007-2010-0024163)
*교신저자 : 장동식(jang@korea.ac.kr)
접수일 11년 06월 10일 수정일 (1차 11년 07월 06일, 2차 11년 07월 27일) 게재확정일 11년 08월 11일

1. 서론

1.1 연구의 필요성

특허 출원을 위한 명세서를 작성하거나 기술발명 및 유사기술 정보를 수집하는 활동에 있어서 선행기술조사는 반드시 수반이 된다. 현재 선행기술조사는 각 국가별 특허청에서 공지하는 등록된 특허정보를 대상으로 해당 기술과 동일하거나 유사한 특허를 검색하고 분석하는 방법으로 이루어지고 있다. 검색 방법으로는 찾고자하는 기술에 대한 키워드검색방법과 키워드를 조합한 검색 식을 이용한 검색방법이 사용되고 있다. 키워드 및 검색 식을 작성하기 위해서는 그 특허에 대한 세부기술 내용파악을 필요로 한다. 일반적으로, 특허검색은 확장 키워드와 확장 검색 식을 찾는 작업 과정을 반복한다. 검색 결과에서 확장 키워드를 찾고 확장 검색 식을 작성하기 위해서는 검색결과를 검토하는 스크리닝 작업이 반복된다[1]. 이러한 무제한적인 반복 작업은 검색시간을 많이 소요하게 하므로, 검색 효율을 떨어뜨릴 뿐만 아니라 목적에 맞는 특허데이터를 찾기에 어려움이 따른다. 또한 키워드 및 검색 식을 확장하는 과정에서 검색자는 검색결과로 나온 수많은 데이터를 대상으로 주제, 요약, 청구항 등 특허명세서의 주요 부분뿐만 아니라 필요에 의해서는 명세서 전문을 모두 검색하고 내용을 살펴봐야 하는 불편함을 가지고 있다. 또한 검색결과에서 불필요한 잡음데이터를 걸러내고, 검색범위 조정을 위하여 동의어 및 외래어 표기에 대한 다양성 등을 모두 고려하기 위한 노력이 필요하다. 이처럼 현재 일반적으로 사용되고 있는 키워드 중심의 검색은 상당한 노력과 시간을 필요로 하므로 검색 효율이 매우 좋지 못하다.

본 논문은 특허정보 검색 결과를 질의 문서와 유사한 문서 순으로 재 정렬 시키고, 사용자에게 질의 문서와 유사정도 정보를 동시에 제공함으로써 검색 결과분석을 정확하고 빠르게 하고자 한다. 즉, 기존에 사용되고 있는 키워드 및 검색 식 기반의 검색이 아니라, 특허문서의 전체 텍스트를 분석하여 특징치를 찾아내는 내용기반 검색방법을 제안하고 검색결과를 질의문서와 유사한 문서 순으로 나타내는 방안을 제시하고자 한다.

1.2 연구의 목적 및 방법

본 논문은 특허청의 데이터베이스에 있는 특허자료를 검색할 때, 사용자가 질의문서와 유사한 문서 순으로 문서를 검색할 수 있는 알고리즘을 연구하는데 목적이 있다. 질의입력에 의한 검색결과가 단순한 발명의 명칭, 국제특허분류코드(IPC), 출원인, 등록일자, 출원번호별로 정렬되어 나오는 것을 지양하고 목적이 되는 세부기술내용을 기반으로 기존에 공개되었거나 출원 및 등록된 특허문서들을 검색하여 기술내용이 유사한 것들을 찾아내고 유사도 순으로 정렬하는 알고리즘을 제안하고자 한다. 즉, 본 논문에서는 검색된 결과 데이터들에 대해서 질의문서와의 유사정도의 정보를 추가적으로 제공하여 관련 기술을 분류하는 결과검토 작업에 소요되는 시간과 노력을 줄이고자 한다.

본 논문에서는 전처리 과정으로 텍스트 분석과정을 통해 특허문서에서의 불필요한 단어를 제거하고, TF-IDF (Term Frequency Inverse Document Frequency) 알고리즘을 활용하여, 각 특허문서별 핵심단어들을 추출하고 이에 대한 가중치를 부여한다. 이러한 가중치를 이용하여 특허문서집합DB에서 질의문서와 유사한 문서들을 찾아낸다. 검색결과는 질의문서와 유사한 순으로 내림차순 정렬하여 나타낸다. 실험데이터로는 기술 군별로 반도체 웨이퍼(Wafer), 터치(Touch) 기술에 대한 데이터를 이용한다. 유사문서검색결과를 일반검색엔진과 본 논문에서 제안하는 방법을 검색결과정확도(Precision) 실험을 통해 성능 비교를 한다. 본 논문은 서론에 이어 제2장에서는 관련 연구를 기술하였고, 제3장에서는 제안된 알고리즘 설계를, 그리고 제4장에서는 실험 및 결과를 다루고, 마지막으로 제5장에서는 결론 및 향후과제를 기술한다.

2. 관련 연구

2.1 기존연구

특허정보검색 목적은 크게 특허권 존재여부검색, 신기술 개발을 위한 지식검색, 특허등록 가능성에 대한 조사, 특허권 무효를 위한 자료 조사로 구분할 수 있다[1]. 검색방법으로는 문헌 번호를 이용한 검색 방법, 특허맵을 이용한 조사방법, 선행기술조사 방법 등 다양한 방법이 존재한다. 최근 지식재산권의 중요성이 강조됨에 따라, 특허문서에 대한 사회적 관심이 높아졌고, 특허 분야 종사자가 아닌 일반 특허정보검색 이용자들이 늘어나고 있다[2]. 하지만, 특허 문서는 일반적인 문서와는 달리 기술적인 내용을 담고 있기 때문에 몇 개의 단어를 이용한 키워드검색에 익숙한 일반사용자가 기술적 내용을 담고 있는 대량의 특허문서들 중에서 자신의 검색목적에 맞는 결과를 찾아내기가 쉽지 않다. 실제적으로 특허정보검색에 익숙한 검색자의 경우는 핵심키워드로 이루어진 검색 식을 주로 이용하여 특허문서를 검색하는 반면, 검색에 익숙하지 않은 일반 검색 이용자의 경우는 특허문서의 기술적

내용 파악에 대한 어려움 때문에 본인이 필요로 하는 정보를 찾기 위한 효과적인 검색 식 작성에 많은 어려움을 겪는다. 매년 특허출원 및 등록 건수의 증가로 인한 검색 대상이 되는 데이터의 양 또한 매년 기하급수적으로 늘어나고 있는 환경은 특허검색을 더욱 어렵게 하고 있다[3]. 최근 이러한 문제들을 해결하고 특허정보검색의 효율을 높이기 위한 연구가 이루어지고 있다. 크게 검색 키워드에 관한 연구와 검색 결과의 군집화에 대한 연구가 주를 이루고 있다. 국내에서는 백종범(2009) 등이 키워드 불일치에 의한 정보 누락을 줄이기 위한 대체어 후보 추출 방법에 관한 연구를 하였고[4], 손기준(2005) 등은 특허 문헌 검색에서 복합명사의 재출현 양상과 복합 명사 역할 변화를 이용한 검색어의 가중치 부여 방법에 관한 연구를 하였다[5]. 그리고 김한기(2007) 등은 대량의 특허 문서에서 국제 특허 분류(IPC) 정보를 이용한 검색 결과 클러스터링에 관한 연구를 하였다[6]. 국외에서는 Loh Han Tong(2006) 등이 TRIZ이론을 이용한 국제특허분류코드(IPC)분류를 대체 할 수 있는 자동화 분류 방법에 관한 연구를 하였고[7], Yen-Liang Chen(2010) 등은 국제특허분류코드(IPC) 기반 벡터스페이스 모델을 이용한 특허검색에 관한 연구를 하였다[8]. 그리고 Kuei-Kuei Lai(2003)은 특허 인용을 이용한 새로운 특허 분류시스템 개발에 관한 연구를 하였다[9]. 앞서 설명한 관련 연구에 대한 설명은 다음과 같다. 먼저, 검색 키워드와 관련된 연구는 사용자가 사용하는 키워드를 대체할 수 있는 대체 키워드를 추천 제시함으로써 키워드 불일치 등의 이유로 정보누락을 최소화 하고자 하는 방법에 대한 연구이다. 이와 같이 키워드 확장에 관한 연구는 검색결과 누락의 최소화에는 도움이 되지만, 특허 검색자가 목적으로 하는 기술적 내용 파악에는 도움이 되지 않는다. 오히려 검색결과의 수가 많아지기 때문에 검색자가 검토해야하는 특허문서의 양이 늘게 되어 효과적인 검색이 이루어질 수가 없게 된다. 다음으로, 특허검색 결과 분류에 관한 연구는 국제 특허 분류(IPC)정보, 출원인, 등록일자, 출원번호 등의 서지적 정보를 이용하여 검색결과를 분류하여 나타내준다. 이와 같이 서지적 정보를 이용한 검색결과의 분류는 검색자가 검색범위를 제한할 수 있도록 하여 검색자가 검토해야 하는 문서의 양을 줄이는 효과를 준다. 하지만, 서지적 정보를 이용한 분류는 문서가 담고 있는 내용을 바탕으로 분류되는 것이 아니기 때문에 검색자가 원하는 정보와 유사한 문서를 찾기 위해서는 검색과 결과 분류, 검토과정을 여러 번 반복해서 수행해야만 하는 불편을 가지고 있다.

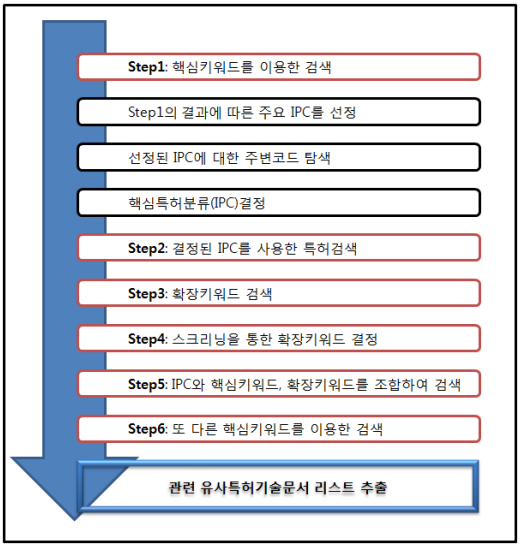
본 논문에서는 현재 특허정보검색의 효율을 높이기 위해 이루어지고 있는 연구를 응용하여 특허문서에 담긴 내용을 기반으로 유사특허문서를 검색하고, 검색결과를 유사한 내용을 담고 있는 문서를 우선 나타낼 수 있는 방안을 제안하고자 한다. 즉, 본 논문은 키워드를 이용하는 검색방법이나 문서 분류 방법을 사용하는 것이 아니라 문서자체를 질의로 사용하여 특허문서 전체의 텍스트를 분석하는 내용기반 검색으로 특징치를 찾아내고, 이를 기반으로 유사문서를 유사도 순으로 랭킹 시키는 알고리즘을 제안한다. 이러한 내용기반 검색은 유사 특허권 여부 검색, 신기술 개발을 위한 지식검색, 특허등록 가능성에 대한 조사, 특허권 무효를 위한 자료조사 등 다양한 목적에 부합하는 조사 결과를 손쉽게 줄 수 있는 장점을 지닌다. 또한 특허 검색에 익숙한 검색자는 물론 검색에 익숙하지 않는 일반 검색이용자 모두에게 검색의 편의성을 줄 수 있고, 또한 내용을 기반으로 문서 검색이 이루어지기 때문에 검색자가 특허 문서를 일일이 분석할 필요가 없고, 키워드를 사용한 검색결과범위보다 훨씬 광범위한 범위에서 검색을 할 수 있게 되며, 최종적으로 유사한 문서 순으로 랭킹된 결과를 나타내기 때문에 검색에 관한 적합성 검토를 해야 하는 검색결과 문서의 수가 줄어드는 효과가 있다.

2.2 스크리닝 기법(Screening)

특허정보검색에 있어서 스크리닝(Screening)이란 특허 집단문서에서 검색자가 필요로 하는 정보를 찾기 위하여 검색된 문서들을 빠르게 읽고 내용을 파악하여 필요한 정보를 얻어내는 작업이다. 일반적으로 작업 대상이 되는 특허문서의 수가 많기 때문에 정보를 추출해내는 작업시간이 많이 필요하다. 검색 영역은 일반적으로 특허문서의 제목(Title), 요약(Abstrate), 대표도면의 내용정보를 위주로 검색이 이루어진다. 특허정보검색에서 스크리닝 작업은 일반적으로 질의문서분석과정, 국제특허분류코드(IPC) 선정과정, 키워드 확장과정, 국제특허분류코드(IPC)와 확장키워드 조합으로 이루어진 검색 식의 결과 분석과정, 최종 유사특허문서 리스트들에 대한 분석과정 등에서 사용된다. 그림 1은 윕스(WIPS)에서 제공하는 STEP 단계별 검색 방법을 나타내고 있다[10]. 이 방법은 특허정보조사에 있어서 일반적으로 사용되는 주제별 검색 방법이다.

STEP1에서는 검색자가 특허문서의 핵심키워드를 찾아내어 1차 특허검색을 실시하는 단계이다. 1차 검색결과를 클러스터링 및 검토 작업을 통하여 검색자의 검색 주제에 맞는 국제특허분류코드(IPC)정보를 선정하고 검색대상이 될 국제특허분류(IPC)를 결정한다. 다음으로 STEP2에서는 결정된 IPC코드별로 전체 특허를 추출해내는 단계이다. STEP3에서는 STEP2에서 검색된 데이터를

대상으로 확장키워드를 찾기 위한 스크리닝(Screening)작업이 수행된다. 스크리닝 대상 건수가 많고, 기술내용별 유사특허문서별로 나열되지 않기 때문에 검색결과에 대한 검토 노력이 많이 요구된다. STEP4는 핵심키워드와 STEP3에서 찾은 확장키워드를 조합하여 최종 검색 식에 사용되는 키워드 검색 식을 작성한다. STEP5에서는 STEP1과 STEP4에서 선정된 주요 IPC코드 검색 식과 확장키워드 검색 식을 조합하여 최종 검색 식을 완성한다. STEP6에서는 검색자가 검색하고자 하는 특허문서의 또 다른 핵심키워드를 찾아내어 위의 STEP2, STEP3, STEP4를 반복하여 최종 검색 식을 완성한다. 최종 검색 식들을 조합하여 특허문서 주제와 관련한 유사특허문서 리스트를 추출하게 된다. 이러한 결과 리스트에서 검색자가 검색주제와 관련된 유사 특허를 찾기 위하여 스크리닝 작업이 필요하다. STEP 단계별 검색 방법은 검색자에게 직관적인 검색방법론을 제공해 줄 수 있는 장점을 지니지만, 검색자가 특허문서구성에 대해 전반적으로 이해해야하며, 특허문서의 기술적인 내용을 파악하여 확장키워드를 찾고, 확장 검색 식을 작성하는 능력이 필요하다. 이 같은 과정과 절차는 특허검색에 익숙하지 않는 사용자에게는 결코 쉽지만은 않다. 더구나 특허검색에 익숙한 이용자라 할지라도 STEP 단계별 검색 방법이 일반적으로 검토대상이 되는 문서의 수가 많기 때문에 검색작업에 소요되는 시간이 많이 필요로 하게 되므로 효과적인 검색이 이루어지고 있다고 보기 힘들다.



[그림 1] 핵심키워드를 활용한 검색방법
[Fig. 1] The retrieval method using core keyword

2.3 TF-IDF

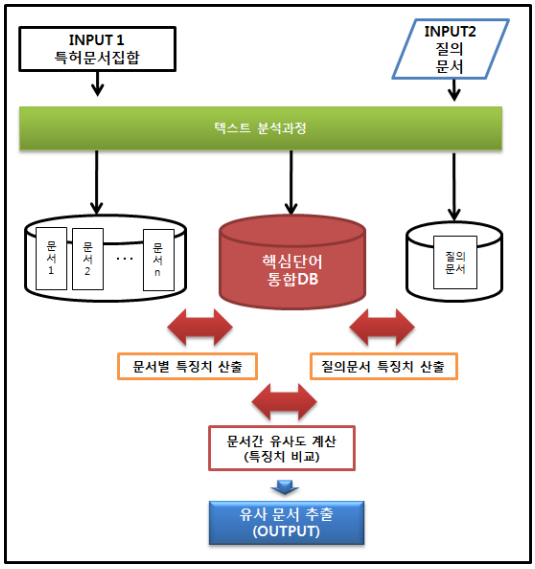
특허 문서 분석을 위해서는 문자로 표현된 특허문서를 분석이 가능한 형태로 변환하는 과정인 텍스트마이닝 과정이 요구된다[11]. TF-IDF는 문서에 포함된 단어들에 가중치 부여를 통하여 문서의 핵심단어를 추출해내는 용도로 사용된다. 여기서 TF(단어빈도수, Term Frequency)는 문서 내에 출현하는 모든 단어를 대상으로 각 단어들이 해당 문서에서 출현하는 빈도를 나타내는 값이다. TF값이 높을수록 해당문서에서 특정단어가 차지하는 중요도가 높다고 할 수 있다. DF(문서 빈도수, Document Frquency)는 문서 집합에서 출현하는 단어의 빈도를 나타낸다. DF 값이 높을수록 해당단어는 문서 집합 내에서 흔하게 사용되어진 단어라고 할 수 있다. 같은 단어일지라도 DF값은 문서집합들의 특성에 따라서 달라질 수 있다[12-13]. 예를 들어 국제 특허분류코드(IPC) 분류 D섹션(섬유)의 특허문서 집합의 경우, “섬유”에 해당하는 특허문서집합이기 때문에 “회로”라는 단어는 잘나오지 않는 단어이다. 따라서 IPC D섹션에서 “회로”라는 단어의 DF값은 작게 나오며 해당 문서를 대표할 수 있는 핵심어가 된다. 반면, 국제특허분류코드(IPC) 분류 H섹션(전기)인 특허문서집합의 경우에는 “회로”라는 단어가 특허문서집합에서 공통으로 자주 등장하는 흔한 단어가 되어버리게 되므로 DF값은 높아지고, 문서집합 내에서 문서를 구분할 수 있는 핵심어가 될 수 없다. 이와 같이 텍스트마이닝에서 이용되는 가중치는 하나의 특허문서 내에 존재하는 단어의 빈도수(TF)와 문서 빈도수(DF)의 역수 값인 IDF(Inverse Document Frequency)가 주로 이용된다[13].

3. 제안된 알고리즘

3.1 제안된 알고리즘

본 논문에서 제안하는 방법은 특허문서가 저장되어 있는 특허청의 데이터베이스에서 특허정보조사 목적에 따라서 다양한 특허문서집합을 생성하고, 특허문서집합에서 질의문서와 유사한 특허문서를 찾아 유사도순으로 나열하고 그 유사도 정도를 수치화하여 보여주는 방법이다. 즉, 내용을 기반으로 하는 검색방법으로 검색결과를 질의문서와 유사한 문서 순으로 찾는 방법이다. 본 논문에서 제안한 알고리즘은 문서 자체를 질의로 사용한다는 것이 특징이다. 다음 그림 2는 본 논문에서 제안하는 유사 특허문서 검색 알고리즘이다. 첫 번째 과정은 검색대상이 되는 특허문서집합과 질의문서를 구분하여 입력한다. 이

때 키워드 입력이 아닌, 문서자체를 입력하는 방법을 사용한다. 두 번째 과정으로, 입력된 문서에 대한 텍스트 분석과정을 거쳐 각 문서별 핵심단어를 찾아내고, 핵심단어 통합DB를 생성한다. 동시에 핵심단어에 대한 가중치를 산출한다. 3.2 텍스트 분석에서 그림 2에 나타나 있는 텍스트 분석과정에 관해 자세히 설명하도록 한다. 세 번째 과정으로, 핵심단어 통합DB와 각 문서별 핵심단어를 비교하여 각 문서별 특징치를 산출해 낸다. 이때 문서별 특징치는 핵심단어별 가중치를 이용하여 구한다. 네 번째 과정으로, 질의문서 특징치와 특허문서집합의 각 문서별 특징치를 비교하여 유사도를 계산한다. 마지막으로, 계산된 유사도를 이용하여 질의문서와 유사한 문서 순으로 랭킹한 뒤, 그 결과를 내림차순 정렬하여 나타낸다.

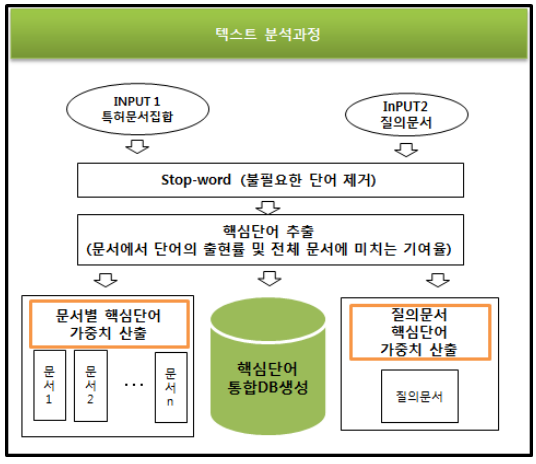


[그림 2] 제안된 특허문서 검색 알고리즘
[Fig. 2] The retrieval algorithm of the proposed patent documents

3.2 텍스트 분석

본 논문에서 제안하는 알고리즘은 내용을 기반으로 하는 검색 방법으로 특허문서 전체를 입력하여 전체 텍스트 분석과정을 거쳐 문서의 특징치를 찾고 문서 간 특징치 비교를 통하여 유사한 문서를 찾아낸다. 다음 그림 3는 텍스트 분석과정을 나타낸다.

특허문서집합과 질의문서가 입력되면 먼저, Stop-word를 이용하여 각 문서의 내용에 기여하지 않는 불필요한 단어를 제거한다. 다음으로, 각 문서별로 사용된 단어의 빈도수 측정을 통하여 각 문서에서의 해당 단어의 출현율을 계산한다. 동시에 해당 단어가 전체 문서에 미치는 기여율을 계산한다. 다음으로, 단어의 출현률과 전체 문서 기여율을 이용하여 각 문서별 핵심단어를 찾아내고, 핵심단어에 대한 가중치를 산출한다. 특허문서집합에 포함된 문서와 질의문서 전체를 통합하는 핵심단어 통합DB를 생성한다. 텍스트분석과정의 결과로 특허문서집합의 각 문서별 핵심단어 가중치와 질의문서의 핵심단어 가중치, 핵심단어 통합DB를 생성한다.



[그림 3] 제안된 알고리즘에서 텍스트 분석과정
[Fig. 3] The text analysis process in the proposed algorithm

3.2.1 Stop-word

본 논문에서는 특허문서에서 자주 사용되는 단어가 문서의 기술내용을 담고 있다는 가정을 하고 있다. 하지만 일반적으로 특허문서에는 관사, 조사, 접속사, 부사, 전치사 등과 같이 기술문서내용과는 관계가 없지만 문서 내에서 많이 사용되는 단어도 포함되어 있다. 따라서 특허문서의 내용을 담고 있지 않는 불필요한 단어를 먼저 제거하는 데이터 전처리 작업이 선행되어야 한다[13]. 표 1은 불필요한 단어에 대한 리스트를 나타낸다.

[표 1] Stop-word 리스트
[Table 1] Stop-word list

Stop-word List	
a-z	a, about, above, across, after, again, also, although, ..., younger, youngest, your, you're, yours, yourself, yourselves, you've

3.2.2 핵심단어 가중치 산출

각 문서별 핵심단어 가중치(W_k)는 각 특허문서에서 사용된 단어의 출현률을 이용하여 추출한다. 단어별 출현률은 특허문서에 출현하는 단어의 빈도값(Term Frequency)을 이용하여 구한다. 다음 식(1)은 단어별 출현률을 계산하는 방법을 나타낸다. 즉, 각 문서별 핵심단어 가중치(W_k)는 단어별 출현률($tf_{i,j}$)값을 이용한다.

$$tf_{i,j} = \frac{f_{i,j}}{\sum_{i=1}^{m} f_{i,j}} \quad (1)$$

$tf_{i,j}$: 문서 j에 대한 단어 i의 출현률

$f_{i,j}$: 문서 j에 사용된 단어 i의 빈도수

m : 문서 j에 있는 단어 수

3.2.3 핵심단어 통합DB생성

핵심단어 통합DB 리스트 생성은 특허문서집합문서들과 질의문서를 대상으로 텍스트 분석과정을 거쳐 추출한 핵심단어들을 하나의 통합 리스트로 작성한다. 핵심단어 통합DB 가중치(W_t)는 각 특허문서에서 사용된 각 단어들의 평균 출현률과 해당단어가 전체 문서에 미치는 기여율을 이용하여 추출한다. 해당단어가 전체 문서에 미치는 기여율은 그 단어가 출현하는 문서빈도수의 역수값(Inverse Document Frequency)을 이용하여 구한다. 다음 식(2)는 해당단어가 전체 문서에 미치는 기여율을 계산하는 방법을 나타낸다.

$$idf_{i,k} = \log \frac{|D|}{|d : t_i \in d|} \quad (2)$$

$idf_{i,k}$: 전체문서 집합 k에 대한 단어 i의 기여율

|D|: 특허문서의 총 개수

$|t \in d|$: 해당단어를 포함한 문서의 개수

핵심단어 통합DB 가중치(W_t)는 각 특허문서에서 사용된 각 단어들의 평균 출현률과 해당단어가 전체 문서에 미치는 기여율을 곱한 값으로 산출한다. 여기서, 각 단어의 평균 출현률 값은 핵심단어 통합DB 리스트에서 중복되는 핵심단어들에 대한 가중치의 합을 중복되는 핵심단어들의 수로 나누어 계산한다. 만약, 리스트에서 중복이 되지 않는 핵심단어는 자신의 가중치 값이 평균 출현률이 된다. 다음 식(3)는 핵심단어 통합DB 가중치(W_t)를 계산하는 방법을 나타낸다.

$$\text{가중치}(W_t) = \overline{tf^*_{i,k}} \times idf_i \quad (3)$$

$\overline{tf^*_{i,k}}$: 핵심단어i의 평균 출현률

단어의 출현률 값이 높을수록 해당문서에서 특정단어가 차지하는 중요도가 높다고 할 수 있으며, 문서 집합에서 출현하는 단어의 빈도 값이 낮을수록 해당 단어는 문서 집합 내에서 기여율이 높다고 할 수 있다.

3.3 문서별 특징치 산출

문서별 특징치 C_p는 텍스트 분석과정을 통해 얻어진 각 문서별 핵심단어 가중치 값 W_k와 핵심단어 통합DB 가중치 값 W_t의 곱의 합으로 얻어진다. 표 2는 본 논문에서 사용한 반도체 웨이퍼 기술 특허문서에 대한 문서별 특징치 산출 관한 예이다.

[표 2] 특징치 산출의 예

[Table 2] The example of characteristic values calculation

핵심단어 통합DB 리스트		
No	핵심 단어	가중치 (W_t)
1	liquid	0.036
2	cleaning	0.029
3	sensor	0.028
4	heating	0.025
...	...	...

반도체 웨이퍼 기술 특허 문서(i)		
No	핵심 단어	가중치 (W_k)
1	heating	0.038
2	wafer	0.036
3	metal	0.034
4	sensor	0.031
...	...	...

핵심단어 통합DB 리스트와 비교 문서 핵심단어 리스트에서 동시에 나타나는 단어가 “heating", "sensor", ... 이므로, 이 문서의 경우 특징치(C_p) 산출하는 방법은 식(4)와 같다.

$$C_p = (0.028*0.031)+(0.025*0.038)+... \quad (4)$$

3.4 문서간 유사도 계산

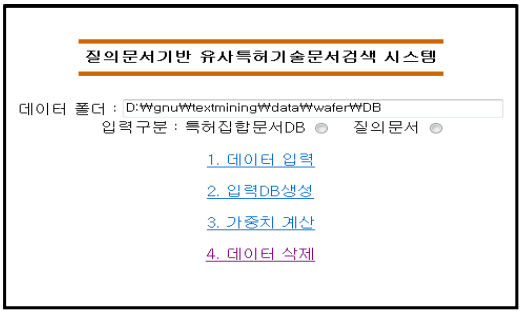
문서간의 유사도는 각 문서별 특징치를 비교하여 유사정도를 측정한다. 유사정도는 질의문서의 특징치를 기준으로 하여 각 문서간의 특징치 차이를 이용하여 나타낸다. 즉 질의문서가 가지는 특징치와 가장 유사한 값을 가지는 문서일수록 질의문서와 유사한 문서로 판단한다. 다음 식(5)는 문서간 유사도 계산방법을 나타낸다.

$$D(X, Y) = |X - Y| \quad (5)$$

위의 문서 간 거리계산값을 이용하여 질의 문서와 유사한 문서 순으로 랭킹하여 내림차순 정렬하여 질의문서와 유사한 문서 순으로 나타낼 수 있다.

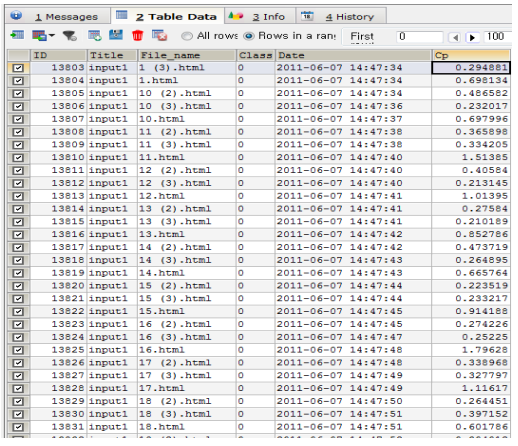
3.5 시스템 구성

다음 그림 4는 질의문서기반 유사특허문서검색 시스템을 구현한 것이다.



[그림 4] 유사특허문서검색시스템
[Fig. 4] The similarities patent document retrieval system

시스템은 웹 기반으로 구축하였으며 DB구축 및 질의문서 입력방법은 텍스트로 된 파일이 저장된 폴더를 지정하는 방법으로 하였다. 유사특허문서검색 시스템에서 입력된 데이터는 MySQL를 이용하여 저장, 분석, 관리하도록 하였다. 그림 5는 특허문서집합과 질의문서가 입력되어진 모습을 나타낸다. 문서가 입력되면 문서별 분석이 이루어지고 문서별 핵심단어에 가중치부여가 이루어진 후 DB에 저장되고 관리되어 진다.


1 Messages | 2 Table Data | 3 Info | 4 History

ID	Title	File_name	Class	Date	Cp
13803	input1	1 (3).html	0	2011-06-07 14:47:34	0.294881
13804	input1	1.html	0	2011-06-07 14:47:34	0.698134
13805	input1	10 (2).html	0	2011-06-07 14:47:34	0.486582
13806	input1	10 (3).html	0	2011-06-07 14:47:36	0.232017
13807	input1	10.html	0	2011-06-07 14:47:37	0.697996
13808	input1	11 (2).html	0	2011-06-07 14:47:38	0.365898
13809	input1	11 (3).html	0	2011-06-07 14:47:38	0.334205
13810	input1	11.html	0	2011-06-07 14:47:40	1.51385
13811	input1	12 (2).html	0	2011-06-07 14:47:40	0.40584
13812	input1	12 (3).html	0	2011-06-07 14:47:40	0.213145
13813	input1	12.html	0	2011-06-07 14:47:41	1.01395
13814	input1	13 (2).html	0	2011-06-07 14:47:41	0.27584
13815	input1	13 (3).html	0	2011-06-07 14:47:41	0.210189
13816	input1	13.html	0	2011-06-07 14:47:42	0.852786
13817	input1	14 (2).html	0	2011-06-07 14:47:42	0.473719
13818	input1	14 (3).html	0	2011-06-07 14:47:43	0.264895
13819	input1	14.html	0	2011-06-07 14:47:43	0.665764
13820	input1	15 (2).html	0	2011-06-07 14:47:44	0.223519
13821	input1	15 (3).html	0	2011-06-07 14:47:44	0.233217
13822	input1	15.html	0	2011-06-07 14:47:45	0.914188
13823	input1	16 (2).html	0	2011-06-07 14:47:45	0.274226
13824	input1	16 (3).html	0	2011-06-07 14:47:47	0.25225
13825	input1	16.html	0	2011-06-07 14:47:48	1.79628
13826	input1	17 (2).html	0	2011-06-07 14:47:48	0.338968
13827	input1	17 (3).html	0	2011-06-07 14:47:49	0.327797
13828	input1	17.html	0	2011-06-07 14:47:49	1.11617
13829	input1	18 (2).html	0	2011-06-07 14:47:50	0.264451
13830	input1	18 (3).html	0	2011-06-07 14:47:51	0.397152
13831	input1	18.html	0	2011-06-07 14:47:51	0.601786

[그림 5] 특허문서 데이터베이스
[Fig. 5] The database of patent documents

4. 실험 및 결과

4.1 실험 데이터

본 논문에서는 미국에 출원되어 있는 반도체 웨이퍼(Wafer)와 터치(Touch) 기술에 관한 특허문서를 이용하여 실험하였다. 검색대상은 2001년 3월 이후 공개되어 있는 특허문서를 대상으로 하였으며, 특허 검색사이트인 WIPS(http://www.wips.co.kr)에서 국제특허분류코드(IPC) 및 기술별 핵심키워드를 이용한 검색 식으로 특허문서집합을 검색하였다. 핵심 키워드를 이용하여 반도체 웨이퍼 기술은 Cleaning, Heating, Testing의 세 가지 세부 기술 군으로 한정지어 검색하였고, 터치 기술은 Capacitance, Infrared, Resistance의 세 가지 세부 기술 군으로 한정지어 검색하였다. 각각의 검색된 상위 100개의 문서를 이용해 일반검색엔진과 제안된 알고리즘을 비교 분석하였다.

4.2 반도체 웨이퍼

식(6)을 이용하여 반도체 웨이퍼 기술을 Cleaning, Heating, Testing 기술의 키워드를 이용해 검색한 결과 총 4958건의 문서가 검색되었다. 잡음데이터는 위 세 가지 기술 군에 포함되지 않은 데이터를 말한다. 편의를 위하여 각 표에서 Cleaning 기술은 C, Heating기술은 H, Testing 기술은 T, Noise는 N으로 표기하였다.

((wafer and semiconductor) and (clean* or heat* or test*)).KEY. AND ((B08B-003* or H01L-021* or B24B-027* or H04N-007* or F27D-021* or G01R-031*)).IPC. (6)

각 문서를 모두 확인하여 스크리닝 기법을 이용해 세부 기술 군으로 나누어야 하므로 실험의 편의성을 위하여 상위 100개의 문서를 이용하여 실험하였다. 상위 100개의 검색된 문서들은 표 3과 같다.

[표 3] 검색결과
[Table 3] The Retrieval result

NO	발명의 명칭	분류
1	Manufacturing method ...	H
2	TEST SYSTEM AND ...	T
3	METHODS AND APPARATUS ...	C
4	UV AND REDUCING ...	C
5	SEMICONDUCTOR WAFER ...	T
...	...	
100	IN-LINE DEPTH ...	T

다음으로 상위 100개의 특허문서들에 대하여 세부 기술구성을 알아보기 위해 스크리닝 기법을 이용하여 검색결과를 분석하였다. 표 4는 상위 100개의 특허문서집합 구성을 나타낸다.

[표 4] 데이터 구성
[Table 4] The Data organization

영역 (Category)	반도체웨이퍼 기술에 관한 특허문서 집합 (개수)
C	10개
H	25개
T	30개
N	35개
Total	100개

스크리닝 기법을 이용한 분석 결과 상위 100개의 검색결과 내에는 Cleaning 10개, Heating 25개, Testing 30개, Noise 35개로 구성되어 있다.

표 5는 검색결과에 대한 기술문서별로 차지하는 비중을 비교한 표이다.

[표 5] 검색결과 비중 비교
[Table 5] The compare of retrieval result rate

방법	구분	질의 문서	결과 (개수)				비율
			C	H	T	N	
일반검색엔진	상위10%	C	5	1	3	1	50%
		H	5	1	3	1	10%
		T	5	1	3	1	30%
	상위20%	C	6	5	5	4	60%
		H	6	5	5	4	50%
		T	6	5	5	4	50%
	상위30%	C	7	8	8	7	70%
		H	7	8	8	7	80%
		T	7	8	8	7	80%
제안된 알고리즘	상위10%	C	8	1	0	1	80%
		H	0	10	0	0	100%
		T	0	0	10	0	100%
	상위20%	C	9	3	2	6	45%
		H	0	17	2	1	85%
		T	0	0	19	1	95%
	상위30%	C	10	6	5	9	33%
		H	0	21	2	7	70%
		T	1	4	22	3	73%

일반검색엔진 방법의 경우는 이미 식(6)을 이용해 검색된 100개의 문서 중 Cleaning, Heating, Testing 각각의 키워드를 이용한 검색 식으로 질의한 결과이며, 제안된 알고리즘은 Cleaning, Heating, Testing 각각에 대한 임의의 특허문서를 질의로 사용한 것이다. 먼저, 일반검색엔진 방법을 사용한 검색 결과 중 상위 10개에 대하여, Cleaning 기술을 질의했을 경우 상위 10개중에서 Cleaning기술은 5개로 50%의 비율을 보였다. Heating 기술을 질의 했을 경우는 상위 10개중에서 Heating에 관한 검색결과는 1개로 10%의 비율을 보였다. Testing 기술의 경우는 상위 10개중에서 Testing에 관한 검색결과가 3개로 30%의 비율을 보였다. 그 밖에 세 개의 기술에 속하지 않는 Noise 문서는 1개로 10%를 차지하고 있다. 다음으로, 제안된 알고리즘을 사용한 검색 결과 중 상위 10개의 경우 Cleaning 기술을 질의 했을 경우 Cleaning기술은 8개, Heating은 1개, Testing은 0개, Noise는 1개로 80%의 비율을 보였다. Heating 기술을 질의 했을 경우는 Cleaning기술은 0개, Heating은 10개, Testing은 0개, Noise는 0개로 100% 비율을 보였다. Testing 기술을 질의했을 경우는 Cleaning기술은 0개, Heating은 0개, Testing은 10개, Noise는 0개로 100% 비율을 보였다. 제안된 알고리즘을 사용하여 검색한 상위 10% 대한 검색 결과가 일반검색엔진을 사용한 검색 결과와 비교하여 성능이 향상된 것을 알 수 있다. Cleaning, Heating, Testing 각각의 질의에 대하여 Cleaning 질의의 경우에는 질의문서와 유사한 문서가 차지하는 비중이 50%에서 80%로 높아졌고, Heating 질의의 경우에는 10%에서 100%로, Testing 질의의 경우에는 30%에서 100%로 유사문서가 차지하는 비중이 높아졌다. 이처럼 제안된 알고리즘을 사용함으로써 질의문서와 유사한 문서를 상위로 랭킹 시키는 효과를 가지고 있음을 알 수 있다. 이에 반하여 상위 20% 대한 검색결과 중 Cleaning 질의의 경우에는 일반검색엔진에서 50%의 비중을 차지하는 것으로 보였으나, 제안된 알고리즘을 사용한 결과가 수치상으로는 45%로 낮아진 점을 볼 수 있다. 이와 비슷하게 상위30%에서는 세부기술 3개의 질의 모두에서 수치적으로 유사문서의 포함 비율이 수치적으로 낮아지는 결과를 볼 수 있다. 하지만 이러한 결과는 제안된 알고리즘을 사용한 검색결과가 질의문서와 유사한 순으로 나열하는 특징 때문에 유사한 문서가 상위10% 이내로 정렬 및 집중되는 현상 때문에 나타난 결과라고 할 수 있다. 특히 반도체 웨이퍼 기술에서 Cleaning의 경우는 사용된 관련 특허문서 10개중 8개가 상위 10%에 모두 포함되어 있다. 이처럼 검색결과의 포함범위가 좁을수록 검색정확도가 높아지고, 검색결과의 포함범위가 커질수록 검색정확도가 낮아지게 되는 현상을 보이는 것이다. 따라서 검색결과 정확도는 검색결과 포함범위를 좁게한 상위 10%에 대한 정확도 비교에 비중을 두어야 하겠다.

표 6은 검색결과 상위 10%, 20%, 30%에 대하여 일반 검색엔진과 제안된 알고리즘의 검색 정확도를 비교한 것이다. 표 6 에서 관련문서총수란 검색DB에 존재하는 질의 문서와 유사한 특허 문서의 총수를 나타낸다. 즉, 특허 문서DB에 존재하는 각각의 기술별 문서의 수를 나타낸다.

[표 6] 검색 정확도 비교
[Table 6] The compare of retrieval accuracy

방법	구분	질의 문서	관련 문서 총수	검색된 유사 문서수	검색정확도
일반검색 엔진	상위10%	C	10	5	50%
		H	25	1	4%
		T	30	3	10%
	상위20%	C	10	6	60%
		H	25	5	20%
		T	30	5	17%
	상위30%	C	10	7	70%
		H	25	8	32%
		T	30	8	27%
제안된 알고리즘	상위10%	C	10	8	80%
		H	25	10	40%
		T	30	10	33%
	상위20%	C	10	9	90%
		H	25	17	68%
		T	30	19	63%
	상위30%	C	10	10	100%
		H	25	21	84%
		T	30	22	73%

Cleaning 기술의 경우 관련문서총수는 10개, Heating 기술의 경우 관련문서총수는 25개, Testing기술의 경우 관련문서총수는 30개가 된다. 관련문서총수는 검색결과의 정확도 조사 범위가 10%, 20%, 30%로 커지더라도 수의 변화는 생기지 않는다. 표 6에서 검색된 유사 문서의 수는 일반 검색엔진과 제안된 알고리즘을 이용한 검색결과 범위의 각 10%, 20%, 30% 이내에 존재하는 유사문서의 수를 나타낸다. 식 (7)은 검색정확도 계산방법을 나타낸다.

$$\text{검색정확도} = \frac{\text{유사한 문서갯수}}{\text{관련문서 총 갯수}} * 100 \qquad (7)$$

일반검색엔진에서 Heating 및 Testing 기술문서 질의의 경우 상위 30%에서 각각 32%, 27%의 정확도를 나타낸다. 이 결과는 관련 질의 문서와 유사한 문서가 DB문서에서 분포가 넓게 퍼져 존재하고 있음을 나타낸다. 즉, 유사문서의 검색결과 분포가 집중되어 있지 않는 것을 알 수 있다. 반면, 제안된 알고리즘을 사용한 경우 상위 30%에서 Heating, Testing문서의 결과가 각각 84%, 73%의 검색정확도를 나타내고 유사문서의 검색결과 분포가 집중되어 있음을 알 수 있다. 상위10%범위에서 제안된 알고리즘을 사용하였을 때 각 질의에 대한 검색결과 정확도 향상 정도를 살펴보면, Cleaning의 경우 검색정확도가 일반검색엔진과 비교하여 50%에서 80%로, Heating의 경우 4%에서 40%로, Testing의 경우 10%에서 33%로 향상되었음을 알 수 있다.

4.3 터치

식(8)을 이용하여 터치기술을 Capacitance, Infrared, Resistance 기술의 키워드를 이용해 검색한 결과 총 225건의 문서가 검색되었다. Noise는 위 세 가지 기술 군에 포함되지 않은 데이터를 말한다. 편의를 위하여 각 표에서 Capacitance기술은 C, Infrared기술은 I, Resistance 기술은 R, Noise는 N으로 표기하였다.

(((touch and panel and (capacitance or infrared or resistance)).KEY. AND (G06F-003*).IPC.)) (8)

표 7은 검색 식(8)을 이용하여 검색한 상위 100개 문서를 나타낸다.

[표 7] 검색결과
[Table 7] The retrieval result

NO	발명의 명칭	분류
1	Touch Panel and Display...	R
2	DISPLAY DEVICE AND...	I
3	TOUCH SCREEN AND...	C
4	Capacitive Touch Screen...	N
5	TOUCH PANEL APPARATUS...	N
...	...	...
100	GHOST RESOLUTION FOR	C

표 8은 터치 기술의 상위 100개 데이터 구성을 나타낸다. 스크리닝 기법을 이용한 분석 결과 상위 100개의 검색결과 내에는 Capacitance 28개, Infrared 14개, Resistance 12개, Noise 46개로 구성되어 있다.

[표 8] 데이터 구성
[Table 8] The Data organization

영역 (Category)	터치 기술에 관한 특허문서 집합 (개수)
C	28개
I	14개
R	12개
N	46개
Total	100개

표 9은 검색결과에 대한 문서별로 차지하는 비중을 비교한 표이다.

[표 9] 검색결과 비중 비교
[Table 9] The compare of retrieval result rate

방법	구분	질의 문서	결과 (개수)				비율
			C	I	R	N	
일반 검색 엔진	상위10%	C	2	4	2	2	20%
		I	2	4	2	2	40%
		R	2	4	2	2	20%
	상위20%	C	8	4	2	6	40%
		I	8	4	2	6	20%
		R	8	4	2	6	10%
	상위30%	C	12	4	3	11	40%
		I	12	4	3	11	13%
		R	12	4	3	11	10%
제안된 알고리즘	상위10%	C	9	0	0	1	90%
		I	0	10	0	0	100%
		R	0	0	10	0	100%
	상위20%	C	18	0	0	2	90%
		I	3	14	0	3	70%
		R	0	0	12	8	60%
	상위30%	C	19	0	0	11	63%
		I	9	14	1	6	47%
		R	7	0	12	11	40%

먼저, 일반검색엔진 방법을 사용한 검색 결과의 경우, 상위 10개 중에서 각 Capacitance 기술, Infrared 기술, Resistance 기술이 차지하는 비율은 20%, 40%, 20%의 비율을 각각 보였다. 그 밖에 세 개의 기술에 속하지 않는 Noise 문서는 2개로 20%를 차지하고 있다. 다음으로, 제안된 알고리즘을 사용한 검색 결과의 경우, 상위 10개 중에서 Capacitance 기술, Infrared 기술, Resistance 기술이 차지하는 비율은 90%, 100%, 100%의 비율을 각각 보였다. 표 10은 일반검색엔진과 제안된 알고리즘 간의 검색 정확도를 비교한 결과이다.

[표 10] 검색 정확도 비교
[Table 10] The compare of retrieval accuracy

방법	구분	질의 문서	관련 문서 총수	검색된 유사 문서수	검색정 확도
일반 검색 엔진	상위 10%	C	28	2	7%
		I	14	4	29%
		R	12	2	17%
	상위 20%	C	28	8	29%
		I	14	4	29%
		R	12	2	17%
	상위 30%	C	28	12	43%
		I	14	4	29%
		R	12	3	25%
제안된 알고리즘	상위 10%	C	28	9	32%
		I	14	10	71%
		R	12	10	83%
	상위 20%	C	28	18	64%
		I	14	14	100%
		R	12	12	100%
	상위 30%	C	28	19	68%
		I	14	14	100%
		R	12	12	100%

일반검색엔진에서 Capacitance, Infrared, Resistance 질의의 경우 상위 10%에서 각각 7%, 29%, 17%의 정확도를 나타내었다. 반면, 제안된 알고리즘을 사용한 경우 Capacitance, Infrared, Resistance문서의 결과가 각각 32%, 71%, 83%의 검색정확도를 나타내었다. 상위10%범위에서 제안된 알고리즘을 사용하였을 때 검색결과 정확도가 일반검색엔진에 비하여 모두 향상되었음을 알 수 있다.

4.4 실험 결과 비교

표 11은 상위 10개 검색결과 범위에 포함되어 있는 질의문서와 유사한 문서의 비중을 비교한 표이다. 반도체 웨이퍼와 터치 기술 둘 다 상위 10개 문서에 대한 평균 유사문서 비중도가 일반 검색엔진 방법 보다 제안된 알고리즘이 높게 측정되었다. 상위 20개, 상위 30개의 경우에는 낮은 비중도를 보였으나, 검색에 있어 상위 10개의 검색 결과가 더 중요하므로, 나머지 상위 20개, 상위 30개의 평균 비중도는 본 실험에서 참고사항으로만 기재한 것이다. 세부 기술 군별로 유사문서 비중이 다르게 나온 이유는 데이터 구성상 세부 기술 군에 포함되지 않는 Noise데이터수가 다르기 때문이다.

[표 11] 상위 10개 문서에 대한 평균 비중 비교표
[Table 11] The average rating table about top 10 documents

구분	반도체 웨이퍼	터치
일반검색엔진	30%	27%
제안된 알고리즘	93%	97%

5. 결론 및 향후과제

본 논문에서는 키워드가 아닌 특허문서자체를 질의(Query)로 사용하였으며, 내용 기반 검색 알고리즘을 제안하였다. 검색결과는 질의문서와 유사한 문서 순으로 랭킹하여 나타내는 방법으로 나타내었다. 본 논문에서 제안된 방법은 특허문서에서 자주 등장하는 핵심키워드가 특허문서의 기술적 내용을 담고 있다고 가정하고 있다. 문서에는 자주 등장하지만 내용과는 관련이 없는 단어들을 먼저 제거하고 텍스트 분석을 이용한 핵심키워드를 찾는 과정을 통해 검색 정확도를 높였다. 검색결과정확도(Precision)측정을 위하여 일반검색엔진에서 검색 식을 이용하는 방법과 제안된 알고리즘을 이용하는 방법을 비교하는 실험을 수행하고 결과를 얻었다. 본 논문에서 제안한 내용 기반 검색방법은 특허정보조사 결과를 쉽게 얻을 수 있는 장점을 가진다. 즉, 특허문서 전체 텍스트를 검색에 그대로 이용할 수 있기 때문에 검색자가 특허 문서를 일일이 분석하는 수고를 덜 수 있고 유사한 문서 순으로 결과 검토가 이루어 질 수 있기 때문에 검색 이용자들에게 검색 편의성을 제공할 수 있다. 검색결과를 질의문서와 유사한 문서 순으로 검색결과 리스트 상위에 나타나게 함으로써 결과 검토에 소요되는 시간과 노력을 줄일 수 있어 검색의 효율을 높일 수 있다. 또한 각각의 검색결과 데이터들이 질의문서와 얼마나 유사한지 정도를 수치 값으로 제공할 수 있게 됨에 따라 유사정도에 대한 정량적인 정보도 줄 수 있으며, 검색을 하는 단계에서 유사한 문서를 우선 검색할 수 있게 하는 기회 등 특허정보검색에 대한 다양한 활용이 이루어 질 것으로 기대된다. 그리고 특허문서 검색 및 분류에 관한 연구 분야에서 키워드가 아닌 문서자체를 질의로 사용하여 특허문서의 전체 텍스트를 분석하는 내용 기반 검색방법의 가능성을 찾았다는 점에 있어 본 연구가 학술적으로 기여하길 기대한다. 본 논문의 한계점은 특허문서의 특성상 문서를 작성하는 작성자에 의해서 전략적으로 해당 기술에 대한 핵심어사용이 되지 않았다고 가정했을 때, 해당 문서가 검색에서 제외될 수 있는 문제점을 가지고 있다. 따라서 동일한 의미를 나타내지만 표현이 다른 유사단어 추출방법에 대한 연구를 추가적으로 수행하여 핵심어 추출 성능을 향상시킬 예정이다. 또한 향후 유사 문서 검색성능을 높이기 위해 문서 간 유사도를 측정하는 다양한 알고리즘을 비교 분석하는 연구를 수행하고자 한다.

References

[1] Korea Intellectual Property Office, Patent and Information Analysis, Korea Intellectual Property Office, December, 2007.

[2] http://www.kipris.or.kr/kor/use/use_1.jsp

[3] KIPO, “The understanding of Intellectual Property”, pp. 128-138, December, 2009.

[4] J. B. Baik, S. M. Kim and S. W. Lee, “Extracting Alternative Word Candidates for Patent Information Search”, Korean Institute of Information Scientists and Engineers, Vol. 15, Issue. 2, pp. 299-303, April, 2009.

[5] K. J. Son and S. J. Lee, "Weighting Methods for compound Nouns in Patent Retrieval System", Korean Institute of Information Scientists and Engineers, Vol. 31, Issue. 1, pp. 895-897, April, 2004.

[6] H. G. Kim, S. H. Lee and Y. H. Mook, "Patent Search System Using IPC Clustering", Korea Contents Association, vol. 5, Issue. 2, pp. 103-106, November, 2007.

[7] L. H. Tong, H. Cong and S. Lixiang, "Automatic classification of patent documents for TRIZ users", World Patent Information, Vol. 28, Issue. 1, pp. 6-13, March, 2006.

[8] Y. L. Chen and Y. T. Chiu, "An IPC-based vector space model for patent retrieval", Information Processing & Management, Vol. 47, Issue. 3, pp. 309-322, May, 2011.

[9] K. K. Lai and S. J. Wu, "Using the patent co-citation approach to establish a new patent classification system", Information Processing & Management, Vol. 41, Issue. 2, pp. 313-330, March, 2005.

[10] WIPS, http://search.wips.co.kr/

[11] I. Feinerer, K. Hornik and D. Meyer. "Text mining infrastructure in R", Journal of Statistical Software, Vol. 25, Issue. 5, pp. 1-54, March 2008.

[12] WIKIPEDIA, http://en.wikipedia.org/wiki/Vector_space_model

[13] G. Salton and M. J. McGill, “Introduction to modern information retrieval”, McGraw-Hill, 1983.

고 광 수(Gwang-su Go) **[준회원]**

- 2009년 6월 : 고려대학교 전자 및 정보공학부 (이학사)
- 2011년 2월 ~ 현재 : 고려대학교 산업경영공학과 석사과정

<관심분야>
특허 정보 분석, 지식관리, 패턴인식

박 상 성(Sang-Sung Park) **[정회원]**

- 2006년 2월 : 고려대학교 산업시스템공학과 (공학박사)
- 2006년 5월 ~ 현재 : 고려대학교 BK21 사업단 연구교수

<관심분야>
컴퓨터 비전, 패턴인식, 전문가시스템응용, 지식관리

정 원 교(Won-Kyo Jung) **[정회원]**

- 2007년 2월 : 경희대학교 산업공학과 (공학사)
- 2009년 2월 : 고려대학교 정보경영공학부 (공학석사)
- 2009년 3월 ~ 현재 : 고려대학교 정보경영공학부 박사과정

<관심분야>
객체지향응용, 프레임웍, 정보시스템

장 동 식(Dong-Sik Jang) **[정회원]**

- 1979년 : 고려대학교 산업공학과 (공학사)
- 1985년 : 텍사스 주립대학 산업공학과 (공학석사)
- 1988년 : 텍사스 A&M 산업공학과 (공학박사)
- 1989년 ~ 현재 : 고려대학교 정보경영공학부 교수

<관심분야>
Computer Vision, 최적화이론, 컴퓨터 알고리즘

신 영 근(Young-Geun Shin) **[준회원]**

- 2005년 2월 : 고려대학교 산업시스템정보공학과 (공학사)
- 2005년 9월 ~ 현재 : 고려대학교 산업시스템정보공학과 석 박사 통합과정

<관심분야>
패턴인식, 스케줄링, 인공지능