# 문서의 의미적 구조정보를 이용한 특허 문서 분류

김재호, 최기선
한국과학기술원 전산학과
전문용어언어공학연구센터/언어자원은행
{jjaeh, kschoi}@world.kaist.ac.kr

# Patent Document Categorization based on Semantic Structural Information

Jae-Ho Kim, Key-Sun Choi
KAIST Computer Science Dept.
KORTERM / BOLA

**요 약**

특허 검색은 수많은 특허 문서 중에서 특정 해당분야의 문서집합 내에서 검색을 수행하기 때문에 정확한 특허 분류에 크게 의존하게 된다. 이러한 특허 분류의 중요성에 덧붙여, 특허 문서의 수가 빠르게 증가하게 되면서 특허를 자동으로 분류하려는 요구가 더욱 필요하게 되었다. 특허문서는 일반문서와는 달리 구조화되어 있기 때문에 특허분류를 하기 위해서는 이러한 점이 고려되어야 한다. 본 논문에서는 k-NN 방법을 이용하여 일본어 특허 문서를 자동으로 분류하는 방법을 제안한다. 훈련집합으로부터 유사문서를 검색할 때, 구조화되어 있는 특허 문서의 특징을 이용한다. 문서 전체가 아닌 [기존 기술], [응용 분야], [해결하고자 하는 문제], [문제를 해결하려는 방법] 등의 세분화된 요소끼리 비교하여 유사성을 계산한다. 특허 문서에는 사용자가 정의한 많은 의미요소가 있기 때문에 먼저 이들을 군집화한 후에 이용한다. 실험 결과 제안한 방법이 특허문서를 그대로 이용하는 것보다는 74%, 특허문서에 나타난 <요약>, <청구항>, <상세한 설명>의 큰 구조정보를 이용하는 것보다는 4%의 성능 향상을 가져왔다.

## 1 서론

특허 제출자는 특허를 제출하기에 앞서 먼저 관련된 기존 특허를 미리 검색하고, 특허 조사관은 제출된 특허가 새로운 것인지 기존 특허를 검색한다. 수많은 특허 문서 중에서 특정 해당분야의 문서집합 내에서 검색을 수행하기 때문에 특허 검색은 정확한 특허 분류에 크게 의존하게 된다. 이러한 특허 분류의 중요성에 덧붙여, 특허 문서의 수가 빠르게 증가하게 되면서 특허를 자동으로 분류하려는 요구가 더욱 필요하게 되었다.

그 동안 k-NN [1, 2], 결정 트리 [3], 베이시안 분류 [4, 5], SVM [6], 신경망 [7, 8] 등 다양한 기계학습 기법과 통계적 방법들이 문서 자동 분류에 적용되어 왔다. 특허도 일종의 문서이기 때문에 특허 분류에 이러한 기법들이 적용 가능하다. 하지만 특허 문서는 일반 문서와는 다른 몇 가지 특징을 가지고 있기 때문에 그 특징들이 고려되어야 한다. 특허 문서의 특징을 요약하면 다음과 같다 [9].

1. *청구항, 발명의 목적, 효과, 구현 등으로 내용이 구조화되어 있다.*
2. *발명의 범위를 넓히기 위하여 청구항에는 모호하고 일반적인 용어를 사용한다.*
3. *많은 기술 용어를 포함한다. 다른 특허에서는 사용하지 않는 용어를 사용하거나 정의하기도 한다.*
4. *문서의 길이가 다양하다. 일본어 어떤 특허 문서는 일본어 3만 자를 포함한다.*

위 특징 중에서 본 논문에서는 첫 번째 특징에 초점을 맞추어 일본어 특허 문서를 대상으로 특허 분류를 수행한다.

본 논문에서 대상으로 하고 있는 일본어 특허 문서는 표 1과 같이 <서지정보>, <요약>, <청구항>, <상세한 설명>, <도면의 설명>, <도면> 이렇게 6개의 큰 영역으로 이루어져 있다. <요약>과 <상세한 설명>은 [기존 기술], [이용 분야], [해결하려는 문제], [해결하려는 수단] 등의 세분화된 요소로 구성되어 있다. [기존 기술]과 [응용 분야]는 기술적 배경과 기술분야에 관련된 정보를 포함하고 있기 때문에 다른 부분보다 분류에 도움이 될 수 있다. [발명의 목적]과 [해결하고자 하는 문제]는 특허 문서를 대표하여 <요약>에 주로 사용되기 때문에 <청구항>와 더불어 중요하다고 볼 수 있다. 그러므로 이러한 세분화된 요소를 분류의 자질로 고려한다면, 특허 분류에서 좋은 성능을 얻을 수 있을 것이다.

**표 1. 일본어 특허 문서의 전체 구조**

| <DOCNO>PATENT-JA-UPA-1995-000001</DOCNO> | |
|---|---|
| **<서지정보>** | **<SDO BIJ>** |
| [공개일] | (43)【公開日】平成７年（１９９５）１月６日 |
| [발명의 명칭] | (54)【発明の名称】スラリ散布を行う土壌作業機 |
| | ..... |
| **<요약>** | **<SDO ABJ>** |
| [목적] | 【目的】 スラリの処理と土壌作業を ..... |
| [구성] | 【構成】 トラクタとスラリを積載し ..... |
| **<청구항>** | **<SDO CLJ>** |
| [청구항1] | 【請求項１】 バキウムカーを牽引し ..... |
| [청구항2] | 【請求項２】 トラクタに対して..... |
| **<상세한 설명>** | **<SDO DEJ>** |
| [산업상의 이용분야] | 【産業上の利用分野】本発明はスラリ散布を行う土壌作業機に関し、..... |
| [발명이 해결하려는 문제] | 【発明が解決しようとする課題】このようなスラリを圃場に供給する..... |
| [문제를 해결하려는 수단] | 【課題を解決するための手段】上述のような目的を達成するために、..... |
| [작용] | 【作用】本発明のスラリ散布を行う ..... |
| [실시예] | 【実施例】以下、本発明を採用した ..... |
| [발명의 효과] | 【発明の効果】以上の説明から明..... |
| **<도면의 설명>** | **<SDO EDJ>** |
| [도1] | 【図１】本発明のスラリ散布を..... |
| **<도면>** | **<SDO DRJ>** |
| [도1] | 【図１】..... |

본 논문에서는 주어진 특허 문서를 그와 유사한 특허 문서의 분류코드에 따라 분류한다. 훈련집합으로부터 유사문서를 검색할 때, 전체 문서가 아닌 앞에서 설명한 같은 세분화된 요소끼리 비교하여 유사성을 계산한다. 특허 문서에는 사용자가 정의한 많은 세분화된 요소가 있기 때문에 먼저 이들을 군집화한 후에 이용한다.

## 2 관련연구

이전까지의 연구에서 특허의 구조적 특성을 적절히 잘 이용한 방법은 없었다.

[10]은 k-NN에 기반하여 특허를 US 특허 코드로 분류하는 도구를 개발하였다. 그는 문서전체가 아닌 제목, 요약, 청구항 그리고 설명부분의 첫 20 줄만을 색인하여 분류에 이용하였다. [11]은 특허 분류에 Winnow 알고리즘 [12]을 사용하였다. 학습자질로 전체문서 대신 요약부분만을 이용하였으나 더 좋은 성능을 보이지는 못하였다. [13]은 특허 문서의 다양한 부분을 색인한 후 분류에 이용하였다. 제목부분 (a), 청구항 부분 (b), 그리고 제목, 고안자, 출원자, 발명기관, 요약, 설명 부분에서 처음 300 단어 (c), 마지막으로 제목, 고안자, 출원자, 요약 부분 (d) 이렇게 4가지 부분을 각각 색인하였다. 실험적으로 (c)가 분류 정확률에서 가장 좋은 성능을 보였다.

기존 연구들은 대부분 특허 문서 구조의 의미적 분석 없이 특허의 큰 영역정보나 문서의 앞 일부분만을 이용하였을 뿐이다.

## 3 특허 분류 방법

### 3.1 전체 시스템 구조

그림 1은 특허 분류 시스템의 전체 구조를 나타낸다. 시스템은 문서 색인, 문서 검색, 분류 세 과정으로 이루어져 있다.

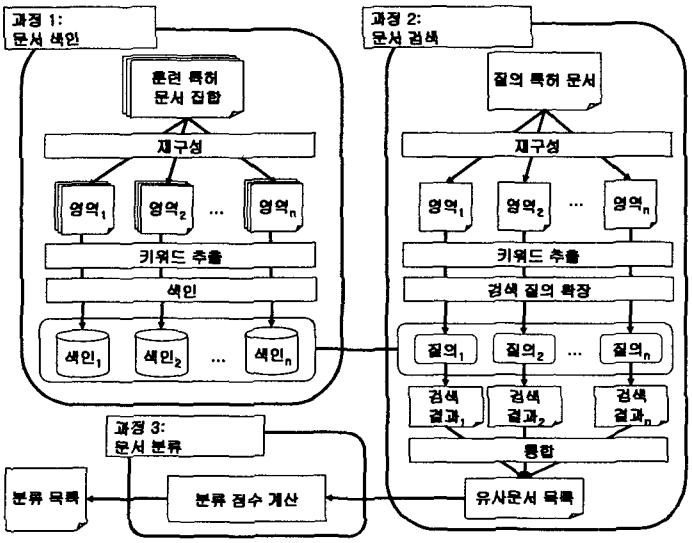


**그림 1. 특허 분류 시스템의 전체 구조**

첫 번째 문서 색인 과정에서는 분류할 질의문서에 대한 유사문서를 검색하기 위하여 훈련집합의 문서를 색인한다. 색인에 앞서 먼저, 훈련 집합의 모든 특허 문서를 미리 정의한 K개의 의미에 따라 재구성하고 각 의미 별로 나눈다. 나뉘어진 각 부분에서 키워드를 추출하여 검색을 위한 색인 파일을 각각 생성한다. 본 논문에서는 문서 색인과 검색에 Lemur 툴킷 [14]을 사용하였다.

두 번째 문서 검색 과정에서는 앞에서 만든 색인 파일을 이용하여 분류할 질의문서에 대한 유사문서를 검색한다. 먼저, 분류할 대상 문서도 훈련문서와 마찬가지로 K개의 의미에 따라 재구성한 후 키워드를 추출하여 K개의 검색 질의로 만든다. K개의 질의를 K개의 색인 파일에서 검색한 결과를 합쳐서 질의에 대한 유사문서 목록을 생성한다. 마지막으로 세 번째 분류 과정에서는 앞에서 검색된 유사문서들의 분류코드를 이용하여 질의 문서의 분류코드를 제시한다. 이때, 분류코드 별 점수는 검색 과정에서 구한 유사문서의 유사도 점수와 순위가 고려되어 계산된다.

### 3.2 문서 색인

이 절에서는 분류할 질의문서에 대한 유사문서를 검색하기 위하여 훈련집합의 문서를 문서의 구조적 정보를 이용하여 색인하는 과정을 설명한다.

색인에 앞서, 먼저 특허 문서의 구조를 살펴볼 필요가 있다. 본 논문에서 대상으로 하고 있는 일본어 특허 문서는 서론에서 설명한 바와 같이 <서지정보>, <요약>, <청

구항>, <상세한 설명>, <도면의 설명>, <도면> 이렇게 6개의 큰 영역으로 이루어져 있고, <요약>과 <상세한 설명>은 [기존 기술], [이용 분야], [해결하려는 문제], [해결하려는 수단] 등의 세분화된 요소로 구성되어 있다. 여기서 큰 영역의 제목은 고정된 제목인 반면, 세분화된 요소의 제목은 사용자가 정의하고 쓰는 제목이기 때문에, 사용자가 부여한 의미가 들어 있다고 볼 수 있다. 본 논문에서는 이것을 사용자 태그라 부르기로 한다.

같은 내용이라 하더라도 글 쓰는 사람에 따라 다르게 제목을 붙일 수 있기 때문에 사용자 태그의 종류는 매우 다양하다. 실제로 1993년 일본어 특허 문서 347,227건에서 사용자 태그를 추출한 결과, 3,516 종류의 사용자 태그가 추출되었다. 이 사용자 태그를 이용하기 위해서는 이것들을 군집화하여 몇 개로 줄여야 한다. 표 2는 고빈도의 사용자 태그의 예를 보여준다.

**표 2. 1993년 특허문서에서 추출한 상위 10개의 사용자 태그**

| 순위 | 빈도 | 사용자 태그 (일본어) | 사용자 태그 (한국어) |
|---|---|---|---|
| 1 | 346,157 | 実施例 | 실시예 |
| 2 | 335,300 | 構成 | 구성 |
| 3 | 330,757 | 産業上の利用分野 | 산업상의 이용분야 |
| 4 | 311,015 | 従来の技術 | 종래의 기술 |
| 5 | 310,276 | 課題を解決するための手段 | 문제를 해결하고자 하는 수단 |
| 6 | 309,026 | 目的 | 목적 |
| 7 | 307,602 | 発明の効果 | 발명의 효과 |
| 8 | 306,350 | 発明が解決しようとする課題 | 발명이 해결하고자 하는 문제 |
| 9 | 243,012 | 作用 | 작용 |
| 10 | 176,676 | 表 | 예 |

사용자 태그에 나타나는 중심어에 의해서 사용자 태그들을 군집화할 수 있다. 먼저 사용자 태그의 마지막 명사는 중심어라는 간단한 규칙을 이용하여 사용자 태그에서 중심어를 추출하여 그 빈도수로 정렬한다. 추출된 1,475개의 중심어 중에서 100개의 고빈도 중심어를 수작업으로 군집화한다. 본 논문에서는 중심어를 "기술분야", "목적", "해결방법", "청구", "설명", "예" 6개의 의미 태그로 분류한다.

100개의 중심어에 의해 1,940종류의 사용자 태그가 분류된다. 이는 누적빈도로 보았을 때, 사용자 태그 전체 빈도의 99.86%에 해당하는 수이다. 표 3은 6개의 의미 태그로 분류된 사용자 태그의 예를 보여 준다.

이때, "**課題を解決するための手段及び作用** (과제를 해결하려는 수단 및 작용)"와 같이 등위 접속사로 연결된 사용자 태그는 "해결방법" 과 "설명" 으로 다중 분류가 가능하게 한다.

6개의 의미 태그 내에서 주로 사용되는 기술 패턴을 구축한다면, 사용자 태그 제목 정보가 없어도 기계학습을 통해 해당 내용을 분류할 수 있을 것이다. 그러나 본 논문에서 사용자 태그의 중심어 비교 외의 다른 처리는 하지 않았다. 중심어로 분류된 1,940개 외의 사용자 태그는 무시되었다.

**표 3. 분류된 사용자 태그의 예 (굵은 글씨는 중심어)**

| 의미 태그 | 사용자 태그의 예 |
|---|---|
| 기술분야 | 産業上の**利用分野** (산업상의 **이용분야**)<br>従来の**技術** (종래의 **기술**)<br>発明の**背景** (발명의 **배경**) |
| 목적 | 発明の**名称** (발명의 **명칭**)<br>発明の**目的** (발명의 **목적**)<br>発明が解決しようとする**課題**<br>(발명이 해결하고자 하는 **과제**) |
| 해결방법 | 問題点を解決するための**手段**<br>(문제점을 해결하려는 **수단**)<br>課題を解決するための手段及び作用<br>(과제를 해결하려는 **수단** 및 **작용**) |
| 청구 | <청구의 범위> 부분 안에 있는 모든 사용자 태그 |
| 설명 | 発明の**効果** (발명의 **효과**)<br>課題を解決するための手段及び作用<br>(과제를 해결하려는 **수단** 및 **작용**)<br>発明の具体的**説明** (발명의 구체적인 **설명**) |
| 예 | **実施例** (실시**예**), 第**実施例** (제실시**예**)<br>**参考例** (참고**예**), **実験例** (실험**예**) |

이렇게 구해진 6개 의미 태그 별로 내용을 모아 그림 2와 같이 특허 문서를 재구성한다.

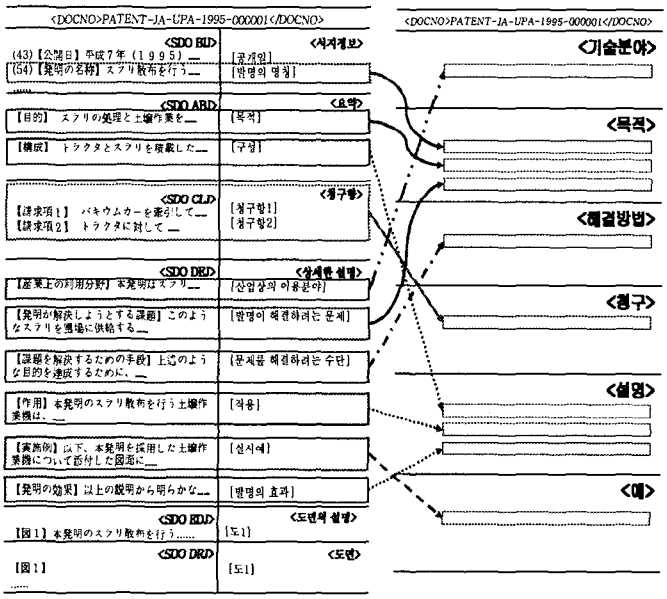


**그림 2. 의미 태그에 기반한 특허 문서의 재구성**

어떤 부분은 지워지기도 하고, 어떤 부분은 다중 분류로 인하여 중복되어 여러 군데에 들어가기도 한다. 각 의미 태그 내의 내용에서 키워드를 추출하여 검색을 위한 색인 파일을 각각 생성한다. 키워드는 단일 명사로 한정하였다.

### 3.3 문서 검색

이 절에서는 앞에서 만든 색인 파일을 이용하여 분류할 질의문서에 대한 유사문서를 검색하는 과정을 설명한다. 분류할 대상 문서도 역시 6개의 의미 태그에 따라 재구성한 후 단일 명사를 키워드로 추출하여 6개의 질의를 만든다. 이때, 질의문서 내 tf (term frequency)만을 키워드 별 가중치로 사용한다.

이 때 검색의 정확도를 높이기 위해서 불필요한 단어는 검색 질의에서 제거한다. 1993년 347,227개의 문서에서의 고빈도 df (document frequency)값을 가지는 명사 500개 중 수작업으로 67개를 불용어로 모았다. **こと** (것), **発明** (발명), **目的** (목적), **問題** (문제), **課題** (과제), **請求** (청구), **記載** (기재) 등이 그 예이다.

유사 문서를 검색할 때, 문서 전체가 아닌 같은 의미 태그 별 내용(의미 영역이라 부르기로 한다)을 비교한다. 기술분야가 같고, 해결하려는 문제와 해결 방법이 같으면 유사한 문서로 본다는 가정에서 나온 것이다. 그림 3과 같이 질의문서와 대상 문서 사이의 같은 의미 영역끼리 비교하여 유사문서를 검색한다.

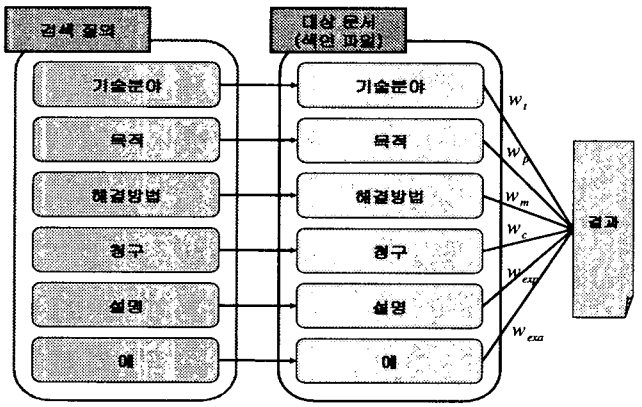


**그림 3. 유사 문서 검색 방법 (1대 1 매핑)**

그러나 이렇게 1대 1 매핑만 하게 되면 다음과 같은 이유로 성능이 더 떨어질 수도 있다.

1. 특허의 청구범위를 넓히기 위하여 청구항에 사용되는 단어들은 모호하고 일반적인 용어가 주로 사용된다. 그래서 청구 영역끼리 비교하면 재현율이 떨어질 수 있다.
2. 사용자가 정의한 사용자 태그를 100% 신뢰할 수 없다. 사용자는 "[해결하고자 하는 문제]"라고 쓰고서는 해결하는 방법에 대해서도 같이 기술할 수도 있다.
3. 본 방법의 의미 태그 분류를 100% 신뢰할 수 없다. 중심어를 기준으로 사용자 태그를 군집화하였다고 하지만 오류는 존재하기 마련이다. "과제의 설명"은 "목표"로 분류되어야 하지만 본 방법에 의하면 "설명"으로 분류된다.

이와 같은 이유로 그림 4그림 4와 같은 비교를 허용한다.

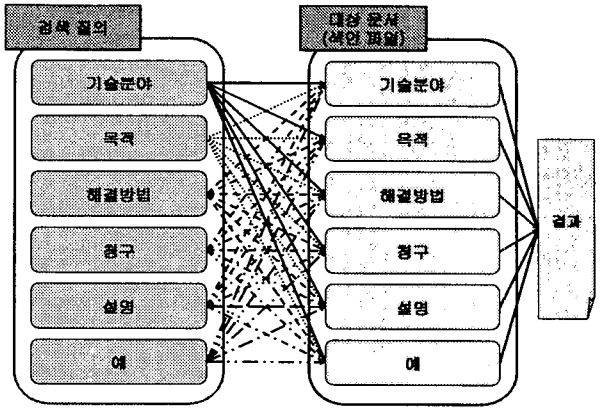


**그림 4. 유사 문서 검색 방법 (N 대 N 매핑)**

교차 비교로 나온 36개의 결과를 합하여 하나의 문서 목록을 생성한다. 합산할 때 개발용 평가 집합에서 학습한 가중치 값을 부여하여 합산하도록 한다. 최종적으로 N개의 문서가 유사문서로 검색된다. 본 논문에서는 N을 200으로 잡았다. 이렇게 교차 비교를 허용함으로써 앞 단계에서 발생하는 의미 태그 별 분류 오류로부터의 복귀를 허용하게 된다.

### 3.4 문서 분류

마지막으로 분류 과정에서는 앞에서 검색된 유사 문서들을 이용하여 질의 문서의 분류코드를 제시한다. 분류코드 별 점수를 계산할 때, 식 (1)과 같이 유사문서의 유사도 점수와 순위를 고려한다.

$$Score_{category}(c) = \sum_{\{d \mid c \in categories\ of\ doc\ d\}} Score_{doc}(d) \times weight_{doc}(d) \qquad \text{식 (1)}$$

$$weight_{doc}(d) = \begin{cases} 1 & rank(d) \le k \\ \alpha & k < rank(d) \le N \ (0 \le \alpha < 1) \end{cases} \qquad \text{식 (2)}$$

$Score_{doc}(d)$ 은 유사문서로 검색된 문서 d 의 유사도 점수이다. $rank(d)$ 는 문서 d가 유사문서로 검색된 순위이다. 문서 가중치 $weight_{doc}(d)$ 는 문서가 k등 이내일 때는 1을, k등보다 크고 N(=200)등 이내일 때는 $\alpha$ 값을 받게 된다. 문서 유사도 점수와 가중치가 곱해진 값이 해당 문서의 분류코드 c 별로 합산이 되어 최종적으로 분류코드 점수 $Score_{category}(c)$ 가 계산되고, 시스템은 이 값을 순위화하여 최종 분류코드 목록을 제시한다.

## 4 실험 및 결과

본 논문에서 제시한 특허 분류 방법을 검증하기 위하여 NTCIR5 특허 분류 태스크 [15]에서 제공된 평가 집합으로 성능을 평가한다. 특허 분류 태스크에는 주제 분류[1]와 F-term 분류[2]가 있는데 본 논문에서는 주제 분류만 실험하였다.

질의 집합은 2,008개의 문서로 구성되어 있으며, 각 질의 문서당 2,520개의 가능한 주제 중에서 상위 100개의 주제를 분류 결과로 제시한다. 보통 각 질의 문서는 한 두 개의 주제로 분류된다. 훈련 집합으로 1993년부터 1997년까지의 170만 건의 특허 문서가 제공되었다.

재현율과 정확률은 각 질의문서 별로 계산이 되어 MAP (Mean Average Precision)값으로 요약된다. 이것은 문서 검색에서의 TREC 스타일의 평가 방법이다. 여기에서는 질의는 질의 문서로, 검색된 문서는 부여된 분류코드로 하여 MAP가 계산되며, 모든 질의 문서에 대한 MAP값의 평균값으로 분류시스템의 성능을 평가한다.

### 4.1 개발용 실험집합에서의 실험 결과

본 실험에 앞서 본 논문에서 제안한 방법을 작은 개발용 실험집합에서 검증하고 교차 비교시의 가중치도 학습하여 구한다. 개발용 실험집합은 다음과 같이 구성되었다. 질의 문서는 1995년 문서집합 중 임의로 1,000개를 선택하였으며, 문서검색을 위한 훈련집합은 1993년 문서집합에서 앞부분의 50,000건을 선택하였다.

본 논문에서 제안한 방법에 대한 비교를 위하여 문서 내 의미적 구조정보를 전혀 이용하지 않고 전체문서를 그대로 이용하는 베이스라인 시스템을 구축하였다. 즉, 질의문서의 문서전체로 검색질의를 만들고, 훈련집합의 문서전체를 대상으로 검색하여 분류를 수행하였다. 베이스라인 시스템의 MAP는 0.2939가 나왔다. 이 때, 문서분류 과정에서 식 (2)의 k는 10, $\alpha$는 0.1로 설정하여, 검색된 유사 문서 상위 10개를 중심으로 분류를 하게 하였다.

다음 표 4는 의미 태그에 기반하여 검색질의와 색인 파일을 만들어 교차 비교 검색을 수행한 결과를 나타낸다. 베이스라인보다 좋은 성능을 보인 결과는 음영을 넣어 표시하였다.

**표 4. 의미 태그 별 교차 비교를 통한 분류 결과 (MAP)**

| 대상 / 질의 | 기술 분야 | 목적 | 해결 방법 | 청구 | 설명 | 예 |
|---|---|---|---|---|---|---|
| 기술분야 | **0.4149** | 0.3691 | 0.2711 | 0.3719 | 0.2360 | 0.3022 |
| 목적 | **0.4041** | **0.4180** | 0.2897 | 0.3567 | 0.2881 | 0.3080 |
| 해결방법 | 0.3385 | 0.3033 | 0.2839 | 0.3592 | 0.2028 | 0.2885 |
| 청구 | 0.3768 | 0.3542 | 0.3390 | **0.4201** | 0.2432 | 0.3182 |
| 설명 | 0.3402 | 0.3291 | 0.2523 | 0.3245 | 0.2241 | 0.2741 |
| 예 | 0.2607 | 0.1933 | 0.1330 | 0.2512 | 0.1097 | 0.2283 |

[1] 주제 분류에서는 각 특허에 대하여 기술적 분야와 같은 하나 이상의 주제를 결정한다.

[2] F-term 분류에서는 특정 주제 내에서 각 특허에 대하여 하나 이상의 F-term (목적, 용도, 구조, 재료, 제법, 제어 수단 등의 기술적 관점에서의 분류 코드)을 제시한다.

첫 번째 줄의 첫 번재 열은 기술분야 질의로 기술분야 색인파일을 검색하여 분류한 것에 대한 MAP값을 나타낸다. 특정 조합들은 베이스라인보다 좋은 성능을 보임을 알 수 있다. 대체로 "기술분야", "목적", "청구" 영역이 다른 영역보다 유사문서를 검색하는 데 더 적합한 정보를 가지고 있음을 알 수 있다.

위 교차 비교 시 가중치를 부여하여 통합한 최종 결과의 성능은 표 5와 같다.

**표 5. 교차 비교 결과를 합친 최종 분류 결과**

| 대상 / 질의 | 가중치 $w_{t1}$ (1,1,1,1,1,1) | 가중치 $w_{t2}$ (1,1,0,1,0,0) | 가중치 $w_{t3}$ (a,b,c,d,e,f) |
|---|---|---|---|
| 기술분야 | 0.4496 | 0.4581 | 0.4641 |
| 목적 | 0.4655 | 0.4683 | 0.4694 |
| 해결방법 | 0.4102 | 0.3988 | 0.4243 |
| 청구항 | 0.4578 | 0.4513 | 0.4608 |
| 설명 | 0.4182 | 0.4139 | 0.4242 |
| 예제 | 0.3403 | 0.3409 | 0.3744 |
| 가중치 $w_{q1}$ (1,1,1,1,1,1) | 0.4590 | 0.4662 | 0.4908 |
| 가중치 $w_{q3}$ (1,1,0,1,0,0) | **0.5011** | **0.5104** | **0.5126** |

위 표의 첫 여섯 행은 각 질의 별로 6개의 대상에 대한 검색결과를 가중치를 다르게 부여하여 합한 결과이다. 첫 번째 열은 가중치를 모두 동일하게 주고 6개의 결과를 모두 더해서 합친 것이고 ($w_{t1}$), 두 번째 열은 성능이 상대적으로 높게 나온 "기술분야", "목적", "청구"만을 합친 것이다 ($w_{t2}$). 세 번째 열은 표 4의 MAP 결과에 비례하게 가중치를 부여하고 합친 것이다 ($w_{t3}$). 예를 들어, "목적" 질의와 "목적" 대상 사이의 검색 가중치는 다음과 같이 계산된다.

$$w_{22} = \frac{0.4180}{0.4041+0.4180+0.2897+0.3567+0.2881+0.3080} = 0.2025$$

질의는 "기술분야", "목적", "청구" 만을 사용하는 $w_{q2}$가 성능이 더 높게 나왔으며, 결과를 동일한 가중치를 주어 합하는 $w_{t1}$보다는 다른 가중치를 주어 합치는 $w_{t2}$, $w_{t3}$가 성능이 높게 나왔음을 알 수 있다. 이 결과는 베이스라인 시스템의 결과보다 성능이 74% 향상된 결과이다.

특허 문서를 구조적 사용자 태그에 기반하여 재구성하는 본 논문의 방법이 효과적임을 보이기 위하여 비교 실험을 수행하였다. 특허 문서의 고정된 큰 구조 영역 (<요약>, <청구항>, <상세한 설명>) 정보만을 이용한 실험으로, 결과는 표 6과 같다.

**표 6. 특허 문서의 고정된 큰 구조 정보를 이용한 실험**

| 질의 \ 대상 | <요약> | <청구항> | <상세한 설명> | 가중치 (1,1,1) | 가중치 (1,1,0) |
|---|---|---|---|---|---|
| <요약> | 0.4631 | 0.4055 | 0.4038 | 0.4815 | 0.4578 |
| <청구항> | 0.4546 | 0.4201 | 0.3788 | 0.4741 | 0.4614 |
| <상세한 설명> | 0.4449 | 0.2095 | 0.2914 | 0.4218 | 0.3738 |
| 가중치 (1,1,1) | 0.4946 | 0.3270 | 0.3565 | 0.4731 | 0.4550 |
| 가중치 (1,1,0) | 0.4838 | 0.4319 | 0.4057 | 0.4918 | 0.4826 |

<상세한 설명> 부분을 이용함으로 인하여 성능이 더 떨어짐을 볼 수 있었으며, 위 실험의 최고 성능 0.4946보다 본 논문에서 제안한 방법의 성능 0.5126이 더 좋음을 볼 수 있었다. <상세한 설명>과 <요약> 부분에서 유용한 부분만을 추출하여 이용하는 것이 효과적임을 알 수 있다.

본 논문에서는 유사문서를 검색하여 그 문서들의 분류코드에 의하여 질의문서의 분류코드를 결정하였다. 유사문서로 몇 개를 보는 것이 좋은 지를 실험을 통해 검증하였다. 검색 방법은 베이스라인과 질의 가중치는 $w_{q2}$로 고정하고 대상 가중치를 달리한 방법을 사용하였다. 표 7과 표 8이 그 결과를 나타내며, $\alpha$를 0.1이라도 부여하여 k개 이후의 문서라도 점수를 작게 하여 고려하는 것이 성능이 좋았으며, k는 5에서 20 정도에서 좋은 성능을 보였다.

**표 7. k에 따른 분류 결과의 변화 실험 ($\alpha$=0.1)**

| 방법 \ k | 1 | 5 | 10 | 20 | 50 | 100 |
|---|---|---|---|---|---|---|
| Baseline | 0.2941 | 0.3033 | 0.2939 | 0.2923 | 0.2796 | 0.2801 |
| $w_{q2}$ $w_{t1}$ | 0.4976 | 0.5046 | 0.5011 | 0.5015 | 0.5073 | 0.5000 |
| $w_{q2}$ $w_{t2}$ | 0.4986 | 0.5100 | 0.5104 | 0.5181 | 0.5168 | 0.5105 |
| $w_{q2}$ $w_{t3}$ | 0.5056 | 0.5117 | 0.5126 | 0.5174 | 0.5167 | 0.5111 |

**표 8. k에 따른 분류 결과의 변화 실험 ($\alpha$=0)**

| 방법 \ k | 1 | 5 | 10 | 20 | 50 | 100 |
|---|---|---|---|---|---|---|
| Baseline | 0.1727 | 0.2492 | 0.2631 | 0.2790 | 0.2762 | 0.2801 |
| $w_{q2}$ $w_{t1}$ | 0.3295 | 0.4399 | 0.4742 | 0.4898 | 0.5070 | 0.5000 |
| $w_{q2}$ $w_{t2}$ | 0.3224 | 0.4502 | 0.4828 | 0.5092 | 0.5154 | 0.5105 |
| $w_{q2}$ $w_{t3}$ | 0.3303 | 0.4525 | 0.4911 | 0.5092 | 0.5146 | 0.5111 |

### 4.2 평가용 실험집합에서의 실험 결과

본 논문의 제안 방법은 NTCIR5 특허분류 태스크 참가 결과에 의해 평가될 수 있다. 31개의 시스템이 "주제 분류" 태스크에 결과를 제출하였으며, 본 논문 저자의 팀에서는 k 값을 달리 하여 (10, 20, 30, 50, 100) 5개의 다른 결과를 제출하였다. 표 9는 제출된 결과에 대한 평가를 보여준다. 시스템 ID ft001과 ft005가 본 논문 저자 팀의 결과이다.

**표 9. NTCIR5 특허 분류 태스크의 주제 분류 평가 결과**

| 순위 | 시스템 ID | 질의 문서 | 방법 | 문서 자질 | MAP |
|---|---|---|---|---|---|
| 1 | ft001 | some part of patent document | KNN | word | 0.6872 |
| 5 | ft005 | some part of patent document | KNN | word | 0.6666 |
| 6 | ft018 | full text of patent document | Naïve Bayes | noun phrases | 0.6591 |
| 9 | ft028 | PAJ Abstract | KNN | N-gram characters | 0.6192 |
| 19 | ft010 | Technical field, prior art and subject to be solved | Simple vector-space model | Words | 0.4886 |
| 22 | ft016 | Claim | Simple vector-space model | Words | 0.4279 |
| 23 | ft012 | Abstract in patent application | Simple vector-space model | Words | 0.424 |
| 29 | ft006 | some part of patent documents | MEM | word | 0.3776 |

지금 이 순간 아직 NTCIR5 워크샵이 열리지 않았기 때문에 다른 팀의 방법을 알지 못한다. 다만 결과를 제출할 때 적어낸 주석을 통해서 추정할 수 있다. 예를 들면, 시스템 ID ft018은 질의문서 전체로부터 검색 질의를 생성하였으며, Naïve Bayes 모델을 이용하여 분류를 수행하였음을 추정할 수 있다. 시스템 ID ft010은 본 논문의 방법과 같이 사용자 태그를 사용한 것처럼 보이나 좋은 성능을 보이지는 못하였다.

## 5 결론

본 논문에서는 주어진 특허 문서를 그와 유사한 특허 문서를 검색하여 그 문서의 분류코드에 따라 자동으로 분류하는 방법을 제안하였다. 특허 문서는 구조화되어 있다는 특징을 이용하여 훈련집합으로부터 유사문서를 검색할 때 문서 전체가 아닌 세분화된 같은 의미 영역을 비교하였다. 특허 문서의 의미적 구조 정보가 특허 분류에 있어 중요한 자질임을 실험을 통하여 증명하였다.

앞으로 본 논문에서 사용한 "기술분야", "목적", "해결방법", "청구", "설명", "예"의 6개 의미 영역이 충분한 지에 대한 검증 작업이 필요하다. 구현과정에서 "해결방법"과 "설명"의 구분이 경계가 명확하지 않은 문제점이 발견되었으며, 그로 인해 실험에서 그 두 영역은 좋은 결과를 보이지 못한 것 같다. 특허 문서를 재구성할 의미 영역에 대한 고찰이 더 필요할 것이다.

또한 특허 문서의 의미적 구조 정보가 특허분류에서 유용한 자질임을 SVM 등의 다른 기계학습 방법의 실험을 통해 증명해 보일 것이다.

## 참고 문헌


[1] W. Lam and C.Y. Ho. (1998). Using a generalized instance set for automatic text categorization. In Proceedings of the 21st Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), 81-89.

[2] Y. Yang. (1999). An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1(1/2): 67-88.

[3] C. Apte, F. Damerau, and S. Weiss. (1998). Text mining with decision rules and decision trees. In Proceedings of the Conference on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web.

[4] L. Douglas Baker and Andrew K. Mccallum. Distributional clustering of words for text categorization. (1998). In Proceedings of the 21st Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), 96-103.

[5] A. McCallum and K. Nigam. (1998). A comparison of event models for naïve bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization.

[6] Thorsten Joachims. (2002). Learning to classify text using support vector machines. Kluwer Academic Publishers.

[7] E. Wiener, J.O. Pedersen, and A.S. Weigend. A neural network approach to topic spotting. (1995). In Proceedings of the Fourth Annual Symposium on Document Analysis and Information retrieval (SDAIR'95), 317-332.

[8] H.T. Ng, W.B. Goh, and K.L. Low. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In 20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), 67-73.

[9] Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Akihiko Takano. (2003). Overview of Patent Retrieval Task at NTCIR-3, Proceedings of the Third NTCIR Workshop.

[10] L. S. Larkey. (1999). A Patent Search and Classification System, Proc. DL-99, 4th ACM Conference on Digital Libraries, 179-187.

[11] Cornelis H.A. Koster, Marc Seutter and Jean Beney. (2003). Multi-Classification of Patent Applications with Winnow, Proceedings PSI 2003, Springer LNCS 2890, 545-554.

[12] Grove, A., N. Littlestone, and D. Schuurmans. (2001). General convergence results for linear discriminant updates. Machine Learning 43(3), 173-210.

[13] C. J. Fall, A. Törcsvári, K. Benzineb and G. Karetka. (2003). Automated categorization in the international patent classification, ACM SIGIR Forum, 37 (1), Association for Computing Machinery.

[14] The Lemur toolkit for language modeling in information retrieval. http://www.lemurproject.org/

[15] NTCIR-5 Patent Retrieval Task. http://www.slis.tsukuba.ac.jp/~fujii/ntcir5/cfp-en.html