빈발항목과 연관규칙을 이용한 특허문서 자동분류에 관한 연구

박 래 정
정보기반팀

Ⅰ. 서론

1. 연구 배경 및 목적

한미 FTA(Free Trade Agreement) 체결(2007. 4. 2)로 특허, 상표, 저작권 등으로 구성되는 지적재산권에 대한 권리의 기한 연장에 따라 특허 보호가 강화됨으로써 국가 산업 경쟁력과 직결되는 특허정보에 대한 관심이 어느 때보다 요구되는 시기라 하겠다. 특허정보란 산업재산권과 관련된 정보로서 특허 출원된 기술 내용 및 권리로 주장된 사항, 출원인 및 발명자 등의 인적사항, 기타 서지사항 등에 대한 정보를 의미한다. 산업의 고도화, 복잡화, 다양화됨에 따라 엄청난 특허기술 정보량이 쏟아지고 있는데 기업이 변화하고 있는 산업 사회에서 생존하기 위해서 이러한 정보를 시기 적절하게 기업경영전략에 반영하지 않으면 안 된다. 현재 우리나라를 포함한 미국, 일본, 유럽과 같은 주요국의 특허청은 이러한 특허정보를 인터넷상에서 검색할 수 있는 검색 사이트를 운영하고 있으며, 이외에도 상업적 목적으로 개발된 다수의 유료 검색 사이트들이 운영되고 있다.

그중에서도 IPC(International Patent Classification) 분류체계에 의한 특허분류시스템은 미국, 유럽, 국제기구(WIPO-PCT), 일본 등 국제협약(IPC)에 의해 운영하고 있는 분류체계 방식이다. 하지만 현재까지 대한민국에서는 IPC 분류시 기계에 의한 자동분류 시스템보다는 사람에 의한 1:1 분류체계 방식을 가지고 있어 적지 않은 시간이 걸리고 있는 실정이며 그 건수가 상당하여 청구항을 비롯한 전체적인 상세설명에 대한 이해가 필요하여 분류자에게도 적지않은 스트레스를 주고 있다.

본 논문에서는 이미 분류체계가 완성된 문서의 A에서 H까지의 8개 클래스별 청구1항과 발명의 명칭을 문서별로 추출하여 불용어제거, 형태소 분석기를 이용 추출한 키워드 데이터 조합에서 발생한 빈발항목과 연관규칙의 집합을 이용하여 특허문서를 자동으로 분류, 추천해 주는 기법에 대하여 연구를 수행한다.

2. 연구방법

본 연구의 목적은 A-H까지의 8개 클래스의 발명의 명칭과 대표청구항으로 접근된 문서들의 텍스트들을 데이터마이닝 기법을 이용하여 각 클래스와 연관된 빈발항목과 연관규칙생성을 통하여 IPC 자동분류를 추천하는데 있다. 이를 달성하기 위하여 다음과 같이 연구를 수행한다.

① 자료의 선정 및 수집
- 발명의 명칭, 대표청구항를 분석하기 위한 자료로 2009년 등록된 이미 분류가 끝난 각 클래스별 1,000개의 한국등록특허를 중심으로 자료를 수집한다.

② 자료 분석 및 방향 선정
- 각 클래스별로 800개의 문서를 추출하여 특허단어 불용어를 제거하고 형태소분석기를 통하여 핵심단어를 추출하고 데이터마이닝 알고리즘(Apriori)를 통하여 빈발항목과 연관규칙을 생성한다.
- 마이닝 결과로 나온 빈발항목을 IPC자동분류가 가능하도록 적용하는 방안을 구상한다.

③ 특허문서 자동분류 추천 기법 제시
- 클래스별로 생성된 지지도에 따른 빈발항목중 1-Max(n)개까지의 Itemset중 가중치를 부여하여 새로운 문서에 대한 매칭도를 산술적으로 계산하여 특허문서를

자동으로 분류하는 기법을 제시한다.

- 제안한 기법의 검증을 통하여 검증데이터를 이용하여 문선단위의 연관성 분석방법과 비교를 통하여 검증한 결과를 보여준다.

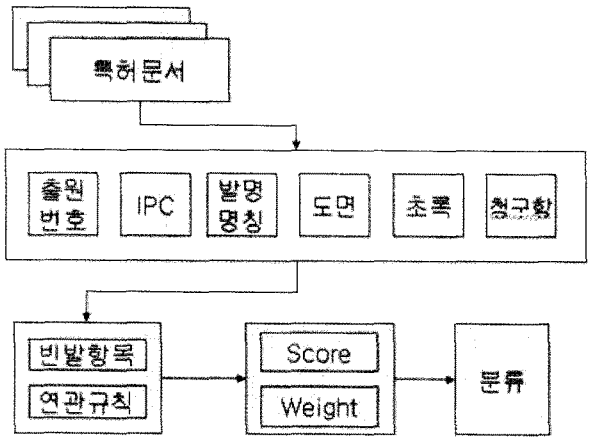
II. 특허문서 분류 방법

특허출원한 특허데이터의 최종 등록여부를 결정하는 단계에서 방식심사, 심사청구를 통하여 국제특허분류(IPC)의 결정단계가 우선한다. 이후 담당심사관이 결정되고 심사를 통하여 이후 출원공개, 실체심사, 특허결정, 등록공고의 과정을 거치게 된다. 하지만 현재 국제특허분류(IPC)는 사람에 의해 한건 한건씩 분류가 이루어지고 있다.

특허데이터 명세서의 속성은 출원번호, 공개번호, 공개일자, 등록번호, 등록일자등의 서지정보와 발명의 명칭, 초록, 청구항, 상세설명등으로 이루어져 있으며 실제로 발명의 명칭과 청구1항은 IPC 분류의 기본정보가 된다.

특허데이터의 일반적인 웹상의 데이터나 일상생활에서 사용하는 데이터와는 조금은 다른 특징을 가지고 있다. 사용자가 직관적으로 알아보기 쉽고 분석을 용이하게 하기 위해 적절한 처리과정을 거친 정제된 데이터가 필요하다. 이를 구하기 위하여 등록특허 데이터의 발명의 명칭과 청구1항에 대하여 데이터 변환과 데이터 정제 및 보정 단계를 수행한다.

등록특허파일을 추출하여 하나의 문서를 하나의 트랜잭션으로 만든다. 각 트랜잭션별로 참조한 실제문서에서 출원번호, IPC, 발명의 명칭, 청구1항을 추출하여 하나의 트랜잭션으로 구성한다. 그 후 각 문서의 불용어를 제거하고 형태소분서기를 통하여 하나의 문서를 하나의 트랜잭션으로 구성한다. 이때 각 트랜잭션별로 문서에서 단어를 추출한 키워드가 중복되지 않게 구성한다.



[그림 1] 구현 시스템 구성도

1. 특허문서 분석

1.1 데이터 범위

테스트용으로 사용된 특허데이터는 한국특허정보원에서 운영중인 KIPRIS 검색시스템을 이용한 것으로 2010년 1월, 2월에 등록된 데이터를 대상으로 한다.

1.2 데이터 가공

특허데이터는 1년에 20만여건이 출원되고 있으며 등록여부에 상관없이 IPC 분류는 이루어진다. 또한 IPC 분류기준은 발명의 명칭과 청구1항에 의존하고 있지만 실제로 분류자의 주관적인 판단에 의해 분류가 되고 있는 실정이다. 하지만 본 연구는 학문적이론에 근거하여 분류체계를 마련하고자 한다. 즉 특허데이터의 전체 항목중에서 불필요한 단어(조사나 특정단어)들을 제거하는 과정을 거치고, 본 연구에 필요한 연관규칙 탐사를 위해 데이터 모델을 구축한다. 관심있는 항목들은 〈출원번호〉,〈IPC분류〉,〈발명의 명칭〉,〈청구1항〉,〈초록〉,〈핵심키워드〉이다.

하나의 트랜잭션이 자연스럽게 정의되는 쇼핑몰의 장바구니 분석과는 달리 특허데이터를 가지고서는 트랜잭션을 정의하기가 자연스럽지 않다. 본 논문에서는 IPC 분류기준이 되는 발명의 명칭과 청구1항을 기준으로 모델링을 완성하고 한 트랜잭션으로 정의하였다. 즉 하나의 트랜잭션에는 〈출원번호〉,〈IPC분류〉,〈발명의 명칭〉,〈청구1항〉,〈초록〉,〈핵심키워드〉으로 구성되며 이들항목은 XML형태의 하나의공보형태로 되어있기 때문에 Java XML Sax Parser를 통하여 파싱과정을 거쳤으며 알아보기 쉽고 분석에 용이하게 하기위해 적절한 변환과정이 필요하였다.

[표 1] 한국등록특허의 추출XML 속성 정보

1. 서지사항	〈KR_Application_Number〉〈/KR_Application_Number〉
2. 요약	〈KR_abstract〉〈/KR_abstract〉
3. 대표도	〈KR_draw〉〈/KR_draw〉
4. 특허청구의 범위	〈KR_cliaims〉〈/KR_cliaims〉
5. 명세서	〈KR_description〉〈/KR_description〉
6. 도면	〈KR_figure〉〈/KR_figure〉

[표 2] 한국등록특허의 추출XML 결과

출원번호	IPC	발명의명칭	청구1항	핵심키워드	기술분야
---	---	---	---	---	---

1.3 특허문서 데이터 정제 및 보정

특허데이터는 서지사항 뿐아니라 초록, 청구항, 상세설명, 기술정보등 다양한 정보가 있기 때문에 필요한 데이터만 선별하는 정제작업을 한다. 본 논문에서는 발명의 명칭과 청구1항, 핵심키워드, 기술정보를 기준으로 트랜잭션을 정의 한다. 또한 특허데이터의 속성상 "제조방법", "것", "1", "2", "방법", "제조", "것", "본", "내", "이", "수", "상기", "발명", "용", "등", "사", "포함", "이" 등의 불용어를 사전에 제거하는 전처리 과정을 거친다. 추출한 단어그룹의 개수가 20개를 초과한 데이터에 대해서만 의미있는 중요한 문서라고 정의하며 또한 핵심키워드로 명시된 단어를 추출하여 단어그룹에 추가한다. 데이터의 특성상 같은의미의 단어가 다르게 표기되어있는 경우가 있어 유사의미의 단어에 대해 대표어로 표기하는 보정작업을 실시한다. [표 3]는 단어의 개수가 20개이상이고 대표로어 보정작업이 완료된 문서의 하나의 트랜잭션으로 처리한 테이블을 나타내었다.

[표 3] 트랜잭션별 문서접근 내역 테이블

TID	출원번호	IPC	발명의명칭	청구1항	핵심키워드
트랜잭션번호	---	---	---	---	---

2. 형태소 분석 및 키워드 추출

추출한 키워드그룹의 개수가 20개를 초과한 데이터에 대해서만 의미있는 중요한 문서라고 정의하고 하나의 트랜잭션으로 나타내고 이 문서목록에 있는 문서 전문을 한글 형태소분석기를 적용하여 키워드를 도출한다. 한글 형태소분석기를 적용한 결과데이터에서 Score가 100이상인 형태소는 체언으로 명사, 대명사, 수사, 의존명사를 포함하므로 Score가 100이상으로 구분된 형태소를 키워드로 추출한다. 각 형태소는 하나의 문서 안에서도 중복되고 다른 문서에서도 중복이 되므로 하나의 트랜잭션내에서는 키워드가 중복이 되지 않게 추출한다. 본 논문의 실험에 이용한 Apriori알고리즘의 구현은 오라클(10g) 데이터베이스의 한 행에 하나의 트랜잭션 항목집합으로 구성하여 varchar2데이터 타입으로 생성하여 적용하였으며 255자이하로 구성하였다. 하나의 키워드가 분류의 기준이 될 수있다고 판단하고 키워드의 개수에는 제한을 두지 않았다. 본 논문에서는 한 트랜잭션내 키워드가 중복되지 않는 키워드 리스트를 중복배제키워드라 정의 한다. [표 4]은 각 트랜잭션에 포함된 문서를 한글형태소분석기를 적용하여 추출한 중복배제키워드리스트 결과 테이블이다.

[표 4] 한글형태소분석기 중복배제 키워드리스트 추출

T1	키워드1	키워드2	키워드3	키워드4	키워드5	키워드6	키워드7	키워드8
T2	키워드1	키워드2	키워드3	키워드4	키워드5	키워드6	키워드7	키워드8
~	~	~	~	~	~	~	~	~
-	-	-	-	-	-	-	-	-

3. 빈발항목집합 추출 및 연관규칙생성

[표 1]에서 나온 각 트랜잭션별 키워드들을 대상으로 초록, 청구1항, 기술분야에 대해서 연관성 분석 알고리즘인 Apriori 알고리즘을 적용하여 분석한다. 이때 최소지지도를 변경해가며 각 지지도별(Support Degree) 빈발항목집합들을 구한다. 본 논문에서는 0.5%의 지지도로 정의하였다. 빈발항목집합에서 최소지지도 보다 높은 모든 항목집합이 들어있으므로 항목간의 중복성이 존재한다. 본 논문에서는 키워드간의 연관성 연구를 목적으로 하므로 각 트랜잭션의 유일한 키워드가 존재하여야 한다. 중복성이 존재하는 빈발항목집합(Frequent Item)으로부터 자신 이외에 다른 빈발항목집합에 포함되지 않는 최대빈발항목 집합인 MFI를 구한다.

3.1 Apriori 알고리즘을 이용한 연관성 분석

[표 5] 데이터마이닝 Apriori 알고리즘

단계 0. 최소지지도 smin을 정한다.
k=1
C_1 =[{i_1 },{i_2 },...,{im}]
L_1 ={c∈C_1 | supp(c) ≥ smin

단계 1. k=k+1
Lk-1로부터 Ck 형성 (apriori-gen 함수)
단계 1-1. (join) Lk-1의 집합들을 접합하여 k- 항목 집합군을 형성한다.
C= Lk-1 * Lk-1
단계 1-2. (prune) C의 (k-1)- 항목 부분집합이 Lk-1에 속하지 않을 때 이를 모두 제거한 후 Ck를 형성한다.
Ck=Φ이면 Stop.

단계 2. Ck의 집합 중 지지도가 최소지지도 이상인 것을 모아 Lk를 생성한다.
Lk={c∈Ck | supp(c) ≥ smin}

Lk : 후보 k-항목집합
Ck : 빈발 K-항목집합

연관규칙 마이닝 알고리즘인 Apriori는 두단계를 통하여 연관성분석을 하는데, 첫 번째 단계는 최소의 지지도(minimum support)이상의 발생지지도(transaction)를 가지는 조합을 찾아 빈발단어 항목을 구성한다. 두 번째 단계는 데이터베이스로부터 연관 규칙을 생성하기 위하여 빈발항목집합(L)에 대해서 빈발항목집합의 모든 공집합이 아닌 부분집합들을 찾는다. 각각의 그러한 부분집합(A)에 대하여, 만약 Support(A)에 대한 Support(L)의 비율이 적어도 최소 신뢰도(minimum confidence)이상이면 A->(L-A)의 형태의 규칙을 출력한다. 이 규칙의 지지도는 support(L)이고, 신뢰도는 support(L)/support(A)이다. Apriori 알고리즘에서 후보집합의 생성은 Apriori-gen을 상용하여 새로운 후보항목집합을 만들게 함으로써 후보항목의 수를 줄일 수 있다. 이에 따라 연관 규칙을 찾는 시간이 감소된다.

[표 4]의 트랜잭션별 키워드 결과를 Apriori 알고리즘을 최소지지도별로 조정하며 적용하여 [표 6]와 같이 각 최소지지도별(Support degree) 빈발항목집합 (Frequent Item-set)들을 구하고 [표 7]와 같이 연관규칙을 생성한다.

[표 6] 최소지지도별(Support degree) 빈발항목집합 (Frequent Itemset)

키워드1(지지도)
키워드1, 키워드2(지지도)
키워드2, 키워드3(지지도)
키워드1, 키워드2, 키워드3, 키워드4(지지도)
XXX,XXX,XXX,XXX,XXX(support)
XXX,XXX,XXX,XXX,XXX(support)

[표 7] 지지도와 신뢰도 기반의 연관규칙 생성

키워드1 <- 키워드1,키워드3(0.5, 100.0)
키워드3 <- 키워드5,키워드6(0.5, 100.0)
키워드4 <- 키워드5,키워드7(0.5, 100.0)
키워드6 <- 키워드5,키워드8(0.5, 100.0)
키워드7 <- 키워드6,키워드8(0.6, 84.6)
키워드8 <- 키워드6,키워드9(0.6, 84.6)

4. 특허문서 자동분류 방법

4.1 빈발항목과 연관규칙생성에 의한 정확도 계산

특허데이터의 키워드는 문서의 내용을 대표하는 단어로써 정확한 키워드를 추출하는 것은 특허문서 자동분류 체계의 효율성을 극대화 시킨다. 따라서 단순히 하나의 문서마다 문서 전문에 존재하는 키워드에 대해 Apriori 알고리즘을 적용하여 빈발항목을 추출하는 방법의 단점을 극복하기 위하여 본 논문에서는 연관규칙 룰셋을 이용하여 지지도와 신뢰도를 기반으로 하여 가중치를 주는 방법을 제안한다. 이방법을 빈발항목과 연관규칙에 의한 정확도 계산방법이라고 정의한다.

4.2 클래스간 키워드 중복제거 및 가중치 부여에 의한 정확도 계산

특허문서에서는 중복키워드는 반드시 고려해야할 대상이다. 즉 키워드의 출현횟수가 늘어났다는것은 그만큼 특허문서의 중요키워드일 확률이 높아진다. 한문서내에서의 중복키워드도 중요하지만 다른 클래스들과의 키워드 중복도 반드시 고려되어야 한다. 고유의 키워드가 될수도 있고 이중키워드가 될 수도 있기 때문이다.

각 클래스간에 발생한 중복키워드와 중복키워드항목집합의 단어들의 연관규칙셋의 중복제거를 통하여 정확도를 높여 기존의 단어와 차별화를 할 수 있는 분석방법을 가중치 기반 키워드추출방법이라 정의한다.

[표 8] 빈발항목 중복키워드 추출 및 제거 개념도

A클래스	B클래스	C클래스	D클래스
A	A2	A3	A4
A,B(X)	A,B(X)	A,B(X)	A,B(X)
A1,C1,D1	A1,C1,D1	A3,C3,D3	A4,C4,D4
A,C,D,F	A1,C2,D2,F2	A3,C3,D3,F3	A4,C4,D4,F4

A1,C1,D1처럼 A클래스와 B클래스의 교집합인 항목집합인 경우 A클래스인지 B클래스인지 확실히 구분해 줄 수 있는 가중치가 필요하다. 본 논문에서는 가중치를 원래 가지고 있던 값의 2배의 수치를 기준으로 한다.

[표 9] 가중치 개념도

A클래스	B클래스
A,B(X)	A,B(X)
A1,C1,D1	A1,C1,D1
A1,C1,D1,F1	A1,C1,D1,F2

A1,C1,D1의 (A1의 출현개수+C1의 출현개수+D1의 출현개수)를 합하여 정확도를 계산하여 A클래스인지 B클래스인지에 대한 구분을 명확히 한다.

4.3 정확도 계산에 의한 특허문서분류 기준항목

본 논문 실험에서 사용하는 데이터를 두 개 부류로 나누었는데, 본 논문에서 제안하는 빈발항목집합과 연관규칙생성에 의한 정확도계산 방법과 클래스간 키워드 중복제거 및 가중치 부여에 의한 정확도 계산방법으로 적용할 문서전문을 실험데이터라 정의하며, 정확도 계산방법의 효율성을 검증 및 비교하기 위한 특허문서전문을 검증데이터라 정의한다. 즉 실험데이터는 빈발항목집합을 구하는데 사용된 이미 IPC부여가 끝난 2010년 등록된 10,000건중 8,000건이고 검증데이터는 최대빈발항목집합을 평가 및 검증하기 위해 이미 IPC부여가 되었지만 아직 부여되지 않았다고 가정하는 즉 이미 답을 알고 있는 나머지 2,000건이다.

실험데이터의 한 빈발항목집합 항목집합의 항목들이 검증데이터의 한 트랜잭션 키워드 항목들에 속할 때 이 빈발항목집합 항목은 완전매치(Complete match)한다고 정의한다. 또한 빈발항목집합과 검증용데이터간에 매치되는 포인트를 구하기 위하여 특허문서의 초록, 청구1항, 기술배경에 대하여 다음과 같이 빈발항목 매치도, 연관규칙룰 매치도를 구한다.

[정의 1] 빈발항목 집합 및 아이템셋의 가중치 설정

검증데이터의 트랜잭션에 완전매치되는 빈발항목집합을 아이템셋별로 추출한다. 아이템셋의 가중치는 2~3가지 종류로 하여 결정한다. 본 논문에서는 가중치를 1, 3, 9, 27, 81과 1, 5, 10, 300, 500의 방법으로 2가지로 정의하였다. FIC를 빈발항목집합수라 정의하고 Sup를 지지도, Conf는 신뢰도, FIA는 빈발항목집합 가중치라 정의한다.

[정의 2] 검증데이터 매치수

빈발항목집합 키워드들을 모두 포함하는 검증데이터의 트랜잭션 수를 검증데이터 매치수라 정의한다.

$$Score1=\text{각 } (FIC \times Sup \times FIA)\text{의 합}$$

실험데이터로 구한 각 지지도별 빈발항목집합의 항목집합은 키워드들과 키워드의 지지도로 구성되며, 이들 키워드들 간의 연관성의 척도는 빈발항목집합의 단일 항목집합의 키워드들과 검증데이터의 단일 트랜잭션에 나오는 키워드와 비교하여 판정한다. 본 논문에서는 최소지지도별 연관성분석으로 나온 빈발항목집합과 검증데이터와의 매치도 판정의 정확성을 높이기 위하여 빈발항목집합수와 검증데이터매치수를 곱한 정확도를 본 논문의 비교기준항목으로 제안한다.

[정의 3] 연관규칙룰셋에 의한 신뢰도, 지지도기반의 매치도

검증데이터를 기준으로 연관규칙룰셋의 지지도, 신뢰도를 기반으로 하여 연관성 척도를 수치화한 값을 나타내는데, 검증데이터의 각 트랜잭션에 매치되는 것으로 표현하는 식(1)은 다음과 같이 계산되어진다.

$$Score2= \Sigma (FIC \times sup \times confi \times FIA)$$

[정의 4] 클래스간 키워드 중복제거 및 가중치 부여

클래스간 발생한 중복키워드에 대하여 전체에 발생한 키워드는 일괄삭제하고 부분적으로 발생한 키워드에 대해서는 그 구분을 확실히 하기 위하여 중복이 발생한 클래스에 2배의 가산점을 부여하여 정확도를 높인다.

Weight = IF(Each class Score[A1,C1,D1] 〉 Each class Score[A1,C1,D1])
TRUE A Score × 2 , FALSE B Score × 2

[정의 5] 초록, 청구1항, 기술배경의 각 항목에 합에 의한 매치도 계산

특허문서의 분류는 어느 한 항목에 의존하는 것이 아니라서 각 항목의 합을 계산하여 정확도를 높이는 방법이 중요한다. 특허문서의 중요한 특징 중의 하나인 이것은 정확도를 높이는 가장 좋은 방법 중의 하나가 아닌가 하다.

각 초록, 청구1항, 기술배경
Category devide = (Score + Weight)

Ⅲ. 실 험

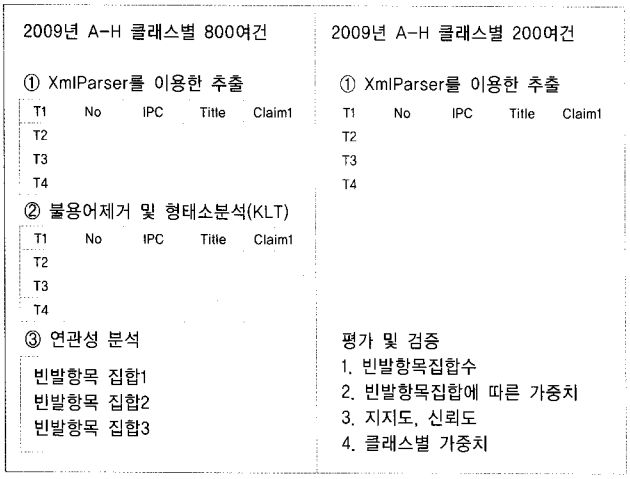
1. 실험대상 및 실험방법

실험은 2010년 대한민국 특허로 등록된 등록특허 A-H 클래스별 10,000여건을 선정하여 8,000여건을 검증데이터로 나머지 2,000여건을 실험데이터로 실험하였다.

실험을 위한 프로그램 작성은 전처리 과정 중 불용어 제거 및 매칭프로그램은 Visual Basic 6.0을 사용하였으며 연관성 분석을 위한 Apriori알고리즘은 공개프로그램을 사용하였으며 데이터베이스는 오라클 10g를 사용하였고 한글 형태소 분석을 위해서 국민대학교 컴퓨터학부 강승식 교수가 만든 한국어 분석 모듈 KLT2.0.0을 이용하였다.

본 논문에서는 특허데이터의 IPC 자동분류의 효율성을 검증하기 위하여 데이터마이닝 알고리즘인 Apriori의 적용이 얼마나 실효성을 갖는지를 검증하기 위한 것이며 80:20의 검증방법으로 검증데이터와 실험데이터를 비교하였다. 추출된 지지도별 키워드간의 연관성 정도를 측정하기 위하여 검증용 데이터를 이용하여 일정간격이 최소지지도별로 빈발항목집합의 개수, 지지도와 신뢰도를 이용한 연관규칙룰셋의 매치도, 키워드가중치 부여에 의한 매치도등을 구한다. 또한 초록, 청구1항, 기술배경의 3가지 항목에 대한 Score1, Score2, Weight 등을 각각 구하여 정확도를 높인다. 본 실험을 위한 프로세스는 [표 10]과 같다.

[표 10] 특허문서 분류 프로세스 개념도



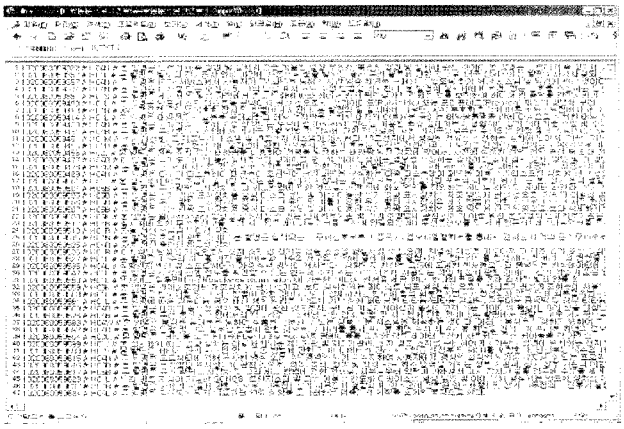
2. 실험과정 및 결과

실험을 위하여 2009년 IPC분류가 완성되었으며 등록된 특허데이터의 변화 및 정제/보정 처리한 실험데이터에서 10,000개의 트랜잭션을 생성하였고 A-H까지의 각 1,000여개의 데이터가 된다. 이중 80%인 각 클래스별 800개를 검증데이터로 사용하였으며 200개를 실험데이터로 선정하였으며 여러 번의 실험을 통해 랜덤하게 데이터가 구성되었다.

2.1 XmlParser를 이용한 특정 태그 데이터 추출

[표 11] 한국등록특허 input 의 태그의 속성 Detail

<KR_ApplicationNumber>10-1999-0016635</KR_ApplicationNumber>
<KR_InventionTitle>음극선관</KR_InventionTitle>
<Claim n=" 1" ><P align=" JUSTIFIED" indent=" 14" >적어도 형광체스크린을 내면에 갖는 패널, 이 패널에 연이어지는 퍼넬, 이 퍼넬의 소직경측의 단에 접합되며 상기 형광체스크린에 대향하여 전자총이 장착되는 네크, 및 상기 네크측의 단에서 패널측으로 뻗은 영역에 편향요크 장착영역을 갖는 음극선관으로서,</P>
<P align=" JUSTIFIED" indent=" 14" >상기 편향요크 장착영역을 상기 네크와의 연결부로부터 적어도 편향요크의 스크린측단까지로 할 때, 상기 네크와의 연결부로부터 편향요크가 장착되는 끝단부까지의 관축을 일정한 간격으로 n등분하여 나누며 그 등분된 각 위치에서 관축에 직각인 단축방향의 두께를 Tv, 장축방향의 두께를 Th, 대각방향의 두께를 Td라 할 때, $0.5 \leq Td/Tv \leq 0.85$, 또한 $0.5 \leq Td/Th \leq 0.85$로 하고, 상기 편향요크 장착영역과 퍼넬부의 변곡점에서의 수직측 두께를 Tvt, 대각측 두께를 Tdt, 수평측 두께를 Tht라 하고, 편향 각이 최대가 되는 관축상의 위치인 편향기준위치에서의 수직측 두께를 Tv<SB>L</SB>, 대각측 두께를 Td<SB>L</SB>, 수평측 두께를 Th<SB>L</SB>이라 할 때, $Tdt/Tvt \geq Td<SB>L</SB>/Tv<SB>L</SB>$, 또한 $Tdt/Tht \geq Td<SB>L</SB>/Th<SB>L</SB>$인 것을 특징으로 하는 음극선관. </P>
</Claim>



[그림 2] 한국등록특허 output 파일형태를 나타냄

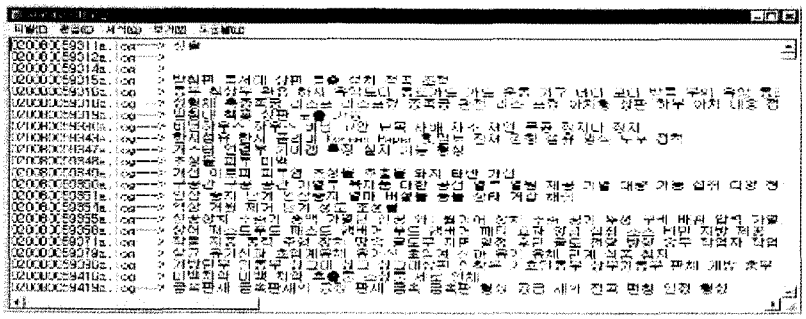
Eclipse를 이용하여 JAVA Xml SAX Parser를 이용하여 등록공보 전체에서 출원번호, 발명의 명칭, 청구1항, 핵심키워드를 추출한다.

[표 12] 한국등록특허 input 의 태그의 속성 Detail

No:	Freq	Score	Term	Loc1	Loc2	Loc3	Loc4
1:	2	1000	억제용	107	124		
2:	2	872	조성물	108	125		
3:	3	770	비만		106	118	123
4:	1	331	A81K		102		
5:	2	331	억제		107	124	
6:	1	321	유효성분	116			
7:	1	238	61K		102		
8:	1	238	치료		119		
9:	1	208	발명		105		
10:	1	200	사포닌	115			
11:	1	144	예방		121		
12:	1	119	유효		116		
13:	1	119	성분		116		
14:	1	119	A		102		
15:	1	93	포함		117		
16:	1	59	1020060060680		106		
17:	1	51	함소		113		
18:	1	48	효과적	122			
19:	1	29	추출		114		
20:	2	12	것		110	127	

Score 100이상 의미있는 키워드

Score 100이하로 불용어 처리



[그림 3] 형태소 분석기 및 불용어 제거 프로그램 수행 후 A클래스(4,561셋)

2.2 불용어 제거 및 형태소 분석

각 클래스별로 파싱된 결과물들을 각각의 트랜잭션으로 정의하고 불용어 제거 및 형태소 분석을 통하여 클래스별 각 문서의 트랜잭션별로 키워드를 추출한다.

1,000건 구분된 각각의 트랜잭션을 형태소 분석단계에서 한 건으로 분리한 후 형태소 분석 후 내부적으로 추출된 Score 100 이상의 데이터만 추출하고 그 후 특허분야에서 공통적으로 발생하는 단어에 대해 불용어를 제거한다.

불용어는 특허데이터의 속성상 제조방법, 것, 1, 2, 방법, 제조, 것, 본, 내, 이, 수 , 상기, 발명, 용, 등, 사, 포함, 이 등으로 정의한다. 또한 Score 100이하의 단어들을 연관성에 지장을 주지 않은 키워드라고 정의한다.

2.3 최소지지도 결정

실험데이터에서 IPC자동분류시 의미있는 빈발항목 개수를 추출하기 위해 최소지지도를 정해야 한다. 최소지지도를 결정하기 위해 최소지지도를 0.5에서 10까지 0.5, 1, 2, 3, 5, 10의 간격으로 변경시켜가며 해당 최소지지도 이상의 빈발항목들로 트랜잭션으로 구성하였다. 최소지지도별 빈발항목개수는 [표 13]과 같은데 지지도가 0.5% 이하로 내려가면 1만여 건 이상의 빈발항목이 발생하여 실제로 매칭도 계산시 어려움을 겪게 된다. 이리하여 실험에서는 지지도별로 실험하였다.

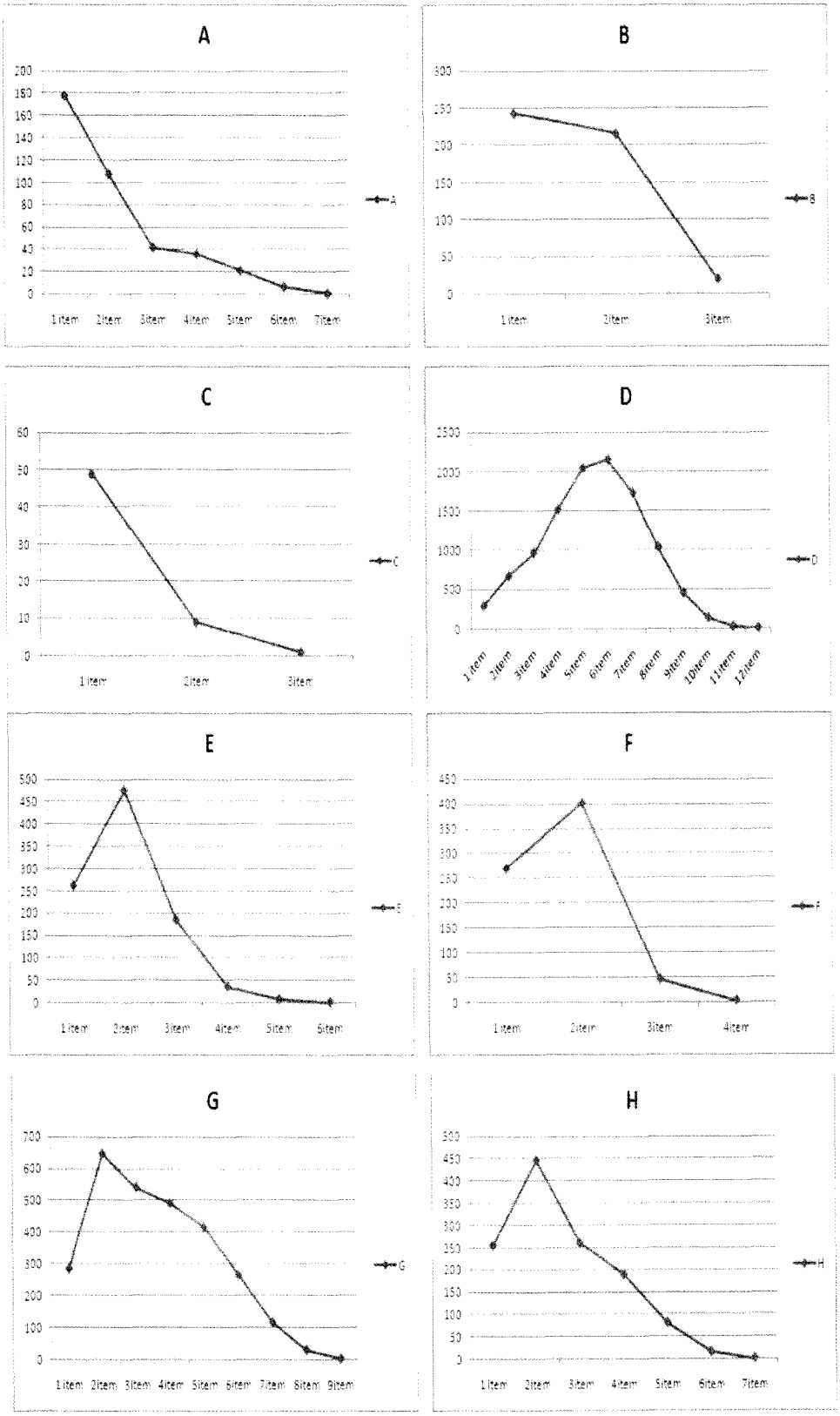
[표 13] 최소지지도별 클래스 빈발항목 개수(단위(지지도=%, 빈발항목=개수)

지지도 / 클래스	지지도0.5	지지도1	지지도2	지지도3	지지도5	지지도10
A클래스	51,450,538	393	88	30	15	1
B클래스		479	127	61	22	5
C클래스	12,588	59	13	4	1	0
D클래스		11,001	329	66	23	8
E클래스		968	218	102	38	12
F클래스	8,254	720	203	91	36	7
G클래스	320,178	2,794	293	131	49	12
H클래스	11,292	1,249	271	122	40	10

[표 14] 최소지지도 1%에서 클래스별 빈발항목 item개수

클래스	빈발개수	1item	2item	3item	4item	5item	6item	7item
A클래스	393	178	108	42	36	21	7	1
B클래스	479	243	216	20				
C클래스	59	49	9	1				
D클래스	11001	284	668	965	1515	2042	2151	1729
E클래스	968	262	476	185	36	8	1	
F클래스	720	269	401	46	4			
G클래스	2794	285	646	540	490	416	264	117
H클래스	1249	255	447	260	189	81	16	1

실험결과 최소지지도가 1% 영역에서 MFIC와 SMFIC가 가장 높은 것으로 실험되었다.



[그림 4] A-H클래스별 빈발항목 Itemset 개수]

2.4 빈발항목집합에 의한 정확도 계산

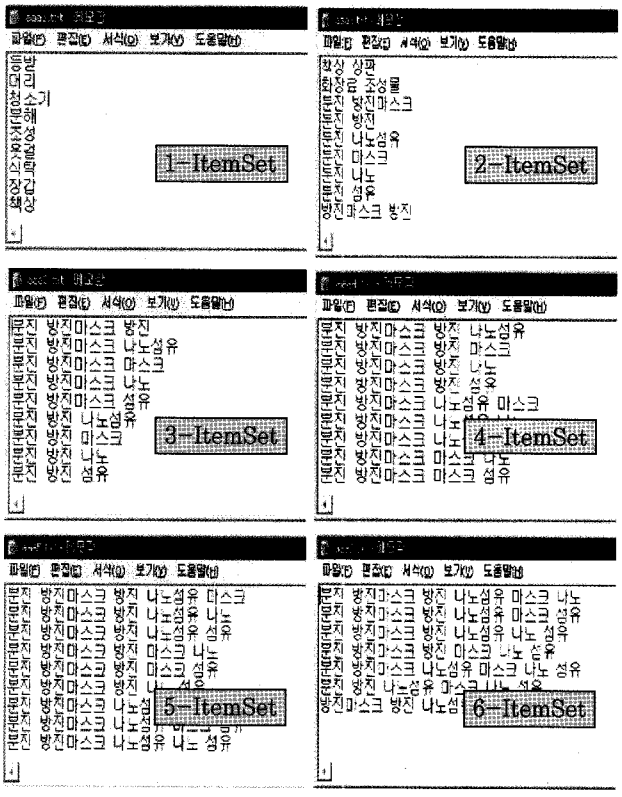
최대빈발항목 개수에 따른 연관성을 분석한다. itemset 1부터 itemset Max(n)까지의 빈발항목의 출현정도를 검증데이터와의 매치도를 조사하여 매치포인트가 높은 쪽으로 분류된다고 정의한다. 빈발항목 Item에 따른 매칭 카운트 가중치는 n과 n-1의 관계에서 n-1의 2배수의 1.5배가 n의 값이로 정의한다. 즉 n-1의 2개 매칭되고 n이 한번 매칭된다고 했을때 n의 값이 더욱 의미 있는 값이 되는것이다. 식으로 나타내면 다음과 같다

$$\text{Item n의 가중치} = [\ \text{Item(n-1)의 가중치} \times 2] \times 1.5$$

[표 15] 빈발항목 Item에 따른 매칭 카운트 가중치

Item	1	2	3	4	5	6	7	8	9	10	11	12
점수	1	3	9	27	81	243	729	2,187	6,561	19,683	59,049	177,147

A-H클래스별로 추출된 Itemset 1부터 ItemsetMax(n)까지의 빈발항목을 1-n개까지의 파일로 저장하여 트랜잭션 단위로 가중치를 이용하여 검증데이터 200개에 적용하여 매칭포인트를 계산하면 다음과 같은 결론을 알 수 있다.

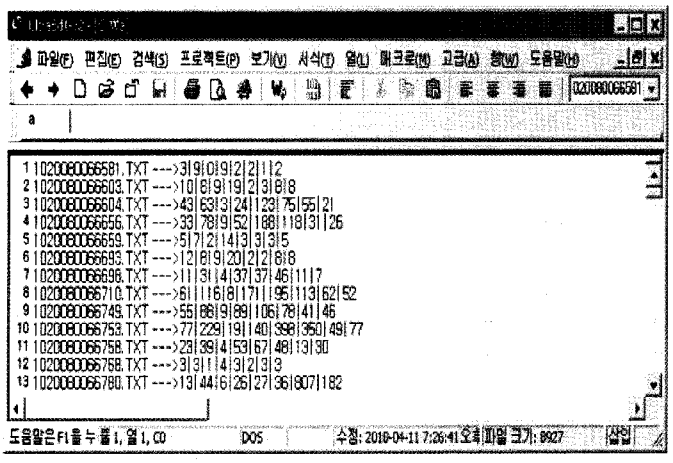


[그림 5] ItemSet 1부터 ItemSet Max(n)까지의 파일구조

매칭도를 결정하는 최종 매칭카운트 식은 다음과 같다.

$$\text{빈발항목집합} = \Sigma\ (\text{빈발항목(n)} * m * \text{트랜잭션 개수})$$

m=가중치 개수, n는 itemset



[그림 6] ItemSet 1부터 ItemSet Max(n)까지의 ItemSet매칭 Output #1

	A	B	C	D	E	F	G	H	초대값	일치여부
1	196	5	1	5	3	3	3	3	196	A
2	196	5	1	5	3	3	4	4	196	A
3	202	5	1	6	2	3	3	5	202	A
4	196	5	1	5	3	3	3	3	196	A
5	4	4	0	4	4	4	2	2	4	A
6	40	25	11	38	11	10	30	27	40	A
7	42	13	9	19	20	7	15	9	42	A
8	6	2	1	3	1	1	3	2	6	A
9	69	2	1	6	2	2	3	4	69	A
10	41	32	2	26	35	35	18	23	41	A
11	32	14	6	17	11	10	9	8	32	A
12	167	8	5	10	7	4	7	7	167	A
13	194	3	2	5	2	1	3	5	194	A
14	218	7	3	9	7	5	6	9	218	A
15	171	5	4	12	3	3	5	5	171	A
16	17	9	3	12	6	6	4	3	17	A

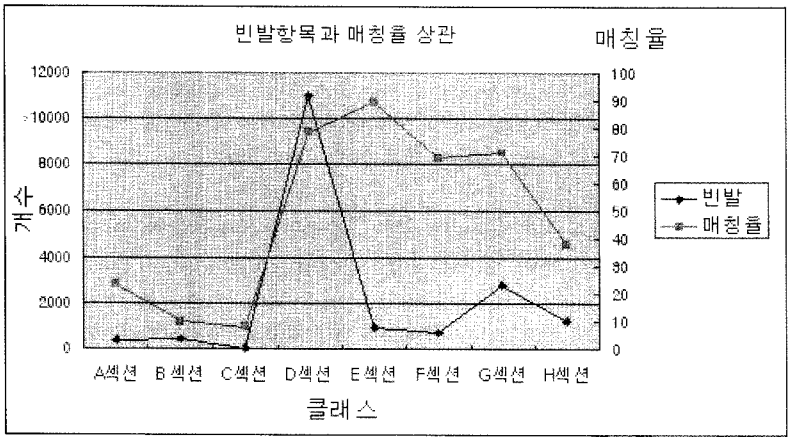
[그림 7] ItemSet 1부터 ItemSet Max(n)까지의 ItemSet매칭 Output #3

[표 16] ItemSet 1부터 ItemSet Max(n)까지의 클래스별 최종결과(방식 1)

클래스	빈발	매칭율	1item	2item	3item	4item	5item	6item	7item
A클래스	393	23%	178	108	42	36	21	7	1
B클래스	479	10%	243	216	20				
C클래스	59	8%	49	9	1				
D클래스	11001	78%	284	668	965	1515	2042	2151	1729
E클래스	968	89%	262	476	185	36	8	1	
F클래스	720	69%	269	401	46	4			
G클래스	2794	71%	285	646	540	490	416	264	117
H클래스	1249	38%	255	447	260	189	81	16	1

[표 17] ItemSet 1부터 ItemSet Max(n)까지의 클래스별 최종결과(방식2)

클래스	빈발	매칭율	1item	2item	3item	4item	5item	6item	7item
A클래스	393	23%	178	108	42	36	21	7	1
B클래스	479	10%	243	216	20				
C클래스	59	8%	49	9	1				
D클래스	11001	78%	284	668	965	1515	2042	2151	1729
E클래스	968	89%	262	476	185	36	8	1	
F클래스	720	69%	269	401	46	4			
G클래스	2794	71%	285	646	540	490	416	264	117
H클래스	1249	38%	255	447	260	189	81	16	1



[그림 8] 빈발항목개수과 매칭율의 상관도 분석 그래프

Ⅳ. 결론 및 향후연구

본 논문에서는 특허문서에서 초록, 청구1항, 기술분야의 특정주제어와 키워드를 추출하는 방법으로 빈발항목에 의한 정확도 제시, 연관규칙에 의한 정확도 제시, 특정 클래스간에 존재하는 중복키워드군에 대한 가중치에 의한 정확도 등 3가지 정확도 계산 방법을 제시하였다. 부가적으로 초록, 청구1항, 기술분야에서의 각 정확도를 파악한 후 합을 이용하거나 특정부분에 가중치를 주는 방법도 제시하였다.

빈발항목집합에 의한 정확도 계산방법은 불용어 제거 단계에서 특허데이터가 가진 특징 중 기술용어와 범용언어에 대한 구분을 명확히 해야하고 불용어 제거 단계에서의 데이터 정제단계가 큰 의미를 차지하였다. 결과론적으로 보면 불용어를 제거하였을 때와 하지 않았을 때의 정확도가 5% 전후로 차이가 난다. 또한 지지도별 정확도는 지지도가 높아짐에 따라 빈발항목의 개수가 줄어들어 실제 계산에 이용되는 후보군이 적어진다는 단점이 있으며 최소지지도를 기반으로 하고 있는 본 논문에서는 0.5% 이하의 지지도를 이용하였을 시 단어군의 수가 너무 많고 2 이상을 넘었을 때는 너무 적어지는 현상이 존재하여 결국 최소지지도를 1%로 하게 되었다. 또한 형태소 분석시 국민대학교 KL2.0을 사용하였으며 내부 알고리즘에 의한 SCore 의 최소치를 100으로 하였다, 100 이하의 데이터의 경우 1글자 단어가 대부분이며 실제 특허의 의미에 영향을 끼치지 않는다는걸 사전에 정의하였다. A-C 클래스의 경우 빈발항목이 D,G클래스에 비해 상대적으로 적은 편이며 정확도도 높지않다는 실험결과가 나온다. A-C의 경우 A섹션-생활필수품, B섹션-처리조작, 운수, C섹션-화학, 야금의 클래스를 갖는 경우이다. 정확도가 낮은 원인은 범용기술분야가 대부분이며 개인출원이 많은 비중을 차지한다. 개인출원의 경우 기술용어가 아닌 일상생활에서 사용하는 범용언어를 사용하고 있어 그 구분이 모호해지는 현상이 있다. 또한 A-C클래스의 특허문서의 경우 서로 다른 기술분야에 어느정도 무리지어 혼재되어 있는 이상적인 경우가 아닌 유사한 기술분야에 몰려있거나 공통된 기술분야로 그룹화되지 않고 혼재되어 있을 확률이 높은 분야인 것이다. 특허문서의 특성상 IPC 분류만으로는 해결이 되지 않은 부분이며 관련 연구분야에서 클러스터링 방법이 많이 제시되었던 분야이다. 본 논문에서는 클러스터링보다는 연관규칙에 의한 신뢰도를 기반으로하는 정확도 계산방법으로 A-C분야의 정확도를 높이도록 할 것이다. 또한 A클래스의 경우 E클래스로 분류되는 경우가 적지 않은데 이는 아마도 A클래스가 너무 범용적인 부분을 다루고 있는것 같고 E클래스의 기술분야로 종속되는 느낌을 준다. E클래스의 경우 빈발항목의 지지도 기반의 매칭율이 거의 90% 해당하며 빈발항목이 1만개 이상이며 기술분야별 그룹화가 잘 되어 있는 곳이라고 판단할 수 있다.

E클래스의 경우 빈발항목개수가 적은 편인데 상대적으로 정확도가 높은 편이다. 이는 군집화 및 그룹화가 잘 이루어져 있으며 대부분의 특허들이 일정한 특허문서 길이

와 기술용어를 사용하고 있다는 것을 알 수 있다.

A-C클래스의 낮은 정확도를 보강할 방법으로 본 논문에서는 신뢰도, 지지도 기반의 연관규칙룰셋의 정확도를 제시한다. A-C클래스는 범용키워드와 군집화 및 그룹화가 이루어지지 않은 곳이라서 클래스의 레벨을 섹션, 서비섹션, 클래스, 서브클래스까지는 낮추어 이론적인 군집화를 적용하고 그룹화 하도록 하였다.

세 번째 본 논문에서 제시하는 방법은 연관규칙룰셋의 중복키워드집합에 관한 부분이다. A에서 H까지 8개 클래스에 모두 나온 단어들은 당연히 제거해서 처리하는 것이 계산 결과를 얻는데 시간비용이 적게 든다. 문제는 A,C에만 중복된 빈발항목집합과 연관규칙셋인 것이다. 이들의 최종결론은 A로 분류냐 아니면 C로 분류냐 하는 궁극적인 의문을 남긴다. 이 경우 본 논문에서는 빈발항목집합이나 연관규칙룰셋의 부분집합의 매칭개수를 각각 조사하여 그 합이 큰 쪽으로 분류하는 방식을 취하였으며 가중치 개념으로 적용하여 본 결과 5% 이상의 정확도 개선효과를 볼 수 있었다. 이는 곧 핵심키워드인 것이다.

결과적으로 본 논문에서 제시한 방법은 수동으로 분류를 해야하는 시간비용에 대한 개념과 단순히 빈출 키워드 조합의 개수에 의한 매치도를 계산하여 정확도를 분석한것보다는 우수한 결론을 실험을 통해 얻었고 실험결과는 분류시스템에서 특정주제어와 가장 연관성이 높고 가장 많이 분류되는 곳으로 추천해주는 시스템에 이용할 수 있고 핵심키워드는 분류가 끝난 후 검색식의 키워드로 활용할 수 있다.

본 연구에서 제시한 방법은 빈발항목과 연관규칙룰셋, 그리고 중복제거에 의한 가중치 개념 등이 적용된 것으로 전처리 과정이 복잡하며 서브클래스 및 하부구조로 내려갈수록 유사단어, 유사분류가 존재하여 분류의 한계가 있다. 이에 대한 지속적인 해결방안이 마련되어야 하며 실험을 통해 나온 평가수치에 대한 축척방안을 통하여 문서에 대한 평가기준으로 사용하는 방법이 앞으로 연구되어야 한다. 클러스터링 결과에 대하여 의미적으로 평가할 수 있는 기법에 대한 연구가 필요하다.

참 고 문 헌

[1] M. Blosseville, G. Hebrail, M. Monteil, N. Penot., "Automatic document classification: natural language processing, statistical analysis, and expert system techniques used together", SIGIR' 92, 1992.

[2] W. Frakes and R. Baeza-Yates, Information Retrieval, Prentice Hall, 1992.

[3] R. Hoch, "Using IR techniques for text classification in document analysis", SIGIR' 94, 1994.

[4] P. Jacobs, Text-Based Intelligent Systems, Lawrence Erlbaum, 1992.

[5] P. Jacobs, "Using statistical methods to improve knowledge-based news categorization", IEEE Expert, April 1993.

[6] L. Larkey and W. Croft, "Combining classifiers in text categorization", SIGIR' 96, 1996.

[7] D. Lewis, "Evaluation and optimizing autonomous text classification system", SIGIR' 95, 1995.

[8] D. Lewis, R. Schapire, and J. Callan, "Training algorithms for linear text classifiers", SIGIR' 96, 1996.

[9] 강승식, 이하규, "한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능", 한글 및 한국어 정보처리 학술 발표논문집, 1996.

[10] 김재군, 김영환, 김성혁, "한국어 정보검색 연구를 위한 시험용 데이터 모음 KTSET 개발", 한글 및 한국어 정보처리 학술 발표논문집, 1996.

[11] 정영미, "정보검색론", 구미무역 출판부, 1993.

[12] 최동시, 정경택, "카테고리와 키워드의 밀접성 정보에 의한 문서 자동분류 시스템 설계 및 구현", 정보과학회 학술발표 논문집, 10, 1995.

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

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

[15] C.Apte, F.Demerau, and S. Weiss(1998). Text mining with deci-

sion rules and decision trees. In Processings of the Conference on Automated Learing and Discovery, Work-shop 6:Learning from Text and the Web

[16] L.Douglaas Baker and Andrew K.Mccallum.Distributional clustering of words for text categorization (1998). In Proceedings of the 21 Ann Int ACM SIGIR Conference on Research and Development in Information Reterival(SIGIR' 9809)

[17] A.McCallum and K.Nigam(1998). Acomparison of event models for naive bayes text classification. In AAAI-98 Work-shop on Learing for Text Categorization

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

[19] 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 Developmnet in Information Retrival(SIGIR' 97), 67-73

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

[21] 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.

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

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

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