잠재 의미 색인 기법을 이용한 국제 특허 분류

진훈태*
*고려대학교 컴퓨터 정보통신대학원
e-mail:admin@doochang.co.kr

International Patent Classificaton Using Latent Semantic Indexing

Hoon-Tae Jin*
*Graduate School of Computer & Information Technology, Korea University

요 약

본 논문은 기계학습을 통하여 특허문서를 국제 특허 분류(IPC) 기준에 따라 자동으로 분류하는 시스템에 관한 연구로 잠재 의미 색인 기법을 이용하여 분류의 성능을 높일 수 있는 방법을 제안하기 위한 연구이다. 종래 특허문서에 관한 IPC 자동 분류에 관한 연구가 단어 매칭 방식의 색인 기법에 의존해서 이루어진바가 있으나, 현대 기술용어의 발생 속도와 다양성 등을 고려할 때 특허문서들 간의 관련성을 분석하는데 있어서는 단어 자체의 빈도 보다는 용어의 개념에 의한 접근이 보다 효과적일 것이라 판단하여 잠재 의미 색인(LSI) 기법에 의한 분류에 관한 연구를 하게 된 것이다.

실험은 단어 매칭 방식의 색인 기법의 대표적인 자질선택 방법인 정보획득량(IG)과 카이제곱 통계량(CHI)을 이용했을 때의 성능과 잠재 의미 색인 방법을 이용했을 때의 성능을 SVM, kNN 및 Naive Bayes 분류기를 사용하여 분석하고, 그중 가장 성능이 우수하게 나오는 SVM을 사용하여 잠재 의미 색인에서 명사가 해당 용어의 개념적 의미 구조를 구축하는데 기여하는 정도가 어느 정도인지 평가함과 아울러, LSI 기법 이용시 최적의 성능을 나타내는 특이값의 범위를 실험을 통해 비교 분석 하였다.

분석결과 LSI 기법이 단어 매칭 기법(IG, CHI)에 비해 우수한 성능을 보였으며, SVM, Naive Bayes 분류기는 단어 매칭 기법에서는 비슷한 수준을 보였으나, LSI 기법에서는 SVM의 성능이 월등이 우수한 것으로 나왔다. 또한, SVM은 LSI 기법에서 약 3%의 성능 향상을 보였지만 Naive Bayes는 오히려 20%의 성능 저하를 보였다. LSI 기법에서 명사가 잠재적 의미 구조에 미치는 영향은 모든 단어들을 내용어로 한 경우 보다 약 10% 더 향상된 결과를 보여주었고, 특이값의 범위에 따른 성능 분석에 있어서는 30% 수준에 Rank 되는 범위에서 가장 높은 성능의 결과가 나왔다.

1. 서론

특허 문서 분류를 위한 국제기준인 IPC(International Patent Classificaton)는 특허 심사 또는 일반인들을 위해 선행 특허문헌을 검색할 때 효과적인 조사 도구로 활용되기 위해 제안된 것이다. 국내 특허·실용신안의 출원건수는 년간 약 15만 여건에 이르며 앞으로도 계속 증가할 것으로 예상되고, IPC 분류체계가 최상위 8개 섹션부터 최하위 약 7만 여개의 서브그룹에 이르기까지 방대한 체계로 이루어져 있음에도 불구하고, 현 행정 시스템은 이를 수작업으로 분류하고 있는 실정이다. 따라서 특허문서 분류에 대한 자동화의 필요성은 자못 크다 할 것이다. 아울러, IPC 체계로 분류된 특허공보가 대외적으로 공적인 문서라는 점을 감안할 때 분류 결과의 신뢰도를 높일 수 있는 방안은 지속적으로 연구되어야 할 과제라 할 것이다.

문서 분류의 성능을 높이기 위한 자질선택 및 색인기법의 중요성은 새삼 강조할 필요가 없을 것이다. 본 논문에서는 특허문서의 분류를 위한 문서 색인에 있어서 용어들에 대한 개념적 접근 방법이 종래의 단어 매칭 방식의 색인 기법에 비해 어느 정도 유용할 것인지에 관한 연구와 명사가 특허 문헌의 의미 구조를 형성하는데 어느 정도의 기여도를 가지는지 실험을 통해 분석해보고자 한다.

2. 관련연구

2.1 특허문서 분류 관련

특허문서의 IPC 분류와 관련하여 문서 빈도 중심의 자질선택방법과 자질선택의 최적의 비율구간 및 분류기의 성능을 비교한 연구에서 카이제곱통계량과 정보이득이 효율적이고, 자질선택의 비율은 20%~40%의 구간이 효율적이라는 연구[1]가 있었다. 또한, 특허 문서의 경우 특유의 의미적 구조를 가지기 때문에 자질선택 이외에도 영역선택이 중요하다는 연구[2]가 있었다. 특허 문헌 검색에서 빈도 정보만을 이용하는 경우의 한계를 극복하기 위해 자주 출현하는 복합명사의 재출현 양상과 복합명사의 역할변화에 따른 가중치의 부여방법에 관한 연구[3]가 있었고, 특허문서에서 단어의 다의성과 표기의 다양성 등으로 야기되는 검색의 불완전성을 보완하기 위하여, IPC별 색인어를

추출하여 대체어 후보를 자동으로 생성하는 것에 관한 연구[4]가 있었다.

2.2 잠재 의미 분석·색인 관련

잠재의미 분석 또는 색인 기법에 관한 국내의 연구로는 기계번역을 위한 역어의 선택에 있어서 단순히 연어 사전만을 이용한 경우보다 잠재 의미 분석 방법을 이용하는 경우가 최고 15%의 성능을 향상할 수 있었다는 연구[5]가 있었고, 의견 문서 분류에 있어서 색인어를 명사 중심의 내용어 집합으로 한 경우 용언, 부사 등으로 구성된 내용어 집합에 비해 잠재 의미 색인에 더 효과적이라는 연구[6]가 있었다.

잠재 의미 구조를 이용한 색인기법은 기계번역 분야나 의견 문서 분류의 연구에는 이용된 바가 있으나, 특허문서 분류에 관한 종래의 연구들은 문헌의 빈도와 단어 매칭에 의한 기법이 주로 의존한 것으로 조사된다.

3. 단어 매칭에 의한 분류를 위한 색인기법

3.1 정보 획득량(Information Gain)

정보 획득량은 기계학습 분야에서 용어의 유익성에 대한 척도로 자주 채용된다. 문서에서 용어의 출현과 부재를 고려해서 범주 예상을 위해 획득된 정보의 비트수를 측정하는 것이다. 용어t의 정보 획득량은 다음 식으로 정의된다[9].

$\{c_1, c_2, \cdots, c_m\}$을 범주집합이라 할 때,

$$G(t) = -\sum_{i=1}^{m} \Pr(c_i)\log\Pr(c_i) + \Pr(t)\sum_{i=1}^{m} \Pr(c_i|t)\log\Pr(c_i|t) + \Pr(\bar{t})\sum_{i=1}^{m} \Pr(c_i|\bar{t})\log\Pr(c_i|\bar{t})$$

3.2 카이제곱통계량(Chi-square)

카이제곱 통계량은 용어t와 범주c 사이의 독립성의 결여 즉, 의존성을 측정하는 것으로 자유도1인 카이제곱 분포와의 비교를 통해 판단할 수 있다. 2차원 이원 분할표를 이용할 경우 용어의 유익성은 아래와 같은 식으로 표현할 수 있다[9].

$$\chi^2(t,c) = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad 1)$$

문서 빈도에 기반을 둔 자질선택 방법들은 이외 다수 있으나 정보 획득량과 카이제곱 통계량이 가장 좋은 성능을 보이는 것을 평가 된다[7].

4. 잠재 의미에 의한 분류를 위한 색인기법

4.1 잠재 의미 색인(Latent Semantic Indexing)[8]

분류 또는 검색에 있어서 색인된 용어 자체가 아닌 의미 구조를 분석하여 용어와 문서간의 관련성 여부를 파악하는 기법이다.

4.2 단어 매칭에 의한 방식의 문제점

단어 매칭에 의한 분류나 검색은 의미에 대한 고려 없이 데이터에 나타난 단어가 동일한지 여부와 출현 빈도만을 고려한다. 그러나 개념을 설명하는 방식은 매우 다양하고, 동일한 개념을 의미(synonymy, 재현율 문제)하는 동의어가 무수히 많을 뿐만 아니라, 하나의 단어 또한 중의적 의미(polysemy, 정확률 문제)를 가지는 경우가 많아 용어 자체의 일치 여부만을 고려해서는 용어와 문서의 의미상의 관련성을 제대로 파악하기 어렵다.

4.3 LSI의 기본원리

문서에서 용어를 사용해서 표현을 하는데 있어서는 부분적으로 모호한 점이 있다 하더라도 잠재적으로는 의미를 가지기 위한 구조가 있다는 가정 하에서 문서 벡터의 공간적 배치를 통해 비록 일치하는 용어가 없더라도 개념적으로 유의미한 관련성이 있는 지를 분석하는 것이다.

색인을 통해 용어와 문서간의 큰 행렬인 “의미 공간(semantic space)”이 구축되면, 이 공간에서 연관성이 있는 용어들은 서로 가깝게 위치하게 된다. 특이값 분해(SVD)는 데이터에서 주요 연관 패턴을 반영하기 위해 공간을 배치할 수 있게 하고, 작고 덜 영향을 미치는 요소는 무시하게 한다. 만약 데이터와 연관된 주요 패턴과 일치되는 용어가 있다면, 실제 문서에 그 용어가 나타나지 않더라도 그 문서는 용어 가까이에 있게 되므로 용어 근처에서 문서를 찾아 분류하거나 반환하게 되는 것이다. 이때 그 공간에서의 단어의 위치는 색인에서 새로운 형태로 기능을 하게 된다.

4.4 구체적인 방법[9]

(1) 문서 색인

형태소 분석 및 불용어 제거 등 전처리를 거친 후 훈련문서 집합과 검증문서 집합을 행렬로 표현한다.

(3) SVD(singular value decomposition) 연산

SVD는 훈련문서 집합에 대해 이루어진다. SVD는 임의의 장방행렬(rectangular matrix) 또는 정방행렬(square matrix)을 새로운 3개의 행렬로 분해하는 선형 대수학의 수학적 알고리즘이다. 임의의 m × n 행렬 X를 위한, SVD연산은 다음 식에 의해 이루어진다.

$$X = UDV^T \quad 2)$$

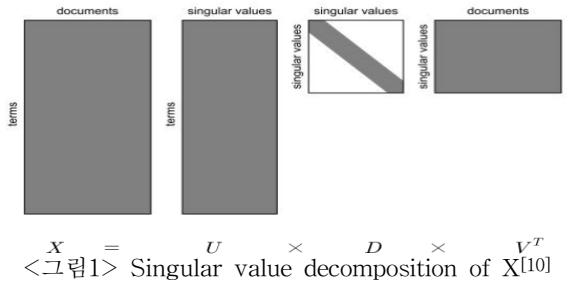
(4) 차원의 축소

<그림1>와 같이 SVD에 의해 원래의 행렬의 분해가 이루어진 후 LSI의 차원의 축소가 이루어진다. 차원의 축소는 특이값들 중 가장 큰 수부터 k개를 선택함으로써 이루

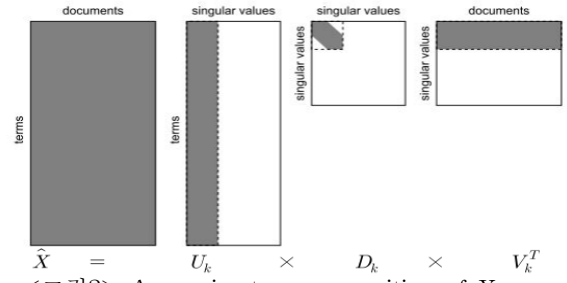
1) N은 전체 문서 수; A는 범주c에 속해있는 문서 중에서 용어t를 포함하고 있는 문서의 수; B는 범주c외 범주에 속해 있는 문서 중에서 용어t를 포함하고 있는 문서의 수; C는 범주c에 속해 있는 문서 중에서 용어t를 포함하지 않는 문서의 수; D는 범주c외 범주에 속해있는 문서 중에서 용어t를 포함하지 않는 문서의 수를 나타낸다.

2) 여기서, $U=XX^T$, $V=X^TX$, 그리고 D는 대각이 X행렬의 특이 값으로 구성된 행렬이다.

어지고, 특이값(singular value)들 중 k개를 제외한 무시할 수 있는 작은 값들은 0으로 세팅되어 행렬의 곱이 이루어지는 동안 U와 V^T에서 무시되어도 되는 행과 열이 결정된다. <그림2>는 결과 행렬을 도시한 것으로 U_k, D_k, 그리고 V_k^T는 이제 k 차원의 "의미 공간"속에서 벡터 좌표를 가지는 새로운 훈련문서 집합이 된다.



<그림1> Singular value decomposition of X[10]



<그림2> Approximate recompoaition of X

(5) 검증집합을 의미공간으로 투영하기(Folding-in)

행렬로 표현된 각 검증집합의 열벡터들은 축소된 차원의 "의미 공간"으로 투영되어야 한다. 이를 위해서는 다음 식을 이용한다.

$$q_i = q_i^T U_k D_k^{-1} \quad 3)$$

(6) 스코어의 할당

훈련집합과 검증집합 모두가 축소된 차원의 "의미 공간" 속에서 표현되면, 각 검증문서 벡터(q_i)와 모든 훈련문서 벡터(x_i) 사이의 거리는 검색에서 널리 사용되는 cosine 유사도 계산식[4]을 이용하여 계산할 수 있으며, 이를 통해 질의와 유사한 문서를 검색하거나, 특정 문서를 스코어가 높은 범주에 할당할 수 있다.

5. 실험 및 평가

5.1 실험환경

본 실험에서는 단어 매칭 분류 방식으로 정보 획득량(IG)와 카이제곱 통계량(CHI)을 이용하여, 잠재 의미 색인(LSI)와 비교하되, 지지벡터기계(SVM), k-최근린법(kNN) 및 나이브 베이즈(Naive Bayes) 분류기를 각각 사용한다. LSI 기법에 있어서 명사의 잠재 의미 구축에 대한 기여도의 평가와 LSI의 적정 자질 수 선택을 위한 특이값(k) 구간을 정하는 실험은 분류기 중 가장 우수한 성능을 보이는 SVM 분류기를 사용한다.

학습을 위한 tool-kit은 뉴질랜드 와이카토 대학에서 개발한 WAKA(Waikato Environment for Knowledge Analysis) ver 3.6.10을 사용한다.

색인 대상 텍스트는 한국특허정보원에서 제공하는 공개 데이터 중 IPC 기준 8개 섹션(A - H)을 기준으로 하되, 분류의 주제를 명확히 하기 위해 각 섹션에 속한 임의의 서브그룹[11]에서 200개씩 1,600개 추출하였다. 검증방법은 교차검증(10 fold) 방법으로 진행하였다. 아래 표1은 본 연구를 위해 추출 대상 서브그룹의 분류 기준에 관한 내용이다. 문서의 대상영역은 제목을 그대로 포함하면서 문서의 특징을 가장 잘 반영하는 요약서(초록)를 대상하였다.

SubGroub	Title
A01P	화합물 또는 조성물의 살생물, 유해 생물 기피, 유해 생물 유인 도는 실물 생장 조절 활성
B29B	성형 재료의 준비 또는 전처리; 조립 또는 예비 성형품의 제조; 플라스틱을 함유하는 폐기물로부터 플라스틱 또는 다른 구성 성분의 회수
C09B	유기 염료 또는 염료 제조에 밀접한 관련이 있는 화합물; 매염제; 레이크
D04B	메리야스 편성
E04D	지붕잇기; 천장; 물받이 홈통; 지붕 공사용 공구
F16J	피스톤; 실린더; 압력용기 일반; 밀봉장치
G01K	온도의 측정; 열량의 측정; 달리 속하지 않는 감온소자
H03L	전자적 진동 또는 펄스 발생기의 자동제어, 기동, 동기 또는 안정화

<표1> 각 서브그룹의 타이틀 및 주제

[Source : WIPO International Patent Classification 2012.01 Version]

5.2 실험결과

(1) 단어 매칭 방식의 색인 기법과 잠재의미 구조의 색인 기법의 비교

초기 자질 추출은 노이즈를 줄이기 위해 클래스 당 최소 빈도를 10으로 하였다. 그 결과 최초 색인 대상 자질의 수는 1271개이다. 평가 척도는 정확률(Precision), 재현율(Recall) 및 F-measure를 이용하여 평가하였다.

Cassifier	measure	CHI	IG	LSI
SVM	Precision	0.875	0.876	0.902
	Recall	0.874	0.875	0.901
	F-measure	0.874	0.875	0.901
KNN	Precision	0.52	0.52	0.708
	Recall	0.494	0.494	0.692
	F-measure	0.486	0.486	0.69
Naïve Bayes	Precision	0.877	0.877	0.779
	Recall	0.873	0.873	0.698
	F-measure	0.873	0.873	0.707

<표2> 분류기 성능 비교

표2에서와 같이 분석결과 LSI기법이 단어 매칭 기법(IG, CHI)에 비해 우수한 성능을 보였으며, SVM, Naive Bayes 분류기는 단어 매칭 기법에서는 비슷한 수준을 보였으나, LSI 기법에서는 SVM의 성능이 월등이 우수한 것으로 나왔다. 또한, SVM은 LSI 기법에서 약 3%의 성능 향상을 보였지만 Naive Bayes는 오히려 20%의 성능 저하를 보였다. 비록 kNN의 경우 LSI기법을 통해 약 40%의 성능향상을 보였으나 상대적으로 매우 낮은 성능을 보이고 있다.

(2) LSI 기법에서 명사의 기여도 평가 및 특이값 범위에 따른 성능 비교

잠재 의미 구조를 위한 명사의 기여도 평가를 위해 형태소 분석기를 이용하여 명사만을 추출한 데이터와 문서 내에 모든 단어를 분리한 데이터를 분리하여 실험하였다.

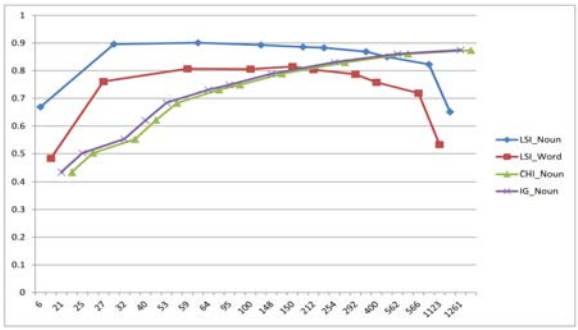
3) q_i 는 검증집합의 열벡터를 나타낸다.

4) $(q_i, x_j) = \frac{q_i \cdot x_j}{|q_i||x_j|}$

마찬가지로 클래스당 최소 빈도를 10으로 하였고, 최초 색인 대상 자질의 수는 1271개이다. 최적의 특이값(k) 범위 즉, 축소된 차원을 정하기 위해 특이 값(singular value)의 내림차순으로 0.1 단위비율(Rank)로 나누어 그에 따른 자질 수와 성능을 비교 분석하였다. 아울러 상대적 비교를 위해 정보 획득량(IG)과 카이제곱 통계량(CHI) 기법을 활용하되, 클래스당 최소 빈도 수를 10단위로 200까지 구분하여 해당 자질 수에 따른 성능을 기록하여 표시하였다.

Rank	Noun		Word Token	
	Number of feature	F-measure	Number of feature	F-measure
0.1	6	0.669	6	0.483
0.2	29	0.896	27	0.761
0.3	63	0.901	59	0.807
0.4	109	0.893	100	0.806
0.5	156	0.886	150	0.815
0.6	219	0.883	212	0.804
0.7	300	0.869	292	0.787
0.8	408	0.85	400	0.758
0.9	575	0.823	566	0.719
0.999	1141	0.651	1123	0.533

<표3> 내용어의 종류 및 자질 수에 따른 성능평가



<그림1> 자질수의 변화에 따른 성능분석

표3과 그림1에 나타난 바와 같이 LSI 기법에서 명사가 잠재적 의미 구조에 미치는 영향은 모든 단어들을 내용어로 한 경우 보다 약 10% 더 향상된 결과를 보여주었고, 특이값(k)의 범위 따른 성능 분석에 있어서는 30% 수준에 Rank 되는 범위에서의 가장 높은 성능 결과가 나왔다.

5.3 평가

LSI 기법은 SVD를 통해 원래의 색인이 새로운 의미 공간으로 구축되고, 용어와 문서의 관련성은 해당 용어의 출현 여부가 아니라 공간에서의 상대적 거리라는 점을 고려할 때 선형모델의 확장인 SVM 분류기가 가장 우수한 성능을 발휘하는 것으로 보인다. 특허문서는 새로운 기술용어들이 많이 발생하고 그에 관한 설명이 다양한 방식으로 이루어진다는 특징을 고려해 단어 매칭에 의한 방법보다 잠재 의미 구조에 의한 방법이 더 적합한 기법으로 보인다. 또한 기술용어가 대부분 명사 또는 복합 명사로 이루어지는 경우가 많기 때문에 특허문서의 LSI 기법적용에 있어서는 명사의 활용이 매우 유용할 것으로 보인다. LSI 기법은 단어 매칭 방식과 달리 매우 낮은 차원에서 높은 성능을 발휘하고, 자질 수의 작은 변동에 매우 민감하게 반응하므로 이를 실험을 통해 잘 정하는 것이 성능 향상에 도움을 준다는 것을 알 수 있었다.

6. 결론 및 향후과제

우리는 잠재 의미 색인 기법은 IPC를 위한 특허문서 자동 분류에 있어서 단어 매칭에 의한 색인 기법에 비해 매우 효과적인 기법이며, 매우 낮은 차원에서 높은 성능을 발휘한다는 것을 실험을 통해 알 수 있었다.

기술의 발달로 인해 새로운 기술용어가 지속적으로 배출되는 환경에서 다른 용어들과의 관계에서 개념을 도출하고 이를 근거로 문서들 간의 관련성을 분석하는 기법은 동일 기술에 대한 다양한 표현들이 존재하는 특허 문서에 적용하기에 매우 유용한 기법임이 틀림없다. 이러한 기법은 향후 분서 분류 분야뿐만 아니라 선행 기술과의 관계에서 신규성, 진보성의 인용참증 적합도를 평가하는 객관적 척도로 활용될 수 있을 것이라 판단된다. 향후 이에 관한 연구를 계획하고자 한다.

참고문헌

[1] 박찬정, 성동수, 이건배 “기계 학습을 이용한 특허 문서의 자동 IPC 분류”, 한국정보기술학회논문지 제10권 제4호 2012. 4.
[2] 김재호, 최기선 “문서의 의미적 구조정보를 이용한 특허 문서 분류”, 한국정보과학회 언어공학연구회 학술발표논문집, pp.28-34, 2005.10.
[3] 손기준, 이상조, “특허 문헌 검색에서 복합명사 가중치 부여 방법”, 한국정보과학회 학술발표논문집, pp.895-897, 2004.
[4] 백종범, 김성민, 이수원, “특허 정보 검색 품질 향상을 위한 대체어 후보 자동 생성 방법”, 정보과학회 논문지: 소프트웨어 및 응용 제36권 제10호, pp861-873, 2009. 10.
[5] 장정호, 김유섭, 장병탁 “잠재의미구조 기반 단어 유사도에 의한 역어 선택”, 한국정보과학회 봄 학술발표논문집, Vol. 29. No1, 2002.
[6] 이지혜, 정영미, “지도적 잠재의미색인(LSI) 기법을 이용한 의견 문서 자동 분류에 관한 실험적 연구”, 정보관리학회지 제26권 제3호, 2009.
[7] 고영준, 서정연, “문서관리를 위한 자동문서범주화에 대한 이론 및 기법”, 정보관리연구, vol.33, no. 2, pp. 19-32, 2002년
[8] Scott Deerwester, Susan T. Dumais, George W. Furnas, Tomas K. Landauer, Richard Harshman, "Indexing by Latent Semantic Analysis", JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 41(6):391-407, 1990.
[9] Abigail R.Razon, Belen D. Calingacion, Rowena Cristina L. Guevara "Automated Essay Content Analysis Using Concept Indexing",http://citeseerx.ist.psu.edu/
[10] Tristan Miller "Essay Assessment with Latent Semantic Analysis",J. EDUCATIONAL COMPUTING RESEARCH, Vol. 29(4) 495-512, 2003.
[11] "INTERNATIONAL PATENT CLASSIFICATION 국제특허분류 가이드 해설", 특허청, p36, 2010년