개인화된 특허 분류 시스템 사례 연구

서형국*° 최광선* 안한준** 최성준*
*솔트룩스 HLT 연구소 **LG전자 특허센터
apollos@saltlux.com kschoi@saltlux.com hjan@lge.com sjchoi@saltlux.com

A Case Study on Personalized Patent Classification System

Hyung-Kook Seo*° Kwang-Sun Choi* Han-Joon Ahn** Sung-Joon Choi*
*HLT Lab, Saltlux
**Patent Center, LG Electronics

요 약

개인화된 특허 분류 시스템은 기존의 자동 분류 및 특허 문서의 특성, 그리고 분류 체계의 개인화를 고려하여 접근해야 한다. 본 논문에서는 개인화된 특허 분류 시스템을 구축하는 데 있어 개인화된 분류 체계 및 모델의 구축, 특히 분류체계 구축에 있어서의 자동화에 초점을 두었다. 우리는 특히 분류체계 구축 자동화에 있어 특허 문서의 기존 분류체계인 IPC 및 문서 클러스터링을 활용하였다. 다음으로 이를 기반으로 한 구축 시스템 사례를 들었다. 구축 후 나타난 정성적 문제점을 분석해보고, 분석 결과를 향후 연구 방향으로 삼고자 한다.

1. 서론

지적 재산권의 중요한 표현 형태인 특허 문서는 점차 그 건수가 증가하고 있다. 현재 국내 연간 출원 특허 건수는 1990년 약 25,000건 수준에서 2004년 137,000여 건으로 약 5.4배 수준으로 증가되었다고 한다[1]. 이렇게 급증하고 있는 특허 문서들 가운데 관심 분야의 특허만을 선택하여 접근하기 위해 많은 기술들이 활용되고 있다. 여기에 각종 텍스트마이닝 기법이 다양하게 활용되고 있다.

특허 문서 분류는 이미 정해져 있는 분류 체계에 대하여 분류 대상 특허 문서들이 어느 하 위 분류 체계에 속하는지에 대한 발견을 자동화할 수 있다는 점에서 매우 유용한 텍스트마이닝 기법이다. 특히 이러한 분류를 가능하게 하는 분류 체계가 단일 레벨이 아닌 계층형일 경우 그 편의성은 더욱 증대될 수 있을 것이다.

특허 문서 분류 체계는 일반적으로 전체 특허를 염두에 둔 분류 체계를 두고, 이 분류 체계 내 각 클래스를 대표할 수 있는 특허 문서들을 학습 문서 집합으로 하여 기계 학습 절차를 통해 구축할 수 있다. 그러나 보통 특정 개인이 관심을 가지는 분류 체계는 그 전체 분류체계 가운데 일부이거나, 혹은 관심사에 따라 재구성된 것일 것이다. 즉, 개인화가 필요한 것이다. 개인화된 특허 문서 분류는 개인이 원하는 분류 체계를 구축하는 데에서 출발한다.

이와 같이 개인화된 특허 문서 분류체계를 구축하고자 할 때 발생하는 문제점은 다음과 같다. 개인은 관심사에 속하는 문서들을 쉽게 선택할 수 있으나, 선택한 문서들을 분류 가능하도록 범주화하는 데 어려움을 느낄 수 있다. 즉, 관심 분야의 문서들을 선택한 후, 이들을 분류체계화 하는 데 시스템의 도움을 입은 자동화가 필요할 수 있다.

이와 관련한 관련 선행 연구 사례로는, 개인화된 분류 체계 구축에 대한 [4]와 같은 접근법이 있다. 여기에서는 키워드들을 이용하여 분류 체계를

정의하고, 학습문서는 키워드를 이용한 검색엔진 질의 결과를 이용하여 구성된다. 이에 대하여 개인화된 분류 체계는 평면적으로 구성되거나, 일반화되어 정의된 분류 체계의 부분집합을 선택하여 활용하였다 (후자의 경우가 더 좋은 결과를 얻었다고 한다). 그러나 이 방법의 경우 개인화된 모델이 전체 모델의 부분집합이라는 점, 그리고 개별 사용자의 개입이 최소화되어 있다는 점에서 문제를 가지고 있다.

우리는 개인화된 특허 분류 시스템 및 그 자동화를 위하여 다음과 같은 사항을 출발점으로 하였다. 첫째는 특허 기술의 분류체계인 IPC (International Patent Classification, 국제특허분류체계)를 응용하여 분류 체계를 자동으로 구축하는 것이다. 그리고 둘째로는 문서 클러스터링을 통해 분류 체계를 자동적으로 구축하는 것이다.

2. 접근 방법

우리가 검토해 본 자동화된 개인화 분류 체계 구축의 첫 번째 접근 방법은 IPC를 이용한 분류 체계 구축이다.

각국의 모든 특허 문서는 특허 기술에 따른 국가별 분류체계를 가지고 있다. 그리고 이 가운데서 국제적으로 공통적으로 쓰이는 분류 체계가 바로 IPC이다. 그리고 IPC는 역시 계층형 분류체계이므로, 의도한 계층형 분류 체계를 이루는 데에도 도움이 될 수 있다. IPC의 예는 [표 1]과 같다. IPC는 최상위 분류체계는 섹션(A), 그 다음은 클래스(01) 및 서브클래스(G), 그 하위에 메인 그룹(001) 및 서브그룹(04)으로 나누어지는 계층형 체계이다.

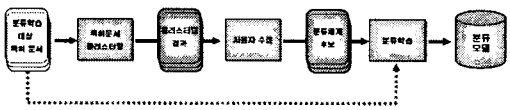
[표 1] IPC의 예

A	01	G	-	001	/	04

이를 이용하여 계층화된 분류체계를 구성하는 방법은 다음과 같다. IPC의 각 구성 단위 (섹션, 클래스, 서브클래스, 메인 그룹, 서브 그룹)는 IPC 분류 체계의 각 구성 레벨로 볼 수 있다. 즉 최상위는 섹션별, 다음 레벨은 클래스별, 그 다음은 서브클래스별과 같은 분류 체계에 속해 있는 것이므로, 대상 문서의 IPC에 따라 레벨을 조정하여 분류 체계를 구축할 수 있을 것이다.

그러나 IPC를 이용한 분류 체계 자동 구축은 다음과 같은 문제점이 있다. 관심 대상 문서들이 서로 다른 기술 분야에 어느 정도 무리지어 혼재되어 있다면 매우 이상적인 경우이겠으나, 이와는 반대로 유사한 기술 분야에 몰려 있거나, 역으로 문서들이 공통된 기술 분야별로 그룹화되어 있지 않고 혼재되어 있을 가능성이 훨씬 높기 때문이다. 경우에 따라서는 IPC가 동일한 문서만으로 분류 체계를 구성해야 하거나, 그와 반대로 IPC 사이의 공통성을 아예 찾지 못하는 경우가 발생할 수 있는 것이다. 이러한 경우 IPC만으로는 자동화된 분류체계 구축이 어려울 수 있다.

이러한 문제를 완화하기 위한 두 번째 접근 방법은 문서 클러스터링을 이용한 분류 체계 구축 지원이다. 문서 클러스터링 기법을 이용하여 대상 문서 집합에 대하여 적절한 깊이의 계층형 클러스터링을 실시하고, 그 결과 혹은 사용자 개입에 의해 다소 수정된 클러스터링 결과를 분류 체계로 삼아 자동 분류 학습을 실시, 개인화된 분류체계를 구성하는 것이다. 즉 이를 통해 구성된 개인화된 분류 시스템 개념은 아래의 [그림 1]과 같이 표현할 수 있을 것이다.



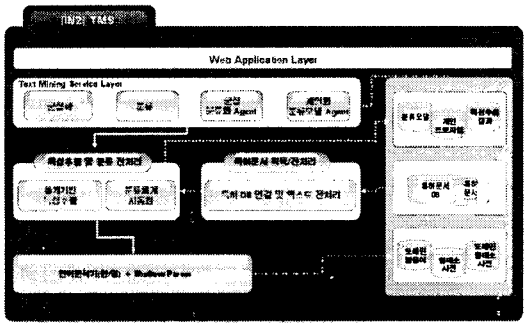
[그림 1] 클러스터링을 이용한 개인화 분류 절차

분류 학습 대상이 될 특허 문서들을 문서 클러스터링을 통해 계층형으로 클러스터링하여 분류 체계 후보로 삼는다. 그 다음 이 분류 체계 후보를 사용자가 수정할 수 있도록 하여 분류 체계에 사용자 의도가 반영될 수 있도록 한다. 마지막으로 이 분류 체계 후보를 대상으로 분류 학습을 실행하여 개인화된 분류 모델을 얻어낸다.

위에서 제기한 두 가지 방법을 함께 응용하면 다음과 같은 방법이 가능할 것이다. 먼저 최상위 계층에 대한 분류는 IPC를 이용하거나, 이에 상응하는 명확한 분류 체계를 이용하여 분류한다. (실제 구축 사례에서는 특허 DB별 분류를 최상위로 가져갔다) 그리고 그 하위 레벨부터는 문서 클러스터링을 이용하여 자동화된 분류 체계를 얻어 사용자 개입 및 분류 학습 과정을 통해 개인화된 분류 모델을 얻는 것이다.

3. 시스템 구축 사례

2장에서 우리는 개인화 분류 체계 구축을 위한 접근 방법을 살펴보았다. 이러한 접근법을 이용하여 우리가 구성한 시스템의 구조도는 아래의 [그림 2]와 같다.



[그림 2] 구현 시스템 구조도

구현 시스템은 크게 3개의 서비스 계층 및 1개의 프리젠테이션 계층으로 나누어졌다. 각각의 서비스는 JDK 1.4를 이용한 java 응용 프로그램으로 작성되었다. (언어 분석 등 일부 모듈은 속도를 위하여 플랫폼 독립적인 C++ 코드로 작성되었고, JNI를 통하여 java와 연동하였다)

텍스트마이닝의 기본이 되는 언어분석 서비스가 가장 아래의 서비스 계층이고, 그 위에 텍스트 전처리 및 특성추출 모듈이 다음 단의 서비스 계층이다. 그 상위 계층에는 주요 텍스트마이닝 서비스(문서 클러스터링, 문서 분류) 및 이 서비스를 응용한 에이전트 스타일의 서비스가 위치한다.

마지막으로 이들 서비스 계층 상위에 웹 어플리케이션을 이용한 사용자 UI 계층이 존재하여 사용자와의 상호작용에 이용된다.

여기에서 우리가 구축한 시스템의 텍스트마이닝 서비스 계층의 네 가지 서비스 및 구현 방법론을 좀 더 자세히 살펴보면 다음과 같다.

문서 클러스터링 서비스

문서 클러스터링 서비스를 제공한다. 문서 클러스터링 알고리즘은 가장 간편한 k-Means 클러스터링 알고리즘을 이용했다. 이때 k-Means 클러스터링 알고리즘은 문서의 특성값 사이의 코사인 유사도를 이용하여 결정하였다[2].

기본 클러스터링 알고리즘을 이용하여 다단계 클러스터링을 구현하였고, 그 절차는 다음과 같다.

- 한 범주(Category)에 속할 수 있는 문서 건수의 최대값을 정한다. 이를 C라고 하자.
- 현재 단계의 대상 문서 수 N개에 대한 적절한 클러스터 개수 k를 다음의 식을 이용하여 구한다. (이는 통계학에서 계급의 수를 구하기 위해 사용되는 Sturges 공식을 변형한 것이다)

$$k = \log_2 N = \frac{\log N}{\log 2}$$

- k개의 클러스터가 나오도록 클러스터링을 기본 알고리즘을 통해 수행하고, 각 결과 클러스터의 문서 건수가 C건을 초과할 때 해당 클러스터에 대해 위의 과정을 재적용한다.
- 모든 결과 클러스터의 문서 건수가 K건 미만이 되면 종료한다.

그리고 추가적으로 기존의 특허 분류체계 혹은 그에 상응하는 최상위 분류체계를 활용하기 위해 클러스터링 서비스에서 제공하는 기능이 있다.

특허 문서의 메타데이터 차원에서 제공할 수 있는 정보, 즉 IPC 정보, 특허문서가 속하는 데이터베이스 (예: 한국 특허 DB, 번역된 일본 특허 DB, 혹은 PCT 특허 DB)에 대한 식별자 등을 기본으로 하여 최상위 수준 혹은 2-3단계 수준에 대한 범주화를 실시할 수 있다. (이에 대한 서비스만으로 클러스터링 서비스를 제공하는 국내 업체 사례도 있다)

문서 자동분류 서비스

크게 두 가지 하위 서비스를 가진다. 분류 체계별 샘플 문서를 이용하여 분류 모델을 만드는 분류 학습 모듈과, 분류 학습 모델을 이용하여 대상 문서를 분류하는 분류 모듈로 나눌 수 있다.

분류 학습 및 분류 모듈 방법론으로는 SVM(Support Vector Machine)을 이용하였다. SVM 모델 최적화를 위한 각종 Parameter는 전체 특허 문서에 대한 분류체계를 구성하는 데 이용한 Parameter를 동일하게 이용하였다. (이에 대한 개인화된 최적 Parameter 모델은 추후 연구 과제로 남겨두었다)

군집 분류화 Agent 서비스

군집 분류화 Agent 서비스는 클러스터링을 통해 자동으로 생성된 분류 체계를 분류 모델로 분류 학습하는 모듈이다. 이를 위해서는 Web 응용프로그램을 이용한 사전 절차가 필요하다. 이는 다음과 같다.

- 군집 분류화 이전에 개인화된 분류에 사용될 대상 문서는 Web Application 단에서 검색 서비스 등을 통하여 선택될 수 있다.
- 대상 문서에 대하여 클러스터링 서비스를 이용하여 클러스터링을 수행한다. 클러스터링 수행 결과는 [그림 3]과 같이 사용자에게 제시된다.

[그림 3] 클러스터링 결과 화면

- 클러스터링 수행 결과를 보고 사용자는 군집 분류화를 선택할 수 있다. 그러나 바로 군집 분류화 에이전트가 실행되는 것은 아니고, 군집 결과를 사용자가 편집할 수 있는 피드백용 페이지를 실행한다.

군집 분류화 Agent는 이 편집 결과를 이용하여 실시간이 아닌 Batch 스케줄링을 통해 분류 모델을 생성한다. 이때 문서 자동분류 서비스의 분류학습 서비스를 이용하여 실제 작업을 수행한다. 그리고 결과로 분류 모델이 정상적으로 생성되었을 때 개인화 분류 Agent가 동작하도록 메시지를 전달한다.

개인화 분류 Agent

개인화 분류 Agent는 군집 분류화 Agent를 통해 생성된 개인화된 분류 모델을 이용하여 개인 프로파일이 가지고 있는 관심 특허 문서를 분류하고 그 결과를 저장하는 Agent이다.

이를 위하여 개인 프로파일은 개인화된 분류 모델, 분류 학습에 사용된 문서 및 사용되지 않은 관심 문서에 대한 정보를 가지고 있다.

특정 개인 프로파일에 대하여 새로운 개인화 분류 모델이 만들어졌을 때 개인화 분류 Agent는 다시 실행되어 새로운 분류 모델에 맞추어 관심 문서로 등록된 문서들을 재분류하고, 그 결과를 관련 저장소에 기록한다.

4. 평가 및 향후 연구과제

본 시스템은 현재 LG전자의 특허정보검색 시스템 내에 적용, 구현되어 있다. 이 시스템은 특허 검색 결과를 개인 프로파일 내 관심 문서로 등록할 수 있고, 본 시스템이 제공하는 개인화된 특허 분류 서비스를 이용하여 분류 모델을 생성, 새로운 관심 문서 및 기존 관심 문서를 분류하여 사용자가 원하는 특허 정보를 좀 더 체계적으로 접근할 수 있도록 돕고 있다.

[표 2] 사용 기록을 통한 정량적 분석

순번	전체 문서수	분류성공문서수	분류성공률 (%)
1	789	735	93.16
2	286	93	32.52
3	100	100	100.00
4	200	200	100.00
5	2227	1197	53.75
6	325	274	84.31
7	53	35	66.04
8	201	100	49.75
9	240	240	100.00
10	200	12	6.00
11	409	382	93.40
12	601	446	74.21
13	39	11	28.21
14	113	4	3.54
15	598	32	5.35
16	140	2	1.43
17	8037	6089	75.76
18	30	5	16.67
19	1311	291	22.20
20	275	213	77.45
21	339	338	99.71
22	414	333	80.43
23	4	4	100.00
24	78	27	34.62
25	176	91	51.70
26	152	4	2.63
27	2863	1396	48.76
28	237	57	24.05
평균 성공률			54.49

전체 프로젝트 건수	56건
분류모델 구성 성공건수	28건
성공률	50%

이 시스템을 통하여 사용자들은 다양한 반응을 보이고 있다. 그 중 가장 관심이 가는 사항은, 분야에 따른 사용자의 만족도의 편차가 크다는 것이다. 만족도를 정량화하기는 쉽지 않지만, 사용 기록을 통해 간단히 정량적으로 분석하면 [표 2]와 같다. 이 자료는 2월에서 6월 사이의 사용 기록을 분석하였다.

분류 모델 구성 실패는 본 시스템을 이용한 분류 모델 구성에 실패했다는 것을 의미하고, 분류 성공률은 구성된 분류 모델을 이용하여 대상 특허가 얼마나 성공적으로 분류되었는지의 비율이다. 이는 구성된 분류 모델이 얼마나 유용한지에 대한 하나의 지표가 될 수 있다고 볼 수 있다. 이 분석 결과에 따른 본 서비스의 분류 모델 구성 성공 건수는 50%대 (28/56) 이고, 이 구성 성공 건수만을 기준으로 했을 때 사용자가 얻은 분류 성공률은 평균적으로 54% 대에 머무르고 있다.

이에 대한 원인은 아마 다음과 같은 이유이지 않을까 생각되고, 각각 추가적인 연구 과제로 생각해 볼 수 있을 것이다.

- 언어분석 서비스에서 제공하는 형태소 분석 결과에서 많은 오류가 발생하였다. 이는 특허 분야에서 특히 빈번하게 나타나는 신규 용어의 문제와 맞물려 있다. 그리고 화학식 등의 경우 언어분석으로는 해결하기 힘든 문제점이 발생하고 있다. 이를 위해서는 특허 분야에 초점을 둔 언어자원의 수집 및 최적화가 필요하다고 판단된다.

- 그리고 개인화 분류 모델에서 샘플로 쓰인 문서들을 분류 모델에 넣고 다시 테스트해 보았을 때의 정확률 및 재현률이 전체 분류 모델보다 떨어지는 현상을 발견하였다. 이는 전체 분류 모델을 구축하는 데 이용한 파라미터를 그대로 개인화된 분류 모델에 적용함으로 벌어진 문제로 보이고, 이를 자동화하여 보정하는 방법을 다음 연구 과제 중 하나로 생각하고 있다.

- 그리고 사용자가 관심을 가지고 분류하려고 하는 문서들은 대개 매우 유사한 분야의 문서들이다. 지극히 유사한 특성을 가진 문서에 대한 클러스터링 및 분류는 지금의 유사도 기반 클러스터링 방법론 및 분류 모델로는 쉽지 않은 작업이 될 수 있다. 따라서 특허 문서의 특성을 고려한 새로운 접근법이 필요할 수 있다. 이에 대하여 다음과 같은 참고문헌들의 시도가 참고가 될 수 있겠다. [3]에서는 특허 문서에 서의 특성(자질) 선택을 위하여 래퍼(Wrapper)와 필터(Filter)를 상호 보완적으로 결합, 최적의 필터를 자동화하여 찾는 래퍼를 제안하고 있다. 그리고 [5]에서는 웹 검색 결과를 분류해 제공하기 위해 웹 문서의 특성을 고려한 "Fast-feature" 를 제시하고 있다.

- 마지막으로 사용자가 궁극적으로 원하는 분류 모델은 의미 기반의 모델이나, 현 시스템이 자동화하여 제공하는 모델은 통계 기반의 모델이라는 데 가장 큰 간극이 있다고 생각된다. 이를 해소하기 위해 특허 도메인을 분석하여 의미기반의(Semantic) 분류 모델을 도출할 필요가 있다.

참고 문헌

[1] 한국특허정보원 특허정보전략팀, "한국의 특허동향 2005", *Patent21 통권 65호*, pp. 4-22, 2006년 1월.

[2] 김남영, "유사도 기법에 따른 K-means 알고리즘을 이용한 문서 클러스터링 결과 분석", 전북대학교 대학원 석사학위논문, 2003

[3] 정하용, 황금하, 신사임, 최기선, "특허 분류를 위한 효과적인 자질 선택", 한국정보과학회 추계학술대회, 2005

[4] Sun, A., Lim, E.-P. & Ng, W.-K, "Personalized classification for keyword-based category profiles.", Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, pp. 61-74, 2002

[5] Bill Kules, Jack Kustanowitz, Ben Shneiderman, "Categorizing Web Search Results into Meaningful and Stable Categories Using Fast-Feature Techniques", *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries JCDL '06*, 2006