전자 카탈로그 자동분류기 시스템과 그 활용
(System and Utilization for E-Catalog Classifier)

이 익 훈 † 전 종 훈 ††
(Ig-hoon Lee) (Jonghoon Chun)

요 약 정확하게 정의된 전자 카탈로그(또는 상품정보)는 전자상거래 시스템의 핵심기반이다. 전자 카탈로그의 분류정보는 전자 카탈로그 정보 구축을 위한 기반 정보이며, 전자 카탈로그를 이용하는 시스템의 질을 좌우하는 중요 정보이다. 그러나, 정보시스템의 활용이 증가함에 따라, 시스템에서 관리해야 할 전자 카탈로그의 양은 대용량화 되었고, 대용량 전자 카탈로그의 분류 작업은 더욱 복잡하게 되었다. 본 논문에서는 전자 카탈로그를 자동분류하기 위한 자동분류기 시스템을 설명하고 자동분류기를 활용한 기업 정보시스템의 카탈로그 관리 프로세스 개선 구축 경험 및 기업의 전자카탈로그 표준화 작업을 위한 자동분류기 활용방법을 제시한다. 더불어 향후 유사 시스템 구축에 도움이 될 수 있도록 경험으로부터 얻은 자동분류기 시스템 구축 및 활용 이슈를 제시한다.

키워드 : 전자 카탈로그, 자동분류, 상품분류, 전자구매

Abstract A clearly defined e-catalog (or product) information is a key foundation for an e-commerce system. A classification (or categorization) is a core information to build clear e-catalogs, can play an important role in quality of e-commerce systems using e-catalogs. However, as the wide use of online business transactions, the volume of e-catalog information that needs to be managed in a system has become drastically large, and the classification task of such data has become highly complex. In this paper, we present an e-catalog classifier system, and report on our effort to improve an e-catalog management process and to standardize e-catalogs for enterprises by use of automated approach for e-catalog classifier systems. Also we introduce some of the issues that we have experienced in the projects, so that our work may help those who do a similar project in the future.

Key words : e-catalog, automated classification, product classification, e-procurement, classification

1. 서 론

전자 카탈로그는 전자상거래 환경에서 상품 및 서비스에 대한 정보를 담는 문서로서, 효율적인 검색과 유지보수를 위해 카탈로그 분류체계에 따라 정확히 분류된 상태로 관리되어야 한다[1]. 흔히 사용되는 UNSPSC[2], eCl@ss[3], G2B(조달청 물품 분류체계)[4,5] 등과 같은 전자 카탈로그 분류체계의 복잡성으로 인해 전자 카탈로그를 분류 체계에 따라 분류하는 작업은 전문가의 수작업으로 이루어져 왔다. 그러나 정보시스템의 활용이 증가함에 따라 한 시스템에서 관리해야 하는 카탈로그는 대용량이 되었고 그에 따라 분류작업은 더욱 어려운 작업이 되었다.

수작업을 통한 전자 카탈로그의 분류작업은 분류체계 자체가 복잡하다는 점과 카탈로그에 대한 전문지식이 요구된다는 점 때문에 많은 시간과 인력을 필요로 하는 고비용의 작업이다. 카탈로그들의 정확한 분류정보는 또한 전자 카탈로그의 품질과 결정하는 매우 중요한 정보이다. 전자상거래 시장의 성장과 기업의 전자구매시스템의 활성화에 따라 정확하고 빠른 자동화된 카탈로그 분류 방법 및 해당 시스템 구축의 필요성이 커졌다.

본 논문은 전자 카탈로그의 자동분류기 시스템을 소개하고 해당 시스템을 현장에 활용한 결과인 전자카탈

· 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업(IITA-2008-C1090-0801-0031)의 연구 결과로 수행되었음

† 정 회 원 : 프람트 기술연구소 연구소장
ihlee@prompt.co.kr
†† 종신회원 : 명지대학교 컴퓨터공학과 교수
jchun@mju.ac.kr

논문접수 : 2008년 8월 26일
심사완료 : 2008년 10월 21일

Copyright@2008 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지: 컴퓨팅의 실제 및 레터 제14권 제9호(2008.12)

로그 등록관리 프로세스 개선, 전자카탈로그 표준화 프로세스 개선 및 관련 구축 이슈를 구체적으로 제시한다. 본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 전자 카탈로그 자동분류기에 대한 연구에 대하여 설명하고, 3장에서는 자동분류기 시스템을 활용한 카탈로그 관리 프로세스 및 구축 이슈들을 살펴본다. 4장에서는 자동분류 시스템을 이용한 전자카탈로그 표준화 작업 프로세스 및 구축 이슈들을 제시한다. 마지막으로 5장에서 결론을 도출한다.

2. 전자 카탈로그 자동분류기

2.1 텍스트 문서 분류와 전자 카탈로그 분류

전자 카탈로그는 일반 텍스트 문서로 간주할 수 있으므로, 인공지능 분야에서 연구된 텍스트 분류 알고리즘들을 사용해서 분류가 가능하다. 실제로 이러한 시도들이 있어왔다[6,7]. 그러나 전자카탈로그를 분류함에 있어, 기존의 텍스트 분류 알고리즘을 적용하기 보다는 전자 카탈로그의 특성을 파악하고 이를 분류시스템에 적용하면 분류의 정확도를 높일 수 있다.

전자 카탈로그는 일반 텍스트 문서와는 달리 아래와 같은 특성을 갖는다.

- 속성-값 쌍들의 집합
  : 전자 카탈로그는 텍스트 위주로 구성된 문서이지만, 개념적으로는 속성-속성값 쌍들의 집합으로 이루어져 있다. 사진기 전자카탈로그의 경우, '상품명', '액정크기'라는 속성이 있다고 가정하면 '상품명'-'ABC_ 디지털사진기', '액정크기'- '2인치'가 속성-속성값의 쌍이 된다. 즉 전자 카탈로그는 준구조적 문서(Semi-structured document)로 볼 수 있다.
- 공통속성과 개별속성
  : 공통속성은 분류와 상관없이 모든 전자카탈로그에 공통적으로 존재하는 속성들이다. 위의 사진기의 예에서는 '상품명'이 공통속성이 된다. 개별속성은 카탈로그가 속하는 분류에 따라서 존재하는 속성들이다. 위의 사진기 예에서는 '액정크기'가 개별속성이다.
- 동의어와 영문 표기
  : 카탈로그 정보는 일반 텍스트 문서에 비해 '베어링', 'Bearing', 'TV', 'Television', '텔레비전' 등의 동의어 사용이 빈번하고 영문표기를 많이 사용하고 있다.

앞에서 언급한 전자 카탈로그의 특성을 반영한 카탈로그 자동분류 시스템을 개발해야 할 뿐만 아니라, 카탈로그 분류 작업 시에도 그 특성을 고려하고 효율적인 프로세스로 작업을 수행하여야만 정확한 카탈로그 정보를 유지 관리할 수 있다.

2.2 전자 카탈로그 자동분류에 대한 기존 연구

[6]은 정보검색 (Information Retrieval)과 기계 학습(Machine Learning)[8] 분야에서의 기술들인 Vector Space Model, K-nearest neighbor, Naïve Bayes Classifier를 사용해서 전자 카탈로그들을 UNSPSC 분류체계로 분류하고 그 결과를 비교한다. 그리고 그 결과로 Naïve Bayes Classifier를 사용했을 때의 분류 정확도가 가장 높다는 것을 제시한다.

[9]가 제시한 XML 문서를 분류하기 위한 Naïve Bayes Classifier는 속성과 키워드(속성값)을 연결하는 방법으로 속성별 분포를 활용하지만, 카탈로그 속성의 가중치 부여를 통한 효과를 얻지 못한다. [10]은 구조적 멀티미디어 문서의 분류를 위해 Bayesian Network를 이용해 속성별 분포를 활용하지만, [9]와 같은 문제가 있다. [11]에서는 속성의 정규화, 가중치 부여와 같은 효과를 splitting-stacking이라는 방법을 통해 제공하지만, 학습과 분류 모두 두 단계(level-1 classifier, meta classifier)를 거쳐야 한다는 점에서 큰 복잡도를 갖고 있다.

표 1은 기존 연구를 참고하여 분류 알고리즘들에 대해 학습속도, 분류 속도, 문자열 데이타 적용성, Dimension에 대한 Scalability, 분류에 대한 인간지식 활용 가능성의 측면에서 비교한 것이다. 분류 정확도는 데이타 도메인에 따라 다를 수 있으므로 비교 대상에서 제외하였고, 인간 지식 활용 가능성은 알고리즘의 변형을 뜻하므로 알고리즘의 변형 용이성을 비교대상으로 한다.

학습속도 측면에서는 100만 건 이상의 대용량 전자카탈로그를 학습해야 하고, 카탈로그 정보를 변경하는 경우에 거의 실시간으로 학습을 수행해야 하므로 학습속도가 너무 느린 Neural Networks를 활용하는 분류방법은 현장에 적용하기 어렵다. 분류 속도 측면에서 카탈로그의 분류를 실시간으로 추천해야 하므로 모든 방법론에서 튜닝 문제가 존재한다. 문자열 데이타 적용 측면에서 Decision Tree는 카탈로그 정보에 적용하기 어렵다. Dimension에 대한 Scalability 측면에서는 다양한

표 1 기존 분류 알고리즘 비교

	학습 속도	분류 속도	문자열 데이타에 적용	Dimension에 대한 Scalability	인간지식 활용 가능성
Support Vector Machine	보통	좋음	보통	나쁨	나쁨
Neural Networks	나쁨	좋음	보통	나쁨	나쁨
Decision Tree	보통	좋음	나쁨	나쁨	좋음
Bayesian Network	좋음	보통	좋음	보통	보통

상품들의 분류에 적용하기 위해서는 Baysian Network 방법이 가장 쉽게 여러 Dimension에 활용할 수 있는 방법이다. 인간지식 활용성 측면에서, 카탈로그의 분류에 관한 전문지식을 잘 활용한 분류 학습 및 추천 방법은 Decision Tree 방법이고, 비교적 적용이 용이한 방법은 Bayesian Network방법이다. 이러한 5가지 측면을 고려하면, 상대적으로 현장에 구축하고 적용하기 용이한 방법은 Bayesian Network 방법이다.

2.3 전자 카탈로그 자동분류 모델

우리는 전자카탈로그의 특성을 고려한 전자 카탈로그 자동분류 모델에 대한 연구[1,12-14]를 진행하였다. [1]에 자세한 모델 수립 과정이 나와 있으므로 본 논문에서는 간단한 모델 소개만을 하겠다.

Naïve Bayes Classifier는 분류들의 집합 C와 속성들의 집합 $<a_1, a_2, \cdots, a_n>$, 그리고 분류 대상 문서들의 속성값들 $<v_1, v_2, \cdots, v_n>$이 주어졌을 때, 식 (1)에 의해서 가장 높은 사후 확률을 갖는 분류를 선택한다.

$$C_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_i P(a_i = v_i \mid c_j) \quad (1)$$

$$C_{NB} = \arg\max_{c_j \in C} \{P(c_j) \prod_i (\prod_k P(t_{ik} \text{ appears in } a_i \mid c_j))^{\frac{w_i}{|v_i|}}\}$$

$$= \arg\max_{c_j \in C} \{P(c_j) \prod_i (\prod_k \frac{n(c_j, a_i, t_{ik})}{n(c_j, a_i)})^{\frac{w_i}{|v_i|}}\} \quad (2)$$

[1]에서 설명하고 있는 바와 같이 전자 카탈로그에 적용하기 위해 식 (1)을 확장 수정하면 수정된 전자 카탈로그 분류기는 식 (2)와 같다. 식 (2)는 복잡한 속성의 값들을 매칭하기 위한 속성의 확장(속성별 분포와 속성값의 어휘 추출 활용)하는 단계, 긴 속성값이 분류에 미치는 영향을 정규화하는 단계, 각 속성별로 상대적 중요도에 따라 가중치를 주는 속성별 가중치 부여의 단계를 거쳐 확장한 수식이다. 식 (2)에서 t_{ik}는 v_i를 파싱하여 추출한 어휘이다. 즉 $v_i=\{t_{i1}, t_{i2}, \cdots, t_{im}\}$이다. $n(C_j, a_i, t_{ik})$는 t_{ik} 어휘가 C_j 분류의 a_i 속성에서 나타난 횟수이고, $n(C_j, a_i)$는 C_j 분류의 a_i 속성에서 나타난 모든 추출 어휘의 총 합계 빈도수이다.

또한, 전자 카탈로그의 특성을 고려한 카탈로그의 분류명 활용과 개별속성의 활용방법에 대한 자세한 설명은 [1]에서 자세히 설명하고 있다.

2.4 전자 카탈로그 자동분류기

그림 1은 이전 절에서 설명한 전자 카탈로그 자동 분류기 시스템의 간단한 프로세스 개념도이다. 먼저, 전자 카탈로그 분류 정보를 가지고 있는 학습 데이타를 대상으로 분류 학습을 수행하여 카탈로그 분류기를 구축(분류학습 단계)한다. 카탈로그 분류기를 구축한 후에는 신규 카탈로그 등록 관리 시에 신규 정보가 어떤 카탈로그 분류로 분류되어야 하는지를 추천(분류추천 단계) 한다.

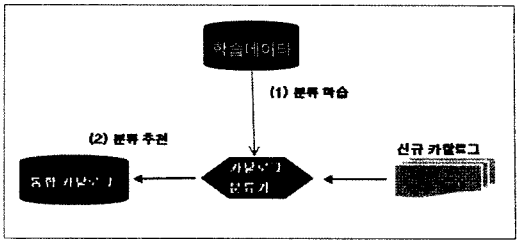

그림 1 전자카탈로그 자동분류기

전자 카탈로그 자동분류기의 분류학습 단계에서는 학습 데이타에 속하는 전자 카탈로그들을 이용해서 두 개의 데이타베이스 테이블 Frequency, Total Frequency 테이블을 만든다. Frequency 테이블의 값은 이전 절에서 설명한 식 (2)에서 $n(C_j, a_i, t_{ik})$에 해당하고, Total Frequency 테이블의 값은 $n(C_j, a_i)$에 해당하는 값을 저장한다.

그림 2는 학습 결과의 예로서 Frequency, Total Frequency 테이블의 예제 값을 보여준다. 그림 2에서 Class_Code는 분류의 번호이고, Attr은 속성번호, Term은 카탈로그의 속성값에서 추출한 어휘, Freq는 해당 어휘의 빈도수 값이다.

전자 카탈로그 자동분류기의 분류추천 단계에서는 분류학습 단계에서 얻은 학습 결과를 활용하여 식 (2)에 해당하는 사후 확률값을 계산하고, top-k에 해당하는 분류를 추천한다. [1]에서 분류 추천을 위한 알고리즘 및 구현방법과 해당 실험결과를 제시하였다.

<Frequency Table>

Class_Code	Attr	Term	Freq
43172410	1	lcd	2
43172410	1	syncmaster	1
43172410	1	100	1
43172410	1	flatron	1
43172410	1	171	1
43172410	2	lcd	2
43172410	2	monitors	2
43172410	3	samsung	1
43172410	3	lg	1
43172410	4	high	1
43172410	4	...	...
43171801	1	xnote	1
43171801	1	lm50	1
43171801	1	dmp2	1
43171801	2	notebook	1
43171801	2	computers	1
43171801	3	lg	1
43171801	4	compact	1
43171801	4	...	...

<Total Frequency Table>

Class_Code	Attr	Total_Freq
43172410	1	6
43172410	2	4
43172410	3	2
43172410	4	8
43171801	1	3
43171801	2	2
43171801	3	1
43171801	4	4

그림 2 카탈로그 자동분류기 학습결과 예

3. 자동분류기를 활용한 전자 카탈로그 관리 프로세스 개선 및 시스템 구축 이슈

3.1 기존의 카탈로그 등록 관리

기존에 수행하던 전자구매 시스템에서의 카탈로그 관리는 수작업에 의해 카탈로그를 분류하는 프로세스를 가지고 있다. 해당 프로세스는 그림 3과 같다.

신규 카탈로그 등록요청은 단건 요청(개별요청)이나 일괄요청으로 이루어지며, 카탈로그 담당자는 요청된 정보의 카탈로그 분류를 수작업에 의해 검색하여 결정하고 해당 카탈로그 정보 표준화 및 추가 정보 입력/보완 작업을 수행한 후 카탈로그로 등록한다. 이때 담당자는 가지고 있는 지식과 더불어 요청 카탈로그 어휘 중에서 카탈로그 분류명과 분류명 유의어를 검색하여 해당 분류를 찾는다.

대기업의 구매부서, 구매대행업무를 하는 기업들 및 조달청에서는 10여명에서 50여명 정도의 인력이 이러한 카탈로그 등록 관리 업무를 담당하고 있다. 몇 년의 경력을 가진 업무 담당자들의 경우에도 특정 분류에 해당하는 카탈로그 정보에 대해서는 어렵지 않게 분류를 결정할 수 있었던 반면, 여전히 어떤 분류가 맞는지 모호한 경우가 많다고 하였으며, 업무의 특성상 전문직이라기 보다는 업무전환이 많아 전문성을 갖추기 힘들다. 카탈로그 분류 작업이 복잡한 것은 많은 업체나 기관들이 활용하고 있는 UNSPSC 등의 분류체계 복잡성(UNSPSC는 분류의 수가 10,000여개 정도임)도 하나의 이유가 되고 있었지만, 이는 전자상거래 및 전자 구매업무의 증가 및 시스템 활용 증대에 따라 피할 수 없는 일이 되었다.

따라서, 카탈로그 등록 관리를 위해 자동분류기를 활용하는 것은 기업의 업무생산성 향상과 기업정보시스템의 기반 데이타 정확성 확보를 위해 매우 중요하다.

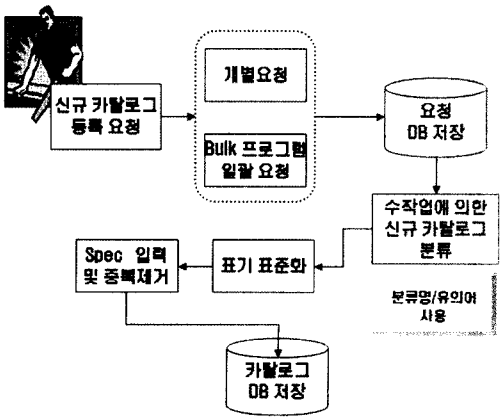

그림 3 카탈로그 등록관리 프로세스

3.2 자동분류기를 이용한 카탈로그 등록 관리

그림 4는 카탈로그 자동분류기를 활용한 카탈로그 등록관리 프로세스를 보여준다. 이 프로세스는 우리가 기업과 조달청에서 전자구매시스템의 카탈로그 등록관리 프로세스 개선 사업을 수행했던 대부분의 사업에서 수행했던 프로세스를 간단하게 정리한 것이다.

자동분류기를 이용한 카탈로그 등록관리에서는 신규 카탈로그 등록 요청 시에 먼저 자동분류기가 요청 정보를 이용하여 분류를 추천하고(그림 4의 1에 해당), 추천된 분류에 따라 해당 분류의 카탈로그 담당자에게 카탈로그 등록 요청 업무가 할당(그림 4의 2에 해당)된다. 수작업에 의해 처리하던 분류작업이 자동분류기에 의해 처리되고 또한 카탈로그 관리 담당자들도 분류 별로 등록 관리 업무를 분업화하여 처리한다. 기존에 카탈로그 분류를 알지 못하여 카탈로그 담당자들이 임의의 모든 카탈로그 등록 요청을 건수 별로 분배하여 처리하던 프로세스와는 많이 다르다. 이러한 개선된 프로세스는 카탈로그 등록 관리 업무의 전문화/분업화 및 업무생산성 향상을 가져온다.

표 2는 이러한 자동분류기를 통한 카탈로그 등록관리 업무와 기존 업무를 비교한 것이다. 표 2에서는 자동분류기를 사용한 카탈로그 분류 업무와 사용하지 않을 경우의 업무를 적용 기술, 카탈로그 분류의 정확도, 업무 개선, 향후 시스템의 활용이 증대될 경우의 모습으로 비교한 것이다. 적용기술 측면에서는 수작업에 의한 업무 시에는 분류명에 대한 유의어가 등록 요청 카탈로그 정보에 있을 경우에 exact match와 데이타베이스 관리 시스템에서 제공하는 단순 검색을 활용하여 분류를 찾는다. 이에 반해 자동분류기는 학습을 통해 자동분류기를 구축하며, 카탈로그 정보에 대해 형태소 분석을 통한 어휘 추출을 하고 그 정보를 활용한다. 이때 대용량 카탈로그 정보에 대한 역 인덱스(inverted index)[15] 기술을 사용하여 분류 추천단계에서 활용한다. 정확도 측

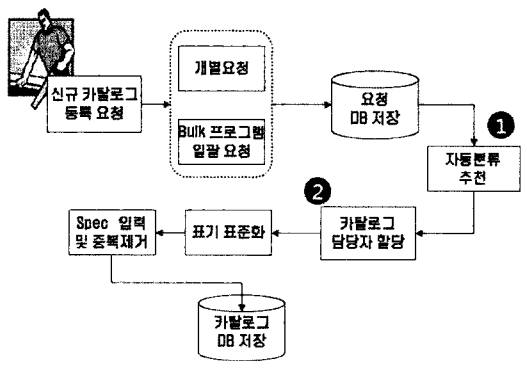

그림 4 자동분류기를 이용한 카탈로그 등록관리

면에서는 조달청과 구매대행업을 하는 기업 A, B사의 시스템 구축사례를 통해 3순위 추천분류까지의 자동분류 정확도이며 기존 방법은 분류명과 유의어를 활용한 exact match 결과이다. Exact match의 경우는 낮은 정확도로 인해 사람이 추가 검색이나 자료조사 같은 수작업에 의해 정확도를 올려야 한다. 업무개선 측면은 이전 절의 프로세스 설명을 참고하기 바란다. 미래 모습 측면에서, 시스템 활용과 카탈로그가 많아질수록 학습기반의 자동분류기는 그 정확도와 일관성이 향상되는 반면, 수작업에 의한 기존 방법은 정확도/일관성이 저하되므로 카탈로그 분류 업무 부하가 증가하고 품질 유지/향상을 위한 비용과 노력은 더욱 증가하게 된다.

자동분류기를 활용한 카탈로그 등록 관리방법은 이전 절에서 제시한 장점을 가지는 반면, 초기 시스템 구축 비용이 필요하고 학습 대상인 기존 카탈로그 정보가 부정확한 경우에는 부정확한 추천으로 인해 카탈로그 분류작업이 더 복잡해질 위험성도 가지고 있다.

3.3 자동분류기를 활용한 카탈로그 등록관리 시스템 구축

자동분류기를 활용한 카탈로그 등록 관리 시스템의 구축 이슈들은 표 3과 같다. 구축 이슈들은 분류학습, 분류추천 단계와 레거시 시스템 연계 이슈로 분리되어 설명한다.

분류학습 단계에서 첫 번째 이슈는 카탈로그 분류체계이다. 자동분류이므로 학습할 분류체계 자체가 정확할 필요가 있으며, 분류별 속성체계와 식별체계 또한 정확할 필요가 있다. 많은 기관이나 기업이 세계적으로 많이 사용하는 UNSPSC를 기반으로 하고 있어서 큰 문제는 없으나, 카탈로그를 정비하지 않고 오랫동안 사용해오던 기업들은 현 상황에 맞는 새로운 분류체계를 새로 개발하고 카탈로그(또는 자재 마스터)들을 표준화하는 작업들을 진행하는 경우도 많다(표준화 작업에 따른 자동분류기 활용은 다음 장에서 소개한다).

분류학습 단계의 두 번째 이슈는 기존 카탈로그의 정확도이다. 자동분류기의 학습 대상이 되는 데이타가 기존 카탈로그 정보이므로, 기존 카탈로그 정보가 심하게 부정확하거나 데이타가 부족한 경우에는 자동분류기의 성능이 저하될 수 있는 요인이 된다.

표 3 자동분류기를 활용한 카탈로그 등록 관리 시스템 구축 이슈

구분	구축 이슈
분류학습 단계	- 카탈로그 분류체계 - 기존 카탈로그의 정확도 - 학습 대상 데이타 선정 - 어휘 사전
분류추천 단계	- 분류 정확도 및 실험 - 추천속도 및 개별추천/일괄추천 프로세스 우선순위
레거시 시스템 연계	- 기준정보 동기화(분류체계, 카탈로그) - 대용량 학습 작업(부분학습지원, 단계적 학습)

분류 학습단계의 세 번째 이슈는 학습 대상 데이타 선정이다. 학습 대상 데이타 선정은 카탈로그 정확도 이슈와 직결되는 이슈이며 학습 단계의 핵심적인 이슈이다. 학습 대상 데이타 선정을 위해서는 아래의 단계를 거친다.

(1) 신뢰할 수 있는 학습 대상 데이타의 범위 선정

기본적으로 전체 기존 카탈로그 데이타가 학습 대상 데이타가 된다. 하지만 자동분류기의 정확도를 높이기 위해, 카탈로그 등록 관리 및 표준화를 수행했던 좀 더 정확한 최근 데이타의 범위(등록일 기준 등)를 정할 수 있다면 해당 기간의 데이타를 학습 대상 데이타로 선정한다. 이때, 학습 대상 데이타의 커버리지(coverage)를 고려하여야 한다. 즉, 학습할 분류의 수가 100개인데, 학습 대상 데이타는 이중 30개 분류에 대해서만 존재한다면(나머지 70개 분류에 대해서는 분류추천을 할 수 없으므로) 이 또한 자동분류기를 제대로 활용할 수 없는 이유가 된다. 우리가 사업을 수행했던 기업들 중 몇몇 기업은 자동분류기를 도입

표 2 자동분류기를 활용한 카탈로그 분류/등록 관리 업무와 기존 업무 방법 비교

	자동분류기 활용	기존 방법
적용 기술	- 인공지능 학습기법 - 자연어 형태소 분석기술 - 대용량 카탈로그 인덱싱 기술	- 분류명 유의어를 이용한 exact match - DBMS가 지원하는 단순검색
정확도	80~95%	30% 이상 기대 어려움
업무 개선	- 개별/일괄 카탈로그 등록 프로세스 반자동화 - 카탈로그 담당자 배정 자동화 - 정확성/신뢰에 의한 업무 부담감소 /생산성향상 - 구매프로세스 lead time 감소	- 낮은 정확도/ 신뢰도로 인한 업무 개선에 한계 - 지속적인 분류명 유의어 관리 부담 - 낮은 정확도와 느린 속도로 인한 분류 확인작업의 반복적 업무 부담
미래 모습	- 학습기법이므로 카탈로그가 많아질수록 정확도/일관성 향상 - 일관적 분류추천으로 카탈로그 정보 표준화에 기여 - 구매, 비용분석, 협력사 평가관리 등의 업무 근간이 되는 기준정보의 품질 향상	- 카탈로그 많아질수록 정확도/일관성 저하되므로 품질유지/향상을 위한 비용과 노력 증가 - 카탈로그 분류 업무 부하 증가 및 업무 생산성 저하

하기 전에 카탈로그 정비 작업을 별도로 수행하고 시스템을 도입하였고 몇몇 기업은 사업 수행과 동시에 일부 카탈로그에 대해 정비를 수행하면서 학습 대상 데이타로 활용하였다.

(2) 비신뢰 학습 대상 데이타의 제거

앞에서 선정한 학습 대상데이타 범위와 더불어, 신뢰할 수 없는 예외 학습 대상 데이타를 제거하는 작업이 필요하다. 대표적인 예는 '기타' 분류이다. 기업에 따라 상황은 달랐지만, '기타'라는 분류가 존재했으며 해당 분류로 인해 자동분류의 성능이 저하되는 경향이 있다. 따라서, '기타'분류에 속하는 카탈로그를 학습할지 말지를 사전에 검토해야 한다. 또한, 일회성 업무를 위한 카탈로그 등록으로 인해 정확하게 분류를 정하지 않는 예외적인 경우에 대해서도 학습대상에서 제외할 필요가 있다.

분류학습 단계의 네 번째 이슈는 어휘사전이다. 학습 단계에서 카탈로그 정보의 속성값들에서 어휘들을 파싱하고 추출할 때 어휘 사전을 이용한다. 따라서, 어휘사전에 등록된 어휘인지 아닌 지에 따라 어휘 추출 결과가 달라지고 분류추천 정확도가 달라진다. 우리는 국립국어원[16] 국어대사전의 말뭉치사전과 더불어, 사업 경험을 통해 자체적으로 어휘를 추가 보완한 어휘사전을 사용하였다.

분류추천 단계에서 첫 번째 구축 이슈는 분류 정확도와 그 실험이다. 자동분류기를 도입했을 때 가장 중요한 것 중의 하나가 분류 추천 정확도이므로 시스템 커스터마이징 작업에서 가장 많은 시간과 노력이 들어가는 작업이 이 부분이며, 이 부분은 학습 단계와 밀접하게 연관된다. 분류 추천 정확도는 1순위 추천 분류의 정확도부터 5순위 분류 추천 정확도를 측정하고 통상적으로 3순위까지의 분류 추천 정확도가 80% 이상 되는 것을 목표로 사업을 진행하였다. 정확도 실험을 위한 실험데이타는 레거시 시스템으로부터 2가지 경우의 데이타를 추출한다. 즉, 레거시 시스템의 기존 카탈로그 데이타의 일부를 실험데이타 집합으로 추출하는 경우와 레거시 시스템의 카탈로그 등록 요청 데이타를 실험 데이타 집합으로 추출하는 경우이다. 기존 카탈로그 데이타로부터 실험 데이타를 추출하는 경우는 이미 분류정보가 무엇인지 정해져 있으므로 바로 정확도를 측정할 수 있다. 그러나 등록 요청 데이타로부터 추출하여 실험하는 경우에는 실제 상황처럼 실험을 할 수 있지만, 요청 카탈로그가 실제 어떤 분류로 등록되어야 하는지 데이타를 구해서 실험데이타로 사용해야 한다. 더불어 실험 데이타 집합의 선택 시에는 분류 커버리지를 고려하여 선택하여야 한다. 학습 대상 데이타 선정과 마찬가지로 적절한 분류 커버리지를 고려하지 않으면 실험 결과를 신뢰하기 어렵다.

분류추천 단계의 두 번째 구축 이슈는 추천 속도이다. 건별 추천이 2-3초 이내에 수행되도록 하는 튜닝 작업과 더불어, 개별 분류추천과 일괄 추천 작업의 프로세스 우선순위 조정 작업이 필요하다. 일괄 분류 추천 작업이 야간에만 이루어지는 경우에는 별 상관이 없지만, 주간 업무 시간 중에도 이루어지는 경우에는 개별 분류 추천의 속도가 영향을 받는다. 따라서, 개별 분류 추천 프로세스의 우선순위를 높여 일괄 추천 작업이 진행 중인 경우에도 개별 추천 작업 속도가 너무 저하되지 않도록 해야 한다.

레거시 시스템 연계에서의 첫 번째 이슈는 기준정보(분류체계정보, 카탈로그 정보)의 연계 및 동기화이다. 분류체계의 변경이 자주 있는 트랜잭션이 아니고, 신규 카탈로그 등록정보가 자동분류기에 실시간으로 학습되지 않아도 업무에는 큰 문제가 되지 않으므로, 기준정보 연계는 보통 일배치(daily batch) 형태로 동기화하도록 구축한다.

레거시 시스템 연계에서 두 번째 이슈는 대용량 학습 작업이다. 초기 시스템 구축 시에 기존 카탈로그 대용량 학습을 하거나 또는 자동분류기 시스템 운영 중간에 기존 카탈로그를 재학습할 필요가 있을 때, 대용량 카탈로그의 학습은 자동분류기 시스템의 부하뿐만 아니라 레거시 시스템에도 부하를 주게 된다(100만 건의 카탈로그를 분류 학습하는데 보통 8시간 정도 소요). 따라서, 대용량 학습 시에 임시 데이타베이스 공간에 기존 카탈로그 정보를 복사한 후, 학습을 수행하는 것이 좋다. 더불어, 자동분류기의 학습 기능에는 특정 분류만 학습하는 부분 학습 기능을 제공하여 부분 재학습과 단계적으로 카탈로그 분류 학습이 가능하게 한다.

4. 자동분류기를 활용한 전자 카탈로그 표준화

4.1 대용량 전자 카탈로그 표준화/정비 작업

3.3절에서 언급한 바와 같이 오랫동안 전자 카탈로그(또는 자재 마스터)를 정비없이 사용하던 기업이나 기관은 현 상황에 맞게 분류체계를 정비 및 카탈로그 정보 정비 작업을 수행한다. 또한 오프라인으로 관리하던 카탈로그 정보를 정보시스템으로 구축하는 경우 등에도 카탈로그 및 분류체계 정비/표준화 작업을 수행한다.

이러한 카탈로그 정비작업은 기업의 상황에 따라 정도의 차이는 있겠지만, 백만 건 내외의 대용량 카탈로그 정보를 표준화/정비해야 한다. 그림 5는 일반적으로 수행하는 분류체계 정비 및 대용량 전자 카탈로그 정비 프로세스를 보여주고 있다. 먼저 카탈로그 표준화를 위한 신 카탈로그 분류체계를 수립한다. 그 다음 수작업에 의해 카탈로그의 분류를 결정하고 해당 분류별 카탈로

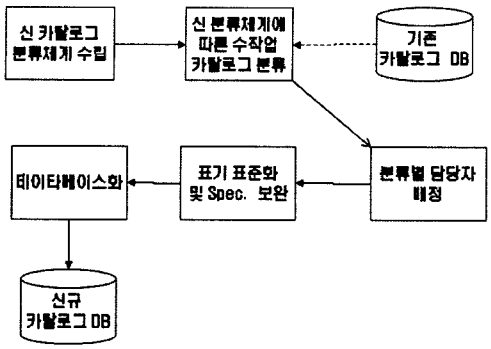

그림 5 대용량 전자 카탈로그 표준화/정비 프로세스

그 담당자가(분류별 담당자가 없이 담당자들이 카탈로그 건수에 따라 업무를 분배 받아 표준화 정비하는 경우도 있음) 해당 카탈로그를 표준화/정비한 후에 데이타베이스화하는 표준화 작업 프로세스를 보여주고 있다.

새로운 카탈로그 분류체계를 수립하고 카탈로그 표준화/정비를 하든 또는 기존 카탈로그를 정확하게 하고 정보를 풍부하게 하기 위해 표준화/정비를 하든 카탈로그를 올바른 분류에 할당하고 해당 카탈로그 정보의 값을 보완/정비하는 것이 중요하다. 기업에서는 카탈로그 분류를 수작업으로 진행하기 위해 10~50여명 이상의 현업 인력이 투입되고 그 작업 기간 또한 만만치 않다. 건설/설비 사업을 하는 A 기업의 경우, 새로운 자재 분류체계를 수립한 후에 400만 여건의 기존 자재 마스터 데이타를 표준화하기 위해 카탈로그 분류 및 표준화 정비 사업을 2년 동안 2단계 사업으로 추진하고 있다.

다음 절에서는 앞에서 설명한 표준화 프로세스를 자동분류기를 이용하여 개선하는 방안에 대해서 설명한다.

4.2 자동분류기를 활용한 대용량 전자 카탈로그 표준화

4.1절에서 설명한 대용량 카탈로그 표준화 프로세스는 자동분류기를 활용할 경우 그림 6과 같은 프로세스로 개선할 수 있다.

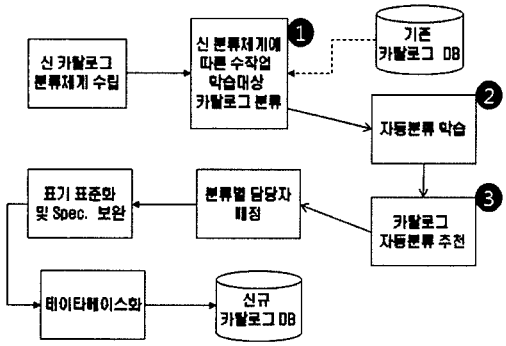

그림 6 자동분류기를 활용한 대용량 전자 카탈로그 표준화/정비 프로세스

자동분류기를 활용하는 프로세스에서도 신 분류체계에 따른 수작업 카탈로그 분류가 필요하다. 하지만, 이 경우에는 학습 대상 데이타만이 필요하므로 수작업 분류작업은 전체 분류 카탈로그 중에서 20~30%의 데이타(통상적으로 학습 대상 데이타는 전체의 30%정도)에 대해서만 수행한다(그림 6의 1작업). 학습 대상 데이타의 생성 이후에는 자동분류 학습작업을 수행하게 되고(그림 6의 2), 이 작업은 몇 시간 이내에 이루어진다. 학습 이후에는 나머지 카탈로그 분류를 위해 자동분류 추천을 수행하고(그림 6의 3), 그에 따라 자동으로 분류별 담당자가 할당된다. 이후 프로세스는 자동분류기를 활용하지 않는 경우와 같다. 자동분류기를 활용하지 않는 경우에는 최초 분류를 결정하기 위해서 임의의 사용자가 임의의 카탈로그에 대해 분류 결정작업을 하고 난 후에야 분류별 담당자가 표준화를 수행했지만, 자동분류기를 활용하는 경우에는 자동분류기에 의해 담당자 할당이 되므로 사용자의 업무 분업화가 더 잘 이루어진다.

더불어, 수작업에 의한 분류작업은 100건의 카탈로그에 대해 5시간 내외의 시간이 걸리지만(기업의 카탈로그 담당자와의 인터뷰를 통한 추정치임), 자동분류기를 활용할 경우에는 100건을 분류하는데 1분 40초 내외의 시간이 소요된다. 따라서 자동분류기를 활용하는 경우 업무 속도로만 보더라도 업무생산성은 100배 이상 향상된다. 그 이외에도 자동분류기를 활용한 대용량 카탈로그 표준화는 표 2의 자동분류기를 활용한 업무 비교에서 언급한 장점들을 가진다. 반면에 자동분류기를 활용한 표준화/정비 업무 개선은 특정 분류의 카탈로그 수가 적거나 신규 상품이어서 정보가 부정확한 경우에는 활용하기 어려우므로 기존 방식대로 수작업에 의한 작업이 필요하다.

자동분류기를 활용한 카탈로그 표준화를 위한 구축 이슈는 신 분류체계를 고려하여 기존 카탈로그에서 학습대상을 선택하는 것이므로 3장의 분류학습 단계 구축 이슈와 크게 다르지 않다. 분류 추천단계는 3장과 구축 이슈가 동일하고, 레거시 시스템 연계이슈도 신 분류체계를 연계한다는 점 외에는 유사하므로 자세한 설명은 생략한다.

5. 결 론

본 논문에서는 전자카탈로그 자동분류를 위한 Naïve Bayes Classifier자동분류기에 대해서 소개하고, 자동분류기를 활용한 기업의 업무 프로세스 개선과 그 구축이슈에 대해서 설명하고 있다.

전자 카탈로그 자동분류를 위한 여러 분류 기법들이 있지만 비교를 통해 Naïve Bayes Classifier가 전자 카탈로그 분류에 적절함을 설명하였다. 또한, 자동분류기

를 활용한 전자 카탈로그 등록 관리 프로세스 개선 방법과 그에 따른 학습단계, 분류추천단계 및 레거시 시스템 연계를 위한 구축 이슈를 설명하였다. 전자 카탈로그 표준화 작업에서도 자동분류기를 활용한 프로세스 개선 방법을 제시하였고 그에 따른 장점과 구축이슈를 제시하였다.

향후 연구로는 어휘 사전 관리를 통해 신조어 등을 추가/삭제 가능한 점진적으로 진화하는 자동분류기 시스템에 대한 연구 등 자동분류기의 정확도를 향상시키기 위한 다양한 연구가 필요하다.

참 고 문 헌

[1] Y. Kim, T. Lee, J. Chun, and S. Lee, "Modified Naïve Bayes Classifier for E-Catalog Classification," DEECS 2006, LNCS 4055, pp. 246-257, 2006.

[2] UNSPSC(The United nations Standard Products and Services Code), http://www.unspsc.org/.

[3] eCl@ss, http://www.eclass-online.com/.

[4] 조달청, http://www.g2b.go.kr.

[5] T. Lee, I. Lee, S. K. Lee, S. Lee, D. Kim, J. Chun, H. Lee, and J. Shim, "Building an Operational Product System," Electronic Commerce Research and Applications, Vol.5/ 1, pp.16-28, 2006.

[6] Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klein, E. Schulten, and D. Fensel, "Golden Bullet: Automated Classification of Product Data in E-Commerce," Business Information System 2002, 2002.

[7] AutoClass, Zycus Inc., http://www.zycus.com/products/autoclass.html

[8] T. M. Mitchell, "Machine Learning," McGraw-Hill International Ed., 1997.

[9] J. Yi and N. Sundaresan, "A classifier for semi-structured documents," 6th ACM SIGKDD, pp. 340-344, 2000.

[10] L. Denoyer and P. Gallinari, "Bayesian Network Model for Semi-structured Document," Classification Information Processing & Management, Vol.40/5, pp.807-827, Elsevier, 2004.

[11] A. Bratko and B. Filipic, "Exploiting Structural Information for Semi-structured Document Categorization," Information Processing & Management, Vol.42/3, pp.679-694, Elsevier, 2006.

[12] 김기륭, "전자카탈로그 자동분류기에 대한 연구", 서울대학교 석사학위 논문, 2003.

[13] 김현철, 이익훈, 이상구, "분류체계 버전정보를 이용한 확장 자동분류 모델", KDBC 2004.

[14] 서광훈, 이경종, 김현철, 이태희, 이상구, "Naïve-Bayesian Classifier를 이용한 전자 카탈로그 자동분류 시스템", 춘계정보과학회, 2004.

[15] R. Ramakrishnan and J. Gehrke, Database Management Systems, 3rd Edition, McGraw-Hill, 2003.

[16] 국립국어원, http://www.korean.go.kr/.

이 익 훈

1996년 서울시립대학교 전산통계학과 학사. 1998년 서울시립대학교 전산통계학과 석사. 2005년 서울대학교 전기전자컴퓨터공학부 공학박사. 2002년 미국 Georgetown Univ. ISIS Center, Researcher. 2003년~현재 주식회사 프람트 기술연구소 연구소장. 2008년~현재 서울대학교 컴퓨터연구소 객원연구원. 관심분야는 데이타베이스, e-비즈니스 시스템, Enterprise SW, 전자상거래, 임베디드시스템

전 종 훈

1995년~현재 명지대학교 컴퓨터공학과 교수. 2001년~현재 (주)프람트 대표이사 사장. 2004년~2005년 University of Denver 방문연구원. 1992년~1995년 University of Central Oklahoma 조교수. 1986년~1992년 Northwestern University 컴퓨터공학과 석사, 박사. 1982년~1986년 University of Denver 전산과학과 학사. 관심분야는 전자상거래 솔루션, 데이타베이스, 의료정보 시스템