Naïve-Bayesian Classifier 를 이용한 전자 카탈로그 자동 분류 시스템

서광훈°, 이경종, 김현철, 이태희, 이상구
서울대학교 전기컴퓨터공학부
{longago°, kjlee, jjanggu, thlee, sglee}@europa.snu.ac.kr

Extending Naïve Bayesian Classifier for Catalog Classification Systems

Kwang-hun Seo°, Kyung-jong Lee, Hyun-chul Kim, Taehee Lee, Sang-goo Lee
School of Computer Science & Engineering, Seoul National University

요 약

B2B Market Place 상에서의 거래에서 나타나는 주요한 특징은 다품종 및 대량의 물품 거래가 *n:n* 거래 관계에 놓여있다는 점과 거래자가 원활한 거래 및 기업 내 관리를 위해 각자의 전자 카탈로그를 이용한 거래를 원한다는 점이다. 하지만 개별적인 전자 카탈로그 사용과 미흡한 표준안은 전자 카탈로그 상호 연계의 걸림돌이 되어 시장 형성의 걸림돌이 되고 있다. B2B Market Place는 표준 분류체계를 중심으로 거래 대상 상품을 재 분류하여 구매 당사자 간의 거래 대상 물품에 대한 상호 매핑을 지원하는 방법 등으로 이를 충족시키려 하고 있다. 하지만 요청되는 다량의 물품에 대해 매번 분류를 수행해야 하는 고비용의 작업이라는 문제점이 있다. 본 논문에서는 이를 극복하기 위하여 기계학습 기법을 이용한 전자 카탈로그 상품 자동분류기를 모델링하고 이를 구현하는 것에 초점을 두었다. 상품의 속성별로 분류에 끼치는 영향력이 다를 것이라는데 착안하여 전자 카탈로그를 상품 단위로 재 모델링 하였으며 속성별 정보가 풍부하지 못한 점을 극복하기 위하여 속성값을 어휘 단위로 구분한 데이터를 추가 하는 확장 모델을 정의하였다. 또한 해당 모델을 학습시키기 위한 알고리즘으로는 속성별로 다른 가중치를 부여 할 수 있도록 확장된 Naïve Bayesian Classifier를 고안하였다. 그리고 이를 B2B Market Place 상의 실 데이터에 적용하여 고안된 모델의 유효성을 검증하였다.

1. 서 론

인터넷의 발전과 함께 전자 상거래에 참여하는 대상이 큰 폭으로 늘고 있다. 특히 B2B Market Place 의 등장은 전자 상거래가 단일 품목 혹은 업종 간의 단순 거래가 아닌 다 업종에서 다양한 상품 정보를 교환하고 거래 프로세스를 진행한다는 점에서 전자 상거래의 양적, 질적 성장을 예고 하고 있다.

일반적인 B2C, B2B 거래와 달리 B2B Market Place의 경우 1:1 혹은 1:*n* 거래 관계가 아닌 다품종, 다량의 물품 거래가 *n:n* 거래 관계에 놓여있다[1]. 거래는 주로 기업의 물품을 조달하거나 판매를 담당하는 부서에 의해 요청되며 이들은 개별적인 상품분류체계를 지닌 경우가 많다. 또한 이들은 다량의 물품에 대한 원활한 거래 및 기업 내 관리를 위해 각자의 전자 카탈로그(분류체계를 포함한 상품정보)를 이용한 거래를 원한다. 하지만 개별적인 전자 카탈로그 사용과 미흡한 표준안은 전자 카탈로그 상호 연계의 걸림돌이 되어 전자 상거래 시장 형성의 걸림돌이 되고 있다.

B2B Market Place는 이를 극복하기 위해 전자 카탈로그간의 상호 매핑을 지원하는 거래 시스템을 하나의 목표로 하고 있다. 이를 위해 UNSPSC와 같은 표준 분류체계를 중심으로 요청자의 거래 대상 상품을 다시 분류하여 당사자 간의 거래 대상 물품에 대한 상호 매핑을 지원하는 방법 등을 사용하고 있다. 이를 통해 거래 당사자간의 욕구를 충족시키는 동시에 자동화된 거래 프로세스를 구축할 수 있는 효과를 얻을 수 있다. 하지만 요청되는 개개의 물품에 대해 매번 분류를 수행해야 하며 이러한 분류는 전문가에 수작업에 의해서 대부분 이루어지고 있어 고비용을 요구하는 작업으로 성장의 걸림돌이 되고 있다. 따라서 이를 자동화 하기 위한 방안 모색이 필요하다.

본 논문에서는 이러한 문제점을 극복하기 위하여 기계학습 기법을 이용한 전자 카탈로그 상품 자동분류기를 모델링하고 이를 구현하는 것에 초점을 두고 다음과 같은 연구를 진행하였다.

우선 기계 학습의 적용을 위해 전자 카탈로그를 상품 단위로 재 모델링 하였으며 여기에서 전자 카탈로그의 정보가 상품의 주요 속성 정보로 이루어져 있는 것에 착안 하여 기계 학습에 있어 속성별 정보를 이용하는 방안을 고려해 보았다. 또한 속성별 정보가 풍부하지 못한 점을 극복하기 위하여 속성값을 어휘 단위로 구분한 데이터를 추가 하는 확장 모델을 정의하였다.

해당 모델을 학습시키기 위한 알고리즘으로는 문서 분류기 연구[2]와 관련된 선행 연구[3][4]에서 좋은 성과를 보인 Naïve Bayesian Classifier를 채택하였으며, 속성별로 해당 상품 분류에 끼치는 영향력이 다를 것이라는 점에 착안 하여 속성별로 다른 가중치를 부여 할 수 있도록 확장된 모델을 고안하였다. 그리고 이를 의 실 데이터를 이용하여 고안된 모델의 유효성을 검증하였다.

이어지는 2장에서는 관련 연구 사례를, 3장에서는 데이터 및 분류기 모델의 정의를, 4장에서는 실험을 다룬다.

2. 관련 연구

본 연구와 근사한 선행 연구로는 전자 카탈로그에 관한 연구[3][4][5]가 있다.

[3]의 연구에서는 다양한 기계학습 기법(k-Nearest Neighbor,

Vector Space Model, Naïve Bayesian Classifier)을 이용하여 상품 분류 표준안인 UNSPSC로의 상품 분류를 수행하였으며 이들간의 성능 평가를 도출하고 있으며 Naïve Bayesian Classifier 에 대해 높은 평가를 주고 있다.

[4]의 연구에서는 두 카탈로그간의 통합문제를 다루고 있으며 Naïve Bayesian Classifier를 확장하여 적용하고 있으며 [5]의 연구에서는 자동 분류를 위한 전자 카탈로그 데이터 모델을 정립하고 이를 Neural Network 을 이용한 기계 학습을 통해 구현하고 있다.

본 연구에서는 상품 정보를 모두 단일한 것으로 보고 있는 선행 연구와 달리 속성별로 해당 상품의 분류에 끼치는 영향력이 틀리 다는 점에 착안하여 속성을 구분한 데이터 모델의 사용과 속성별 가중치를 부여하였다는 점에서 차별성을 지니고 있다.

3. 전자 카탈로그 상품 자동 분류기 모델

3.1. 기본 Data Model

전자 카탈로그는 상품 개개의 정보와 상품에 대한 분류체계 정보를 지닌다. 여기에서는 기 분류된 전자 카탈로그의 정보를 이용하여 기계학습 시킨 다음 학습된 정보를 통해 새로 유입된 상품 정보를 기존의 전자 카탈로그에 맞추어 분류하는 것을 목적으로 하고 있으며 편의상 분류 체계 정보에 존재하는 계층구조 등의 정보는 다루지 않기로 하였으며 한 상품은 단일한 Class에만 속한다고 가정하였다. 따라서 UNSPSC, HS 와 같은 특정 상품 분류체계와 무관하게 분류체계 정보를 상품에 대한 단순한 하나의 '상품 분류 코드'라는 속성으로 취급하기로 하고 학습 시 분류를 구분하는 인식자(identifier) 역할에만 한정하기로 하였다. 이에 따라 전자 카탈로그 정보를 다음과 같이 단순화하였다.

Def 1. Data Model

$CATALOGS = \{C_1, C_2, \ldots, C_n\}$

$C_i = \{ P_1, P_2, \ldots, P_n \}$

$P_i = \{ (a, v) \mid a \in Attribute, v \in Value \}$

CATALOGS 는 전체 전자 카탈로그, C_i 는 하나의 상품 분류 Class, P_i 는 단위 상품 정보를 의미하며 단위 상품 정보는 속성(Attribute) 와 속성값(Value) 의 순서쌍의 집합으로 정의된다. 또한 Attribute와 Value에는 특정한 한계가 없는 것으로 정의 하였다.

상품 정보를 단순한 Value 의 나열로 단순화하지 않고 속성-속성값의 순서쌍 집합으로 둔 이유는 속성별로 속성값이 해당 상품의 분류에 끼치는 영향이 다를 것이라는 것에 착안하여 모든 Value 가 동등하지 않다는 가정을 하기 위해서이다. 이는 전자카탈로그가 일반 문서와 달리 속성별로 지니는 속성값의 차이가 크며 각각의 속성이 상품을 분류하는데 기여하는 정도가 다를 것이라는 점에서 착안하였다. 예를 들면 '제조사'의 경우 동일 업체가 다양한 분류의 제품을 생산할 수 있으므로 '제품명'과 같은 기여도를 지니지 못하지만 '**전자'와 같은 제조사명을 지니는 경우 '전자제품' 에 관련된 제품일 확률이 통계적으로도 높아 기여도가 없다고 할 수 없다는 점을 들 수 있다.

3.2. 확장된 Data Model

전자 카탈로그가 가지는 속성값은 일반 문서와 같이 풍부한 값을 가지지 못한다. 또한 유사한 데이터라 할지라도 학습된 데이터에 정확하게 일치하는 정보가 없을 경우 해당 상품을 분류하지 못하는 한계를 지닌다. 예를 들면 'XX전자 24인치 TV'와 'XY전자 20인치 TV' 의 항목으로 학습된 경우 'XX 전자 20인치 TV'를 같은 Class로 분류해 내지 못한다. 하지만 예제와 같이 일반적으로 속성값은 어휘 집합 (bag-of-words)으로 구성 된다는 점에 착안하여 학습 시 전체 속성값만을 대상으로 하지 않고 주요 명사 단위로 분리된 정보까지 포함시키면 이러한 문제점을 극복할 수 있을 것이라 생각되었다.

문서 분류에서의 Keyword 와 같이 상품명 등의 속성값들도 개개의 속성을 대표하는 어휘들을 포함한 어휘 집합으로 구성되어 있다. 앞선 예에서 'TV'가 그러한 예이다. 따라서 학습 및 분류 시 속성값 그 자체 이외에도 작은 어휘 단위로 파싱하여 작은 단위의 Keyword도 분류기준으로 포함되게 확장할 필요가 있다. 이를 반영하기 Data Model을 다음과 같이 확장하였다.

Def 2. Extended Product Data Model

$P_i = \{ (a, Voc(v)) \mid a \in Attribute, v \in Value \}$

Voc() 함수는 *v*에서 유효한 어휘만을 추출하는 함수로 어휘집합을 산출한다. Voc() 함수는 다양한 방식의 파싱 함수 적용이 가능하다고 가정하였다.

(예) 속성 *a*: 제품명
속성값 *v* : 'XX 전자 20인치 TV'
단순 단어 파싱
Voc(*v*) → {'XX 전자','20인치','TV' }
명사사전을 이용한 우측 및 긴 단어 중심 파싱
Voc(*v*) → {'XX 전자','20인치 TV','TV' }

3.3. 자동 분류기: Naïve Bayesian Classifier

Naïve Bayesian Classifier는 기계학습 분야에서 많이 응용되어 온 자동분류 알고리즘으로 문서분류기에 대한 연구에서 높은 성능을 보이고 있다[4]. 또한 전자 카탈로그의 기계학습을 통한 자동 분류연구에서 k-Nearest Neighbor, Vector Space Model 과의 비교에서 Naïve Bayesian Classifier가 가장 높은 분류 정확도를 보이고 있으며[2] 카탈로그 Integration 연구에서도 해당 분류기를 확장하여 좋은 성과를 얻은 바 있다[3]. 본 연구에서도 이를 배경으로 Naïve Bayesian Classifier를 기반으로 한 전자 카탈로그 상품 자동 분류기를 구축하였다.

Naïve Bayesian Classifier는 주어진 학습 데이터를 이용하여 다항 생성자를 추정한 다음, 분류 대상을 포함될 확률이 가장 높은 Class로 분류하는 것으로 개별 속성값들이 서로 독립적이라는 가정하에 다음과 같은 식으로 분류를 결정한다.[5]

$$v_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_i P(v_i \mid c_j)$$

c 는 Class, v 는 Value를 의미한다.

어휘를 기반으로 한 문서 분류기의 경우 일반적으로 다음과 같이 방식으로 해당 식을 추정하여 사용한다.

$$P(c_j) \leftarrow \frac{|docs_j|}{|Examples|}, \quad P(v_i \mid c_j) \leftarrow \frac{n_k + 1}{n + |Voca|}$$

*Examples*는 학습대상 문서들의 수, $docs_j$는 클래스 Cj에 포함될 문서들의 수, n_k 는 $docs_j$에서 키워드 v_i가 반복해서 나타난 횟수, n 은 $docs_j$에 포함된 문서들에서 추출된 키워드의 수, *Voca*는 분류를 대표하기 위해 *Examples*에서 선출한 모든 키워드의 수를 의미한다.

기존의 연구 [2][3]에서는 카탈로그에 있는 모든 속성값이 독립이며 분류에 동등한 영향력을 끼칠 것으로 가정하고 위의 방식을 그대로 반영하였다. 하지만, 앞서 이야기 한 바와 같이 전자 카탈로그의 상품 정보는 각각의 속성은 서로 다른 도메인을 지니고 있으며 분류에 끼치는 영향도 상이할 것으로 추정된다. 이는 전자 카탈로그에서의 상품 분류가 가지는 문서 분류와의 주요한 차이점이다. 따라서 본 연구에서는 이러한 차이점에 착안하여 각 속성별로 서로 다른 가중치를 줄 수 있도록 기존의 Naïve Bayesian

Classifier를 다음과 같이 확장하였다. 단, 모든 속성값들이 독립이라는 가정은 같다. 그리고 앞서 정의한 ***Def 2. Extended Product Data Model*** 에 따라 클래스 C_j에 속하는 상품군에서 속성 a_k에 해당하는 속성값 V는 $Voc(V)$로 파싱되어 개개의 속성값에 대한 키워드 v_i로 변경하여 사용한다.

Def 3. Classifier Model

$$v_{NB} = \arg\max_{c_j \in C} \{P(c_j)\prod_i P((a_k, v_i) \mid c_j)\}$$

$$= \arg\max_{c_j \in C} \{\prod_i P(v_i \mid (c_j, a_k)P(a_k)\}$$

a_k 는 해당 v_i가 속한 Attribute를 의미하며 각 Attribute 별 가중치를 부여하기 위해 다음과 같이 구현하였다.

Def 4. Weighted Classifier

$$P(v_i \mid c_j, a_k)P(a_k) \leftarrow \frac{n_{(a_k,i)}+1}{n_{a_k}+|Voca_{a_k}|} \times w_{a_k}$$

n_{a_k} 는 학습 데이타 내 전체 상품에서, 클래스 c_j 에 포함된 상품군의 속성 a_k 에 해당하는 모든 속성값 키워드의 총 개수를 의미한다. $Voca_{a_k}$ 는 속성 a_k 에 대하여 분류를 대표하기 위해 상품에서 선출한 모든 속성값 키워드의 수를 의미하며, $n_{(a_k,i)}$ 는 클래스 c_j 내에서 속성 a_k 에 속한 속성값 키워드 v_i가 의 출현 빈도수 (frequency)를, w_{a_k} 는 속성 a_k 에 대한 개별 가중치를 나타낸다.

4. 실험 및 결과

4.1 실험 환경

학습 및 실험 데이터는 UNSPSC 분류를 따르는 IMK[6]에서 사용하고 있는 실재 상품 데이터 중 UNSPSC의 한 Segment를 선정해 이용하였다. 그리고 어휘 파싱은 우측 및 긴 단어 중심의 파싱을 수행하였고 사전은 기 구축된 상품 사전을 이용하였으며 자연어 명사 사전을 이용한 실험을 추가 하였다. 그리고 해당 속성이 없을 경우에는 매우 낮은 값으로 연산되게 하는 별도 가중치를 추가 하여 연산하였다. 그 외에 분류 정확도의 향상을 위해 자연어 사전을 이용하여 상세한 어휘 파싱을 통해 확장된 데이터 모델로 분류 수행 시 정확성이 높은 어휘(단순 공백으로만 파싱된 어휘)를 가지는 데이터에 대해서는 높은 가중치를 별도로 더 부가하는 실험을 추가 하였다.

4.2 실험 결과 및 평가

< 표 1. 가중치 변화에 따른 결과 정확도 >

	상품명	규격	제조사	추가정보	결과(정확도)
1	1	1	1	1	22%
2	4	1	1	1	75%
3	4	0	1	1	80%

'상품명', '규격', '제조사', '추가정보'는 속성에 해당하며 표 내의 수치는 부여한 가중치에 해당한다. 분류코드 5XXXXXXX 에 해당 하는 기 구축 상품데이타 5600여개의 75%를 학습데이터 25%를 분류 대상 데이터로 이용하고 어휘 사전은 상품 사전 이용하였다.

표 1 은 부여한 가중치에 따른 결과의 정확도를 보여 준다. 실험 1에서는 모든 속성이 같은 가중치를 가졌을 때의 경우로 상품명 속성만을 사용했을 때의 결과 (약 70% 이하)에 비해 나쁜 결과를 보이고 있다. 이는 각각의 정보가 가지는 영향력의 정도가 서로 다른 것을 동일하게 취급할 경우 분류에 악영향을 끼침을 보여준다. (같은 제조사명이 서로 다른 클래스에 반복 되는 경우 등)

단일 속성만을 이용해서 분류했을 때 가장 높은 결과를 보인 상품명 속성에 가중치를 더 부여할 경우 정확도의 향상을 보였다. (실험 2,3)

< 표 2. 실 사용 데이터 대상 실험 및 추가 가중치 부여 결과 >

	상품명	제조사	추가정보	결과(정확도)
1	4	1	1	59%
2	4	1	1	71%
3	10:1	1	1	85%

분류코드 5XXXXXXX 에 해당 하는 기구축 상품 5600여개를 학습데이터로 이용하여 실제 고객의 요청 데이터 5000여건을 분류한 경우이며 실험 2, 3은 어휘 사전은 상품 사전 외에 자연어 명사 사전을 추가 이용한 경우이다. 상품명의 10:1은 단순히 공백에 의해 파싱된 어휘에 10의 가중치를 주고 그 외에 명사 사전의 단어 중심으로 파싱된 어휘에 매칭하는 경우에는 1의 가중치를 주었음을 의미한다

표 2에서는 실재 고객들이 요청한 데이터를 대상으로 한 실험 결과로 자연어 어휘 사전을 통한 상세 파싱 시 정확도 향상(실험 1,2의 비교)과 학습데이터와 정확하게 같은 상품명에 대해 추가적인 가중치를 부여한 경우의 정확도 향상(실험 2,3의 비교)을 보여준다. 이외에 IMK의 데이터 약 21만건을 대상으로 한 실험에서도 유사한 정도의 정확도를 얻었다.

5. 결론

본 연구는 B2B Market Place 상에서의 효율적인 상품분류시스템 구축과 관련하여 다음과 같은 연구를 진행하였다.

- 자동 분류기 도입을 위한 전자 카탈로그 데이터 모델을 정의하고 어휘 확장을 통한 데이터 모델 확장을 꾀하였다.
- 전자 카탈로그의 특성에 따라 Naïve Bayesian Classifier가 속성별로 서로 다른 가중치를 부여 할 수 있도록 확장 하였다.
- 실 데이터를 이용한 실험을 통해 확장된 모델의 유효성을 검증하였다.

그 외에도 유사어 혹은 온톨로지를 통한 전자 카탈로그 어휘 확장, 다른 기계학습 알고리즘의 도입, 카탈로그의 계층구조 고려 등의 앞으로 수행하여야 할 다양한 연구 과제들이 남아 있다.

주요 참고 문헌

1. Dongkyu Kim, Jaebum Kim, and Sang-goo Lee, "Catalog Integration for Electronic Commerce through Category Hierarchy Merging Technique", 12th International Workshop on Research Issues on Data Engineering, Engineering E-Commerce/E-Business Systems, 2002
2. D. Koller and M. Sahami: Hierarchically classifying documents using very few words. In Proceedings of the International Conference on Machine Learning, vol 14, Morgan-Kaufmann, July 1997.
3. Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klein, E. Schulten, and D. Fensel, "GoldenBullet: Automated Classification of Product Data in E-commerce" , Proceedings of Business Information System 2002, 2002
4. R. Agrawal and R. Srikant, "On Integrating Catalogs", The 10th International World Wide Web Conference, 2001
5. 김기룡, "전자카탈로그 자동 분류기에 대한 연구", 서울대학교, 2003.8
6. T. M. Mitchell, "Machine Learning", McGraw-Hill International Ed., 1997
7. 삼성 iMarketKorea (http://www.imarketkorea.co.kr)

-본 연구는 정보통신부의 대학IT연구센터(ITRC) 지원을 받아 수행되었음