전자 카탈로그 자동분류에서 Naïve-Bayesian Classifier 데이터 모델 확장

김성환° 김현철 이태희 이상구
서울대학교 대학원 전기·컴퓨터 공학부
{ringknocker°, hckim, thlee, sglee}@europa.snu.ac.kr

Extending Data Model of Naïve-Bayesian Classifier in e-Catalog Classification

Sung-hwan Kim° , Hyun-chul Kim, Tae-hee Lee, Sang-goo Lee
School of Computer Science & Engineering, Seoul National University

요 약

인터넷 환경에서의 B2B Market Place의 출현은 판매자와 구매자와의 다자간 거래를 가능하게 하였다. 이러한 기반에서 상품정보를 포함하는 전자 카탈로그의 활용은 나날이 증가하고 있다. 그러나 동일한 상품에 대한 분류체계와 기준이 다르므로 전자카탈로그에 대한 재분류는 고비용을 초래하는 필수 불가결한 문제로 남게 되었다.

본 연구에서는 이러한 문제를 해결하기 위해 기계학습 기법을 이용한 Naïve Bayesian classifier 모델을 사용하였다. 학습 데이터를 생성해야 하는 Naïve Bayesian 알고리즘 적용 시 전자 카탈로그는 일반 문서보다 상대적으로 학습 정보가 적으므로 데이터 모델의 확장을 통해 학습 정보를 생성하여 이러한 단점을 보완하였다. 전자 카탈로그 자동분류에 있어서 효과적이고 풍부한 양의 학습 데이터를 생성하는 것이 분류 정확도 향상에 중요한 영향을 미침을 실험을 통해 확인하였다.

1. 서 론

1.1 동 기

보편화된 인터넷 환경에서의 B2B Market Place의 출현은 여러 공급자와 구매자가 Market Place를 통해 서로의 정보를 탐색하고, 정보를 교환하는 다자간의 N 대 N의 상거래를 가능하게 하였다. 이러한 다자간의 거래에 있어서 상품 정보를 포함하는 전자 카탈로그는 그 활용과 중요성이 나날이 증대되고 있다[1]. 그러나 분류체계가 상이하고 엄격한 표준안이 사용되고 있지 않는 문제로 인해 전자 카탈로그의 분류 시 고비용이 초래되어 전자 카탈로그의 폭넓은 사용이 제한을 받고 있다.

이러한 재분류시의 고비용 문제를 해결하기 위해 상이한 분류체계의 전자 카탈로그를 사용자의 요구에 맞게 재분류하기 위한 여러 연구가 진행되고 있다.

본 논문에서는 이러한 문제를 해결하기 위해 [1]에서 제안한 기계학습을 이용한 전자카탈로그 자동분류기를 사용하여 데이터 모델 확장 방법에 따른 상품 분류 정확도를 실험을 통해 확인하였다.

1.2 본문 구성

2장에서는 전자 카탈로그 자동분류시스템에 관한선행 연구를, 3장에서는 전자 카탈로그 자동분류 모델과 확장 데이터 모델을, 4장에서는 실험 및 결과를, 5장에서는 결론 및 향후 연구를 다룬다.

본 논문은 정보 통신부의 대학 IT 연구 센터 ITRC(Information Technology Research Center) 지원을 받아 수행되었음

2. 관련 연구

본 연구와 관련된 선행 연구로는 전자카탈로그에 대한 데이터 모델 및 적용에 대한 연구[2][3]과 기계학습기법을 이용한[4] Naïve Bayesian Classifier를 기반으로 확장된 데이터 모델을 사용하고, 속성별로 서로 다른 가중치를 부여하여 분류 정확도를 향상한 [5][6]연구가 있다. [6]은 [5]의 자동분류 모델을 기반으로 하여 중심 분류체계의 버전정보를 이용한 자동분류 모델을 제안하였다. 본 연구에서는 선행연구에서 사용한 확장된 데이터 모델 구축 시 어휘 확장 방법에 따른 분류 정확도의 변화를 자동분류기를 사용한 실험을 통해 확인해 보았다.

3. 전자 카탈로그 자동분류 모델

3.1. Naïve-Bayesian Classifier

본 연구에서는 문서분류에서 좋은 성능을 내는 Naïve Bayesian을 이용해서 자동분류기를 구현하였다[7].

Naïve Bayesian 분류기는 주어진 학습 데이터를 이용하여 다항 생성자를 추정한 다음, 분류 대상이 포함될 확률이 가장 높은 클래스로 분류하는 것으로 개별 속성

값들이 서로 독립적이라는 가정 하에 다음과 같은 확률 식으로 분류를 결정한다[8].

Bayes' Theorem (1)

$$P_r(C_j|P_x) = \frac{P_r(C_j)P_r(p_x|C_j)}{P_r(p_x)} \quad -①$$

•$P_r(C_j|P_x)$는 상품(p_x)가 상품군(C_j)에 포함될 확률

Bayes' Theorem (2) - ①식에 대한 개별 설명

$$P_r(C_j) = \frac{|C_j|}{|total\ product|}$$

$$P_r(p_x|C_j) = \prod_i P_r(v_i|C_j)$$

$$P_r(v_i|C_j) = \frac{n_k + 1}{n_{c_j} + |voca|}$$

|total product| → Catalog에 속해있는 모든 상품의 수
$|C_j|$ → j번째 상품군에 속해 있는 모든 상품들의 수
|voca|→ Catalogs 에 속해 있는 모든 상품군들의 Value수

다음은 확장 데이터 모델을 이용한
확장 Naïve Bayesian 분류기 모델이다[8].

Def 1. Extending Classifier Data Model

$$V_{NB} = \{\arg \max P_r(C_j)\prod_i P_r((a_k, Voc(v_i))|C_j)\ \}$$

다음은 가중치 정보를 포함하는
확장 Naïve Bayesian 분류기 모델이다[5][6].

Def 2. Weighted Classifier

$$P_r((a_k, Voc(v_i)\ |\ C_j) = \frac{n_{(a_k,l)} + 1}{n_{a_k} + |voca_{a_k}|} \times w_k$$

3.2. 기본 데이터 모델

전자 카탈로그 자동분류 모델에서 제시하는 기본 데이터 모델은 아래와 같이 크게 전자 카탈로그, 상품군, 상품으로 정의한다. 이는 [5][6]에서 사용한 기본 데이터 모델과 동일하며 속성과 속성 값의 쌍으로 상품정보를 표현하였다.

Def 3. Basic Data Model

$Catalogs\ \{C_1, C_2, ..., C_n\}$

$C_j = \{P_1, P_2, ..., P_m\}$

$P_k = \{(a,v)\ |\ a \in Attribute,\ v \in Value\ \}$

Catalogs → 전체 전자 카탈로그,
C_j → 하나의 상품분류 Class, P_k→ 단위 상품 정보

단위 상품 정보는 속성(Attribute)과 속성 값(Value)의 순서쌍의 집합으로 정의 된다. 또한 Attribute와 Value에는 특정한 한계가 없는 것으로 정의하였다.

3.3. 확장 데이터 모델

3.3.1 확장 데이터 모델

전자 카탈로그는 분류를 위한 정보가 문서 분류와 비교해 보았을 때 풍부하지 않다[5][9]. 전자 카탈로그 분류는 상대적으로 적은 정보를 가지고 분류를 해야 하기 때문에 유사한 데이터라 할지라도 학습된 데이터에 정확하게 일치하지 않으면 분류를 하지 못하는 한계를 가지고 있다.
예를 들면 '양가죽', '소가죽' 의 항목으로 학습된 경우 '염소가죽'을 같은 Class로 분류해 내지 못한다. 하지만 속성 값을 어휘 집합(bag-of-words)으로 구성하여 학습 시 전체 속성 값만을 대상으로 하지 않고 주요 명사 단위로 분리된 정보까지 포함시켜서 학습 및 분류 속성 값 그 자체 이외에도 작은 어휘 단위로 파싱하여 작은 단위 Keyword도 분류기준으로 포함되게 확장하였다.
이를 반영하기 위해 Data Model을 다음과 같이 확장하였다[5].

Def 4. Extending Product Data Model

$$P_k = \{(a, Voc(v))\ |\ a \in Attribute,\ v \in Value\ \}$$

상품 (P_k)는 하나의 상품군(C_j)에 속하는 k번째 상품이다. 속성(a)는 상품(P_k)에 속하는 속성(attribute)을 나타낸다. 어휘 셋(Voc(v))은 상품 (P_k)의 상품명을 어휘사전과 동의어 사전을 통해서 파싱 한 어휘 셋을 나타낸다.
(예) 양가죽 = {(단위 파싱, 양), (단위 파싱, 가죽), (전체 파싱, 양가죽)}

3.3.2. 데이터 모델 확장 방법

본 연구에서는 데이터 모델 확장을 위해 다음의 두 가지 방법을 사용한다. 첫 번째는 '명사 사전을 이용한 단순 명사 파싱'이고, 두 번째는 '명사 사전을 이용한 우측 및 긴 단어 중심 파싱'이다.
먼저 '명사 사전을 이용한 단순 파싱'을 살펴보면 입력한 상품 어휘 명을 2음절에서 전체 어휘 음절수까지의 연속된 음절로 분해한 후 2음절에서부터 12음절 사이의 명사사전과 비교하여 사전과 일치하는 어휘가 발생 시 파싱을 하는 것을 말한다. 다음으로 우측 및 긴 단어 중심 파싱은 명사사전과 일치하는 파싱 발생 시 좌측을 기준으로 동일 부분에서 발생하는 파싱 어휘 중에서 가장 긴 단어를 파싱하고 가장 긴 단어 바로 다음의 우측으로 넘어가서 파싱을 하는 것을 말한다. 이는 보다 긴 음절의 파싱 단어가 의미적으로 유효하다고 판단하고 파싱을 하는 것을 말한다. 두 방법 모두 입력 어휘의 Full Name은 전체 파싱어휘로 처리하며 1음절 명사어휘의 비교는 하지 않는다.
두 가지 방법에 대한 예를 들면 다음과 같다.

■ 데이터 모델 확장 예

속성 : 부분 파싱, 전체 파싱
속성 값 : '**항공기연료**'

○ 명사 사전을 이용한 단순 명사 파싱
- 어휘 분해(○:파싱발생,×파싱 발생 없음)
  [항공(○),공기(○),기연(×),연료(○),
  항공기(○), 공기연(×), 기연료(×)
  항공기연(×), 공기연료(×),
  항공기연료(○, 전체 파싱)]
- 파싱어휘 → {항공, 공기, 항공기, 연료,
  항공기연료}
- 파싱 어휘 수 : 5개

○ 명사 사전을 이용한 우측 및 긴 단어 중심 파싱
  [항공(×,2음절 부분 파싱) < 항공기(○,3음절 부분 파싱), 연료(○), 항공기연료(○, 전체 파싱)]
- 파싱어휘 → {항공기, 연료, 항공기연료}
- 파싱 어휘 수 : 3개

위 예에서 볼 수 있듯이 명사 사전을 이용한 단순 명사 파싱에서 생성되는 파싱 어휘수가 우측 및 긴 단어 중심 파싱에서 생성되는 어휘 수보다 더 많다. 또한 단순 명사 파싱에서 생성되는 어휘는 우측 및 긴 단어 파싱에서 생성되는 어휘를 포함하는 것을 확인할 수 있다. 단순 명사 파싱은 가능한 한 많은 파싱을 수행함으로써 의미적 관계가 적은 단순 명사도 어휘 셋에 포함 시켰다. 위의 두 가지 데이터 모델 확장 방법에 따른 분류 정확도의 값을 자동분류기를 통한 실험을 통해서 비교 확인하였다.

4. 실험 및 결과

4.1. 실험 환경

학습 및 실험 데이터는 UNSPSC Code v7.0401의 Segment를 이용 하였으며 속성 값은 상품명만 사용하였다. 사전은 기 구축된 상품사전을 이용하였다. 속성은 해당 어휘의 전체파싱(가중치 10), 부분파싱(가중치 2) 두 가지만 구분하였다.(가중치는 ***Def 2*** 의 w_k를 의미한다)

4.2 실험 결과 및 평가

<표 1. 데이터 모델 확장 방법에 따른 분류 정확도>

확 장 방 법	분류 정확도
단순 단어 파싱	88.22%
우측 및 긴 단어 중심 파싱	82.21%
• UNSPSC의 Segment를 임의 추출하여 1751개의 상품 데이터 중 75%를 학습 데이터로 25%를 분류 대상 데이터로 이용하고 어휘 사전은 상품 사전을 이용하였다	

표 1.에서 볼 수 있듯이 우측 및 긴 단어 중심 파싱보다 단순 단어 파싱을 사용한 확장 모델에서 분류정확도가 더 높게 나타났다.
이를 통해 단순 단어 파싱으로 생성된 단순 어휘가 분류 정확도 향상의 주요 원인으로 작용하였다는 것을 확인할 수 있다.

5. 결론 및 향후 연구

본 연구에서는 학습데이터를 통한 전자 카탈로그 재분류 시 카탈로그에서 얻을 수 있는 최소한의 정보(상품명)만을 가지고 이를 확장 데이터 모델을 통해 분류 정보를 생성하여 재분류를 하였다.
문서 분류와 비교하여 상대적으로 적은 정보를 가지고 분류를 해야 하는 Naïve-Bayesian Classifier를 이용한 전자 카탈로그 분류에 있어서 어휘 확장은 학습정보 부족 문제를 해결하는 중요한 방법이 되었다. 기존 연구에서는 어휘 확장 시 유효 어휘만을 파싱 하였으나 본 연구에서는 유효 어휘 확장 시 파싱되는 모든 어휘들을 포함하는 단순 단어 파싱과 단순 어휘를 가급적 포함하지 않는 우측 및 긴 단어 파싱을 비교 실험하여 단순 파싱 시 분류 정확도가 더 높게 나타나는 것을 확인함으로써 어휘 확장 시 포함되는 단순 어휘가 분류정확도를 향상시킨다는 것을 확인하였다. 향후 어휘 확장 시 단순 파싱어휘 포함비율에 따른 분류정확도 변화와 의미적 관계를 유지하면서 학습정보를 더욱 확장시킬 수 있는 방안에 대한 연구가 필요하겠다.

6. 참고 문헌

[1] http://www.nso.go.kr 대한민국 통계청
[2] Authur Keller, "Smart catalogs and virtual catalogs",Computer Science Department, Stanford University, 1995
[3] Sherif Danish, "Building Database-driven Electronic Catalogs", SIGMOD Record, Vol.27, No4.1998
[4] T. M. Mitchell, "Machine Learning", McGraw-Hill International Ed,1997
[5] 서광훈, 이경종, 김현철, 이태희, 이상구 "Naïve-Bayesian Classifier를 이용한 전자 카탈로그 자동분류 시스템",춘계 정보 과학회, 2004
[6] 김현철, 이익훈, 이상구 "분류체계 버전정보를 이용한 확장 자동분류 모델", KDBC, 2004
[7] Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Kle in, E. Schulten, and D. Fensel, "Golden Bullet in a Nutshell", the 15th International FLAIRS Conference, 2002
[8] R. Agrawal and R. Srikant, "On Integrating Catalogs", The 10th International World Wide Web Conference, 2001
[9] 김기룡, "전자카탈로그 자동 분류기에 대한 연구", 서울대학교 석사학위 논문, 2003