온톨로지 기반 웹 문서 분류

송무희° 임수연 민도식 강동진* 이상조
경북대학교 컴퓨터공학과, 경북대학교 정보전산원*
mhsong@knu.ac.kr° {nadalsy, dosik79}@hotmail.com {djkang*, sjlee}@knu.ac.kr

Ontology-Based Document Classification

Mu-hee Song° Soo-yeon Lim Do-sik Min Dong-jin Kang* Sang-jo Lee
{Dept. of Computer Engineering, Information Technology Services*} Kyungpook National University

요 약

본 논문에서는 웹 문서들이 가지는 용어 정보들과 어휘들의 의미구조를 계층적 형태로 표현한 온톨로지 기반 자동 문서분류 방법을 제안한다. 문서 분류는 문서들을 가장 잘 표현할 수 있는 자질들을 정하고 이러한 자질들을 통해 미리 정의된 2개 이상의 카테고리에 문서의 내용을 파악하여 가장 관련이 있는 카테고리로 할당하는 것이다. 본 논문에서는 웹 문서에서 추출한 용어 정보들의 유사도와 온톨로지 카테고리의 유사도를 계산하여 웹 문서를 분류하며, 문서 분류를 위한 실험데이터나 학습과정 없이 바로 실시간으로 문서분류가 이루어지며, 결과적으로 문서들이 가지는 고유한 의미와 관계의 식별을 통하여 보다 더 정확하게 문서분류를 가능하게 해준다.

1. 서 론

급속도로 발전하는 인터넷의 사용증가 추세에 맞추어 웹상에서 볼 수 있는 전자문서의 양은 엄청나게 증가하고 있다. 이러한 전자문서가 양적으로 크게 늘어남에 따라 사람이 수많은 정보를 일일이 분류하는 것은 매우 힘들어 졌다. 이에 따라 문서를 알맞게 정해진 카테고리로 분류하는 것을 도와주는 도구에 대한 필요성이 점차 커지고 있다.

이러한 문제의 해결책으로 본 논문에서는 온톨로지를 이용한 자동 문서분류 방법을 제안하고자 한다. 일반적으로 웹 문서들은 다음과 같은 특징들을 가지게 된다. 첫째, 웹사이트를 단위로 하여 구성된다. 즉 웹사이트는 하나의 주제를 담고 있는 여러 웹 문서들로 이루어지며, 둘째, 각 웹 문서는 어떤 웹사이트의 일부분으로 존재한다. 마지막으로 일반적인 웹사이트는 특정주제와 관련 있는 개인이나 단체가 운영한다. 따라서 본 논문에서는 웹 문서들의 이러한 특징들로 인하여 용어 정보들을 추출하기가 용이하다고 간주한다. 본 논문에서는 웹 문서들이 가지는 용어 정보들과 단어들의 의미구조를 계층적 형태로 표현한 온톨로지를 바탕으로 유사도(similarity)를 계산하여 웹 문서를 분류하게 되며, 결과적으로 문서들이 가지는 의미론적 내용과 관계의 식별을 바탕으로 보다 더 정확하게 문서분류를 가능하게 해준다. 본 논문에서 적용되는 온톨로지는 개념(concept)과 개념에 대한 특징(feature), 개념간의 관계(relation) 그리고 문서 분류를 위한 제약조건(contraint)들이 계층적으로 이루어지며, 이러한 온톨로지의 계층적인 구조를 문서 분류에 적용하는 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 통하여 문서분류와 온톨로지에 대한 배경지식을 알아보고, 3장에서는 본 논문에서 제안한 온톨로지를 이용한 문서분류에 대해 살펴보며, 마지막으로 4장에서는 결론 및 향후 연구방향을 제시한다.

2. 관련연구 및 배경지식

많은 양의 문서를 효율적으로 관리, 검색하기 위한 문서 분류 모델에 관한 연구는 이미 오래 전부터 계속되어 왔다. 이 장에서는 문서분류와 관련된 기존 연구와 본 논문의 근간을 이루는 온톨로지에 대해서 알아본다.

2.1 관련연구

문서 분류는 문서들을 가장 잘 표현할 수 있는 자질들을 정하고 이러한 자질들을 통해 미리 정의된 2개 이상의 카테고리에 문서의 내용을 파악하여 가장 관련이 있는 카테고리로 할당하는 것이다. 문서 분류를 위한 대표적인 모델로는 크게 학습 문서들에서 나타나는 범주간의 구별된 규칙을 이용하여 전문가가 찾아주거나 학습을 통해 추출된 규칙을 이용하여 문서를 분류하는 규칙기반 모델[1], 학습문서에서 자질을 추출하여 이를 확률적인 접근방법으로 사용한 베이지언 확률 모델[2], 기계학습 방법을 이용한 지지벡터기반(Support Vector Machine:SVM)[3], 그리고 정보 검색 관점에서 분류할 문서를 질의로 보고 이와 유사한 문서를 찾는 방법인 K-최근린법[4] 등이 있다. 그러나 이러한 방법들은 문서분류의 정확도를 어느 정도 보장하지만 그림 1에서와 같이 미리 규칙을 위한 학습과정이 필요하며, 그에 따른

학습데이터가 반드시 필요하다[2].

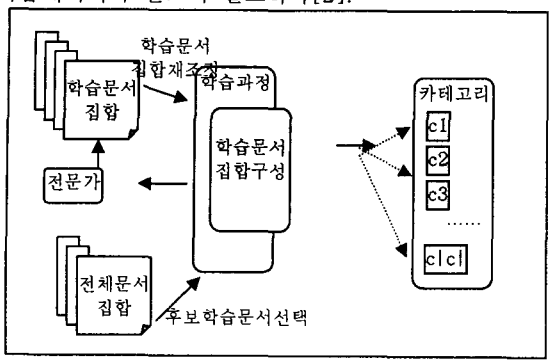

그림 1 기존 문서분류 시스템

본 논문에서는 이러한 과정 없이 바로 실시간으로 문서분류를 할 수 있는 온톨로지 기반 자동문서분류 방법을 제안한다.

2.2 온톨로지

최근 온톨로지 기술은 시멘틱 웹을 구현하기 위한 중요한 요소로서 웹을 기반으로 한 온톨로지에 대한 연구는 그 응용 범위가 갈수록 확산되고 있다. 온톨로지도 넓은 의미에서는 데이터베이스라고 할 수 있지만 데이터보다는 복잡한 형태의 지식과 관련되어 있다는 의미에서는 지식베이스라고 부르기도 한다. 하지만 엄격히 말해 온톨로지는 지식내용과 절차적 추론과정을 포함하는 포괄적 의미의 지식보다는 용어 사이의 '개념'적 관계에 국한되어 있기 때문에 지식베이스와 구별되는 또 다른 형태의 데이터베이스라고 할 수 있다. 즉, 온톨로지는 특정 영역에서 공통적으로 사용되는 어휘들의 집합을 개념적으로 표현하는 방법이다[5].

3. 온톨로지를 이용한 문서분류

3.1 본 논문에서의 온톨로지 구조

본 논문에서는 온톨로지를 "어휘들에 대해서 일정영역의 개념적 예들을 한 곳으로 집합시킨 하나의 독립된 집합체"로 정의하기로 한다. 물론 여기에는 단순한 어휘들의 집합이 아니라 간단한 규칙들과 의미적 연관관계를 가진 단어들의 집합을 의미한다. 온톨로지는 어휘의 정의를 다른 어휘와의 논리적 관계뿐만 아니라 가장 기본적(primitive) 어휘부터 파악해 나가는(bottom-out) 구조를 통해 나타낸다. 그래서 본 논문에서는 온톨로지가 가장 기본적인 어휘에서 출발한다는 점에서 온톨로지의 구조를 의미적 계층구조로 보고, 웹 문서의 분류에 적용하기로 한다.

- 온톨로지의 구성요소
  - 개념 : 특정 도메인에서 사용되는 일반화된 용어
  - 특징 : 개체를 설명하기 위해 사용되는 특징,
  - 관계 : 개념들간의 관련 유형을 정의하여 연결시키는 것
  - 제약조건 : 각 개체가 활성화 되기 위해서 필요한 조건

- 온톨로지의 관계유형
  - 동등(E-R) : 개념간의 의미가 동일한 경우
  - 상속(is-a) : class와 instance 가 같은 경우
  - 부분(has-a) : A has a B 로 A가 B를 부분으로 가진경우(has-part, part-of)

3.2 문서 분류를 위한 온톨로지 구축

본 논문에서는 문서분류실험을 위해 "경제" 도메인에 대한 온톨로지를 구축하였다. 온톨로지의 구축 순서는 첫째, 문서집합에서 높은 출현빈도를 가진 단어들은 다른 많은 단어들과 유기적으로 연결되어 있다고 가정한다. 둘째, 이들 단어들을 이용하여 기초적인 네트워크를 구축한다. 셋째, 선택된 단어들과 관련이 있는 단어들을 네트워크에 추가함으로써 온톨로지를 확장해 나간다. 이렇게 구축된 온톨로지와 수집한 웹 문서들에서 용어 정보들을 추출하여 이들간의 유사도를 계산하여 분류가 시작된다.

그림 2는 구축된 온톨로지 구조의 예이다.

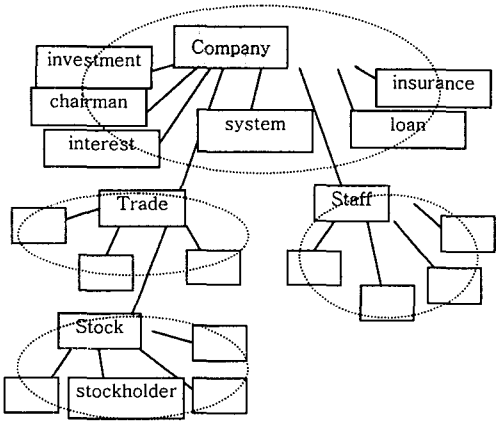

그림 2 온톨로지 구조 예

3.3 온톨로지를 이용한 문서 분류

웹 문서를 분류하는 과정은 문서 안에서 중심이 되는 단어를 찾아내는 과정과 추출된 단어를 이용하여 개념계층(온톨로지)상의 노드에 매핑하는 과정으로 구성된다.

문서에서 단어를 추출하기 위한 방법으로는 전처리 단계로서 불용어제거와 스테밍 처리, 그리고 정보검색 측정치 tf*idf에 기초를 두고 있다. tf*idf는 역문헌 빈도수를 단어의 빈도수와 같이 적용함으로서 그 문서를 대표하는 단어들을 효율적으로 찾아주는 알고리즘이다.

tf(i,j)가 문서 $d_i \in D^*$, i=1,2,3…,N 에서 용어 j의 용어빈도이고, df(j)가 얼마나 많은 문서 용어 j가 나타나는지를 계산하는 용어 j의 문서 빈도일 때 문서에서 용어 j의 tf*idf (term frequency/inverted document frequency)는 다음과 같이 수식(1)로 정의된다.

$$tf * idf(i,j) = tf(i,j) \times \log\left(\frac{N}{df(j)}\right) \qquad (1)$$

tf*idf는 너무 빈번히, 혹은 너무 드물게 나타나는 용어들은 그렇지 않은 용어들 보다 낮게 랭크 되고 그래서 분류 결과에 좋은 영향을 미치게 된다. 용어선택에서 전

처리 과정을 거친 문서 집합으로부터 문서 하나에 포함된 모든 용어 목록을 만든다. 그래서 문서 선택은 W(j)를 최대화하는 용어 j를 선택하고 가장 적절한 용어에 대한 tf*idf값인 tf*idf(i,j) 를 포함하고 있는 문서 d_j 대한 다음 수식(2)와 같은 벡터를 나타낸다.

$$W(j)=\sum_{i=1}^{N} tf * idf(i,j) \qquad (2)$$

분류를 위한 유사도 계산은 수식(3)을 이용하였으며 [6] 문서는 가장 큰 유사도를 가지는 하나의 노드에 할당하게 되므로 한 문서는 최종 하나의 클래스로 분류하게 된다.

$$Sim(Node,d)=\frac{\sum_{i=0}^{N} freq_{i,d}/\max_{l,d}}{N}\times\frac{V_d}{V} \qquad (3)$$

여기서 N은 한 노드에서의 총 특징의 수이며, $freq_{i,d}$는 문서 d에서 매칭되는 특징 i 의 빈도수를, $max_{l,d}$는 문서 d에 의해 가장 많이 매칭되는 특징의 빈도수를 나타낸다. V는 제약조건의 수를, V_d는 문서 d에 의해서 만족되는 제약조건의 수를 말한다. 문서 분류 과정은 관계의 사용이 "is-a", "has-a", "part-of", "has-part" 일 경우에만 일어나며, 다른 노드와 관련이 있을 경우 관련된 노드를 분류과정에서 포함시켜 유사도 계산을 행하게 된다. 이로써 문서의 보다 더 정확한 분류가 가능하게 된다.

전체적인 웹문서 분류 과정은 그림 3와 같다.

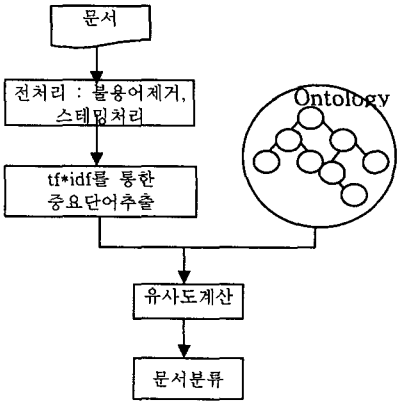

그림 3 온톨로지를 이용한 웹 문서의 분류 과정

문서 분류의 진행 절차는 그림 4에서와 같이 온톨로지의 루트 노드에서부터 시작하여 하위 노드의 방향으로 진행되며, 그 처리과정은 첫째, 추출된 단어에서 자질벡터를 계산하고, 둘째, 계산된 자질벡터에 의해 추출된 단어들의 분류가 이루어지며, 셋째, 분류된 단어들로부터 카테고리 벡터를 계산한다. 그 다음으로 문서에서 추출된 단어들의 자질벡터와 카테고리 벡터의 유사도를 계산하고, 문서는 가장 유사하다고 판단된 노드에 매핑되게 된다. 이 방법을 통해 적은 수의 노드 지식으로 온톨로지를 표현 할 수 있으며 계층적인 분류를 통해 좀 더 정확한 분류가 가능하게 된다.

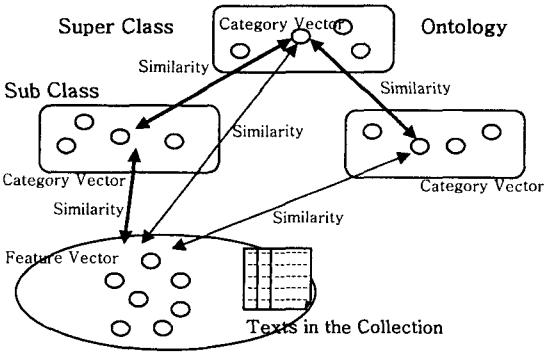

그림 4 온톨로지를 이용한 문서분류 예

4. 결론 및 향후 과제

본 논문에서는 온톨로지를 이용하여 웹 문서들이 가지는 의미적 관계를 개념 구조로 표현하고, 또한 구축된 온톨로지를 이용하여 웹 문서를 분류하는 것이다. 본 논문에서 제안한 부분은 문서에서 추출된 용어 정보를 바탕으로 온톨로지 구조를 분류된 카테고리로 보고, 각 유사도를 계산하여 결과값이 높은 순서대로 정렬하여 문서 분류가 이루어지게 된다. 본 논문에서 제안한 문서 분류 방법은 학습과정, 실험데이터 없이 바로 실시간으로 분류가 이루어진다는 점에서 그 의미가 있다고 할 수 있다.

향후 과제로 본 논문에서 제안된 방법을 검증할 실험 절차와 그에 따른 성능 평가가 뒤따라야 할 것이며, 그리하여 좀 더 효율적인 온톨로지 표현과 문서 분류 방법에 대한 연구가 진행되어야 할 것이다. 또한 본 논문에서 제안한 분류정보를 정보검색에 활용하여 효율을 높이는 방안과 웹 문서에서부터 의미적 개념, 관계를 자동으로 추출하는 방법을 계속 연구, 진행하고자 한다.

참 고 문 헌

[1] Chidanand Apt, Fred Damerau, and Sholom M. Weis, "Towards Language Independent Automated Learning of Text Categorization models," proc. of the 17th annual international ACM-SIGIR, 1994.

[2] 김제욱, 김한준, 이상구, "베이지언 문서분류시스템을 위한 능동적 학습기반의 학습문서집합 구성방법", 2002.12 정보과학회 논문지 제29권 제12호.

[3] Mart A. Hearst, "Support Vector Machines," IEEE Information Systems, 13(4):18~28, 1998.

[4] Yiming Yang and Xin Liu, "A Re-examination of Text Categorization Methods", Proc. Of the 22th annual International ACM-SIGIR, 1999.

[5] T.R Gruber, "Towards Principles for the Design of Ontologies used for Knowledge Sharing," International Journal of Human-Computer Studies, 1995.

[6] 정현섭, 양재영, 최중민, "개인화 된 웹 네비게이션을 위한 온톨로지 기반 추천 에이전트", 2003.2 정보과학회 논문지, 제30권 제1호.