온톨로지의 개념구조에 의한 웹페이지의 의미적 분류

송무희° 임수연 박승배 강동진* 이상조
경북대학교 대학원 컴퓨터공학과, 경북대학교 정보전산원*
mhsong@mail.knu.ac.kr°, nadalsy@hanmail.net, {seongbae, djkang*, sjlee}@mail.knu.ac.kr

Semantic Classification of Web Pages using Ontology Concept Structure

Muhee Song° Sooyeon Lim Seongbae Park Dongjin Kang Sangjo Lee
Daegu, Kyungpook National University

요 약

본 논문에서는 온톨로지의 개념구조를 이용한 웹페이지의 의미적 분류방법을 제안한다. 웹 문서들이 가지는 용어 정보들과 어휘들 간의 개념 구조를 파악하여 온톨로지를 확장시키면서 이를 문서분류에 적용하여 의미적 분류가 이루어지게 한다. 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 온톨로지와 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 온톨로지의 개념구조를 이용한 웹페이지의 의미적 분류방법을 제안한다. 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. 서론에 이어 2장에서는 관련 연구를 통하여 문서분류와 온톨로지에 대한 배경지식을 알아보고, 3장에서는 본 논문에서 제안한 온톨로지의 개념구조를 이용한 웹페이지의 의미적 분류에 대해 살펴보며, 4장에서는 실험 및 평가, 그리고 마지막으로 5장에서는 결론 및 향후 연구 방향을 제시한다.

2. 관련연구

많은 양의 문서를 효율적으로 관리, 검색하기 위한 문서 분류 모델에 관한 연구는 이미 오래 전부터 계속되어 왔다. 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 단점은 문서분류의 정확도를 어느 정도 보장하지만 미리 규칙을 위한 학습과정이 필요하며, 그에 따른 학습 데이터가 반드시 필요하다. 그리고 최근에 들어 온톨로지를 기반으로 하여 웹페이지를 분류하려는 많은 연구가 진행되고 있다. Prabowo[5]는 듀이 십진분류법(DDC)과 미 의회 도서관(LCC)을 고려하여 웹페이지를 분류했는데 이 과정에서 온톨로지를 구축하였다. 이 방법은 분류를 위해 시소러스나 사전을 이용하지 않고 온톨로지를 사용했다는 점과 웹페이지의 텍스트로부터 온톨로지를 구축했다는 점, 그리고 온톨로지와 분류체계의 연관관계를 구성한 점 등이 특징적이라고 할 수 있다. 하지만 이 접근법의 단점은 비록 표준 분류 방법을 따르고 있지만 분류에 대한 사용자의 복잡한 요구가 적절히 반영되기 어렵다는 것이다.

본 논문에서는 개념에 대한 계층적 구조를 갖는 온톨로지를 이용하여 웹페이지의 의미적인 정보를 반영한 분류방법을 제안한다. 본 논문에서 제안하는 방법은 미리 규칙을 위한 학습과정 없이 바로 실시간으로 문서 분류가 이루어지며, 또한 단순히 키워드와 용어정보만을 가지고 분류하는 것이 아니라, 그 의미적 기반에 의하여 결과적으로 문서들이 가지는 고유한 의

미와 관계의 식별을 통해 보다 더 정확하게 문서분류를 가능하게 해준다.

3. 온톨로지의 개념구조를 이용한 문서의 의미적 분류

3.1 본 논문에서의 온톨로지 개념구조

온톨로지는 개념적이고 술어적인 혼란을 감소시키는 것이 목적이다. 이것은 주어진 응용도메인의 특성을 나타내는 관련개념들의 집합과 정의 그리고 그들 간의 관계로 이루어진다.

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.

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[7]. 그림 1은 온톨로지의 구축 과정을 나타낸 것이다.

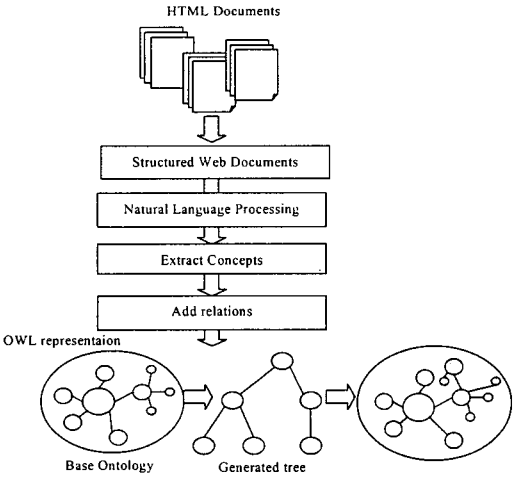

그림 1 온톨로지의 구축과정

3.3 온톨로지의 개념정보를 이용한 웹페이지의 분류

웹 문서를 분류하는 과정은 문서 안에서 중심이 되는 단어를 찾아내는 과정과 추출된 단어를 이용하여 개념 계층(온톨로지) 상의 노드에 매핑 하는 과정으로 구성 된다. 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_jD*, 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.

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

tf*idf는 너무 빈번히, 혹은 너무 드물게 나타나는 용어들은 그렇지 않은 용어들 보다 낮게 랭크되고 그래서 분류 결과에 좋은 영향을 미치게 된다. 용어선택에서 전처리 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*idf 값 인 tf*idf(i,j) 를 포함하고 있는 문서 d_j 대한 다음 수식(2)와 같은 벡터를 나타낸다.

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

분류를 위한 유사도 계산은 수식(3)을 이용하였으며[8] 문서는 가장 큰 유사도를 가지는 하나의 노드에 할당하게 되므로 한 문서는 최종 하나의 클래스로 분류하게 된다.

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

여기서 N은 한 노드에서의 총 특징의 수이며, $freq_{i,d}$는 문서 d에서 매칭되는 특징 i 의 빈도수를, $max_{i,d}$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 2와 같다.

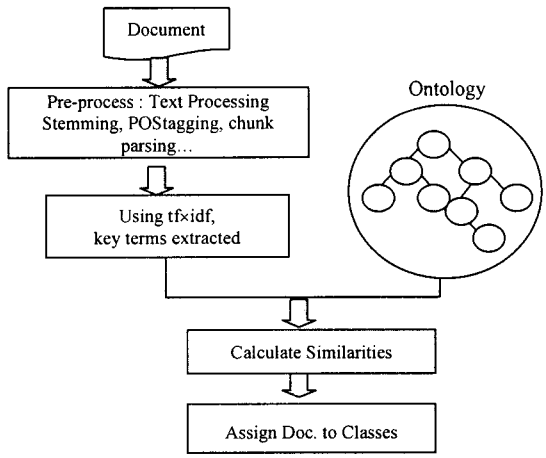

그림 2 Ontology를 이용한 웹페이지 분류

4. 실험 및 평가

문서분류는 문서들의 내용을 파악하여 문서가 속한 정확한 클래스를 정하는 작업으로 사전에 정해져 있는 클래스들 중에서 어떤 클래스에 문서가 속하는지 판단하는 것이다. 본 논문의 실험을 위하여 데이터는 표1에서와 같이 웹상의 디렉토리 서비스를 제공하는 야후검색사이트의 분류클래스에 따른 경제 클래스의 웹문서를 받아서 사용하였다.

표 1 문서분류에 사용된 웹문서 개수

Class no.	Class name	Number of documents
1	Cooperatives	620
2	Employment	1,685
3	Finance	750
4	Marketing	680
5	Organizations	650
6	Trade	850
	Total	5,235

제안하는 시스템의 분류성능을 평가하기 위하여 표준 정보검색 측정법인 정확율, 재현율, F1등을 이용하여 평가되었으며 평가식들은 다음과 같이 정의 된다.

$$\text{정확율} = \frac{a}{a+b}, \tag{4}$$

$$\text{재현율} = \frac{a}{a+c}, \tag{5}$$

$$F1 = \frac{2}{\frac{1}{\text{정확율}} + \frac{1}{\text{재현율}}}, \tag{6}$$

여기에서 a, b, c 값은 표 2에 정의 되어있다. 표3에 보이는 것처럼 시스템 분류와 전문가 판정 사이의 관계가 4가지 값을 사용하여 표현하였다. F1 측정법은 정확도와 재현율에서 일정의 평균치를 나타낸다.

표 2 표3에서 사용된 a,b,c 파라미터들의 정의

값	의미
a	시스템과 전문가 모두 할당된 범주와 일치
b	시스템은 할당된 범주와 불일치하고 전문가는 일치
c	전문가는 할당된 범주와 불일치하고 시스템은 일치
d	시스템과 전문가 모두 할당된 범주와 불일치

표 3 분류정확도를 계산하기 위한 결정 행렬

전문가	시스템	
	Yes	No
Yes	a	b
No	c	d

정확율, 재현율, F1-measure 측정식을 이용한 각각의 결과는 표4에 나타내었다.

표 4 분류 결과

Class no.	정확율(%)	재현율(%)	F1(%)
1	77.21	93.84	84.72
2	92.48	94.16	93.31
3	93.93	95.38	94.65
4	91.17	95.38	93.23
5	91.30	96.92	94.03
6	91.97	96.92	94.38
Average	89.68	95.43	92.39

5. 결론 및 향후과제

본 논문에서는 온톨로지의 개념정보를 이용하여 웹페이지의 의미적 분류방법을 제안한다. 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 온톨로지의 개념구조를 분류된 카테고리로 보고, 각 유사도를 계산하여 결과 값이 높은 순서대로 정렬하여 문서 분류가 이루어지게 된다. 본 논문에서 제안한 문서 분류 방법은 학습과정, 실험데이터 없이 바로 실시간으로 분류가 진행이 되며, 온톨로지의 개념정보를 이용하여 의미적 분류가 이루어진다는 점에서 그 의의가 있다고 할 수 있다. 결과적으로 문서들이 가지는 고유한 의미와 관계의 식별을 통하여 보다 더 정확하게 문서분류를 가능하게 해준다.

향후 과제로는 본 논문에서 제안한 방법과 다른 문서 분류 방법과의 비교검토가 뒤따라야 할 것이며, 좀 더 효율적인 온톨로지 표현으로 웹 페이지에서부터 의미적 개념-관계를 자동으로 추출하는 방법, 웹 정보 추출에서 가장 큰 문제로 떠오르는 다양한 형태의 정보출처와 빈번히 바뀌는 정보원에 대하여 이를 효과적으로 대처하기 위한 연구가 진행되어야 할 것이다.

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 17thannual 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] R.Prabowo, M.Jackson, P.Burden and H.Knoell, "Ontology-Based Automatic Classification for the WEB Pages: Design, Implementation an Evaluation," Proc.of 3rd International Conference, Singapore, pp.182-191, 2002.
[6] T.R Gruber, "Towards Principles for the Design of Ontologies used for Knowledge Sharing," International Journal of Human-Computer Studies, 1995.
[7] 임수연, 송무희, 이상조, "전문용어의 처리에 의한 도메인 온톨로지의 구축", 정보과학회 논문지(B), 제31권 3호, pp.353-360, 2004
[8] 정현섭, 양재영, 최중민, "개인화 된 웹 네비게이션을 위한 온톨로지 기반 추천 에이전트", 2003.2 정보과학회 논문지, 제30권 제1호.