한국산업정보학회논문지 제9권 제3호(2004. 9)

온톨로지 기반의 웹 문서 자동 주제 식별

(Automatic Topic Identification Based on the Ontology for Web Documents)

최 인 대*, 남 인 길**, 부 기 동***
(In-Dae Choi, In-Gil Nam, Ki-Dong Bu)

요 약 본 연구의 목적은 온톨로지 계층구조에 정의된 키워드들 간의 연관성을 참조함으로써 주어진 텍스트의 주제를 식별할 수 있는 방법을 개발하는 것이다. 텍스트의 중요한 문장들로부터 추출된 키워드들은 계층구조에 존재하는 개념들에 사상된다. 모든 단어가 사상되면, 대응되는 개념들은 한 개의 단일 개념으로 일반화 되며, 그 단일 개념이 텍스트의 주제가 된다. 본 연구는 온톨로지와 단어 빈도를 사용해서 신뢰성과 정확도를 향상시키기 위한 지식 베이스와 통계적 접근을 병행한 하이브리드 방식의 접근 방식으로서 성능을 향상시켰다. 실험 결과 제안한 방법이 기존의 지식 베이스만을 사용한 방법보다 성능이 우수함을 보였다.

핵심주제어 : 자동주제식별, 온톨로지

Abstract The goal of this research is to develop a method of identifying a topic of a given text by looking at relationship of keywords defined in an ontology hierarchy. The keywords which are extracted from important sentences of the given text are mapped onto their correspond concepts which exist in the hierarchy. After all the words are mapped, the correspond concepts will be generalized into one single concept. The single concept will most likely be the topic of text. Our research have an approach that promotes both satisfaction in term of robustness and accuracy using ontologies and word frequency. So, this attempts are done in what they call as a hybrid approach. We try to take the challenge by using knowledge-statistical base approach. Experimental results show that proposed method outperforms the existing method using knowledge-base only.

Key Words : Automatic Topic Identification, Ontology

1. 서 론

최근 인터넷의 발달로 정보검색(information retrieval), 질의응답(question answering), 자동주제식별(automatic topic identification) 등의 분야에서 웹 기반의 텍스트 문서 정보처리 방식이 중요하게 다루어지고 있다. 특히 산재된 문서 정보에 대해 필터링 기능을 수행할 수 있는 자동 주제 식별은 텍스트 처리에 대한 비용과 시간을 절약하고, 정보검색과 정보 요약(information summarization) 등의 분야에서 기반 기술로 적용할 수 있으므로 활용의 가치가 높다[1].

텍스트 정보처리 분야에서는 애플리케이션 간에

* 도림거창전문대학, 컴퓨터정보시스템과
** 대구대학교, 정보통신대학 컴퓨터·IT공학부
*** 경일대학교, 컴퓨터공학부

정보를 효율적으로 공유하고 재사용을 용이하게 해주며, 웹 기반의 지식 처리를 가능하게 해 주는 온톨로지(ontology)의 구축 및 활용에 대한 관심이 높아지고 있다[2]. 온톨로지는 개념 모델링(conceptual modeling)을 통해 구조화된 자료로 표현하고 접근할 수 있다는 점에서는 의미 데이터베이스(semantic database)와 유사하다고 할 수 있으며, 온톨로지 구현 언어[3,4,5]의 발전으로 웹상에서 상호 교환이 용이한 전자문서 형태의 의미 데이터베이스 혹은 메타 데이터베이스로 활용할 수 있게 됨에 따라 시멘틱 웹(semantic web)을 구현하기 위한 중요한 수단으로 사용되게 되었다.

자동 주제 식별 분야에서는 온톨로지를 의미 분석을 위한 지식 데이터베이스로 사용함으로써 정확률을 향상시킬 수 있다. 즉, 대상 문서에서 높은 빈도수로 출현하는 중심어들 중에서도 다른 단어와의 의미적인 연관 관계에 의해 그 중요성이 달라질 수 있으며, 이러한 단어간의 의미적 연관성을 검사하는데 온톨로지를 지식 데이터베이스로 활용할 수 있다. 온톨로지의 계층구조는 텍스트에서 추출된 단어 간의 관계를 분석하고 일반화하는 용도로 사용하게 된다.

본 연구에서는 자동 주제 식별에 있어 온톨로지를 구축하고 활용하였으며, 중요 단어 선정과 온톨로지 매핑시 가중치 계산에 있어 빈도수를 검사하는 통계적 방법을 병행하였다. 온톨로지는 국내 검색엔진인 네이버(Naver)[6]의 검색 분류 체계를 이용하여 구축하였다. 네이버는 웹상에서 활용 가능한 주제의 대부분을 갖고 있고 인터넷에서 정보를 쉽게 찾을 수 있도록 도와주는 국재의 가장 유명한 포털 사이트 중의 하나이다. 본 논문에서 제안한 방법은 50개의 뉴스 기사를 대상으로 정확률 측면에서 기존의 방법과 성능을 비교 분석하였다.

2. 자동 주제 식별 개요

2.1 관련 연구

자동 주제 식별 방법은 크게 지식 기반 접근 방법과 통계적 접근 방법, 그리고 두 가지를 병행하는 하이브리드 방식이 있다. 지식 기반 방식은 방대한 지식 데이터베이스를 검색하기 때문에 정확률은 우수하지만 시간 소모가 많다는 단점이 있으며, 통계적 접근 방식은 시간 소모는 적지만 정확도가 떨어진다는 단점이 있다. 이 두 가지 방법 중 장점을 취하는 하이브리드 방식도 가능하지만 처리 과정이 복잡하여 시스템의 구현이 어렵다.

1999년 Sabrina 등[6]은 지식 기반 방식으로서 야후(yahoo) 온톨로지의 매핑에 따른 가중치 기법으로서 제목, 부제목, 문단의 첫 번째 및 두 번째 문장 및 예측 가능한 특정 위치에 나타날 가능이 있는 단어를 가중치를 높게 설정하여 주제를 식별하는 방법을 제안하였다. 즉, 단어를 추출하여 중요 단어 리스트(stop-list)를 만들고 미리 구축된 온톨로지 계층구조에 매핑 함으로서 가중치를 문단의 위치에 따라 차등 부여하는 방법을 사용하였으며 야후라는 검색 분류 체계를 온톨로지로 구축하여 단어 분류 지식 정보로 활용하였다.

그러나 Sabrina 등의 연구에서는 가중치 부여에 있어 문서 내 단어의 위치 정보만을 활용하였기 때문에 정확률이 떨어진다는 단점이 있다. 따라서 본 연구에서는 온톨로지를 이용한 단어 간의 연관성 검사와 빈도수에 의한 통계적 접근 방식을 병행함으로써 정확률을 향상시킬 수 있는 방법을 제안하는 데 주안점을 두었다.

2.2 온톨로지 기반 자동 주제 식별

Sabrina 등의 연구 방법[6]에서는 다음과 같은 절차로 주제를 자동 식별하게 된다. 먼저, 문단 위치에 따른 중요 단어 선정 과정에서 "중요한 단어는 예측 가능한 특정 위치에 나타날 가능성이 높다"라는 일반적 사실과 제목, 부제목, 문단의 첫 번째와 두 번째 문장이 상대적으로 중요하다는 가설하에 이를 통한 중요 단어 리스트를 작성한다. 다음은 온톨로지 매핑을 통한 가중치를 적용하는 단계로서 미리 구축된 야후 온톨로지의 계층 구조에 중요 단어를 매핑 함으로써 문단 위치에 입각한 차별적인 가중치를 부여하고 가중치 전파에 의한 최적 경로를 선정한다.

온톨로지 매핑 과정은 동일한 단어일지라도 다양한 어의를 내포할 가능성이 존재함으로, 동일한 단어에 대해서도 같은 주제에 속하는지를 판단하기 위한 클러스터 식별이 필요하며, 이를 위해 매핑된 단어들의 가중치 누적 연산을 통한 최적 경로를 선정하게 된다.

다음 단계는 최적 경로 상에서 주제를 대표할 수 있는 단일 노드 추출을 위한 일반화 과정으로서 최적 경로에 속하는 개념 노드들의 밀도를 측정하여 가장 높은 비율의 노드를 선택하는 것이다. 이렇게 선택된 하나의 노드가 대상 문서를 대표하는 식별된 주제가 된다.

그러나 Sabrina 등의 연구는 문장의 제목, 부제목, 문단의 첫 번째와 두 번째 문장에 속하는 단어를 중요 단어로 선정하였기 때문에 온톨로지 매핑 시 입력 단어에서 정확률에 대한 오차를 갖게 된다. 아울러 온톨로지 매핑 시 가중치를 부여하는 알고리즘에서도 동일한 방법을 적용하고 있기 때문에 결과적으로 정확률을 저하시키는 원인이 되고 있다.

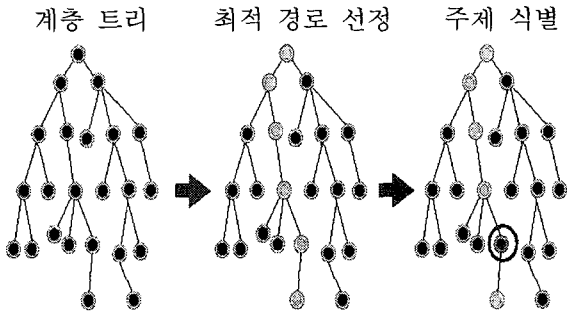
3. 온톨로지 기반 자동 주제 식별 기법의 제안

3.1 제안 개요

본 논문에서 제안한 자동 주제 식별 기법은 기존의 네이버의 문서 분류 체계를 온톨로지로 활용하였으며, 빈도수를 이용한 중요 단어의 추출과 추출한 단어와 온톨로지 노드 간의 매핑을 통한 가중치 부여 시에 빈도수를 반영하여 매핑 된 단어들의 의미적 연관성을 반영함으로써 주제 식별의 정확률을 향상시키는 데 목적이 있다.

온톨로지를 지식 데이터베이스로 활용하는 이유는 단어와 단어 사이의 의미적 연관성을 파악함으로써 정확한 주제를 식별해 내기 위함이다. 예를 들면, "도둑"이라는 단어가 높은 빈도수로 출현해 온톨로지에 매핑 되었을 경우, 주변에 인접한 다른 중요 단어들이 "컴퓨터", "보안", "핵커" 등 컴퓨터 관련 분류 클러스터에서 많은 매치가 일어나면 "도둑" 이라는 단어는 은행 강도 등의 개념 보다는 보안이나 핵커 등 컴퓨터 관련 용어들과 보다 밀접한 연관 관계를 갖고 맺어져 있다고 볼 수 있다. 이 때 클러스터는 온톨로지 계층 구조상에서 부분 트리로 나타나며 연관성 있는 개념들의 집합을 의미한다.

자동 주제 식별은 이러한 단어들의 매핑에 의해서 온톨로지 상에서 중요 클러스터를 식별해 내는 것이 중요한 데,. 이러한 과정은 본 논문에서는 온톨로지 계층 구조상에서 가중치 전파에 위한 최적 경로를 선택하는 문제로 해결하고 있다. 그림 1에서 보는 것처럼 만약 "도둑"이란 용어가 계층 트리 상에서 여러 곳에서 매핑이 일어나게 되면 다른 단어들과의 연관성에 의해 가장 높은 누적 가중치를 갖는 클러스터를 우선 식별해야 한다.

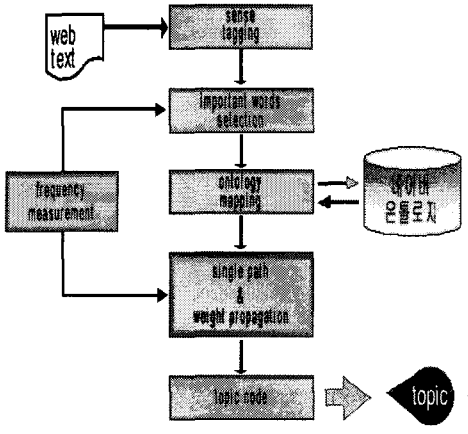


<그림 1> 온톨로지 매핑 절차

이 과정이 그림 1에서 보는 최적 경로의 선정 단계이며, "도둑"이라는 단어가 소속된 클러스터가 "컴퓨터" 관련 분류 클러스터라면 "도둑"은 "강도" 보다는 "컴퓨터" 혹은 "보안"이라는 단어와 보다 밀접한 연관성을 갖게 된다는 것을 의미한다. 따라서 문서의 주제는 가중치가 가장 높은 클러스터 내에 속한다고 보는 것이 타당하며, 이 클러스터 내에서 일반화 과정을 거쳐 선택된 단일 노드가 문서의 주제를 나타내는 개념 노드가 된다.

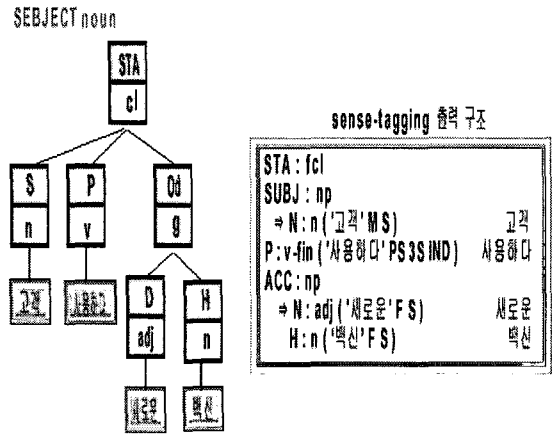
3.2 시스템 구성 및 전처리 과정

본 논문에서 제안한 자동 주제 식별의 처리 과정을 시스템 구성도로 나타내면 그림 2와 같다.



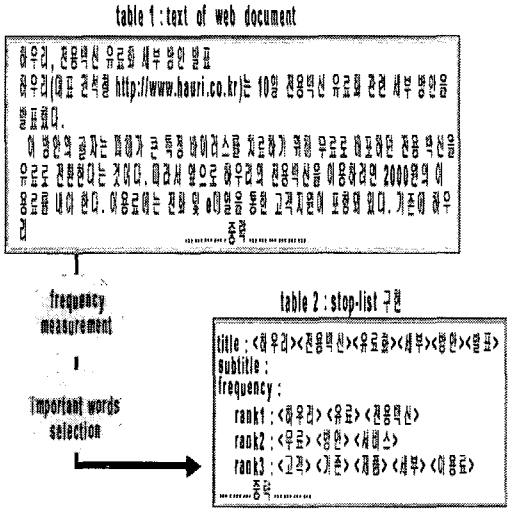
<그림 2> 시스템 구성도

그림에서 보는 바와 같이 주제 식별을 위해서는 먼저 전 처리로서 원문을 파싱하여 분리된 어휘에 대해 품사와 구문 정보를 부가하는 품사 태깅 (sense tagging) 과정을 거쳐야 한다. 본 논문에서 적용한 품사 태깅 파서는 경일대학교 정보검색 연구실에서 개발한 공개 파서를 사용하였으며, 그 파싱 결과는 그림 3과 같은 파스 트리와 구문 정보로 표현 할 수 있다.



<그림 3> 품사 태깅의 결과

전 처리의 다음 단계에서는 품사 태깅의 결과로 얻어진 품사 정보를 이용하여 명사, 형용사, 부사, 동사를 대상으로 빈도수 검사를 통해 순위 내에 포함 되는 단어를 추출하여 중요 단어 리스트를 구성한다.



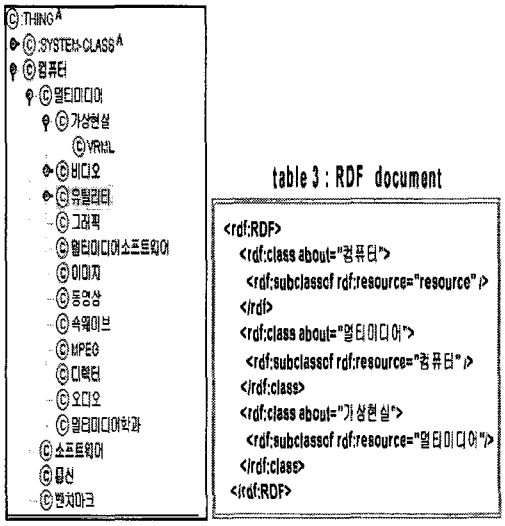
<그림 4> 중요 단어 리스트의 구성 예

본 연구에서는 제목과 부제목에 속하거나 중요 순위 내에 포함되는 명사, 형용사, 동사, 부사를 대상으로 중요 단어 리스트를 구성하였다. 이 과정은 그림 4에서 예로서 보여주는데, 테이블 1은 대상 문으로 주어진 뉴스 기사를 나타내고 테이블 2는 제목, 부제목 그리고 빈도수 검사에서 상위 순위에 속하는 단어를 추출한 중요 단어 리스트의 구성 결과를 나타낸다.

다음은 온톨로지 매핑 단계로서 텍스트에서 추출된 키워드들은 계층구조의 대응되는 개념 노드로 매핑하는 과정이다. 이 때 매핑되는 키워드는 중요도에 따른 대응되는 개념 노드에 가중치를 부여한다. 본 연구에서는 단어의 출현 빈도를 반영하여 가중치를 부여하였다. 다음 단계에는 각 노드에 부여된 가중치를 부모 노드로 전파하면서 가중치를 누적시켜 가장 값이 큰 최적 경로를 찾아내고 마지막 단계에서 일반화 과정을 거쳐 최적 경로에 포함되는 개념 노드들을 한 개의 단일 노드로 일반화 한다.

3.3 온톨로지의 구축

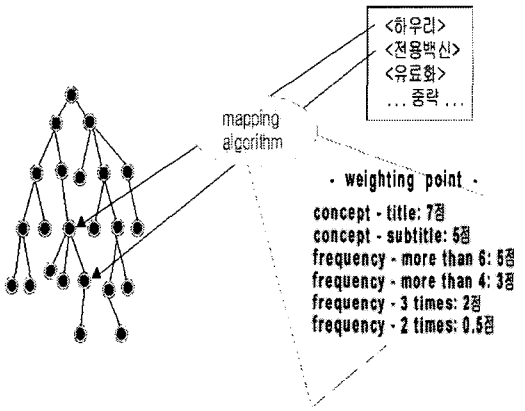
온톨로지 구축을 위한 도구로 널리 사용되는 소프트웨어에는 Amaya, Protégé-2000, Mozilla, SilRI 등이 있다. 본 연구에서는 네이버 온톨로지의 구축을 위해 Stanford 대학에서 개발한 온톨로지 에디터인 Protégé-2000 [6]을 사용하였다.



<그림 5> Protégé-2000에서 RDF 생성의 예

그림 5는 Protégé-2000으로 온톨로지를 구축하는 과정을 예로서 보여준다. 그림에서 좌측 부분은 온톨로지의 구축을 위한 Protégé-2000의 입력창을, 우측 부분은 입력한 자료를 RDF로 변환한 결과를 나타낸다. 구축이 완료된 온톨로지는 지식 베이스로 활용하게 되며 중요 단어 리스트에 포함 되는 단어들을 매핑 시켜 가중치가 높은 클러스터를 식별하는 용도로 사용하게 된다.

온톨로지 매핑 시에는 제목과 부제목, 출현 빈도가 높은 단어에 대해서 순위화된 가중치를 부여해야 하는데, 그림 6에서 알 수 있듯이 본 연구에서는 제목과 부제목에 대해 각각 7점과 5점을 부여하고, 출현 빈도수가 6이상 일 때는 5점, 4이상일 때는 3점을 그리고 빈도수가 3일 경우에는 2점을 빈도수가 2일 경우에는 0.5점의 가중치를 부여하였다.



<그림 6> 온톨로지 매핑과 가중치 부여

부여된 가중치는 온톨로지 계층구조 상에서 전파 알고리즘을 통해 누적치를 계산하여 최적 경로, 즉 가장 높은 가중치를 갖는 클러스터를 선정하게 된다. 이 과정을 수행하기 위해서는 온톨로지 매핑 시에 RDF 문서를 직접 접근하여 단얼부터 매핑된 개념 노드에 가중치를 먼저 기록해 놓아야 한다. RDF의 인스턴스 노드에 가중치를 기록하는 Java 코드는 그림 7과 같다.

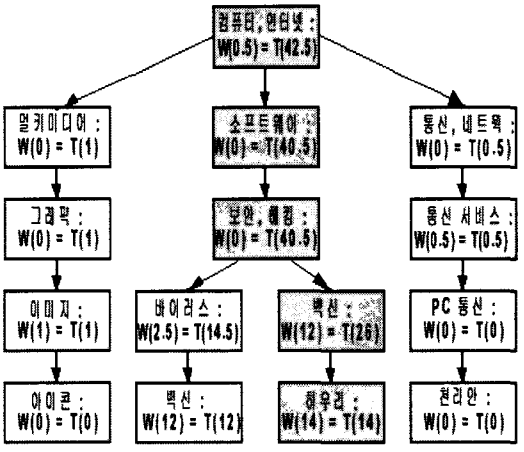
3.4 주제 식별

전술한 바와 같이 중요 단어의 온톨로지 매핑에 의한 가중치 부여가 완료되면, 온톨로지 계층구조 상에서 높은 연관성을 갖고 맺어진 단어들의 클러스터를 찾는 최적 경로 선정 알고리즘을 수행하여야 한다. 이 과정은 부여된 가중치의 누적 값을 전파하여 그 값이 가장 큰 최적 경로를 선택하는 과정으로 볼 수 있다. 누적값의 전파 방식은 자식 노드로부터 누적된 가중치 값에 현재 노드 자신의 가중치 값을 더해서 상위 노드로 전파하고 이와 같은 가중치의 전파 과정을 루트 노드에 도달 할 때 까지 반복해 나간다.

```
for(int count=0;count<nl.getLength();count++)
{NamedNodeMap attrs=nl.item(count).getAttributes();
if(t==attrs.item(2).getNodeName())//t와 attrs의 21번째 속성rdf:label값을 비교한다.
nl.item(count).appendAttributes(weigh).setNodeValue("w");
else
nl.item(count).appendAttributes(weight).setNodeValue("0");
//Node n=new Node();
//if(t==nl.item(count).getAttributes().getnamedItem("rdf:label").getNodeValue())
//nl.item(count).getAttributes().getNamedItem(weight).setNodeValue("w");
//nl.item(count).getAttributes().getNamedItem(weight).setNodeValue("0");}
```

<그림 7> 가중치 부여를 위한 Java 코드

그림 8은 이러한 가중치의 전파 과정을 보여주며 노드에 기록된 w(x) = T(y) 식에서 좌측 항은 가중치를 우측 항은 누적치를 나타낸다. 최적 경로를 선정할 때는 누적치가 가장 큰 경로에 속하는 노



<그림 8> 가중치 전파에 의한 최적 경로 설정

드들을 모으면 최종적인 최적 경로가 되게 된다. 그림 8에서 최적 경로로 설정된 컴퓨터, 인터넷/소프트웨어/보안, 해킹/백신/하우리는 이 논문의 주제가 내포된 클러스터임을 알 수 있다.

이렇게 최적 경로가 선정되면, 자식 노드로부터 누적된 누적 값을 균등 분산한 밀도값에 의해 최적 경로 중 한 개의 노드를 결정하게 된다. 이 과정을 일반화(generalization)라고 하며 선택된 노드가 문서의 주제가 된다. 문서의 주제로 식별된 단어는 나중에 주제문 혹은 주제 단락을 파악하기 위한 이차적 용도로 사용될 수 있다. 그림 9에서는 이러한 단일 노드의 추출 과정을 보여준다. 먼저 단일 노드의 밀도를 계산하는 공식은 다음과 같다.

밀도(D) = 노드의 누적값(T) / 자식 노드 수(n)

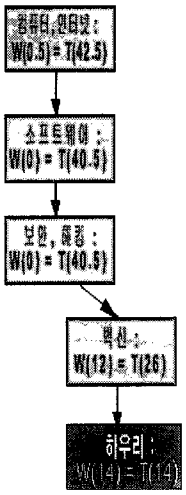
위 수식을 최적 경로 상에 속하는 노드에 적용한 결과는 다음과 같다.

밀도계산 :
컴퓨터, 인터넷(42.5/14 = 3.03)
소프트웨어(40.5/5 = 8.1)
보안, 해킹(40.5/4 = 10.125)
백신(26/1 = 26)
하우리(14)
추가적 계산 :
백신(26) - 하우리(14) = 백신'(12)
백신'(12) < 하우리(14)

끝으로 리프 노드인 "하우리"는 더 이상 자식 노드가 존재 하지 않으므로 자신의 가중치를 밀도값으로 사용하며, 부모 노드인 "백신"의 경우는 리프 노드의 가중치 값을 뺀 값을 밀도값으로 사용하게 된다. 따라서 다음과 같이 추가적인 계산을 실행한 결과 단일 노드로서 최대 밀도값을 갖는 "하우리"가 단일 노드로 선정되며 바로 이 단어가 대상 문서에 대한 주제어가 된다.

4. 성능 평가 및 분석

본 논문에서 제안한 방법의 성능을 평가하기 위해서 디지털 조선일보로부터 컴퓨터 관련 뉴스 기사 50개를 임의 선정하였으며, 전문가의 도움을

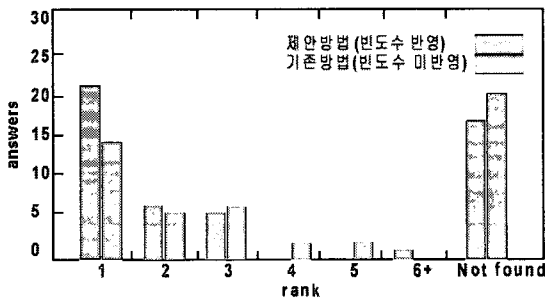


<그림 9> 단일 노드 추출

받아 문서별 정답을 6순위 까지 기록한 정답 세트를 작성하였다.[8] 또한, Protégé-2000을 사용하여 약 5,000 단어 정도의 네이버 온톨로지를 구축하였으며, 빈도수 검사 모듈과 온톨로지 매핑 모듈 그리고 가중치 전파 알고리즘 처리 소프트웨어를 JAVA 언어를 이용하여 구현하였다. 실험에 있어서 비교 대상으로는 Sabrina의 기존 연구를 대상으로 하였으며, 주제 식별의 정확률 측면에서 실험 결과를 분석하였다. 순위별 비교 시에 6순위를 벗어난 응답은 "실패"로 간주 하였다. 정확률을 계산하는 공식은 다음과 같다.

정확률 = 순위내 정답회수 / 실험회수

그림 11은 이러한 실험 결과를 순위별 막대그래프로 보여주며 표 1은 순위내에 포함된 정답에 대한 정확률을 보여준다.



<그림 11> 실험 결과 - 순위별 식별 결과

<표 1> 실험결과 - 정확률

	rank 1	rank 2	rank 3	rank 4	rank 5	rank 6	매치 안됨	정확률 (%)
기존방법	14	5	6	2	2	0	21	58
제안방법	21	6	5	0	0	1	17	66

실험 결과에서 본 논문에서 제안한 방법이 기존의 방법에 비해 순위 1, 순위 2, 순위 6에서 보다 높은 정답률을 보였으며, 전체 정확률에 있어서 66%로 기존의 방법 58% 보다 향상된 성능을 확인할 수 있었다. 성능 향상의 가장 중요한 원인으로서는 중요 단어 선정 시와 가중치 부여 시에 빈도수에 의한 통계적 방법을 병행한 것이라 할 수 있다.

따라서 본 연구는 지식 기반 접근 방법과 통계적 접근 방법을 병행하는 하이브리드 방식이라고 할 수 있다. 그러나 지식 기반 접근 방식으로 최신의 기법인 온톨로지 기반의 주제 식별 방법에 통계적 접근을 병행하는 방식은 지금까지 이루어진 바가 없어 선험적 연구로서 본 연구의 의의를 찾을 수 있다. 그러나 본 연구의 실험 결과가 정확률 측면에서는 향상을 가져 왔지만 지식 기반 접근 방식의 최대 단점인 시간 소모의 비용 문제는 여전히 과제로 남아 있다.

한편, 순위 내 정답 주제와 매치 되지 않는 경우 중 상당수는 네이버 온톨로지의 어휘가 주제 식별을 완벽하게 처리할 수 있을 만큼 충분하지 못해서 발생하는 경우 일 수도 있다. 모든 중요 단어를 매핑 할 만큼 충분하지 못한 어휘는 결국 온톨로지의 원천적 결함 문제를 야기 시키기 때문에 이를 보완하기 위한 지식베이스의 보강 방안도 필요하다. 웹 기반 시소로스로 널리 알려진 WordNet[9] 등 외부 온톨로지를 병행하는 것도 한 가지 방법이 될 수 있을 것이다.

5. 결 론

본 논문에서는 온톨로지를 이용한 지식베이스 접근 방식과 통계적 방법을 병행한 하이브리드 방식의 자동 주제 식별 방법을 제안하였다. 제안한 방법을 구현하기 위해서 Protégé-2000을 사용하여 약 5,000 단어 정도의 네이버 온톨로지를 구축하였으며, 빈도수 검사 모듈과 온톨로지 매핑 모듈 그리고 가중치 전파 알고리즘 처리 소프트웨어를 개발하였다.

제안한 방법의 성능을 실험한 결과, 온톨로지만을 이용한 기존의 방법에 비해 중요 단어 선정과 가중치 부여 시에 빈도수를 반영하는 본 논문의 방법이 정확률 측면에서 우수한 성능을 나타냄을 입증할 수 있었다. 그러나 이 방법은 아직 제안 단계로서 모든 시스템이 구현 된 것은 아니며, 성능을 평가하기 위해서 반드시 필요한 부분만 모듈별로 코딩하였으며 네이버 온톨로지 역시 컴퓨터와 관련된 단어에 한해서만 구축되었다.

본 연구의 최종 목표는 정확률과 처리 비용 면에서 성능이 우수한 전체 시스템을 구축하는 것이다. 그러기 위해서는 단어의 출현 빈도수뿐만 아니라 웹 문서의 소스, 즉 html에 접근하여 태그 정보에 수록된 각 종 마크 업 자료들을 활용하는 방안 등 통계적 자료로 활용할 수 있는 부가적인 정보를 사용함으로써 보다 높은 정확률을 갖도록 하는 것이 중요하다. 또한 정확률을 저하 시키는 가장 큰 원인인 온톨로지 매핑 시 어휘 부족 문제를 해결하기 위한 방법 역시 중요하게 다루어져야 한다. 온톨로지의 어휘 부족 문제는 WordNet과 같은 외부 온톨로지를 사용한다면 상당 부분 해결이 가능할 것으로 보인다. 따라서 본 연구의 향후 과제는 웹 문서의 특성에 맞는 통계적 접근 방식을 보강하고 외부 지식베이스를 보강하여 보다 높은 정확률을 갖는 자동 주제 식별 시스템을 구현하는 것이다.

참 고 문 헌

[1] Lin, C.Y and E.Hovy. Identifying Topics by Position of The Workshop of Intelligent Scalable Text Summarization '97. (1997).

[2] 이상구. "전자신문, 온톨로지에 대한 새로운 접근", 2003. 6. 24.

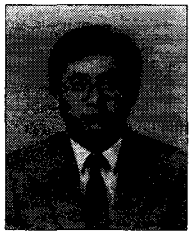
[3] Steffen Staab, Michael Erdmann, Alexander Maedche, Stefan Decker. "An Extensible Approach for Modeling Ontologies in RDF (S), 2000.

[4] Stefan Decker and Sergey Melnik. "The

Semantic Web : The Roles of XML and RDF", 2000
[5] Dan Brickley, R.V. Guha, "Resource Description Framework (RDF) Vocablulay Description Language 1.0: RDF Schema, "Technical Report, W3C Working Draft, http://www.w3.org/TR/rdf-schema, 2003.
[6] Sabrina Tiun, Rosni Abdullah, Tang Enya Kong: Automatic Topic Identification Using Ontology Hierarchy. 'CICLing' 2001: pp. 444-453.
[7] Protégé, "The protégé project, "http://protege.stanford.edu, 2002.
[8] http://home.ewha.ac.kr/~jhkim/project/00invest/8/digitalchosun.htm
[9] Scott, S. and. S. Matwin. Text Classification using WordNet Hypernyms. In the proceeding of Workshop usage of WordNet in Natural language Processing Systems: Montreal, Canada. (1998).

최 인 대 (In-Dae Choi)

1994년 경일대학교 컴퓨터공학부 졸업(전자계산기 공학사)
2004년 대구대학교 대학원 정보통신공학과 멀티미디어 공학전공(공학석사)
2004년~현재 도립거창전문대학강사



남 인 길 (In-Gil Nam)

1978년 경북대학교 전자공학과 졸업(공학사)
1981년 영남대학교 대학원 전자공학과(계산기전공 공학석사)
1992년 경북대학교 대학원 전자공학과(전산공학전공 공학박사)
1980년~1990년 경북산업대학교 전자계산학과 부교수
1980년~현재 대구대학교 컴퓨터·IT공학부 교수
(관심분야: 데이터베이스, 이동컴퓨팅)

부 기 동 (Ki-Dong Bu)

1984년 경북대학교 전자공학과 졸업(전자계산기전공 공학사)
1988년 경북대학교 대학원 전자공학과(전산공학전공 공학석사)
1996년 경북대학교 대학원 전자공학과(전산공학전공 공학박사)
1988~현재 경일대학교 컴퓨터공학부 교수
(관심분야 : 데이터베이스, GIS, 정보검색)