WordNet과 텍스트 코퍼스에 기반한 의미 관계를 활용한 웹 텍스트 조사 기법

이호석, 김영택
뉴미디어학과 공과대학 호서대학교, 컴퓨터신기술연구소 서울대학교
hslee@office.hoseo.ac.kr

A Web Text Mining Technique using Semantic Relations based on WordNet and Text Corpus

Ho Suk Lee, Yung Taek Kim
New Media Dept. Hoseo University, Computer Technology Research Center SNU

요 약

본 논문은 문장 분석에 의하여 의미 관계를 생성하고 의미 네트워크에 의하여 유사한 의미 관계를 고려하는 의미 중심의 웹 텍스트 검색 기법에 대하여 논의한다. 기존의 웹 텍스트 검색은 단어만을 혹은 의미 관계만을 고려한 검색이었다고 할 수 있다. 그러나 문장 분석에 의한 의미 관계의 생성과 의미 네트워크에 의한 유사한 의미 관계의 고려는 기존의 단어 중심 혹은 의미 관계 중심의 검색 한계를 넘어서 유사한 의미 관계를 고려한 좀 더 포괄적이고 계층적인 검색을 가능하게 할 것으로 생각된다.

1. 서 론

많은 양과 종류의 문서가 웹에서 제작되어 사용되고 있으며 앞으로 이 추세는 더욱 증가할 것이다. 이러한 웹 문서의 형태에는 일반 텍스트 형태, HTML 형태, 그리고 XML 형태가 있다. 이러한 형태의 문서 중에서도 일반 텍스트 형태의 문서의 양이 급속도로 증가하고 있다고 한다. 이러한 일반 텍스트는 구조를 가지고 있지 않기 때문에 직접 컴퓨터 프로그램으로 처리하는 것이 어렵다. 따라서 이를 해결하기 위한 텍스트 검색 기법에 관심이 모아지고 있다[1][2].

참고문헌 [3]은 텍스트 형식의 웹 문서로부터 의미 관계(semantic relation)를 찾아내어 검색하는 방법을 논의하고 있다. 단순히 단어 모음(bag)만을 가지고 웹 문서를 검색하는 것은 단어들 간의 의미 관계를 나타낼 수가 없어서 의미적으로 중요한 검색이 불가능하며 때로는 의미적으로 잘못된 검색 결과를 제시할 수가 있다고 한다. 참고문헌 [3]에서는 이러한 문제점을 극복하기 위하여 문장의 의미 관계를 추출하는 방법을 사용하였다. 텍스트 문서에 대하여 대명사 지시(reference) 문제 해결[4], 품사(part-of-speech) 태깅(tagging)[5], 구문 분석[6]을 사용하여 텍스트의 문장에서 의미를 나타내는 명사구(noun phrase)와 동사구(verb phrase) 등을 찾아내고 이를 기반으로 텍스트의 문장을 의미 관계로 변환하여 나타내었다. 의미 관계에는 agent, theme, modifiedBy 세 가지를 정의하여 사용하였다. 다음에 용어(term) 분류체계(taxonomy)를 구성한다. 이 시도의 목적은 비슷한 용어를 사용하여 구성된 의미 관계를 일반화시켜서 통계적으로 의미가 있는 패턴을 찾아서 효과적인 검색을 하기 위한 것이다.

최근에 온톨로지 구축을 위한 목적으로 용어의 분류 체계에 대한 연구가 많이 진행되었다. 이러한 방법에는 문자(symbol)를 이용한 방법과 통계를 이용한 방법이 있다. 문자를 이용한 방법의 단점은 모든 가능한 용어 관계를 추출할 수 없다는 것이다. 통계를 이용한 방법의 단점은 통계적인 수치에 의해서만 용어 분류가 이루어지기 때문에 실질적인 의미를 나타내지 못할 수가 있어서 응용에 직접 사용하기가 어려운 면이 있다는 것이다. 또한 이들 방법이 실용성을 가지기 위해서는 대용량의 코퍼스(corpus)가 필요하다. 다른 방법으로는 WordNet[7]을 사용하는 방법이 있다. 기본적으로 미리 분류된 용어로부터 WordNet 을 사용하여 응용 영역에 적합한 용어 분류를 점진적으로 생성하여 전체 용어의 분류 체계를 구성하는 방법이 있다. 여기에는 용어간의 의미 유사성을 계산할 수 있는 수식이 필요하다.

이렇게 하여 응용에 적합한 용어의 분류체계가 구성되면, 이 분류체계를 기반으로 텍스트를 구성하는 문장들의 의미 관계를 구성한 다음에, 용어들 간의 일반적인 연합 관계(generalized association)에 대한 조사를 시작한다. 여기서 문제가 되는 것은 과잉 일반화(overgeneralization) 문제이다. 용어들 간의 분류 체계를 기반으로 조사를 수행하는 과정에서 분류 체계의 상부 구조로 접근하게 되면 대부분의 경우에 과잉 일반화 문제에 빠지게 된다고 한다. 즉, 구체적인 검색이 이루어지는 것이 아니고 매우 일반적 검색이 이루어져서 적합하지 않은 검색 결과를 생성할 수가 있기 때문이다. 따라서 적절한 수준에서 구체적인 검색 결과만을 선택하여 과잉으로 일반화된 결과가 제시되지 않도록 방지하여야 한다. 참고문헌 [3]은 GP-Close (Close Generalized Pattern Mining) 라는 방법을 고안하여 이 문제를 다루고 있다. 그리고 제안된 방법을 RDF(Resource Description Framework) XML 문서에 적용한 결과를 보여주고 있다.

참고문헌 [8]에서도 유사한 연구를 하였다. 이 논문에서는 직접적으로 현재의 Google 검색 결과와 자신들의 검색 결과를 제시하여, 자신들 방법의 진보된 결과를 제시하고 있다고 주장한다. 참고문헌 [8]에서는 Google에 "Shanghai", "hotel", "Five star" 라는 세 개의 검색 단어를 입력하였을 경우에, 단 하나의 검색 결과만이 상하이에 위

치한 별 5개의 호텔을 나타내고 다른 하나는 대만에 위치한 "Far Eastern Plaza Hotel"을 나타내고 있었다고 한다. 그 이유는 이 호텔에 "Shanghai" 라는 이름을 가진 별 5개짜리의 유명한 레스토랑이 있었기 때문이라고 한다. 따라서 사람이 "Shanghai", "hotel", "Five star" 라고 입력하였을 경우에, 일반적인 상식으로는 상하이에 위치하는 별 5개의 호텔을 의미한다고 할 수 있으나, Google에서는 하나의 결과만이 이러한 검색 결과를 보여주고 나머지는 의미적으로 관련이 거의 없는 검색 결과를 보여주었다고 한다.

이 논문에서는 그 이유가 관계 상실(relation lost)이라고 주장하고 있다. 다시 말하면 Shanghai, hotel, five star라는 용어들 사이의 상식적인 의미 관계를 Google 이 파악하지 못하고 오직 3개의 별개 용어로 파악하고 검색을 하였기 때문에 사람이 원하지 않는 검색 결과가 나왔다는 것이다. 따라서 이러한 상식적인 의미 관계의 설정이 중요한 문제라고 주장하고 있다.

이 문제의 해결을 위하여 저자들은 Microsemantic Web 환경을 구축하였다. 예를 들어, 참고문헌 [8]에서는 "travel" 이라는 온톨로지를 구축하고 여기에 "city"와 "AccommodationRating"에 대한 정의를 기록한 데이터 모델을 구성하여 이 문제를 해결할 수 있다고 주장하고 있다. 즉, RDF를 사용하여 Microsemantic Web 환경을 구축하여 - 즉, 용어들 간의 의미 관계를 구축하여 - 이 문제를 해결하였다고 한다.

참고문헌 [9]는 웹 정보 검색 시스템 19종류에 대한 종합 개관을 설명하고 있다. 이 논문은 검색 시스템을 기능, 사용된 기술, 자동화의 수준 등의 세 가지 관점에서 분류하여 보여주고 있다. 기능적인 관점의 분류에서는 manual, supervised, semi-supervised, unsupervised로 분류하여 설명하고 있다. 기술적인 관점의 분류에서는 학습 방법이 사용되지 않는 경우, 상향식(하향식) 학습 방법이 사용되는 경우, 그리고 패턴/문자열 검색이 사용되는 경우로 분류하여 설명하고 있다. 학습 방법이 사용되지 않는 경우는 검색 토큰 단위가 수동인 경우가 많았으며, 상향식(하향식) 학습 방법이 사용되는 경우에는 단어 단위가 많았으며, 조사 기법이 사용되는 경우에는 태그 단위가 많았다. 자동화 수준의 분류에서는 API 지원과 응용의 관점으로 분류하였다.

참고문헌 [10]에서는 사용자가 XML 문서에 사용자 정의 데이터 타이프를 정의하면 조사와 검색이 용이하다는 것을 주장하고 있으며 제안한 데이터 타이프에 대한 상세한 설명이 제시되어 있다. 참고문헌 [11]은 의미 네트워크와 코퍼스 통계에 기반을 둔 문장의 유사성 측정에 관한 방법을 논의하고 있다. 두 개의 문장에 대한 유사성은 어휘 데이터베이스의 정보(의미 네트워크)와 코퍼스로부터 계산된 통계로부터 계산된다. 어휘 데이터베이스의 사용은 인간의 상식을 모델링하는 것을 가능하게 하였으며 코퍼스 통계의 사용은 다른 도메인에도 적용될 수 있도록 하였다고 한다. 의미 네트워크로는 WordNet[7]을 사용하고 코퍼스는 Brown 코퍼스를 사용하였다. 참고문헌 [12]는 웹 테이블로부터 의미 있는 헤드 정보를 추출하는 기법을 논의하고 있다. HTML 문서에 있는 테이블의 종류를 조사하여 의미를 가진 테이블을 분리하였다. 그리고 테이블로부터 의미를 나타내는데 필요한 데이터(feature)를 추출하여 집합을 구성하였으며 이 데이터들을 기반으로 판단 트리(decision tree)를 구성하여 의미 있는 head 요소를 HTML 문서로부터 추출하여 분류할 수 있도록 하였다.

2. 본 론

2.1 연구 범위

본 논문의 연구는 기본적으로 의미 관계의 생성과 의미 네트워크를 활용한 웹 텍스트 문서의 검색 기법에 관한 것이다. 참고문헌 [3]에서도 구문 분석을 수행하여 의미 관계를 구성하고, WordNet을 사용한 의미 네트워크를 구성하여 용어 체계를 분류하고 이를 기반으로 웹 텍스트 문서를 검색하였다. 그러나 이 논문은 대명사 지시 문제를 해결하고자 하였으며, 품사 부여와 구문 분석에서 몇 가지 중요한 사항들을 언급하지 않았다.

우선 첫째로, 대명사는 하나의 명사를 지시할 수도 있지만 구절을 지시할 수도 있다. 대명사의 지시 문제를 완전하게 해결하기 위해서는 문장과 문단의 의미 관계를 먼저 알아야 한다고 생각한다. 따라서 이것은 의미를 먼저 파악하고 나서 시도할 수 있는 문제라고 생각한다. 두 번째로, 품사 부여는 전자 사전에 수록된 내용을 기반으로도 결정할 수 있다. 그러나 단어에 하나 이상의 품사가 부여될 수도 있다. 품사 부여의 단계에서 완전한 결정을 내릴 수가 없는 경우도 많다. 품사 부여를 완벽하게 하기 위해서는 문맥 정보가 필요하며 때로는 의미 정보가 필요한 경우도 있다. 따라서 사전에 기록된 가능한 모든 품사를 단어에 부여하여야 한다. 세 번째로, 대부분의 경우에 문장에 대한 구문 분석도 하나 이상의 결과를 생성할 수가 있다. 이 경우에도 생성된 모든 구문 분석 결과를 고려하여야 한다. 즉, 대명사 지시 문제는 의미 분석 없이는 해결할 수 없는 경우가 있으며, 품사 부여와 구문 분석의 경우에도 하나 이상의 결과가 생성되기 때문에 부가적인 정보가 없는 상황에서는 모든 경우를 제시하여야 한다는 것이다.

본 논문에서는 대명사 지시 결정 문제는 다루지 않기로 한다. 그리고 품사 부여와 구문 분석의 다중성에 대한 문제만 다루어서 보완하기로 한다. 품사 부여에서 단어의 품사가 여러 개로 될 경우에는 모든 경우를 부여하기로 한다. 그리고 구문 분석에서 여러 개의 구문 분석 결과가 생성될 경우에도 모든 경우를 고려하기로 한다. 그리고 명사구와 동사구이외에 복합 명사(compound noun)를 다루기로 한다.

2.2 복합 명사 처리 방법

복합 명사는 명사가 여러 개 연속하여 앞의 명사가 뒤의 명사를 수식하여 다른 의미의 명사가 되는 것으로 정의할 수 있다. 복합 명사의 경우는 사전에 수록하고 하나의 기호(symbol)를 부여하고 명사 품사를 부여한다. 다음에 문장을 처리할 때에 시스템 내부적으로는 원래의 복합 명사 대신에 부여된 기호를 사용하면 처리하면 될 것이다. 이렇게 하면 구문 분석의 복잡성을 감소시킬 수가 있어서 복합 명사 자체로 인하여 여러 개의 구문 분석 결과가 생성되는 것을 막을 수가 있으며, 또한 의미 관계를 구성할 때도 복잡성을 감소시켜서 여러 개의 의미 관계 구조가 생성되는 것을 막을 수가 있을 것이다. 그러나 검색을 수행할 때에는

기호를 사용하지 않고 원래의 복합 명사를 사용하여서 최대한 원하는 검색 결과를 얻을 수 있도록 해야 할 것이다.

2.3 구문 분석

구문 분석은 참고 문헌 [6]의 파서를 이용해 보는 것도 좋을 것으로 생각한다. 구문 분석의 결과가 여러 개 생성되면 모든 결과에 대하여 의미 관계를 생성한다. 따라서 검색에 사용되어야 하는 것은 문장에 있는 복합 명사들과 생성된 여러 개의 구문 분석 결과가 될 수 있다. 참고 문헌 [3]과 [8]에는 이러한 경우에 대한 논의가 제시되어 있지 않다. 이들 참고 문헌들은 품사 부여와 구문 분석의 결과가 오직 하나만 나온다고 가정하고 논의를 전개하고 있는 것으로 보이나 사실 실험을 해보면 일반적인 문장의 경우에도 2~3개 정도는 보통이고 그 이상 여러 개의 품사 부여와 구문 분석 결과가 생성되는 경우가 많다. 따라서 하나의 문장에 대한 의미 관계도 여러 개를 구성할 수가 있다. 그러나 생성된 여러 개의 의미 관계에 의미 네트워크를 활용하여 확률 값을 계산하여 부여하고 값이 높은 의미 관계부터 검색 대상으로 하면 좋을 것으로 생각한다.

2.4 의미 관계 구성

의미 관계는 일반적으로 그래프 구조로 나타난다. 그래프의 노드(node)는 단어를 나타내고 에지(edge)는 의미 관계를 나타낸다. 다음과 같은 영어 문법이 있다고 가정한다.

```
Sentence -> Noun_phrase + Verb_phrase
Noun_phrase -> Noun
Noun_phrase -> Article + Adjective + Noun
Noun_phrase -> Article + Noun_compound
Verb_phrase -> Verb
Verb_phrase -> Verb + Noun_phrase
Verb_phrase -> Verb + adverb
Verb_phrase -> be + Verb
Prep_phrase -> Prep + Noun_phrase
Prep_phrase -> Prep + Noun_compound
```

그러면 아래 문장 1과 2에 대하여 구문 분석한 결과를 얻을 수 있다.

문장 1: Korea defeated Spain in World Cup final.

문장 2: Korea was defeated by Spain in World Cup final.

문장 1에 대한 구문 분석 결과는 다음과 같다.

```
Sentence -> Noun_phrase + Verb_phrase
Noun_phrase -> Noun
Noun -> Korea
Verb_phrase -> Verb + Noun_phrase
Verb -> defeat
Noun_phrase -> Noun
Noun -> Spain
Prep_phrase -> Prep + Noun_compound
Prep -> in
Noun_compound -> World Cup final
```

그리고 의미 관계는 다음과 같다고 할 수 있다.

```
defeat(subject(Korea), object(Spain))
modifier(in(World Cup final))
```

문장 2에 대한 의미 관계는 다음과 같이 될 수 있다.

```
defeat(subject(Spain), object(Korea))
modifier(in(World Cup final))
```

즉, 웹 조사와 검색은 구문 분석 결과로 생성되는 이러한 의미 관계를 기반으로 수행될 수 있다.

2.5 의미 네트워크 구성

의미 네트워크는 기존의 WordNet을 기반으로 구성된다. WordNet에 문장을 구문 분석하여 추출한 단어들, 예를 들면, defeat, Korea, Spain, World Cup final 등을 더하고 이 단어들과 관련된 단어들을 WordNet에서 선택하여 각 문장마다 의미 네트워크를 구성할 수가 있다. 따라서 문장에 대하여 의미 네트워크를 구성하고 이를 기반으로 의미 관계를 구성하면 검색에 매우 유리한 대상을 확보할 수 있을 것이다. 예를 들어, defeat과 관련이 있는 attack, win, smash, beat, overwhelm 등의 용어를 검색 대상의 용어로 간주할 수 있으며, Korea외에도 South Korea와 North Korea 도 고려할 수 있도록 할 수 있다. World Cup 이외에도 Asian Game 등의 용어도 고려할 수 있도록 할 수 있다. 이 부분은 참고문헌 [11]에서 논의한 어휘 데이터베이스 부분에 해당된다고 할 수 있다.

2.6 검색 대상

검색은 의미 관계를 기반으로 수행될 수 있다. 예를 들어, defeat(subject(Korea),object(Spain)), defeat(subject(Korea), _), 혹은, defeat(subject(Korea), object(object(Spain)) +modifier(in(World Cup final)), defeat(subject(Korea),object(Spain))+modifier(in(_)) 등이 검색의 대상이 될 수 있도록 고려할 수 있다. 위에서 _는 변수를 나타낸다. 이외에도 의미 네트워크로부터 형성된 유사한 의미 관계도 검색의 대상으로 고려할 수 있다. 물론 생성된 유사한 의미 관계에는 적절하게 계산된 확률 값을 부여하여 검색의 우선 순위와 범위를 설정할 수 있도록 하여 검색을 조절할 수 있도록 하여야 할 것이다.

2.7 일반화 조절 기능

검색의 일반화는 용어 관계를 구성하여 수행할 수 있다. 명사 Korea와 Spain 그리고 복합 명사 World Cup final 에 대하여 수행될 수 있다. 명사 Korea와 Spain에 대하여는 국가이름 범주를 부여하여 분류할 수 있을 것이다. 우선 defeat(subject(Korea), object(Spain))과 관련된 문서들이 검색될 것이다. 그리고 Korea와 Spain 이외에 다른 국가에 해당되는 경우의 defeat(Country1, Country2)에 해당되는 문서가 출력될 수 있을 것이다. 이 경우 너무 일반적인 경우가 되지 않도록 조절할 수 있는 장치가 필요하다. 가령 Korea의 경우에 아시아 국가라는 중간 범주를 설정할 수 있을 것이고, Spain의 경우에 유럽 국가라는 범주를 설정할 수 있을 것이다. 즉, 중간 범주를 설정하고 범주의 범위를 넘어설 때마다 용어와 범주 거리를 계산하고 이 값이 일정 이상의 값을 나타내면 조사나 검색을 중지하는 기능을 사용하면 일정 범주 밖으로 일반화되는 경우를 조절할 수 있을 것이다. 즉, 용어 분류 체계를 나타내는 의미 네트워

크에서 에지를 지나갈 때마다 거리를 계산하여 중요한 의미가 있는 범주 밖으로의 용어 의미 일반화를 제어할 수 있을 것이다. 앞 절의 예에서 설명한 동사 defeat과 관련이 있는 attack, beat, smash, overwhelm 등의 용어에 대하여서도 이와 같은 일반화 관계를 고려할 수 있을 것이다. 이 부분은 의미 관계의 유사성을 계산하는 부분이라고 할 수 있다.

2.8 새로운 논의 사항

본 논문에서는 참고문헌[3][8]에서 고려하지 않은 품사 부여와 구문 분석의 다중성을 고려하여 생성된 모든 문장 분석 결과를 대상으로 검색을 할 것을 논의하였다. 다중성을 고려할 경우 검색 경우의 수가 증가할 가능성은 있다. 그러나 이것은 우선 순위를 설정하여 적절한 범위 내에서 제한하면 해결할 수 있을 것이다. 그리고 의미 네트워크를 사용하여 의미적으로 유사하거나 관련이 있는 용어를 찾아서 추가적으로 의미 관계를 구성하여 검색 대상으로 하여야 한다고 논의하였다. 유사성을 비교하기 위해서는 적절한 유사성 비교 수식을 사용하여야 한다. 참고문헌 [3]은 코사인 측정법(cosine similarity measure)을 사용하였다.

그리고 WordNet에 기반을 둔 의미 네트워크를 구성할 때에 그 구조를 인간의 의식에 비추어서 체계적인 순서로 구조화하여 구성하는 것을 고려해 보는 것이 좋을 것이며, 인터넷에서 접근 가능한 모든 문서에 대하여 의미 네트워크를 구성해 보는 것도 좋을 것이다. 의미 네트워크를 구성할 때에는 확률 방법론[2][13][14][15]를 활용하여 구성하면 좋을 것이다. 이러한 확률 값은 의미 네트워크 상에서 적절한 추론을 가능하게 하고 또한 과잉 일반화 혹은 부적절한 추론과 같은 오류를 조절할 수 있도록 할 것이다.

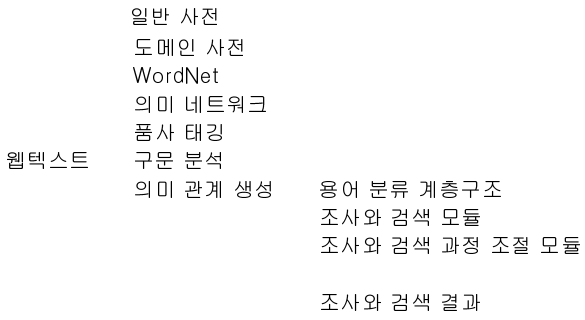
그리고 여러 분야 예를 들어 철학, 정치학, 경제학, 역사, 문학, 과학, 의학, 공학 등 각 분야에서 작성된 문서들에 대하여 용어 분류 체계와 의미 네트워크를 구성할 수가 있을 것으로 생각되며, 그 결과들을 모두 통합하면 위에서 언급한 모든 분야의 문서에 대하여 의미 네트워크를 구성할 수가 있을 것이다. 이러한 시도는 WordNet 이론 체계의 완성을 가져올 수가 있을 것으로 생각되며, 전산 언어학과 전산 심리학의 단단한 기반도 될 수 있을 것으로 생각된다. 왜냐하면 참고문헌 [16]에 의하면 인간의 마음이라는 것이 대부분 의식적인 언어 현상으로 볼 수 있다고 하므로, 감각(sense)을 나타내는 부분과 묘사(description)를 나타내는 부분에 대한 언어의 이해, 저장, 그리고 생성을 처리하면 이러한 부분들에 대한 인간의 마음은 표현할 수가 있을 것이기 때문이다.

2.9 시스템 구조

다음은 시스템을 구성하는 중요 구성 요소이다.

. 사전
. 품사 태깅 모듈
. 구문 분석 모듈
. 의미 관계 생성 모듈
. WordNet
. 의미 네트워크 모듈
. 조사(마이닝) 모듈
. 검색 모듈
. 조사와 검색 과정 조절 모듈
. 사용자 인터페이스

다음은 시스템 구조에 대한 그림이다.



그러나 이 방법의 한 가지 문제점은 모든 텍스트 문서에 대하여 품사 분석과 구문 분석을 실시하기 때문에 분석을 실시하지 않는 방법보다 수행 속도가 느려질 수 있다.

3. 결 론

본 논문에서는 참고문헌 [3][8]에서 고려하지 구문 분석을 고려하여 좀 더 적절한 검색 결과가 출력될 수 있도록 하였다. 즉, 품사 부여와 구문 분석의 다중성을 고려하였으며, 의미 네트워크를 사용하여 의미적으로 연관성이 있는 의미 관계를 고려하여 검색이 좀 더 포괄적이고 계층적으로 수행될 수 있도록 하였다. 앞으로의 연구로는 논문에서 논의하고 제안된 아이디어와 시스템에 대한 구현과 실험이 있어야 할 것이다.

참고 문헌

[1] Jiawei Han, Micheline Kamber, Data Mining concepts and techniques (2nd ed.), Morgan Kaufman 2006.
[2]Pang-Ning Tan, et al., Introduction to Data Mining, Addison-Wesley 2006.
[3] Tao Jiang, Ah-Hwee Tan, Ke Wang, “Mining Generalized Associations of Semantic Relations from Textual Web Content,” IEEE Trans. on Knowledge and Data Engineering, Vol. 19, No. 2, February 2007.
[4] Diana Maynard, et al., "A Survey of Uses of GATE," Research Memo CS-00-06, Dept. of Computer Science, Univ. of Sheffield, July 2000.
[5] Eric Brill, "A Simple Rule-Based Part of Speech Tagger," Proc. Conf. of Applied Natural Language Processing, pp.152-155, 1992.
[6] Michael J. Collins, "A New Statistical Parser Based on Bigram Lexical Dependencies," Proc. of Conf. Assoc. Computational Linguistics, pp.184-191, 1996.
[7] George A. Miller, "Wordnet: A Lexical Database for English," CACM, Vol. 38, No. 11, pp.39-41, 1995.
[8] Yufei Li, et al., “A Relation-Based Search Engine in Semantic Web,”IEEE Trans. on Knowledge and Data Engineering, Vol. 19, No. 2, February 2007.
[9] Chia-Hui Chang, et al., “A Survey of Web Information Extraction Systems,” IEEE Trans. on Knowledge and Data Engineering, Vol. 19, No. 2, February 2007.
[10] Jeff Z. Pan, “A Flexible Ontology Reasoning Architecture for the Semantic Web,” IEEE Trans. on Knowledge and Data Engineering, Vol. 19, No. 2, February 2007.
[11] Li Y., et al., "Sentence similarity based on semantic nets and corpus statistics," IEEE Trans. on Knowledge and Data Engineering, Vol. 18, No. 8, pp.1138-1150, August 2006.
[12] Sung-Won Jung, Hyuk-Chul Kwon, "A scalable hybrid approach for extracting head components from Web tables," IEEE Trans. on Knowledge and Data Engineering, Vol. 18, No. 2, pp.174-187, Feb. 2006.
[13] Richard E. Neapolitan, Learning Bayesian Network, Prentice-Hall, 2004.
[14] George E. P. Box, George C. Tiao, Bayesian Inference in Statistical Analysis, Addison-Wesley Publishing Company, 1973.
[15] Tom Mitchell, Machine Learning, McGraw-Hill, 1997.
[16] Eric R. Kandel, In Search of Memory, W. W. Norton & Company, Inc., 2006.