정답문서집합 자동 구축을 위한 속성 기반 분류 방법

(Attribute-Based Classification Method for Automatic Construction of Answer Set)

오 효 정 † 장 문 수 †† 장 명 길 †††
(Hyo-Jung Oh) (Moon-Su Chang) (Myung-Gil Jang)

요 약 본 논문에서는 사용자에게 보다 유용한 정보를 제공하기 위하여 개념의 활용분야에 따른 속성 분류 기법이라는 새로운 분류 기법을 제안하고, 이를 활용해 정답문서집합 지식베이스를 자동으로 구축하는 방안을 제시한다. 제안된 방법은 범주간의 구분이 유동적인 속성의 특성을 반영하기 위하여 속성 특징(clue)을 활용함으로써 분류 정확도를 높이고, 개념망에 정의된 개념들 사이의 관계를 참조함으로써 지식베이스를 구축하기 위한 노력과 비용을 최소화하여 점진적인 분류기 생성을 가능하게 한다. 실험을 통해 제안된 방법의 정확도와 효율성을 입증하였으며, 정답문서기반 정보검색 시스템을 위한 정답문서집합 구축 과정에 적용시킨 결과를 제시함으로써 방법의 실제 효용성을 보였다.

키워드 : 속성기반분류, 지식베이스 자동구축, 정답문서기반 정보검색

Abstract The main thrust of our talk will be based on our experience in developing and applying an attribute-based classification technique in the context of an operational answer set driven retrieval system. To alleviate the difficulty and reduce the cost of manually constructing and maintaining answer sets, i.e., knowledge base, we have devised a new method of automating the answer document selection process by using the notion of attribute-based classification, which is in and of itself novel. We attempt to explain through experiments how helpful the proposed method is for the knowledge base construction process.

Key words : Attribute-based Classification, Automatic Knowledge base Construction,. Answer set driven IR

1. 서 론

웹의 발달로 사용자가 접할 수 있는 디지털 문서의 양이 급증함에 따라 많은 양의 문서를 체계적으로 관리하고 이를 효율적으로 검색하기 위한 문서 분류 모델에 관한 연구는 이미 오래 전부터 계속되어 왔다. 현재 인터넷 검색 사용자는 수백만 건에 달하는 검색 결과에서 자신이 원하는 정보를 찾기 위해 문서의 주제, 형식뿐만 아니라 문서의 쓰임에 따른 구별을 원하고 있다. 이처럼 사용자의 다양한 요구를 수용하기 위해 새로운 관점의 분류 모델에 대한 연구가 많이 시도되고 있는데, 특히 최근에는 기존의 주제 범주에 따른 문서 분류 모델[1]에 관한 연구뿐만 아니라 하이퍼텍스트의 특성에 따른 분류 모델[2]이나 문서의 형식이나 글의 유형에 따른 장르 분류 모델[3]에 관한 연구도 시도되고 있다. 이와 더불어 문서 자동분류 기술을 정보검색에 활용하여 사용자 질의 의도에 적합한 검색 결과를 제공하는 기법에 관하여 두드러진 연구 결과가 나오고 있다[4]. 특히 사용자가 자주 묻는 질의에 대하여 미리 정답을 할당해 놓아 사용자의 검색 요구에 적절히 활용하고 있다[5]. 본 논문에서는 사용자가 알고 싶어하는 정보(information need)의 종류, 예를 들어 특정 개념의 "정의"나 "전망" 등 그 개념이 실제 웹 문서상에서 사용되고 있는 영역에 따른 관점으로 문서를 분류하는 속성 분류 기법을

† 비 회 원 : 한국전자통신연구원 휴먼정보검색연구팀 연구원
ohj@etri.re.kr
†† 비 회 원 : 서경대학교 소프트웨어학과 교수
cosmos@skuniv.ac.kr
††† 정 회 원 : 한국전자통신연구원 휴먼정보검색연구팀 팀장
mgjang@etri.re.kr
논문접수 : 2002년 12월 11일
심사완료 : 2003년 4월 18일

제시한다.

이와 다른 측면으로, 최근 정보검색 환경은 빠른 속도와 함께 높은 정확도를 요구하는 추세가 이어져, 그 한 방안으로 정답문서기반(answer approach) 정보검색이라는 새로운 검색 방식이 선보여지고 있다[5, 6]. 이를 위해서는 단순히 명사 빈도수에 의한 키워드 색인에 의존하는 기술에서 나아가, 전문가의 지식을 통해 문서를 관리하고 사용자의 질의에 적합한 정답문서를 제시하는 기술이 필요하다. 지식베이스를 구축하는 방법으로는 사용자가 찾고자 예상되는 정보를 지식베이스로 구축하여 제공하는 방법이 있는데, 현재는 애스크지브스(Ask-Jeeves)[5]에서 보는 바와 같이 사용자가 찾고자 예상되는 질문과 답을 정답 네트워크를 통하여 미리 수작업으로 구축하여 제공하고 있는 실정이다. 그러나, 이를 수동으로 구축하는 방법은 시시각각 변하는 정보의 변이를 반영하기 어려울 뿐만 아니라 고비용이 소요된다는 단점이 있다[7].

본 논문에서는 이러한 단점을 해결하기 위해, 속성 분류라는 새로운 관점의 분류 기법을 통해 웹 문서를 인간의 지적 자원인 개념망에 결합시켜 하나의 지식베이스로 구축하는 방법을 제안하고자 한다.

본 논문의 주안점은 정답문서기반 정보검색 시스템을 구축하는데 있어, 사용자가 원하는 정보가 구축되어 있는 지식베이스를 속성 분류라는 새로운 기술을 통해 자동으로 구축하는 방법을 제안하는데 있다. 또한 제안된 방법을 정답문서기반 정보검색 시스템 구축과정에 적용시킨 결과를 제시함으로써 방법의 실제 효용성을 보이기로 한다.

2장에서는 본 논문에서 궁극적으로 구축하고자 하는 지식베이스의 구조와 이를 활용해 사용자가 원하는 정보를 검색하는 정답문서기반 정보검색 시스템에 대해 설명하고, 3장에서는 본 논문에서 제안하는 속성 기반 분류 모델이라는 새로운 분류 방법을 제시한다. 4장에서 제안된 방법에 의한 실험결과를 제안하고, 5장에서 결론과 개선점을 밝힌다.

2. 정답문서집합 기반 정보검색

2.1 정답문서집합 기반 정보검색 시스템

본 논문은 정답문서 기반 정보검색을 위한 정답문서집합 구축 과정에 있어, 속성 분류라는 방법을 활용해 자동화하는데 그 목적이 있다.

정답문서집합 기반 정보검색은 서론에서 언급한 것처럼 전문가의 지식을 통한 정답의 구축과 검색이라는 점에서 기존 웹 페이지 검색과 구별된다. 여기서 정답문서집합이란 웹 문서를 명사 개념어별로 분류하고, 사용자의 관심 영역에 따라 세분해서 나누어 놓은 문서집합을 의미한다. 전문가의 지식을 활용하여 정답문서를 구축하는 데에는 여러 가지 방법[5, 6, 8]이 있지만, 본 논문에서는 인간의 지적 자원인 개념망을 활용하여 수작업과 자동 분류를 혼합하는 하이브리드 구축방법을 통하여 지식베이스를 구축한다.

그림 1은 정답문서집합 기반 정보검색 시스템의 구조를 보여준다. 정답문서집합 기반 검색은 기본적으로 입력된 자연어 질의를 분석하여 얻어지는 개념어와 속성

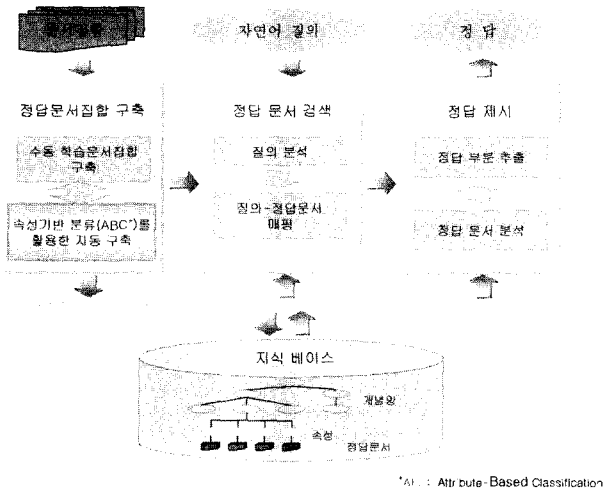

그림 1 정답문서기반 정보검색

의 정보를 통해 기 구축된 지식베이스로부터 정답문서 집합을 찾는 과정이다. 이 때 사용자에게 검색된 정답문서집합의 문서를 개념어와 속성의 특징으로 분석하여 질의의 의도와 일치하거나 가장 유사한 부분을 찾아서 정답으로 제시한다. 예를 들어 설명해 보면, 사용자가 "엔젤투자란 무엇이며, 엔젤투자시 유의할 점은 무엇인가?"라는 질의를 입력했을 때, 사용자가 원하는 개념은 "엔젤투자"이고 궁금한 정보는 "정의"와 "주의사항"이라는 속성임을 분석하여 이에 해당하는 문서를 제시한다. 개념어와 속성에 대한 설명은 2.2절의 지식베이스의 구성요소에서 설명하기로 한다. 본 논문에서 제안하는 속성 기반 분류 방법은 그림 1의 정답문서집합을 자동으로 구축하는 과정에 활용된다.

2.2 지식베이스

본 논문에서 구축하고자 하는 지식베이스는 크게 3부분, 개념망과 속성, 속성에 따른 정답 문서로 구성되어 있다. 일반적으로 사용자들이 인터넷 검색을 통해 요구하는 정보는 인간의 지적 자원인 개념망 상에서의 특정 개념과 그 개념에 대해서 알고 싶어하는 정보의 종류로 매핑될 수 있다[9]. 개념망은 사전에 등재된 표제어를 대상으로 사전의 뜻풀이를 기준으로 개념적인 상하관계를 연결함으로써 인간의 공통적인 개념체계를 표현하고 있다. 이때 특정 개념에 대해 사용자가 알고 싶어하는 정보(information need)의 종류, 즉 개념어의 의미적인 특징을 나타내는 분류 항목을 "속성(attribute)"이라고 정의한다. 속성이란 인터넷에 존재하는 수많은 정보 중에서 그 개념이 사용되고 있는 영역을 의미하는 것으로, 예를 들면 "엔젤투자"라는 개념에 해당하는 속성으로는 "문제점", "전략", "전망", "주의사항" 등이 있을 수 있다. 즉 속성이란 주어진 개념어에 대해 사용자가 궁금해 할 만한 정보의 종류로써 각 개념어마다 다르게 정의된다. 속성은 개념망과 인터넷 상의 정보를 연결시켜 지식베이스를 구축하는 매체로 다음과 같은 특징을 갖는다.

- 개념에 해당하는 문서집합을 대표한다. → 문서를 나누는 기준이 된다.
- 개념어의 부류에 따라 다르다. → 비슷한 개념어는 비슷한 속성을 갖는다.

본 논문에서는 속성이 특정 개념과 관련된 문서를 나누는 기준이 된다는 특징을 문서분류에 적용하여 속성분류라는 새로운 관점의 분류 모델을 제시하고, 이를 정답문서기반 정보검색을 위한 지식베이스 구축 과정에 활용하고자 한다. 그림 2는 개념망 상의 계층에 따른 개념어의 부류를 나타낸 그림으로 "외국인 투자", "공공투자", "분산투자", "엔젤투자" 등은 같은 부류에 속한 개

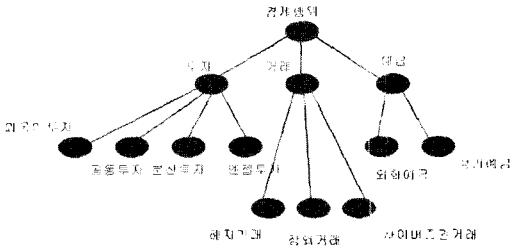

그림 2 개념망 상의 계층에 따른 개념 부류

표 1 개념에 의해 정의된 속성

개념어 \ 속성	가입대상	관련기관	종류	목적	문제점	장점	전략	전망	절차	정의	정책	현황	주의사항	시험문제	상담사례	금융상품	속성수
경제행위	O		O		O		O			O		O	O				7
투자		O			O		O	O	O	O	O	O	O			O	10
엔젤투자	O	O		O	O	O	O	O	O	O	O	O					11
공동투자					O		O	O		O	O	O	O				7
분산투자				O	O	O				O				O		O	6
외국인투자	O	O	O		O		O	O	O	O	O	O	O		O		12
거래					O	O	O			O		O	O	O			7
헤지거래			O	O	O	O	O			O		O	O	O	O		10
장외거래					O	O	O			O		O	O	O			7
사이버증권거래		O			O	O	O	O	O		O	O	O				9
예금	O		O	O		O		O		O		O	O		O	O	10
외화예금	O		O	O		O		O				O	O		O	O	9
정기예금	O		O			O				O		O	O		O	O	8
속성할당 개념수	6	1	6	5	10	9	9	7	1	11	5	12	11	4	5	5	

념어들이다. 표 1은 개념망에 정의된 개념어가 갖는 속성, 즉 <개념, 속성> 쌍을 나타내는 것으로, "외국인 투자"에 대한 내용을 포함하는 문서들은 "정의"나 "전략", "주의사항"이라는 범주로 분류될 수 있다. 또한 위에서 나열한 투자의 자식 노드들이 비슷한 속성을 갖고 있음을 보여준다. 즉, "투자"의 자식 노드(child node)인 "외국인 투자", "공동투자", "분산투자", "엔젤투자"는 같은 부류로 비슷한 속성을 공유하게 된다.

그림 2와 표 1을 종합해 보면, 개념어에 속하는 문서들은 속성으로 구분되고, 같은 부류의 개념군들은 비슷한 속성을 공유하게 된다. 예를 들면, 예금류에 해당하는 개념군은 "금융상품", "상담사례" 등의 속성을 공유하고, 투자류의 개념어들은 "전망", "정책" 등의 속성을 공유한다. 본 논문에서는 이처럼 비슷한 속성을 공유하는 개념어들의 관계를 "유사속성관계(α-relation)"라 정의하고, 이를 ETRI 개념망에 기술하였다[1]. 현재 구축되어 있는 ETRI 경제 개념망은 15,000여개의 개념어로 구성되어 있는데, 이렇게 많은 개념어마다 관련된 속성을 정의하는 것은 거의 불가능 한 일이다. 그러므로 "유사속성관계"를 통해 대표되는 개념어에 해당하는 속성만을 정의함으로써 이러한 문제를 해결할 수 있다. 본 논문에서는 "유사속성관계"를 활용하여 적은 양의 학습문서를 구축하고도 전체 지식베이스 구축이 가능하도록 하는 방법을 제안한다.

3. 정답문서집합 구축을 위한 속성 기반 분류 방법

3.1 속성기반분류 모델

일반적인 자동문서분류는 미리 구축된 정교한 문서집합을 통해 분류기를 구축하고, 한번 구축된 분류체계로 대량의 문서집합을 할당하기 때문에, 학습시 사용된 학습문서집합의 성격에 따라 분류기의 성능이 많은 영향을 받게 된다. 그러나, 본 논문에서 구축하고자 하는 지식베이스의 경우에는 정의된 분류체계 즉 속성이 수시로 변하는 특성을 가지며, 또한 이들 범주가 내용상으로 차이가 작은 것도 존재한다. 또한, 정교한 학습 문서를 위해서는 수작업으로 학습 문서를 구축해야 하기 때문에, 전체 분류 대상 문서에 대해서 학습을 위한 문서집합의 양이 절대적으로 작을 수밖에 없다. 이러한 문제들은 지식베이스의 자동구축의 성능을 저하시키는 역할을 하게 되므로, 성능을 보완하기 위한 기법이 필요하다. 본 논문에서는 자동분류를 통한 정답문서 구축의 성능을 높이기 위하여 다음과 같은 특징을 활용한다.

- 속성 특징의 활용 : 속성의 특성을 반영하는 단서(clue) 패턴(pattern)
- 속성 변별력이 없는 단어 선별
- 개념망의 "유사속성관계(α relation)" 활용

속성 특징 활용과 속성을 구별하는데 도움이 되지 않는 단어를 선별하는 과정은 분류기의 성능을 향상시켜 자동구축의 신뢰도(effectiveness)를 높이고, "유사속성관계"는 적은 양의 학습문서를 구축하고도 전체 개념망에 대해 분류가 가능하도록 자동구축 대상의 범위(coverage)를 보장한다.

3.1.1 속성 특징 활용

각 개념어에 해당하는 문서들은 그 활용 분야와 사용자의 요구에 따라 속성으로 분류된다. 그러나, 속성은 자동분류 관점에서 보는 일반적인 분류체계와는 다른 성격을 갖는다. 일반적인 분류체계, 예를 들어 주제별 분류체계는 각 범주들이 문서를 구성하는 핵심어들에 의해 내용상으로 뚜렷이 구분되는(disjoint) 성격을 갖는 반면, 정답문서 분류체계에서의 속성은 그 범주가 내용상으로 구별되는 것이 아니라 개념어의 쓰임에 따라 구별된다. 그러므로, 각 범주간 중복되는 문서가 학습문서로 제시되거나, 문서에 나타나는 용어만으로는 해당 범주를 표현하지 못하는 경우가 발생한다.

본 논문에서는 이러한 문제점을 해결하기 위해 수작업으로 구축된 학습문서집합을 대상으로 각 속성의 특징을 표현하는 규칙(clue word)을 정의하고, 이를 기계학습(machine learning)을 통해 구축된 분류기에 반영한다. 속성 규칙은 단어(word), 구(phrase), 문장(sentence), 절(paragraph)로 나뉘어 작성되며, 각 규칙 패턴마다 다른 가중치로 기계학습 분류기에 반영된다. 구축된 속성 규칙은 전체 83개의 속성에 대해 단어 패턴 1,139개, 구 패턴 480, 문장 패턴 499개로 각 속성당 평균 25개의 규칙으로 구축되었고, 이는 기계학습을 통해 구축된 용어 기반 분류기(centroid)의 오류를 보정함으로써 신뢰도(effectiveness)를 향상시키는 역할을 한다. 속성의 특징에 따라 어떤 속성은 속성 특징을 활용한 규칙 기반 분류(Rule-based classification)가 우월한 경우가 있고, 어떤 속성은 기계학습기반 분류가 유리한 경우가 있다. 그러므로, 본 논문에서는 규칙 기반 분류와 기계학습 기반 분류를 혼합하는 방법을 사용하였으며, 이들의 반영비율은 속성마다 다르게 설정하여 사용하였다.

1) ETRI 경제 개념망에서는 같은 부모를 갖은 형제 노드 개념어들과 "동의어 관계"에 있는 개념어들이 서로 같은 속성을 공유하는 것으로 나타났다.

표 2 같은 "유사속성관계" 그룹의 속성 분포

	정의	정책	지급	상담사례	문제점	종류	목적	현황	규정	장점	계산법	협상	속성수
성과급	O	O	O	O	O	O	O	O	O	O			10
시간급	O		O	O					O	O	O		6
기본급		O	O						O			O	4
직무급	O		O		O	O	O	O	O	O			8
능력급	O	O	O		O			O	O	O			7
상여금		O	O	O	O			O	O		O	O	8

3.1.2 속성 변별력이 없는 단어 선별

속성 분류기가 <개념, 속성> 쌍에 대한 범주를 학습할 때, 개념의 의미를 전달하기 위해 사용된 단어는 속성을 구별하는데 도움이 되지 않는다. 본 논문에서 구축하고자 하는 정답문서집합은 먼저 특정 개념에 해당하는 문서를 수집한 후, 수집된 문서에 대해 속성을 분류하는 과정을 거친다. 이 같은 방법으로 특정 개념어에 해당하는 문서를 수집한 후 속성을 정의하였기 때문에, 구축된 학습문서집합의 특성상 특정 개념어와 관련된 단어가 속성과는 관계없이 그 개념어에 속하는 모든 문서에 나타날 수 있다. 그러나, 이러한 단어가 같은 개념에 속하는 여러 속성들에 걸쳐 출현한다면, 각 속성의 특징을 흐리게 할 뿐 아니라 속성간의 차이를 모호하게 한다. 따라서, 속성 분류의 정확도를 높이기 위해 학습문서에서 특정 개념어와 관련된 높은 빈도(frequency)로 여러 속성에 걸쳐 출현하는 단어를 선별하여 이를 제거시켜주는 단계가 필요하다.

3.1.3 "유사속성관계" 활용

지식베이스 구축에 자동문서분류 기법을 활용하기 위해서는 개념망에 존재하는 모든 노드, 즉 모든 개념어에 해당하는 학습문서집합이 구축되어야 한다는 조건이 선행되어야 한다. 그러나, 본 논문에서 대상으로 하는 약 만5천여 개에 달하는 경제 개념어나 5만여 개의 일반 개념어에 대한 학습문서집합을 구축하는 일은 너무 많은 노력을 요할 뿐 아니라 구축된 학습 문서 집합이 웹 문서의 특성상 시간이 지남에 따라 그 활용도가 떨어지게 된다. 본 논문에서 구축할 경제 개념망에서의 <개념, 속성> 쌍은 대략 250,000[2]개로, 이를 수작업으로 구축하는 것은 거의 불가능한 일이다.

이러한 문제점을 해결하기 위해 학습문서집합 없이 자동으로 정답문서를 할당하기 위한 방법이 필요한데, 본 논문에서는 *ETRI* 개념망에 정의된 "유사속성관계(α-relation)"를 활용한다. "유사속성관계"란 비슷한 속성을 가지는 개념어들의 집합을 의미하는 것으로, 개념망상에서 같은 부모를 갖는 형제(sibling)들은 "유사속성관계"로 이루어질 수 있는데, 표 2의 예를 들면 "임금"을 부모로 하는 형제 개념어들("성과급", "시간급", "기본급")은 모두 비슷한 속성을 갖는다.

그러므로, "유사속성관계"로 정의된 그룹 중 대표되는 개념어에 대해서만 학습문서집합을 구축하고 이를 통해 속성 분류기를 생성한다면, 나머지 개념어들에 대해서도 속성에 해당하는 문서를 할당할 수 있게 된다. 이것은 분류기 생성을 위한 노력을 줄이는 역할과 개념망의 자동 할당 범위(coverage)를 넓히는 효과를 가져온다.

학습 문서 집합을 구축하기 위해서는 "유사속성관계"에 있는 개념 그룹 중에서 대표되는 개념어를 선정해 그와 관련된 내용의 문서를 수집한다. 수집된 문서집합을 클러스터링[10]하여 대표 개념어가 가질 수 있는 속성을 정의하는 기반(seed)로 제공함으로써, 학습 대상 개념어의 <개념, 속성> 쌍과 이에 해당하는 문서집합을 정의한다. 구축된 문서집합을 학습해 초기 속성 분류기를 구축하고, 이를 통해 "유사속성관계"에 있는 다른 개념어에 해당하는 문서에 속성을 부여한다. 이때 속성이 부여되지 않은 문서는 미할당 문서로 남겨지고, 이를 다시 클러스터링함으로써 새로운 속성을 선별하게 된다. 새로운 속성으로 부여된 문서집합은 초기 학습문서집합에 추가되어 재학습이 이루어진다. 이러한 과정을 통해 초기 속성 분류기가 점진적으로 보다 정확한 분류기로 개선된다. 이 기법은 "같은 부류의 개념군들은 비슷한 속성을 공유하게 된다"라는 가정을 바탕으로 한 것으로, 이러한 가정이 모든 개념어에 대해서 성립하는 것은 아니지만 학습문서를 모두 구축할 수 없는 대용량 문서분류에서는 충분히 유용하다. 표 2에서 "성과급"과 "시간급"에 대해서만 속성을 정의하더라도 나머지 개념어는 "협상"을 제외한 모든 속성이 분류되며, "기본급"까지 속성을 정의하면 모든 속성에 대해서 분류가 가능하다.

2) 14,700(정답문서 할당 대상 개념어 수) * 18 (한 개념어 당 평균 할당 속성 수)

3.2 정답문서집합 지식베이스 구축

속성 분류 모델을 활용해 지식베이스를 구축하기 위해서는 분류할 속성을 정의하고 학습하는 학습 단계(training stage)와 문서를 할당하는 분류 단계(assigning stage)가 필요하다. 학습 단계는 다시 학습문서집합을 구축하고 분류할 범주, 즉 속성을 정의하는 과정과 구축된 학습문서에 의해 분류기를 생성하는 과정으로 나뉘는데, 자세한 과정은 다음과 같다.

- Step 1: 학습문서집합 생성(Human-generated training data)

  학습 데이타를 구축하기 위한 개념어를 추출한다. 이때 전체 지식베이스 구축에 영향을 가장 많이 미치는 노드를 선택하는데, 여기서는 "유사속성관계"에 있는 개념군의 대표 개념어를 선택한다. 선택된 개념에 적합한 문서집합을 클러스터링하여[10] 해당 개념이 가질 수 있는 속성을 정의하고, 정의된 속성의 특징을 나타내는 단서가 되는 특징 단어나 구, 기타 요소를 추출한다.

- Step 2: 기계학습 분류기 생성(Machine learning classifier)

  Step 1에서 생성된 학습 문서집합을 통해 기계학습 분류기 C_m을 생성한다. 자질 추출 과정 중 개념어의 특성을 나타내는 자질은 제외된다.

- Step 3: 규칙기반 분류기 생성(Rule-based classifier)

  Step 1에서 추출된 속성 특징(rule)을 활용해 규칙기반 분류기 C_r을 생성한다. 이때 step 2에서와 마찬가지로 개념어의 특성을 나타내는 자질은 제외된다.

- Step 4: 속성 분류기 생성(ML + Rule-based = hybrid classifier 생성)

  Step 2, 3에서 생성된 분류기 C_m과 C_r을 다음과 같이 병합해 속성 분류기 C을 생성한다.

$$C = \alpha \cdot C_m + \beta \cdot C_r \quad (1)$$

여기서, α와 β는 기계학습 분류기와 규칙 기반 분류기의 신뢰도를 의미하는 것으로 통합 분류기에 대한 반영 비율을 말한다. 속성의 특성에 따라 기계학습 분류기의 반영비율이 높은 경우가 있고, 규칙 기반 분류기의 신뢰도가 높은 경우도 있다. α와 β는 내부실험 결과를 통해 자동으로 결정된다.

위 과정을 통해 학습된 속성 분류기를 통해 지식베이스를 구축하는 방법은 크게 2가지로 나뉜다. 그 중 하나는 이미 학습된 개념어에 해당하는 문서를 정의되어 있는 속성에 따라 분류하는 경우이고, 다른 하나는 새로운 개념어에 대해서 속성을 정의하고 문서를 할당하는 경우이다. 전자의 경우에는 구축된 분류기를 적용하여 문서를 할당하는 점에서 일반 분류기를 활용하는 과정과 동일하다. 그러나, 학습된 개념어가 아닌 새로운 개념어의 경우에는 다음 부스팅(boosting) 과정[8,11]을 통해 학습문서집합을 자동으로 구축한 뒤, 이를 점진적으로 학습하여 분류에 활용하게 된다. 다음은 정답문서집합을 구축하는 과정이다.

- Step 1: 대상문서집합(D_c)수집

  메타검색기와 문서 필터링[12]을 활용해 주어진 개념어(concept)에 해당하는 문서를 수집하여 대상문서집합 D_c을 구성한다.

- Step 2: D_c에 있는 문서에 속성 부여

  속성 분류기를 활용해 대상문서집합(D_c)에 속성을 부여한다. 이때 속성을 부여할 대상 개념어가 미리 학습된 개념어이거나 "유사속성관계"에 있는 경우에는 미리 정의된 속성 집합을 대상으로 분류하고, 만약 개념망에 미리 정의된 속성 집합이 없는 경우에만 모든 속성을 대상으로 분류하는 것이 정확성과 효율성 측면에서 유리하다.

- Step 3: 모든 개념어에 반복 수행

  Step 1을 반복 수행함으로써 모든 개념어에 대한 점진적인 분류기 생성이 가능하다. 이를 통해 작은 학습문서 집합을 구축하고도 전체 개념어에 해당하는 정답문서를 할당할 수 있게 되어 지식베이스 구축의 자동화를 꾀할 수 있다.

4. 실험 및 분석

정답문서집합 자동 구축을 위한 속성 기반 분류 방법 적용 실험은 다음 두 가지 목적으로 수행하였다.

- 규칙기반 분류기와 기계학습 분류기를 통합한 방법의 효과를 살펴본다.
- "유사속성관계"를 활용한 지식베이스 자동 구축 방법의 유용성을 보인다.

본 논문에서는 속성 분류를 위해 기계학습 방법으로 베이지언(Naive Baysiean) 모델을 사용하며, 자질 추출을 위해서는 EMIM(Expected Mutual Information Measure)을 사용한다[13]. 현재 구축된 지식베이스는 경제 관련 분야의 14,700개 개념어와 83개의 속성, 약 140만 문서로 구성되어 있다. 각 개념어에 대해서 수집된 문서는 평균 43.4개이고, 각 개념어 당 평균 25개의 정답 문서가 할당되어 있으며, 이들 정답문서는 평균 18개의 속성에 분류되어 있다. 실험을 위해 약 2% 정도의 문서를 수동 구축하였으며 이중 4,950개의 문서는 학습

문서로 사용하였다. 실험에 사용한 문서는 4,599개이고, 개념어 120개, 속성 83개를 사용하였다. 실험에 대한 평가 방법으로는 재현율(recall)와 정확율(precision)을 함께 이용하여 성능을 나타내는 F-score[14]를 사용한다.

4.1 통합 방법(Hybrid Method) 효과 실험

통합 방법 효과 실험[실험 1]의 주된 목적은 지식베이스를 자동으로 구축하기 위한 속성 분류기의 성능을 평가하는 실험으로, 속성 분류기의 최적의 조건을 찾는데 있다. 실험은 기계학습 분류의 성능을 비교 기준(baseline)으로 하고, 규칙기반 분류, 이를 병합한 분류기의 성능을 비교하기로 한다. 통합 분류기의 성능을 평가하기 위해 식 (1)에서 α와 β를 모두 1로 한 경우, α와 β를 각각 2와 5로 설정한 경우, 속성에 따라 α와 β를 다르게 설정한 경우로 나누어 실험하였다. [실험 1]을 통해 찾은 최적치는 [실험 2]에서 사용된다.

표 3의 실험 결과에 나타난 것처럼, 규칙만 활용한 경우(.3596)에 비해서는 기계학습 결과만 활용한 경우(.5193)가, 기계학습 결과만 활용한 경우에 비해서는 이 둘을 복합적으로 활용한 경우(.5789)의 성능이 우수함을 알 수 있다. 이는 속성을 표현한 규칙이 기계학습을 통해 구축된 용어 기반 분류기(centroid)의 오류를 보정함으로써 정확도(effectiveness)를 향상시키는 역할을 하고 있음을 의미한다.

한편, 기계학습의 결과에 비해 규칙과 기계학습을 결합한 방법(hybrid categorization)의 결과의 성능 차(11.4%)가 크게 두드러지지 않는 이유는 규칙만 활용한 결과와 기계학습만 활용한 결과가 상쇄하는 경우가 발생하기 때문이다. 예를 들면, "정의"라는 속성은 규칙만 활용한 경우가 월등하게 나타나고, "상품"이라는 속성은 기계학습 결과만 활용한 경우가 월등하게 나타난다. 그러므로, 규칙과 기계학습의 결과를 적절히 조합하는 방안이 필요하다. 본 논문에서는 규칙과 기계학습의 반영 비율을 내부 학습 결과에 따라 α와 β를 다르게 설정하여 사용하였다. 실험한 결과, 비교 기준(.5193)에 비해 최고 24.7%(.6478)의 성능향상을 보였다.

4.2 "유사속성관계(α-relation)"의 활용 실험

"유사속성관계" 활용 실험[실험 2]은 개념망에 정의된 "유사속성관계"를 활용하여 점진적으로 분류기를 생성하는 방법의 효과를 검증하는 실험으로, 지식베이스 자동 구축 방법의 효용성을 증명하는 실험이다. 실험 방법은 문서집합을 모든 속성 집합(U)을 대상으로 분류하는 방법과 개념망에 정의된 "유사속성관계"를 참조하여 해당 속성 집합(A)에만 문서를 할당하는 방법으로 나누어 실험한다. 마지막으로 두 번째 방법을 확장하여 해당 속성 집합에 대해 분류를 실행한 후, 미할당 문서로 남겨진 문서에 대해 추가 속성 집합(A'=U-A)을 대상으로 분류해 보았다.

표 4는 학습하지 않은 개념어 9개에 대해 새로운 정답문서집합을 구축한 경우를 실험한 결과로, "유사속성관계"를 활용하지 않은 경우와 비교하고 있다.

표 3 [실험 1] 규칙기반, 기계학습기반, 통합 방법의 성능 비교

속성 할당 실험 방법	Precision	Recall	F-score	개 선
규칙기반 분류	.3502	.3990	.3596	- 30.75%
기계학습기반 분류	.4580	.5806	.5193	비교 기준(Baseline)
통합 방법(1:1)	.4291	.4847	.4569	12.0%
통합 방법(2:5)	.4941	.6637	.5789	+ 11.4%
통합 방법(α_i:β_i)	.6091	.6865	.6478	+24.7%

표 4 [실험 2] "유사 속성 관계" 활용에 따른 속성 할당 비교

속성 할당 실험 방법	대상 속성수	할당 속성수	Pre.	Recall	F-score	시간
모든 속성(U) 대상 (α-relations 미활용)	83	42	.5025	.4662	.4835	4
해당속성(A) 대상 (α-relations 활용)	21	19	.6020	.6696	.6358 (+31.4%)	1
해당속성(A) 대상 (α-relations 활용) + 추가 속성 부여	83	29	.5828	.6992	.6410 (+32.6%)	1.7

성능 평가를 위해 미리 수작업을 해본 결과, 9개의 개념어가 갖는 속성은 기존에 정의된 속성 21개에 새로운 속성 6개를 포함한 28개였고 정답문서 수는 579개였다. 표 4를 보면, "유사속성관계"를 활용하지 않은 경우(.4835)에는 현재 학습된 모든 속성(집합 U) 83개를 대상으로 분류한 결과 42개의 속성을 할당하고 있다. 반면, "유사속성관계"를 활용하여 기 정의된 속성(집합 A) 21개를 대상으로 분류한 경우(.6358)에는 19개의 속성을 할당하였으며 성능은 31.4% 향상되었다. 또한 해당 속성만을 활용한 경우가 그렇지 않은 경우에 비해 4배 정도 빠른 속도를 보임으로써, 제안한 분류기가 성능과 속도 면에서 모두 우수함을 알 수 있다.

한편, "유사속성관계"를 활용한 경우에는 기존에 정의되지 않은 새로운 속성 6개를 할당하지 못하는 문제가 발생한다. 이러한 문제점은 먼저 "유사속성관계"를 통해 속성을 부여하고, 임계치 기법(thresholding strategies) [15]을 활용해 분류 확률값이 임계치 이하로 나타난 문서는 미할당 문서로 정의한 후, 이에 대해 추가 속성(집합 A') 분류를 실시한다면 해결될 수 있다. 실험 결과(.6410) 모든 속성을 동시에 적용한 경우(.4835)에 비해 32.6%의 성능이 향상되었으며, 속도는 2.3배 이상 빨라졌다. 반면 해당 속성만을 할당한 경우와 성능 차이가 거의 없는데, 그 이유는 문서 분류 기법의 특성상 미할당 문서로 정의하는 경우가 드물어 추가 속성을 대상으로 분류할 기회가 적어지기 때문이다.

5. 결 론

본 논문에서는 사용자에게 보다 지능적인 정보를 제공하기 위하여 개념의 활용분야에 따른 속성 분류 기법이라는 새로운 분류 기법을 제안하고, 이를 활용해 지식베이스를 자동으로 구축하는 방안을 제시한다. 제안된 방법은 범주간의 구분이 유동적인 속성의 특성을 반영하기 위하여 속성 특징(clue)을 활용함으로써 분류 정확도를 높이고, 개념망에 정의된 개념들 사이의 관계를 참조함으로써 점진적인 분류기 생성을 가능하게 한다. 실험한 결과, 기계학습으로 생성된 분류기에 속성 특징을 반영한 경우에 24.7%의 성능 향상을 보였으며, 개념망에 정의된 지식을 활용한 경우가 그렇지 않은 경우에 비해 32.6%의 정확도 향상과 400%의 속도 향상을 얻을 수 있었다.

그러나, 제안한 속성 분류 방법은 개념망에 미리 정의된 지식을 활용한다는 점에서 분류기 구축과정에서 요구되는 정보의 확보가 어렵다는 제약이 있다. 또한 속성이라는 범주는 시간이 변함에 따라 그 정의가 달라지며 범주의 구별이 모호하게 되는 경우도 발생하게 된다. 향후 연구 방향으로는 이러한 속성의 특성을 분류기에 제때 반영하기 위한 방법이 필요하다.

참 고 문 헌

[1] Fabrizio Sebastiani, "Machine Learning in Automatic Text Categorization," ACM Computing Surveys, 34(1):1-47, 2002.

[2] Oh, H. J., Myaeng, S. H., Lee, M. H., "A Practical Hypertext Categorization Method using Links and Incrementally Available Class Information," Proc. of the 23rd annual international ACM-SIGIR '2000, pp 264-271, Athens, Greece, 2000.

[3] Yong-Bae Lee, Sung Hyon Myaeng, "Text Genre Classification with Genre-Revealing and Subject-Revealing Features," Proc. of the 25th annual international ACM-SIGIR '2002, pp 145-150, Tampere, Finland, 2002.

[4] Jeong-Mook Lim, Hyo-Jung Oh, Sung-Hyon Myaeng, and Mann-Ho Lee, "Improving Efficiency with Document Category Information in Link-based Retrieval," Proc. of the international Workshop on IRAL'99, 1999.

[5] Aks Jeeves™, http://www.askjeeves.com

[6] 장명길, 오효정, 장문수 외 3인, "의미기반 정보검색", 정보과학회지, 19(10):7-18, 2001년 10월.

[7] Andrew McCallum, Kamal Nigam, et al., "A Machine Learning Approach to Building Domain-Specific Search Engines," Proc. of the 16th IJCAI Conference, pp 662-667, 1999.

[8] Sanda Harabagiu, Dan Moldovan, et al, "FALCON: Boosting Knowledge for Answer Engines," Proc. of Text Retrieval Conference (TREC-9), November, 2000.

[9] Marius Pasca and Sanda M. Harabagiu, "The Informative Role of WordNet in Open-Domain Question Answering," Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resource, pp 138-143, CMU, Pittsburge PA, June 2001.

[10] C. Aggarwal, S. C. Gates, P.S.Yu., "On the merits of using supervised clustering for building categorization systems," Proc. of the SIGKDD-99 Conference, 1999.

[11] Robert E. Schapire and Yoram Singer, "Boos Texter: A Boosting-based System for Text Categorization," Machine Learning, 39, pp 135-168, Kluwer Academic Publishers, 2000.

[12] 정용교, 신승은, 오효정, 장명길, 서영훈, "Answer set 자동구축을 위한 문서 필터링", 제14회 한글 및 한국어정보처리학회, pp. 253~258, 2002.

[13] David D. Lewis, "Representation and Learning in

Information Retrieval," Ph.D thesis, Dep. of Computer Science, Univ. of Massachusetts, 1992.
[14] Yiming Yang and Xin Liu, "A Re-examination Of Text Categorization Methods," Proc. of the 22th annual international ACM-SIGIR '1999, pp 42-49 Berkeley, USA, 1999.
[15] Yiming Yang, "A Study on Thresholding Strategies for Text Categorization," Proc. of the 24th annual international ACM-SIGIR '2001, pp 137-145, New Orleans, USA, 2001.

오 효 정
1998년 충남대학교 컴퓨터과학과(학사) 2000년 충남대학교 컴퓨터과학과(석사) 2000년~현재 한국전자통신연구원 휴먼정보검색연구팀 연구원. 관심분야는 문서자동분류, 정보검색, 자연어처리, 기계학습

장 문 수
1992년 고려대학교 전자전산공학과 졸업 1994년 고려대학교 전자공학과 석사 2001년 일본 동경공업대학 지능시스템과학전공 박사. 2000년~2003년 한국전자통신연구원. 2003년~현재 서경대학교 소프트웨어학과 전임강사. 관심분야는 정보검색, 자연어처리, 퍼지응용

장 명 길
1988년 부산대학교 계산통계학과 (학사) 1990년 부산대학교 계산통계학과 (석사) 2000년 충남대학교 컴퓨터과학과 (박사) 1990년~1998년 5월 시스템공학연구소 선임연구원. 1998년 6월~현재 한국전자통신연구원. 휴먼정보검색팀 팀장. 관심분야는 자연어처리, 정보검색, 생물정보학