자동분류를 이용한 정답문서집합 구축

장문수[0] 오효정 장명길
한국전자통신연구원 언어공학연구부
{cosmos, ohj, mgjang}@etri.re.kr

Construction of Answer Sets using Automatic Categorization

Moon-Soo Chang[0] Hyo-Jung Oh Myung-Gil Jang
Linguistic Engineering Department, ETRI

요 약

최근의 인터넷 정보검색은 방대한 정보의 수용과 지능적이고 개인화된 검색 결과 요구라는 사뭇 상반된 요구를 만족시켜야 한다. 기계적으로 키워드를 매칭시켜 나오는 문서를 사용자에게 맡기는 식의 검색은 더 이상 환영을 받지 못한다. 우리는 이러한 추세에 맞추어 의미기반 정보검색에 필요한 개념망과 정답문서집합으로 구성된 지식베이스를 제안한 바 있다. 본 논문에서는 방대한 구조의 개념망과 연결되는 정답문서집합을 유동적인 인터넷 환경에 적용하기 위해 자동으로 구축하는 시스템을 제시한다. 자동구축은 문서분류(document categorization) 기술을 활용하여 개념어에 문서를 할당하는 방법과 속성에 문서를 할당하는 방법으로 나누어 이루어진다. 제시한 방법은 실험을 통하여 기본적인 속성 할당에는 상당한 효과가 있는 것으로 판단되었고, 일부 미할당 문서에 대해서는 클러스터링과 같은 다른 알고리즘이 필요하다.

1. 서론

인터넷에 웹이 등장한 이래로 엄청난 양의 정보가 인터넷 상에 쏟아져 나왔다. 이와 함께 인터넷에서의 정보검색은 과거의 기술문헌 검색으로부터 꾸준한 발전을 거듭하여, 오늘날에는 수많은 상용 검색 시스템이 개발되어 인터넷 시장을 주도해왔다. 이러한 정보검색 기술은 10억 페이지 이상의 방대한 웹 문서로부터 사용자가 원하는 정보를 제공하기 위하여 크게 두 가지 방향으로 발전해왔다. 하나는 상용 인터넷 검색엔진의 효시라고 할 수 있는 야후에서 시작한 것으로, 수작업으로 필요한 정보만을 분류하는 디렉토리 검색 방법이고, 또 하나는 과거 소규모 문헌 검색에서부터 사용되어 오던 키워드 색인 검색 방법으로, 대부분의 상용 인터넷 검색엔진에서 웹 페이지 검색 방법으로 사용하고 있다.

이들 기술은 전문가의 수작업과 웹 로봇에 의한 자동 구축이라는 방법론적 차이와 추구하는 바가 정보의 질과 양으로 양분되어 있기 때문에 좀처럼 융화될 수 없는 기술로 제각각 발전해왔다. 그러나, 최근에 와서는 보다 정확하면서 빠른 정보, 보다 개인화된 정보가 요구되는 추세가 이어져, 보다 지능적인 정보를 제공하는 google과 같은 검색엔진이 각광을 받고 있다. 즉, 많은 정보를 제공하면서도 단지 기계적인 색인 기술이 아닌 전문가 혹은 사용자의 노하우가 적용된 기술이 사용자에게 호응을 받게 된 것이다.

우리는 이러한 추세에 맞추어 방대한 인터넷 정보를 인간의 지적 자원인 개념망과 연결시켜 하나의 지식베이스로 구축함으로써 보다 지능적인 정보검색이 가능하게 하는 방법을 제시하였다[1]. 이 지식베이스는 개념망과 속성으로 연결된 정답문서집합으로 구성되어 사용자의 질의를 개념망을 이용하여 해석하고, 해석결과와 정답문서집합을 연계하여 정보를 제공할 수 있게 한다. 본 논문에서는 상기한 지식베이스가 실제로 방대한 인터넷 정보를 지속적으로 수용할 수 있도록 하기 위해 자동분류기법(automatic categorization)을 이용한 정답문서집합의 자동구축 방안을 제안한다.

2장에서는 지식베이스의 구조와 활용에 관하여 살펴보고, 3장에서 제안된 지식베이스에 적합한 자동분류 방법과 이것을 적용한 구축 시스템에 관하여 설명한

다. 4장에서는 제안된 방법에 의한 실험결과를 제시하고, 5장에서 결론과 개선점을 밝힌다.

2. 연구배경

우리가 제안한 지식베이스는 의미기반 정보검색 소프트웨어 개발[2]의 일환으로 구축된 것으로, 사용자 질의의 의미적인 분석을 토대로 사용자의 요구에 근접한 문서를 제공하기 위하여 지식 구조와 웹문서를 직접 연결해보자는 의도에서 출발하였다. 여기서 지식구조는 한국어 명사 개념망인 ETRI 개념망이며, 여기에 연결되는 웹문서를 정답문서집합[1]이라고 한다. 정답문서집합은 개념망 상의 개념어가 웹 상에서 어떤 주제에 대해서 사용되고 있는가에 따라 웹문서를 분류해 놓은 문서집합이며, 이들을 분류하는 항목을 속성이라고 한다. 따라서, 속성은 개념어의 실생활에서의 활용범위를 나타내며, 정답문서집합의 분류 기준이 된다.

ETRI 개념망은 사전의 뜻풀이를 중심으로 한국어 명사의 의미적인 상하관계를 나타낸 개념분류 체계이며, 국어학 전문가의 지식을 기반으로 구축되었다. 그리고, 정보검색이라는 실용성에 목적을 두고 개발되어 1차적으로 경제분야를 중심으로 개념관계를 정의하고 있다[2]. 따라서, 경제분야에 있어서 국어사전에 없는 경제용어들도 상당수 포함하고 있으며, 경제에 관한 사용자 질의에 나타나는 대부분의 경제용어를 포괄하고 있다. 그림 1은 개념망과 정답문서집합으로 구성되는 지식베이스의 구조를 도식화해서 나타내었다.

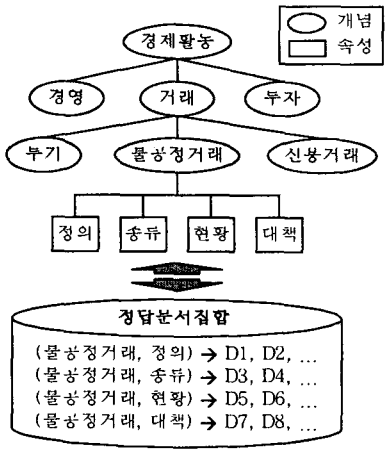

그림 1. 지식베이스 구조

그림 1에서 "불공정거래"라는 개념어는 "거래"라는 상위어를 가지며, "거래"는 "경제활동"에 속한다. 그리고, "불공정거래"는 "정의", "종류", "현황", "대책" 등의 속성을 가진다. 각 속성은 복수개의 정답문서를 가지게 되고, "불공정거래에 대한 대책은 무엇인가?"와 같은 질문에 대해 정답을 제공하게 된다.

위에서 설명한 지식베이스를 기반으로 하는 정보검색은 AskJeeves[3]의 정답문서 제공과 유사한 양식으로 사용자의 자연어 질의를 입력받아 정답문서집합을 제공한다. 질의/응답(Question-Answering)의 기초 기술을 이용하는 AskJeeves는 미리 전문가들에 의해 수작업으로 분야별 정답문서를 구축하고, 사용자 질의문의 키워드와 패턴으로부터 분야를 추정하고 해당 분야의 정답문서를 제공하고 있다. 그러나, 우리는 개념망을 이용하여 질의문으로부터 개념어와 속성을 추출하여 정답문서집합를 찾는다[2]. 그리고, 개념망의 상하위어 정보와 그에 따른 속성 정보를 이용하여 사용자 질의와 유사관계에 있는 정보도 함께 제공함으로써 사용자의 만족도를 높인다. 따라서, 이러한 정보를 급변하는 방대한 웹페이지 문서에 대해서 적용하기 위해서는 대량의 정답문서집합의 구축 방법이 제안되어야 한다.

3. 정답문서집합 구축 시스템

2장에서 언급한 바와 같이 정답문서를 연결할 ETRI 개념망은 약 2만개의 엔트리를 가지는 방대한 양의 구조이다. 지식베이스를 구축하는 과정은 각 개념어에 해당하는 문서집합을 수집하고, 문서 집합의 특성을 나타내는 속성을 추출하는 작업이 수반된다. 그러나, 이를 모두 수작업으로 구축한다는 것은 매우 어려운 일일 뿐 아니라, 시간이 지남에 따라 사용자가 원하는 정보가 달라지고 이에 따라 웹 상의 문서는 시시각각으로 변화하고 있기 때문에 해당 정답 문서 집합을 재구축해야 하는 경우가 자주 발생한다[3]. 그러므로, 정답문서집합은 자동으로 구축되고, 정보의 빠른 갱신이 보장되어야만 제안된 지식베이스가 정보검색에서 가치를 가지게 된다.

---

[1] 정답문서집합은 사용자 질의가 나타내는 의도에 적합한 정답을 포함하는 문서라는 뜻에서 명명된 것이다.

[2] 현재 ETRI에서는 금성출판사 뉴에이스 사전의 표제어를 기준으로 10만 단어 수준의 일반단어를 대상으로 개념망을 구축 중이다.

[3] 상용 검색엔진에서는 웹페이지 검색에 사용되는 문서를 일주일에서 한달 간격으로 다시 읽어온다.

3.1 정답문서집합 구축

정답문서집합을 구축하기 위해서는 각 개념어에 대해서 정답문서의 후보를 수집하고, 이들을 분석하여,개념어의 속성을 정의해야 한다. 그리고, 각 속성들에 포함될 정답문서를 결정한다. 이러한 과정은 다음과 같은 일련의 작업을 통해서 이루어진다.

① **문서 필터링** : 특정 개념어와 관련있는 문서를 수집하여 관련도가 높은 문서를 추출한다.
② **속성 정의** : 수집된 문서를 분석하여 주제별로 속성을 정의한다.
③ **정답문서 분류** : 수집된 문서들에 대해서 각 속성별로 문서를 분류한다.
④ **속성 특징 추출** : 각 속성에 할당된 문서를 분석하여, 해당 속성의 특징을 나타내는 속성의 단서(clue), 즉 특징 단어나 구, 기타 요소를 추출한다[4].

정답문서집합을 구축하기 위해 우선적으로 해야 할 일은 해당 개념어와 관련있는 문서를 수집하는 것으로, 우리는 이를 위하여 메타검색기를 활용한다. 그러나, 메타검색기를 통해 수집된 문서는 개념어와 내용상으로 일치하지 않는 문서도 상당수 존재하므로 메타검색 결과에서 개념어에 해당하는 문서만 추출하는 과정이 필요하다. 이것은 문서 필터링(document filtering)과 관련된 내용으로 문서 분류(document categorization) 기술을 활용하여 자동으로 해결할 수 있다.

개념어 별로 수집된 문서를 보고 속성을 정의하는 과정은 문서의 전반적인 내용을 파악하고 통합하는 종합적인 지식이 필요한 부분이므로 현재 기술로는 자동으로 속성을 정의할 수는 없다. 그러므로, 속성은 수집된 문서의 내용을 분석하여, 개념어의 쓰임에 따라, 즉 사용자가 해당 개념어에 대해 알고 싶어하는 분야에 따라 수작업으로 정의한다. 그러나, 모든 개념어에 대해서 수작업으로 속성을 정의할 수는 없으므로, 이것에 대한 대책은 3.2절에서 논의한다.

속성이 정의되면 각 속성에 대해서 정답문서를 분류하게 된다. 앞에서 설명한대로 개념어는 그 쓰임에 있어서 개념망 상의 상위어(parent)에 의해 제약을 받고, 유사하거나 동의 관계에 있는 개념어는 비슷한 쓰임을 갖는다. 다시 말해 같은 상위어에 해당하는 개념어들(sibling)이 가질 수 있는 속성은 유사하며 그 종류 또한 유한하다[5]. 그러므로, 각 개념어에 의해 필터링된

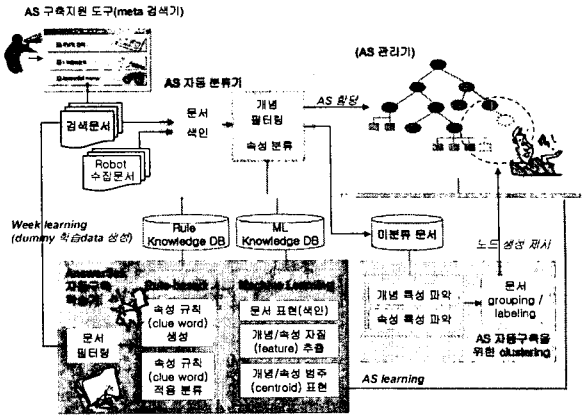

그림 2. 정답문서집합 자동구축 시스템

문서 집합을 정의된 속성 분류 체계(category)로 분류하는 방법을 통해 자동화를 할 수 있다. 또한 기 정의된 속성 분류 체계에 해당하지 않는 문서 집합에 대해 클러스터링(clustering)을 통해 새로운 속성을 제시하는데 도움을 준다. 우리는 이러한 과정을 처리하는 정답문서집합 자동구축 시스템을 제안한다. 그림 2는 시스템의 구성 및 처리과정의 개념도이다. 시스템은 크게 자동구축기와 미분류 문서 처리기로 나뉘고, 자동구축기는 학습기와 분류기로 나누어진다. 학습기와 분류기에 대한 설명은 3.2절에서 하고, 분류기의 성능은 4장 실험에서 설명한다[6].

3.2 정답문서 자동분류

일반적인 자동문서분류는 미리 구축된 정교한 문서집합을 통해 분류기를 구축하고 학습된 분류 체계에 대량의 문서 집합을 할당한다. 그러므로, 학습 시 사용한 문서 집합의 성격에 영향을 많이 받게 된다. 또한 한번 정의된 분류 체계에 대해 수정이 가해지지 않는다. 그러나, 본 논문에서 구축하고자 하는 정답문서집합의 경우에는 정의된 분류 체계(속성)가 수시로 변하는 특성을 가지며, 또한 이들 범주가 내용상으로 차이가 작은 것도 존재한다. 그리고, 정교한 학습 문서를 위해서는 수작업으로 학습 문서를 구축해야 하는데, 전체 분류 대상 문서에 대해서 학습을 위한 문서 집합

[4] 이 특징 요소들은 자동 분류 알고리즘에서 분류를 위한 휴리스틱 규칙으로 사용된다.
[5] 단어의 쓰임은 사용에 따라 무한히 달라질 수 있지만, 개념망 상의 단어는 상위어가 포괄하는 개념의 범위에 의해 그 쓰임이 제한되며, 또 정보검색이라는 실용적인 측면에서 웹문서는 사용자의 주요 관심 영역에서만 존재하므로 이것을 분류하는 속성의 개수는 많지 않다.
[6] 본 논문은 문서분류(document categorization)를 적용한 정답문서집합 구축에 관한 내용으로, 클러스터링을 이용하는 미분류 문서 처리기는 본 논문의 실험에서 제외되었다.

의 양이 절대적으로 작을 수 밖에 없다. 이러한 문제들은 정답문서집합의 자동구축의 성능을 저하시키는 역할을 하게 되므로, 성능을 보완하기 위한 기법이 필요하다. 본 논문에서는 자동분류를 통한 정답문서 구축의 성능을 높이기 위하여 다음 두 단계를 추가한다.

- 개념망의 KIND_OF 관계 활용
- 속성 특징의 활용 : 속성의 특성을 반영하는 단서(clue) 패턴

KIND_OF 관계는 학습문서 구축에 활용하여 자동구축 대상의 범위(coverage)를 보장하고, 속성 특징은 분류기의 성능을 향상시켜 자동구축의 정확도(precision)를 높인다.

가. KIND_OF 관계 활용

정답문서 구축에 자동문서분류 기법을 활용하기 위해서는 개념망에 존재하는 모든 노드, 즉 모든 개념어에 해당하는 학습 문서 집합이 구축되어야 한다는 조건이 선행되어야 한다. 그러나, 현재까지 구축된 2만여 개에 해당하는 개념어나 앞으로 구축할 10만 가량의 개념어에 대한 학습 문서 집합을 구축하는 일은 너무 많은 노력을 요할 뿐 아니라 구축된 학습 문서 집합이 시간이 지남에 따라 그 활용도가 떨어지는 일도 발생하게 된다. 이러한 문제점을 해결하기 위해 학습 문서 집합 없이 자동으로 정답문서를 할당하기 위한 방법이 필요한데, 본 논문에서는 ETRI 개념망에 정의된 KIND_OF 관계를 활용한다. 개념망의 상하위 관계는 KIND_OF 관계(예를 들어, 은행 - 시중은행)로 이루어져 있는데, 상위어에 연결되는 여러 하위어들 중에는 사용범위가 비슷한, 즉 비슷한 속성을 가지는 개념어(예를 들어, 약속어음, 견질어음)를 묶어줄 수 있다. 그러므로, 같은 상위어를 가지는 하위어 중에서 상당수는 그들 중 대표되는 개념어에 대해서만 학습 문서 집합을 구축하고 이를 통해 분류기를 생성한다면, 나머지 개념어들에 대해서도 정답문서를 할당할 수 있게 된다. 이것은 분류기 생성을 위한 노력을 감소하는 역할과 개념망의 자동 할당 범위(coverage)를 넓히는 효과를 가져온다. 표1은 KIND_OF 활용의 효과를 실험한 예로써, "임금"의 하위어들에 대한 속성 분포를 나타낸 것이다. 표1에서 "성과급"과 "시간급"에 대해서만 속성을 정의하더라도 나머지 개념어는 "협상"을 제외한 모든 속성이 분류되며, "기본급"까지 속성을 정의하면 모든 속성에 대해서 분류가 가능하다.

표 1. 같은 KIND_OF 그룹의 속성 분포

	정의	전망	정책	지급	상담사례	문제점	관련도서	종류	목적	판례	현황	규정	장점	계산법	강의계획서	협상	속성개수
성과급	O	O	O	O	O	O	O	O	O	O	O	O	O				13
시간급	O			O	O							O	O	O	O		7
기본급			O	O								O				O	4
직무급	O			O		O		O	O		O	O	O				8
능력급	O		O	O		O	O				O	O	O				8
상여금			O	O	O	O				O	O	O		O		O	9

나. 속성 특징 활용

각 개념어에 해당하는 문서들은 그 활용 분야와 사용자의 요구에 따라 속성으로 분류된다. 즉, 내용적으로 같은 주제를 가지는 문서들은 속성을 통하여 하나로 분류된다. 이러한 같은 유형으로 분류되는 문서들의 특징은 속성의 특성을 나타내도록 해야 하는데, 수작업으로 구축된 정답문서집합을 대상으로 각 속성의 특징을 표현하는 규칙으로 단서 패턴들을 정의하고 이를 기계학습(machine learning)을 통해 구축된 분류기에 반영하도록 한다. 구축된 속성 특징에는 단어(word)나 구(phrase), 절(clause)의 어휘 패턴들로 구성되는데, 각 속성 특징 패턴마다 다른 가중치로 기계학습 분류기에 반영된다. 이는 기계학습을 통해 구축되는 용어 기반 분류기(centroid)의 오류를 보정함으로써 자동분류의 정확도를 향상시키는 역할을 한다.

본 논문에서는 자동분류를 위해 베이지언(Naive Baysiean) 모델을 사용하며, 자질 추출을 위해서는 EMIM(Expected Mutual Information Measure)을 사용한다[4]. 그리고, 여기에 위에서 제안한 두 단계의 기법을 추가하여 자동 구축의 성능을 향상시키고자 한다. 제안한 정답문서집합의 자동 구축 방법에 대하여 4장에서 실험을 통하여 성능을 평가한다.

4. 실험 및 분석

본 논문에서 제안한 정답문서집합 구축 방법의 성능을 평가하기 위해 다음과 같은 2가지 실험을 실시한다.

[실험 1] 정답문서집합 구축을 위한 자동 분류 성능 평가를 위한 실험
[실험 2] 제안된 모델의 특징에 따른 성능 분석을 위한 실험

현재 구축된 정답문서집합은, 개념어에 할당된 문서는 평균 43.4개이고, 각 개념어 당 평균 25개의 정답 문서가 할당되어 있으며, 이들 정답문서는 평균 18개의 속성에 분류되어 있다. 실험에 사용한 문서는 4,539개의 문서와 개념어 120개, 속성 83개를 사용하였다. 실험에 대한 평가 방법으로는 재현율(recall)와 정확율(precision)를 함께 이용하여 성능을 나타내는 F-score[4]를 사용한다.

4.1 실험 1

[실험 1]은 제안된 분류 모델의 성능에 대한 실험으로, 정답문서 집합을 구축하는 순서에 따라 나누어 실시한다. 정답 문서 집합을 구축하는 과정은 먼저 메타 검색기로 수집한 문서 중에서 해당 개념어와 관련있는 문서만을 추출하는 개념 필터링 과정과, 필터링된 문서 집합을 내용과 특징에 따라 속성을 부여하는 속성 분류 과정으로 나뉜다. 표 2은 개념 필터링 성능과 속성 분류 성능을 나누어 비교한 결과이다. 이때 분류는 중 복분류가 가능하도록 하며, 적절한 범주가 없는 경우에는 미할당으로 남겨둔다. 중복분류 방법은 평균기울기 활용과 이상치 제거로 나누어 실험하는데, 평균기울기를 활용하는 방법은 결과 값의 평균기울기와 상위 값의 기울기의 차를 통해 범주를 할당하는 방법이고, 이상치 제거 방법은 결과 값을 정규화하여 상위 이상치로 범주를 결정하는 방법이다[5].

개념 필터링의 경우, 해당 개념어와 관련이 있는지 그렇지 않은지에 따라 결과값의 차이가 크기 때문에, 표 2에서 보는 것처럼 이상치 제거 기법을 활용한 경우의 재현율이 평균 기울기 기법보다 높게 나타난다. 개념 필터링의 경우 정확율보다는 재현율이 더 중요한 역할을 하기 때문에, 이상치 제거 기법을 활용한 중복분류 방법의 성능이 더 좋음을 알 수 있다. 그러므로, 본 논문에서는 개념 필터링 과정시 이상치 제거 기법을 활용한 분류기법을 사용하기로 하고, 이후 속성 분류 실험에서도 이 결과를 이용한다.

속성 분류의 경우를 살펴보면, 개념 필터링 결과와 상반되는 결과를 나타냄을 알 수 있다. 속성 분류 단계에서는 개념 필터링이 된 문서 집합에 속성을 부여하는 역할을 수행하는데, 이때 속성 분류 결과값이 서로 비슷하기 때문에 발생하는 이상치가 적게 된다. 그러므로, 상위 이상치를 활용하는 이상치 제거 기법으로는 속성을 중복할당 하기가 어렵다. 속성 분류의 경우에는 재현율뿐만 아니라 정확율도 중요한 역할을 하기 때문에, 본 논문에서는 평균 기울기 기법을 활용하여 속성을 할당하기로 한다.

표 2. 정답문서집합 자동 할당 결과

실험 방법		Precision	Recall	F-score
개념 필터링	평균기울기	.7913	.7996	.7954
	이상치제거*	.7476	.9767	.8621
속성 분류	평균기울기*	.4941	.6637	.5789
	이상치제거	.3731	.3960	.3945

표 3. 속성 규칙 활용에 따른 속성 할당 비교

속성 할당 실험 방법	Precision	Recall	F-score	비 고
규칙만 활용	.3502	.3990	.3596	Baseline
기계학습 결과만 활용	.4580	.5806	.5193	+44.4%
규칙활용+ 기계학습	.4941	.6637	.5789	+60.98% (+11.4%)

4.2 실험 2

제안된 정답문서집합 자동 구축을 위한 분류 모델은 문서 집합에 일반적인 범주가 아닌 속성을 할당한다는 점에서 일반 분류 모델과 차이가 있다. [실험 2]는 제안된 분류 모델의 특성에 따른 속성 분류 성능을 알아보기 위한 실험으로, 다음과 같은 특징에 따라 비교 실험한다. 실험 결과는 표 3과 표 4에에 나타내었다.

- 속성을 표현하는 규칙(clue) 활용
- 개념망에 정의된 KIND_OF 관계 활용

표 3의 결과를 분석해 보면, 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 작은 이유는 규칙만 활용한 결과와 기계학습 결과만 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 반영비율을 2:5로 사용하였다.

표 4. KIND_OF 관계 활용에 따른 속성 할당 비교

속성 할당 실험 방법	대상 속성	할당 속성	Pre.	Recall	F-score	비고
KIND_OF 미활용	83	42	.5025	.4662	.4835	+43
KIND_OF 활용	22	20	.5050	.5847	.5648	+99

정답문서집합 자동 할당을 위해서는 개념어당 속성이 정의 되어 있어야 한다는 조건이 선행되어야 한다. 그러나 개념망에 존재하는 모든 개념어에 속성을 정의한다는 것은 쉬운 일이 아니다. 본 논문에서는 이러한 문제점을 해결하고 전체 개념망에서 정답문서집합의 자동 할당 범위를 넓히기 위하여, 개념망에 정의된 KIND_OF 관계를 활용하기로 한다. 이를 통해 미리 학습하지 않은 개념어에 대해서도 같은 KIND_OF 관계 집단에 속하는 개념어는 속성을 할당할 수 있게 된다. 표 4는 학습하지 않은 개념어 9개에 대해 새로운 정답문서집합을 구축한 경우를 실험한 결과로, KIND_OF 관계를 활용하지 않은 경우와 비교하고 있다.

성능 평가를 위해 미리 수작업을 해본 결과, 9개의 개념어가 갖는 속성은 기존에 정의된 속성 22개에 새로운 속성 6개를 포함한 28개였고 정답문서 수는 579개였다. 표 4를 보면, KIND_OF 관계를 활용하지 않은 경우에는 현재 학습된 모든 속성 83개를 대상으로 분류한 결과 42개의 속성을 할당하고 있다. 반면, KIND_OF 관계를 활용하여 기 정의된 속성 22개를 대상으로 분류한 경우에는 20개의 속성을 할당하였으며 성능은 16.8% 향상되었다. 또한 수작업 결과와 비교해봤을 때, 사람이 미처 발견하지 못한 정답을 KIND_OF 관계를 활용한 경우에는 43개, KIND_OF 관계를 활용한 경우에는 99개를 제시함으로써 수작업에서 발생하는 오류를 자동 분류에서 보정할 수 있음을 나타내고 있다.

한편, KIND_OF 관계를 활용한 경우에는 기존에 정의되지 않은, 새로운 속성 6개를 할당하지 못하는 문제가 발생한다. 이러한 문제점은 KIND_OF 관계를 통해 먼저 속성을 부여하고, 속성이 부여되지 못한 미할당 문서에 대해 전체 속성 분류를 실시한다면 해결될 수 있다.

5. 결론

우리는 보다 지능적인 정보검색에 대한 요구에 맞추어, 기 제안된 개념망을 이용한 지식베이스에 대해서 문서분류 기술을 사용하여 대용량 정답문서집합을 자동으로 구축하는 방안을 제시하였다. 자동 문서 분류 기술은 명확한 범주를 가지고 상당량의 학습이 요구되는 기술이지만, 거의 무제한인 인터넷 문서를 대상으로 하고, 범주간의 구분이 유동적인 개념어의 속성에 대해서 자동 분류를 하기 위하여, 우리는 개념망의 KIND_OF 관계와 속성 특징을 이용하여 우리의 지식베이스에 맞는 자동 문서 분류 방법을 제안하였다. 우리는 실험을 통하여 제안한 자동 분류가 개념어에 문서를 할당하는 개념 필터링에 효과적임을 확인하였으며, 필터링된 문서에 속성을 할당하는 속성 분류에 있어서도 60%정도의 성능을 보임으로써, 기존 상용 검색엔진의 웹페이지 검색에서 보여주는 검색 결과보다 유용한 것으로 판단되었다.

그러나, 제안한 자동 분류 방법은 KIND_OF 관계의 적용에 있어서 새로운 속성을 만들지 못하는 단점을 가지고 있으며, 모든 문서를 분류하는 문서 분류 기술의 특성상 기존 속성에 적합하지 않는 문서에 대해서 미할당 문서로 적절히 남기지 못하는 경우가 있다.

그리고, 우리는 이러한 미할당 문서에 대한 처리 방안으로 클러스터링과 같은 방법을 통해 새로운 속성을 부여하는 방법을 고안 중에 있다.

6. 참고 문헌

[1] 장문수, 장명길, 김현진, 오효정, 이재성, "인터넷 질의/응답을 위한 지식베이스 구축", *제12회 한글 및 한국어 정보처리 학술대회*, pp.198-202, 2000.

[2] 장명길, 김현진, 장문수, 최재훈, 오효정, 이충희, 허정, "의미기반 정보검색", *정보과학회지 10월호 한글정보처리 특집*, 2001.

[3] Ask Jeeves, http://www.ask.com/docs/about/whatIs Ask Jeeves.html

[4] 오효정, 임정묵, 맹성현, 이만호, "점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 모델", *제 11회 한글 및 한국어 정보처리학술대회*, 1999.

[5] 강현철, 한상태, 최종후 외 "SAS Enterprises Miner를 이용한 데이터마이닝 - 방법론및활용 -", *자유아카데미 출판사*, 1999.