검색엔진의 정확률 향상을 위한 질의어 의미와 사용자 반응 정보의 이용*

Using Query Word Senses and User Feedback to Improve Precision of Search Engine

윤성희(Sung-Hee Yoon)**

초 록

본 논문은 정보검색 시스템의 사용자 질의어와 색인에 기반한 검색 과정에서 나타나는 중의성 해소를 위해 질의어 의미정보와 사용자 피드백을 사용하여 검색 성능을 향상시키는 방법을 소개한다. 의미정보를 이용하여 질의어의 중의성을 해소하는 검색 과정은 검색 결과로서 의미적으로 무관한 많은 문서들을 배제할 수 있다. 이를 위해 검색의 색인이 되는 명사 중심의 의미범주를 기반으로 의미정보 지식베이스를 구축하고, 검색 문서들을 색인어와 해당 의미범주로 분류한다. 검색 과정에서는 사용자의 질의 의미 선택과 정답 문서에 대한 참조 행위를 웹 페이지의 순위 결정에 반영하여 검색 성능을 향상시킬 수 있다.

ABSTRACT

This paper proposes a technique for improving performance using word senses and user feedback in web information retrieval, compared with the retrieval based on ambiguous user query and index. Disambiguation using query word senses can eliminating the irrelevant pages from the search result. According to semantic categories of nouns which are used as index for retrieval, we build the word sense knowledge-base and categorize the web pages. It can improve the precision of retrieval system with user feedback deciding the query sense and information seeking behavior to pages.

키워드: 정보검색, 질의어, 의미 중의성, 사용자 피드백
search engine, query word sense, ambiguity, user feedback

* 이 논문은 상명대학교 2009년 교내연구비 지원의 결과임.
** 상명대학교 컴퓨터소프트웨어공학과 교수(shyoon@smu.ac.kr)
■ 논문접수일자 : 2009년 10월 18일 ■ 최초심사일자 : 2009년 10월 20일 ■ 게재확정일자 : 2009년 10월 30일
■ 정보관리학회지, 26(4): 81-91, 2009. 〔DOI:10.3743/KOSIM.2009.26.4.081〕

1. 서 론

사용자의 질의를 입력으로 하여 대량의 문서들을 검색하고, 사용자의 질의 목적에 관련된 내용을 포함하는 적합 문서들을 선별하고 순위화하여 정답후보 문서집합으로 제시하는 목적의 소프트웨어 도구를 정보검색 시스템이라고 한다(Baeza-Yates Ricardo. and Reberio-Neto Berthier 1999). 검색 대상 문서의 양이 급속하게 증가하고 있는 추세에 따라 정보검색 시스템의 성능 향상에 대한 요구가 함께 높아지고 있으며. 수많은 검색 결과들 중에서 사용자의 질의 의도에 맞는 문서들을 신뢰할 만한 순위로 제공하는 성능 향상에 초점을 두는 기술 개발이 요구되고 있다. 정보검색과 관련되는 기술의 발전을 위해서 미국 NIST(National Institute of Standard and Technology)의 주관으로 TREC 학술대회가 매년 열리고 있으며, 여러 관련 기술 분야별로 전문화된 트랙(track)을 두어 기술발전을 유도하고 있다(TREC: Text Retrieval Conference). 검색 시스템의 성능을 객관적으로 평가하고 비교하기 위하여 테스트 컬렉션(test collection)을 마련하고 제공하며, TREC 학술대회를 통하여 연구 집단들이 개발한 시스템의 성능 비교 및 정보 교환이 가능하도록 하고 있다.

사용자 질의가 갖는 명확한 의도가 검색시스템에 반영되지 못하는 중요한 이유에는 질의어가 갖는 의미 중의성(semantic ambiguity) 문제가 있다. 부적절한 색인어 및 탐색어의 선정을 비롯하여 동형이의어나 다의어와 같은 중의성을 가진 단어들을 질의어나 색인어로 사용할 경우 사용자가 원하지 않는 문서를 검색결과로 제시할 확률이 매우 높다.

본 논문에서는 검색 질의어의 의미 중의성을 해소하기 위해 색인어가 되는 어휘들의 의미 정보를 의미 범주로 분류한 지식베이스를 기반으로 사용자 피드백을 통해 질의어의 의미를 결정하는 방법을 소개한다. 또한, 검색 대상이 되는 웹 문서 집합들에 대해서도 문서의 색인어들에 대한 의미정보를 통해 분류하여 검색과정에서 중의성을 해결할 수 있다. 뿐만 아니라, 검색 과정에서 정답 후보 문서들 중에서 사용자가 보이는 적극적인 검색 행위를 수집하고 계량화하여 사용자의 반응 피드백으로 검색 문서의 순위 결정에 반영함으로서 순위 정확률을 높일 수 있음을 보여준다.

관련 연구들에서 보듯이, 특히 웹 검색에서는 전문적이기 보다는 일반적인 용어들이 질의어로 많이 사용되고 있다. 또한 단일 질의어 중심의 사용자 질의 및 검색 환경에서 입력 질의어의 중의성 해소와 검색 대상 문서들의 의미정보 분류에 기반한 검색 방법은 검색 정확률 측면에서 성능 향상을 크게 기대할 수 있다. 사용자의 입력 질의어가 갖는 의미 중의성을 해결하고, 개별 의미범주에 따라 분류된 문서 집합으로부터 정답 문서를 선택하고, 정답 후보 문서 중에서 사용자가 반응을 보이는 검색행위를 가중치로 지속적으로 반영함으로써 검색 정확도를 높일 수 있음을 실험 결과를 통해 보여주었다.

2. 관련 연구

2.1 사용자 검색 행태 분석

정보검색 시스템의 사용자 검색 행위에 대한 특성을 분석한 여러 연구들이 지속적으로 소개되어 왔다(김성진 2006; 박소연, 이준호 2002). 일반적으로 정보검색 시스템은 입력 질의와 문서에 나타나는 색인어의 형태적 일치를 검사함으로써 적합 문서 여부를 결정하며, 많은 웹 문서들을 주기적으로 방문하고 색인 데이터베이스를 갱신한다. 한편, 대부분의 검색 시스템 사용자들은 복잡한 검색식이나 연산자를 사용하지 않고 매우 단순한 키워드를 검색 질의어로 입력하고, 결과적으로 검색 결과 집합에 대해 충분히 만족하지 못하여 검색 키워드에 대한 사소한 수정과 함께 반복적으로 검색하게 된다. 즉 사용자가 처음 입력한 질의는 적합한 문서를 찾기 위한 초기 시도라고 볼 수 있으며, 이후 검색된 문서의 연관성 평가를 통해 더 관련이 많은 문서를 찾기 위해 재검색하는 것이다.

또한 웹 환경 검색 시스템의 사용자들은 다양한 계층 및 연령층이면서도 트랜잭션 로그 분석을 이용한 연구의 결과를 통해 볼 때에는 "웹 검색에 있어서 사용자의 검색 방식의 단순함"이 가장 두드러진 특징으로 나타나고 있다(김성진 2006; 김희섭, 박용재 2004; 박소연, 이준호 2002). 웹 검색엔진의 탐색관련 특성에 관한 이전의 연구에서는 질의어의 수가 1~3개인 질의가 전체 질의의 84%에 달하는 것으로 분석되었으며, 하나의 질의가 평균 2.2개의 질의어로 구성되었음을 실험을 통해 보이기도 하였다. 이와 같이 사용자의 질의가 단일 어휘이거나 단순한 명사구의 형태를 갖는 많은 경우에 사용자의 질의 의도는 더욱 판별하기 어렵게 된다. 따라서 검색의 정확률 향상을 위해서는 질의어의 의미를 구분하고 초기 검색 결과에 대한 사용자 만족도의 척도가 되는 참조 행위를 반영하는 과정이 포함되어야 한다.

2.2 질의어 중의성 해소를 위한 의미정보

언어처리 분야에서 어휘 의미정보를 체계화하는 방법은 여러 가지가 소개되어 왔다. 대표적인 방법 중 한 가지는 어휘에 대하여 동의어, 반의어, 상위의미, 하위의미 등과 같은 어휘의 연관성을 정의한 사전인 워드넷(WordNet)이다(Moldovan and Mihalcea 1998; Moldovan and Mihalcea 2000). 워드넷은 1985년 프린스턴(Princeton) 대학에서 개발되어 영어권에서 연구에 널리 활용되어 왔다. 워드넷에는 어휘의 의미에 대한 범주 분류가 잘 정의되어 있으며, 단어들 사이의 계층구조와 연관관계가 여러 형태로 표현되어 있다. 단어의 의미관계를 표현하는 리소스로 잘 알려진 워드넷(WordNet)은 의미가 유사한 단어들의 집합(SynSet)간의 연결로써, 단어 하나하나의 개념관계를 표현하고 있어서 유사한 단어들의 집합을 이루고 있으므로 대용어 선택이나 다국어 번역에서의 의미 공유 등에서는 효과적으로 활용되고 있다. 특히 영어권에서는 워드넷을 이용한 지식베이스 관련 연구가 활발히 이루어지고 있다.

실세계 지식을 보완하기 위해 시소러스로부터 검색어들을 추출하여 개념검색어로 구분하여 도식화하고, 사용자의 선택에 따라 검색질의를 확장하는 방법이 소개되기도 하였다(강현

구 2002). 시소러스에서의 상하관계는 사전적인 의미가 아닌 필요에 따른 용례에 의한 관계가 반영되어서, 시소러스가 단어 간의 관계에 대한 기준이 명확하지 않다고 평가되기도 한다(장명길 외 2001).

한편 어휘 개념망 방법을 한국어 명사에 대한 의미 지식베이스 구축에 실험적으로 적용한 사례도 있다(KIBS Korean Information Base System). 단어들의 관계를 설정한다는 점에서 기존의 시소러스와 유사하고, 사전의 뜻풀이를 중심으로 개념어들 간의 국어학적 의미관계를 연결하여, 단어들의 의미 포함관계를 명확하게 나타내고자 하였다. 상하관계에 의존한 개념망을 보완하기 위하여, 동의, 유의, PART-OF, 반의 등의 관계를 추가로 정의하고 있다.

TREC의 의미범주 계층을 활용한 실험의 예도 있다(윤성희, 장혜진 2004; 윤성희, 백선욱 2004). 질의문에 대한 정답을 웹 문서에서 찾아주는 질의응답 시스템에서 단어의 의미정보를 활용하여 질의문장의 유형을 분류하고 있는데, 이때 TREC-10과 TREC-11에서 질의 시스템을 위한 의미범주의 계층을 기반으로 하고 있다.

한편으로는, 자연어처리 기술을 이용한 의미기반 색인과 검색 방법에 대한 연구도 활발히 소개되고 있다(Lee, Park, and Won 1999; Belkin, Kelly, et al. 2003; Salton Gerard 1988; Zhai Chengxiang 1997). 이들 관련 연구들은 어휘, 구절, 문장의 중의성을 해소하기 위해 형태소분석 및 태깅(tagging) 기술 뿐만 아니라 구문분석과 의미분석 등의 자연어처리 기술을 이용하는 방법으로서, 마찬가지로 사용자 질의의도를 추출하고 텍스트와 질의의 의미를 색인에 반영하여 검색성능을 높이는 것이 목적이다. 이러한 방법에서는 자연어처리 기술이 정보검색에서 가장 중요한 과정이 되며, 의미색인 단위의 설정과 추출은 자연어처리 기술의 적용수준에 따라 좌우된다.

3. 검색 시스템 구성

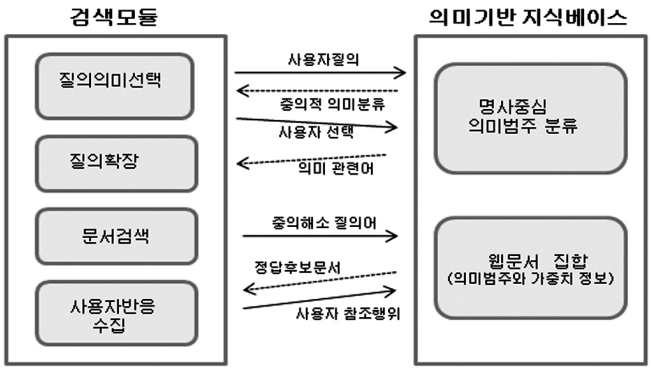
본 연구의 검색시스템의 검색 흐름은 다음의 〈그림 1〉에서와 같이 의미정보를 활용하고 사용자 피드백을 반영하는 과정으로 요약될 수 있다.

의미기반 지식베이스로서 의미범주를 분류하는 지식베이스와 의미 범주별로 분류되고 지속적으로 가중치가 갱신되는 웹 문서 집합을 지식베이스로 갖는다. 검색 과정은 사용자의 질의입력, 의미선택 후 질의확장, 의미 선택된 색인으로 웹 문서 검색, 정답후보 문서에 대한 사용자 반응 수집과 가중치 갱신의 순서로 이어진다.

3.1 질의어 의미 분류와 선택

앞에서 서술하였듯이, 정보검색에서 중의성 문제가 해결되면 정보검색 시스템의 정확도를 크게 향상시킬 수 있다. 이를 위하여 사용자 질의와 색인 문서의 중의성 해소를 위해 우선 단어의 의미정보를 체계화한 지식베이스를 구축하는 것이 필요하다.

본 연구에서는 선행 연구에서 사용한 바 있는 TREC의 의미범주(semantic category) 체계를 기반으로 하여 중의성을 갖는 동형이의어를 중심으로 의미기반 지식베이스를 구성하였다(TREC Text Retrieval Conference). TREC-



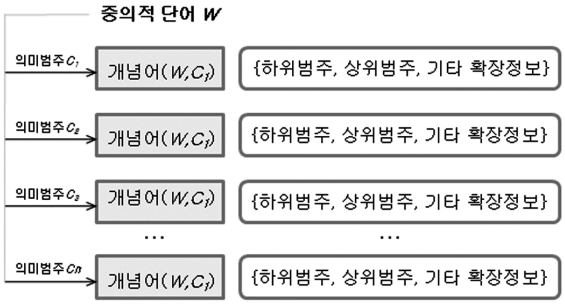
〈그림 1〉 의미정보에 기반한 정보검색

10 및 TREC-11에 참가한 검색시스템들을 참고하여 구축한 의미 범주 분류의 예를 다음 〈표 1〉에서 보이고 있다. 중의성 해소를 위한 의미정보는 중의성을 갖는 단어의 의미범주 분류와 그 사전적 해석으로부터 추출할 수 있는 확장정보들을 포함한다. 지식베이스에는 해당 표제어에 대해 의미범주는 물론 상위범주, 하위범주, 동의어, 유의어, 뜻풀이에서 추출되는 공기(co-occurrence)관계 단어 등의 의미 관련어들을 함께 포함하여, 웹문서의 의미분류 뿐만 아니라 질의확장과 유사도(similarity)에 기반한 순위결정에 사용될 수 있다. 중의성을 갖는 단어가 의미 범주로 분류되는 체계를 〈그림 2〉와 같이 표현할 수 있다.

사용자가 입력하는 질의어의 중의성을 해소하고, 질의어의 의미와 범주를 결정하기 위해 사용자 피드백을 이용하는 인터페이스를 설계하였다. 사용자가 입력한 질의어가 중의성을 갖는다면 의미정보 지식베이스에 복수의 의미범주 항목을 포함하고 있으며, 의미선택기는 각 질의어의 각 의미범주 단위로 동의어, 상위어, 사전적 설명 등을 추출한다. 추출된 복수의 의미범주는 사용자 인터페이스를 통해 사용자에게 전달되고, 사용자는 본인이 의도한 질의의 의미를 선택할 수 있다. 개념이 선택된 질의어는 앞선 연구에서와 같은 방법으로 질의확장 과정을 거치며(윤성희, 장혜진 2004), 사용자 질의어와 선택된 의미범주을 색인으로 조합하

〈표 1〉 TREC의 의미 범주 분류 예

상위 의미 주 예	하위 의미범주 예
Human(사람)	Artist, Politician, Economics, Sports ...
Location(장소)	Place, City, Continent, State, Capital ...
Time(날짜나 시간)	Year, Month, Day, Season, Hour ...
Number(수)	Count, Price, Percent, Weight, Height ...
Organization(조직)	School, Company, Government, Team ...
Object(물체나 사물)	Plant, War, Religion, Organ ...



〈그림 2〉 중의적 단어들의 의미 분류

여 문서 집합을 검색하게 된다. 사용자의 피드백이 없을 경우, 중의적 질의어의 각 의미범주 단위별로 독립적인 웹 문서 검색을 수행한다.

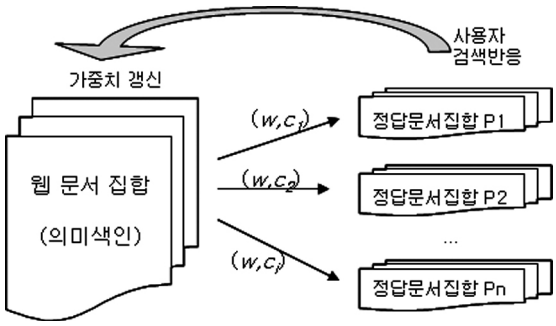
3.2 검색 문서 의미 색인과 사용자 반응 수집

검색 대상 문서의 검색 키(key)가 될 질의어가 중의성을 가질 경우, 전혀 다른 의미를 담고 있는 문서들이 동일한 색인어를 통해 선택될 수 있다. 검색 문서의 색인에 대한 중의성을 해소하기 위해 앞에서 기술한 의미기반 지식베이스를 구축하는 한편, 검색 대상인 문서들을 어떤 주제에 대해서 사용되고 있는가에 따라 개념적으로 구분하여 초기 색인하는 과정이 필요하다. 색인어가 중의성을 갖는 경우, 지식베이스에서 제공하는 의미범주별로 동의어, 유의어, 공기단어 등의 의미관련어를 이용하여 해당 문서에서 색인어가 갖는 의미를 결정한다. 예를 들어, 색인어가 '연기'일 때, '담배'나 '화재' 등이 함께 나타나는 웹 문서는 '연기'의 여러 의미 중에서 '물건이 탈 때 생기는 빛깔 있는 기체'의 의미범주로 색인될 수 있다. 한편, 검색 문서들은 단일한 범주에만 속하지 않을 수 있으므로 검색 문서를 단일한 범주로 결정하는 것은 정보활용 측면에서 위험한 문제를 일으킬 수 있다. 이를 해결하기 위해 〈그림 3〉에서 보이는 바와 같이 검색 문서를 〈색인어, 의미범주〉 단위로 다중 색인하여 검색의 변별력을 높일 수 있다.

검색엔진은 검색 결과인 정답후보문서들을 중요도에 따라 순위로 반영하여 사용자에게 제시함으로써 사용자가 정답문서를 효율적으로 접근할 수 있도록 돕는다. 일반적으로 검색 결과 문서들에 대한 순위결정은 질의어와 웹 페이지 단어들을 형태적으로 비교한 유사도(similarity)를 이용하여 이루어진다. 그러나 앞에서 설명한 바와 같이 의미범주 단위로 색인된 검색 문서들은 색인어와 의미범주의 조합으로 유사도를 계산하고 각 순위 정보를 갖는다.

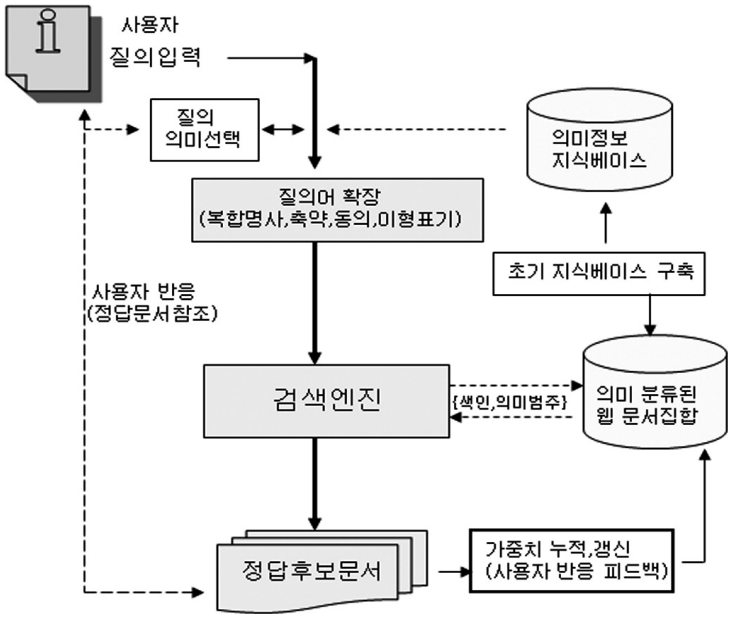
검색 엔진이 색인어의 의미범주 별로 검색하여 제시한 결과 문서들, 즉 정답후보문서들 중에서 사용자의 실제 선택과 참조 행위가 있는 문서들은 사용자 질의에 대한 검색 정확도가 높은 문서라고 볼 수 있다.

사용자의 문서 참조 행위는 해당 문서에 대한 묵시적인 평가라고 할 수 있으며, 이러한 정보선별 행동들을 누적하고 웹 페이지의 가중치에 반영하여 웹 페이지에 개념별 인기도를 부



〈그림 3〉 웹 문서의 의미별 검색

여하고 순위 결정함으로써 검색성능을 향상시킬 수 있다. 검색 결과 문서에 대한 사용자의 참조 행위를 모니터하여 사용자의 묵시적 평가를 웹 페이지의 범주별 가중치에 누적하는 방식으로 문서의 순위에 반영하는 방법이다. 이와 같은 방법은 단어의 형태적 일치와 출현빈도에 의해 순위를 결정하는 기존의 방법들과 달리 질의어의 중의성을 해소하고 사용자의 실질적인 참조 행위를 결합하여 지속적으로 사용자 중심의 문서 가중치를 갱신할 수 있다. 본 연구에서 제안하고 있는 의미정보에 기반한 검색시스템의 전체적인 구성을 다음의 〈그림 4〉에서 보여주고 있다.



〈그림 4〉 의미정보에 기반한 검색 시스템

4. 실험 및 평가

본 연구는 실험을 위해 기존의 실험에서 사용된 바 있는 약 4,500여 문장을 포함하는 문서들과 개인, 학교, 기관 등의 홈페이지들, 학술 내용과 시사 뉴스 등의 내용을 담은 웹 문서들을 대상으로 하였다. 이 문서 집합은 앞선 연구와 실험을 위해 구축된 바 있으며(윤성희, 장혜진 2004; 윤성희, 백선욱 2004), 본 연구의 실험을 위해 시사 뉴스 분야의 웹 문서가 확장되고 중의적 색인어에 대한 의미범주별 색인 작업이 추가되었다.

실험 대상인 검색시스템의 사용자들에게 웹 문서에 대한 검색 질의를 입력하도록 하였으며, 실험에서 약 4천회 이상의 질의 입력을 수집하였다. 그 결과로 본 실험을 위해 검색 대상 문서들과 사용자 질의에 빈번하게 나타나는 중의적 어휘들과 관련 연구에서 사용된 바 있는 어휘들 중에서 10개를 〈표 2〉와 같이 선별하였다. 중의성을 갖는 단어는 각 의미별로 웹 문서에 등장하는 빈도에 차이를 보였다. 웹 페이지에서 높은 빈도로 사용되는 의미의 경우, 검색 결과로도 다수 나타남으로써 정확도는 상대적으로 높게 평가된다. 반면에, 사용자의 질의 의도가 자주 사용되지 않는 의미에 있고, 중의성을 해소하지 않은 채 형태 비교로 검색한다면 상위 결과 중에서 정답문서를 발견하기 어려우며, 결과적으로 검색 정확도는 크게 떨어지게 된다. 실험 예로서, 중의적 단어 '신장'의 경우 '내장기관의 하나', '키', '길게 늘임', '새로 단장함'의 각 의미로 등장하는 실험 웹 문서의 비중이 54%, 13%, 29%, 3.3%로 나타났다. 따라서 중의성을 해소하지 않은 채 '신장'을 질의어로 사용했을 경우, '내장 기관의 하나'의 의미로 사용된 문서가 가장 많이 검색될 것으로 짐작할 수 있으며, 실제로 실험에서 상위 20개 페이지 중에서 19개, 상위 30개 웹 문서 중에서 24개의 문서들이 첫 번째 의미로 사용된 문서였다. 반면에 '새로 단장함'의 의미를 위해 질의어 '신장'을 입력한 경우 정답문서는 상위 30개 문서 중에서 1개에 불과했다. 또한 중의적 단어 '연기'의 경우, 중의성이 해소되지 않은 질의어 '연기'를 입력하였을 경우, 정답 후보 문서 상위 30개 중에서 28개가 '관객 앞에서 연극 따위의 재주를 나타내 보임'의 의미로 사용된 문서들이었다. 실제로 '연기'로 색인된 실험 문서들 중에서 '연기'가 이 의미로 사용된 비중은 49%에 달했다.

〈표 3〉은 중의성을 갖는 사용자의 질의어에 대해서 중의성을 해소하지 않은 채 검색한 경우와 사용자가 의미범주를 선택하고 의미 색인된 웹 문서들을 검색한 경우의 평균 성능을 보여주고 있다. 질의 중의성 해소 후, 의미 분류된

〈표 2〉 실험에 사용된 중의적 검색어 예

중의적 검색어	의미 분류 개수	중의적 검색어	의미 분류 개수
신장	4	지원	3
연기	4	지구	2
기구	2	기관	3
다리	2	신장	4
인도	5	눈	3

〈표 3〉 중의성 해소와 평균 검색 성능

	상위 30개 문서	상위 20개 문서	상위 10개 문서
중의성 처리 전	34%	41%	47%
중의성 해소 후, 의미범주별 검색	58%	77%	83%

문서들을 검색한 경우 매우 높은 비율로 정답 문서를 포함하고 있음을 볼 수 있다. 참고로, 검색시스템 사용자의 행태로 볼 때, 사용자들은 대부분 검색결과 중 상위 20개가량의 문서만을 참조하는 것으로 분석된 바 있다(김성진 2006; 박소연, 이준호 2002). 따라서 상위 30위내의 정답문서의 수는 검색 결과의 성능을 확인하기에 충분하다고 보인다.

사용자의 정답문서에 대한 참조반응을 순위 갱신에 반영하기 위해 각 검색의 결과로 제시되는 정답 후보문서들 중에서 사용자가 상위 정답 문서로 평가하는 것들을 다섯 개 이상 선택적으로 참조하도록 하여 그 반응을 누적하였다. 실험이 진행된 후 검색결과의 문서들 간에 변경되고, 상위랭킹 문서들에 대한 만족도, 즉 상위 10위 내의 문서들에 대한 정답 평가 정도는 크게 향상되어 83%에 달했다. 상위 10위 내의 문서들 중에서 대부분이 사용자가 만족하는 정답 문서들에 해당하는 것들로 평가할 수 있다.

현재까지의 실험은 검색성능의 정확률 향상을 검증하는데 초점을 두었지만, 본 실험에서 사용한 검색 대상 문서집합의 규모를 확대하여 의미적으로 보다 다양한 문서들을 검색하는 사용자들의 자연어 질의를 추가 수집하고, 실험을 계속 확장하고 있다. 중의성을 갖는 검색어들을 다수 수집하고, 각 의미범주 별로 웹 문서에 나타나는 빈도와 검색 결과에 대한 분석이 계속될 예정이다. 또한 두 개 단어 이상의 질의어가 중의성에 해소에 미치는 영향과 오류의 경우에 대한 분석도 필요하다고 보인다. 이와 더불어, 수시로 생성되고 소멸되며 내용이 변경되는 동적인 웹 환경을 위한 유연한 색인 데이터베이스 구축 시스템과 실험적 규모 이상의 의미 지식베이스를 구축할 필요가 있다.

5. 결 론

본 논문에서는 정보검색 시스템의 사용자 질의어와 색인에 의한 문서 검색 과정에서 중요한 문제가 되는 중의성을 해소하고 검색성능을 향상시키기 위해 질의어의 의미정보를 활용하고 사용자의 검색반응을 수집하여 순위에 반영하는 방법을 제안하였다.

질의와 검색 과정에서 중의성 해소 과정은 검색결과들 중에서 의미적으로 무관한 많은 문서들을 배제함으로써 검색 정확률을 높이고 결과적으로 사용자 만족도를 크게 높일 수 있는 매우 중요한 처리 과정이다. 본 연구를 위한 자원으로 검색의 색인어가 되는 명사들을 중심으로 의미정보 체계를 지식베이스로 구축하고, 웹 문서들을 색인어의 의미범주로 분류한다. 검색 과정에서는 사용자의 피드백에 의해 질의어의 정확한 의미를 결정할 수 있으며, 색인어와 의미정보가 함께 색인된 문서 집합에서 질의 의도에 맞는 문서들을 선택할 수 있다. 또한

검색 문서들에 대한 사용자의 참조 행위가 선택된 문서에 대한 정답 평가로 계산되어 순위를 결정하는 가중치에 누적된다.

이와 같이 사용자 피드백을 통해 질의어 중의성을 해소하고 정답 후보 문서에 대한 사용자의 참조 행위를 수집하여 순위 결정에 반영함으로서 검색시스템의 성능이 향상될 수 있음을 실험 결과를 통해 보여주었으며, 앞으로 계속될 연구의 방향을 제시하였다.

참 고 문 헌

강현규. 2002. 개념 검색어 대체를 통해 질의 형식화를 도와주는 개념 마법사의 설계 및 구현.『정보처리학회논문지』, 9-B(4): 437-444.

김성진. 2006. 이용자 중심 웹 정보탐색 연구의 실체이론 분석.『정보관리학회지』, 23(3): 127-146.

김희섭, 박용재. 2004. 정보시스템의 이용자만족지수 모형개발 및 측정.『정보관리학회지』, 21(4).

박상규, 이찬규, 윤경현, 김성희, 이준호. 2007. 검색엔진에서 일간질의어 분포의 정상성에 관한 연구.『정보관리학회지』, 24(4): 255-265.

박소연, 이준호. 2002. 로그 분석을 통한 이용자의 웹 문서 검색 행태에 관한 연구.『정보관리학회지』, 19(3): 111-122.

박의규, 나동열, 장명길. 2005. 문장-질의 유사성을 이용한 웹 정보검색의 성능 향상.『정보과학회 논문지: 소프트웨어 및 응용』, 32(5): 406-415.

윤성희, 장혜진. 2004. 자연어 질의분석과 검색어 확장에 기반한 웹 정보검색.『정보관리학회지』, 21(2): 235-248.

윤성희, 백선욱. 2004. 단어 의미정보를 활용하는 이용자 자연어 질의 유형의 효율적 분류.『정보관리학회지』, 21(4): 251-263.

이용구, 정영미. 2007. 사전 정보를 이용한 단어 중의성 해소 모형에 관한 실험적 연구.『정보관리학회지』, 24(1): 321-342.

이재윤. 2007. 분포유사도를 이용한 문헌 클러스터링의 성능향상에 관한 연구.『정보관리학회지』, 24(4): 267-283.

장명길, 김현진, 장문수, 최재훈, 오효정, 이충희, 허정. 2001. 의미기반 정보검색.『정보과학회지』, 19(10): 7-18.

Baeza-Yates, Ricardo. and Reberio-Neto Berthier. 1999. *Modern Information Retrieval.* Addison Wesley.

Belkin, N. J. et al. 2003. "Query length in interactive information retrieval." *SIGIR*, 2003: 205-212.

Fagan, Joel L. 1987. "Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods." Ph.D. thesis,

Cornell University.

KIBS: Korean Information Base System. 〈http://kibs.kaist.ac.kr〉.

Lee, G., M. Park, and H. Won. 1999. "Using syntactic information in handling natural language queries for extended boolean retrieval model." *Proceedings of the 4th international workshop on information retrieval with Asian language(IRAL99)*: 63-70.

Moldovan, D. and R. Mihalcea. 1998. "A Word Net-Based Interface to Internet Search Engines." *Proceedings of FLAIRS-98*.

Moldovan, D. and R. Mihalcea. 2000. "Using WordNet and Lexical Operators to improve Internet Searches." *IEEE Internet Computing*, 4(1): 34-43.

Perez-Carballo, Jose and Strazalkowski Tomek. 2000. "Natural Language Information Retrieval: progress report." *Information Processing & Management*. 36(1): 155-178.

Salton, Gerard. 1988. *Automatic text processing*. Addison-Wesley publishing company.

TREC(Text Retrieval Conference) 〈http://trec.nist.gov/pubs.html〉.

Won, H., M. Park, and G. Lee. 2000. "Integrated indexing method using compound noun segmentation and noun phrase synthesis." *Journal of KISS: Software and Applications*, 27(1): 84-95.

Zhai, Chengxiang. 1997. "Fast statistical parsing of noun phrases for document indexing." *Fifth Conference on Applied Natural Language Processing*: 312-319.