워드넷 기반 협동적 평가와 하이퍼링크를 이용한 검색엔진의 성능 향상

김 형 일† · 김 준 태††

요 약

본 논문에서는 검색엔진의 성능 향상을 위하여 질의어의 모호성 해결과 새로운 가중치 부여 방식을 제안한다. 일반적인 검색엔진은 질의어의 형태와 같은 것들이 포함되어 있는 웹 페이지를 결과로 보여주는 내용기반 방식을 사용하고 있다. 검색 결과로 나타난 웹 페이지들의 순위를 결정하는데 있어서도 주어진 질의어와 웹 페이지 사이의 키워드 매칭에 의한 내용기반 방식을 사용한다. 이와 같이 질의어의 형태만으로 웹 페이지들과 유사도를 비교한다는 것은 정확한 검색에 많은 장애를 준다. 또한 질의어의 의미에 모호성이 존재할 경우에는 사용자의 의도와 관련 없는 것들이 결과로 나타나기도 한다. 이러한 원인의 발생은 일반적인 검색엔진들이 내용기반 방법을 기반으로 웹 검색에 이용되기 때문이다. 본 논문에서는 질의어에 모호성이 있는 경우 워드넷을 이용하여 모호성을 해결하도록 하는 사용자 인터페이스를 구현였다. 그리고 사용자의 클릭 수를 각 웹 페이지의 가중치에 누적함으로써 다수 사용자의 협동적 평가에 따른 웹 페이지의 중요도가 검색 순위에 반영되도록 하였다. 클릭수의 누적에 있어서 질의어의 의미 카테고리별로 가중치를 구분하여 저장함으로써 보다 세분화된 웹 페이지 가중치 부여 방식을 사용하였다. 그리고 웹 페이지의 하이퍼링크를 웹 페이지의 가중치에 적용하였다. 웹 페이지의 가중치에 하이퍼링크를 적용함으로써 웹 페이지의 대표성을 가중치에 부여하여 가중치에 신뢰도를 증가시켰다. 실험용 검색엔진이 일반 검색엔진에 비해 높은 검색 정확도를 나타내는 것을 실험을 통해 확인하였다.

Improving Performance of Search Engine By Using WordNet-based Collaborative Evaluation and Hyperlink

Hyungil Kim† · Juntae Kim††

ABSTRACT

In this paper, we propose a web page weighting scheme based on WordNet-based collaborative evaluation and hyperlink to improve the precision of web search engine. Generally search engines use keyword matching to decide web page ranking. In the information retrieval from huge data such as the Web, simple word comparison cannot distinguish important documents because there exist too many documents with similar relevancy. In this paper, we implement a WordNet-based user interface that helps to distinguish different senses of query word, and constructed a search engine in which the implicit evaluations by multiple users are reflected in ranking by accumulating the number of clicks. In accumulating click counts, they are stored separately according to senses, so that more accurate search is possible. Weighting of each web page by using collaborative evaluation and hyperlink is reflected in ranking. The experimental results with several keywords show that the precision of proposed system is improved compared to conventional search engines.

키워드 : 정보검색(Information Retrieval), 검색엔진(Search Engine), 워드넷(WordNet)

1. 서 론

초기의 검색엔진은 웹의 정보 공유를 중요시하여 사용자가 원하는 정보를 웹에서 대량으로 추출하여 주는 것만을 고려하였으나, 현재와 같이 방대한 정보가 내재되어 있는 웹에서는 정보의 수량적 접근보다는 사용자가 원하는 정보를 얼마나 정확히 추출할 수 있는가가 중요시 되어야한다. 현재의 웹 검색엔진에서 사용자가 원하는 정보를 검색할 경우 사용자가 획득한 결과 정보는 사용자의 요구에 대한 관련성 측면을 고려할 때 완전한 정보라 할 수 없다. 그래서 사용자는 결과 정보에 대한 이차적 판별을 적용하여 자신에게 적합한 정보만을 추출해야한다. 이러한 문제는 정보의 대량화와 내용기반을 이용한 정보 추출의 수량적 접근으로 발생된다. 이러한 정보 대량화의 역기능은 현재의 웹 검색엔진에서 해결할 문제이며, 사용자가 보다 편리한 검색을 할 수 있는 검색엔진의 설계 또한 시급한 문제이다.

대다수의 검색엔진들은 검색 결과의 순위를 계산하는데

※ 본 논문은 정보통신부 대학기초연구지원 사업의 연구 결과임.
† 준 회 원 : 동국대학교 대학원 컴퓨터공학과
†† 정 회 원 : 동국대학교 컴퓨터공학과 교수
논문접수 : 2004년 3월 4일, 심사완료 : 2004년 5월 25일

있어서 사용자가 사용한 질의어가 특정 웹 페이지에서 많은 분포를 이룰 경우 높은 가중치를 부여하는 내용기반 가중치 방법을 활용하고 있다[7]. 이러한 내용기반 가중치 방식은 사용된 질의어의 의미는 배제되고 해당 질의어 형태가 많이 포함되어 있는 웹 페이지에 높은 가중치를 부여함으로 다의성을 갖는 질의어에 대하여 올바른 가중치를 부여할 수 없다. 예를 들어 임의의 사용자가 "배(의미 : 과일)"에 대하여 알고 싶을 경우 검색엔진을 통해 "배"라는 질의어를 사용하여 검색을 하게 된다. 그러나 웹 검색엔진들은 질의어의 형태만을 고려하여 검색에 이용함으로써 "배(의미 : 과일)"의 의미는 고려하지 않고 "배"라는 어휘 형태가 존재되어 있는 웹 페이지들을 추출하여 사용자에게 결과 정보로 제공한다. 이러한 문제점으로 형태는 같으나 의미가 다른 단어인 "배(의미 : 선박)"라는 단어가 많이 내포되어 있는 웹 페이지도 결과 정보 대상으로 존재성을 갖는다. 또한 키워드 매칭에 의한 내용기반 검색 방식은 질의어와 웹 페이지 사이의 형태적 관련도만 계산함으로 중요한 웹 페이지들을 선별하여 보여주는 역할은 할 수 없다.

이러한 검색엔진의 단점으로 구글(Google)과 다이렉트히트(DirectHit)와 같은 차세대 검색엔진에서는 새로운 가중치 결정 방식을 도입하여 사용자들에게 많은 호응을 얻고 있으며, 기존의 검색엔진을 이용하여 웹 페이지들을 수집한 후 웹 페이지의 가중치를 재조정하여 사용자에 보여주는 메타 검색엔진의 연구 또한 활발히 진행되고 있다[5].

본 논문에서는 현재 검색엔진들의 가중치 결정 방식의 문제점을 해결하기 위한 새로운 가중치 결정 방법과 질의어의 모호성 해결 방안을 제안한다. 본 논문에서 사용한 가중치 결정 방법은 의미 카테고리 구조를 이용한 협동적 평가 방법과 하이퍼링크 가중치이다. 이러한 가중치 결정 방식을 이용함으로써 웹 페이지의 검색 성능을 향상시켰다. 그리고 질의어의 모호성 해결을 위해 워드넷 기반의 사용자 인터페이스를 설계하여 정확한 질의어의 의미를 추출하였다.

본 논문에서는 워드넷 기반 사용자 인터페이스로 질의어의 모호성을 해결할 수 있는 실험용 검색엔진을 구축하였으며, 실험을 통하여 본 논문에서 제시한 협동적 평가 방법과 하이퍼링크 가중치를 사용한 가중치 결정 방법이 웹 검색에서 우수한 성능을 나타냄을 확인하였다.

본 논문의 구성은 2장에서 관련 연구에 대하여 기술하고, 3장에서는 본 연구의 결과물인 실험용 검색엔진에 대하여 설명하며, 4장에서는 본 연구의 실험 방법과 실험 결과에 대해 설명한다. 5장에서는 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구

본 장에서는 전통적 방식의 가중치 적용 방법을 사용한 일반 검색엔진의 순위 결정 방식의 단점으로 나타나게된 차세대 검색엔진들에 대하여 기술하고 차세대 검색엔진으로 각광받고 있는 다이렉트히트(DirectHit)와 구글(Google) 시스템과 웹 페이지 가중치 적용 방법에 대하여 기술한다. 그리고 본 장에서는 질의어의 모호성 해결을 위해 활용한 워드넷의 기본 구조와 실험용 검색엔진에서 적용한 의미 카테고리 구조에 대해 설명한다.

2.1 차세대 검색엔진

현재 웹 정보검색에서 내용기반 방식의 문제점으로 차세대 검색엔진에 대한 활발한 연구가 진행되고 있다. 차세대 검색엔진들은 내용기반 방법의 단점을 보완하기 위하여 새로운 검색 기술의 개발과 내용기반 방식을 탈피한 웹 페이지 가중치 부여 방식에 대해 연구되고 있다. 차세대 검색엔진으로는 구글(Google), 다이렉트히트(DirectHit), 사비서치(Savvy Search), 클레버(Clever), 등을 예로 들 수 있다.

구글은 웹 페이지들의 하이퍼링크를 이용하여 웹 페이지 검색에 이용한다면 사용자에게 중요도가 높은 웹 페이지를 추출하여 줄 수가 있다는 개념에서, 스탠포드대학에서 개발이 진행되어 현재는 상용화되어 있는 검색엔진이다. 하이퍼링크는 웹에 산재된 웹 페이지들을 연결하여 주는 하나의 수단으로 활용되는 도구이다. 이러한 웹 페이지의 연결 구조를 이용하면 하나의 가설을 세울 수 있다. 특정 주제어에 관해서 중요도가 높은 웹 페이지는 다른 웹 페이지들로부터 많은 참조를 받게 된다.

간단한 예를 들어 설명하면, "자바(의미 : 프로그래밍언어)"에 관해 조사가 필요할 경우 사용자들은 검색엔진을 통하여 웹 페이지들을 추출해 낼 것이다. 이때 검색된 웹 페이지들의 하이퍼링크 관계를 조사하게 된다면 많은 웹 페이지들이 java.sun.com을 가리키고 있을 것이다. 이러한 하이퍼링크 정보를 정보검색에 활용하면, 특정 주제어에 관하여 대표성(Representativeness)을 나타내는 웹 페이지를 추출해 낼 수 있게 된다. 이러한 웹 페이지의 연결 구조를 파악하여 정보검색에 이용한 상용화 검색엔진이 구글이다. 구글에서 사용한 웹 페이지의 연결 구조에 의한 가중치 부여방식은 Kleinberg의 HITS 알고리즘에 잘 소개되어 있다[13].

다이렉트히트는 특정 질의어와 일치하는 단어가 웹 페이지에 많이 포함되어 있다고 적합한 웹 페이지라 단정지을 수 없다는 가정에서 개발이 추진되었다. 다이렉트히트에서는 웹 페이지 가중치 결정 방법에 사용자의 반응을 이용함으로써 다수의 사용자가 특정 질의어에 대하여 어떠한 웹 페이지를 많이 참조하는지 알 수 있다. 그리고 다수의 사용자에게서 특정 질의어나 주제어로 많은 참조가 이루어진 웹 페이지는 다른 사용자에게도 중요도가 높은 웹 페이지이다. 이러한 사용자 반응을 웹 페이지의 가중치 결정 방법에 도입함으로써 내용기반 방법의 단점을 보완할 수 있었

으며, 웹 페이지의 인기도를 웹 페이지의 가중치에 반영함으로 부가적 정보 이용의 효과(사용자 관심도)를 얻어 낼 수 있게 된다.

웹 페이지의 순위를 결정하기 위한 가중치 결정 방식은 검색엔진 개발자들의 기술 축적으로 좋은 알고리즘이 많이 개발되어 왔지만, 웹 페이지 가중치 결정 방식에서 사용자들의 웹 페이지에 대한 반응 또한 간과할 수 없는 요소이다. 사용자의 결과 웹 페이지 검색 행위를 분석하면, 사용자들은 검색엔진을 통하여 질의어를 던진 후 검색엔진에서 보여 준 결과 웹 페이지들 중에서 질의어의 의미와 합당한 웹 페이지만을 검색하게 된다는 것을 알 수 있다. 이러한 사용자 의지가 담긴 결과 웹 페이지의 검색을 웹 페이지에 대한 사용자 반응이라 한다. 이러한 사용자의 반응을 이용하면 웹 페이지의 인기도(Popularity)를 결정할 수 있다.

구글은 하이퍼링크 정보를 이용하여 웹 페이지에 가중치를 부여하고, 다이렉트히트는 특정 질의어에 대하여 과거의 사용자로부터 많은 검색이 이루어졌던 웹 페이지에 높은 가중치를 부여하는 방식을 따르고 있다. 이러한 새로운 웹 페이지 가중치 방식의 도입으로 전통적 방식을 따르는 검색엔진보다는 성능을 향상시킬 수는 있으나, 초기의 검색에서 질의어의 형태를 이용한 질의 방식을 따르므로 여전히 질의어의 모호성이 해결되지 못한 상태로 결과 웹 페이지가 추출되는 단점을 내포하고 있다.

2.2 웹 페이지 가중치

정보의 가치는 사용자의 요구에 따라 달라진다[8]. 이러한 사용자의 정보 요구란 해당 정보 매체들의 가치를 결정하는 중요한 요소이다. 이러한 사용자의 정보 요구에 부응하기 위해 특정 행위를 사용자에게 요구하는 검색엔진들도 있으며, 검색엔진을 특정 영역으로 제한한 것들도 있다[10]. 이와 같은 사용자의 요구에 올바른 정보 결과를 제공하기 위허서는 사용자의 요구를 정확히 파악하는 것이 우선시 되어야 하고, 사용자의 요구에 대한 정확한 파악을 위해 사용자 반응에 대한 연구가 선행되어야 한다. 사용자 반응을 이용함으로써 검색엔진에서는 많은 장점을 얻을 수 있다. 이러한 장점을 극대화하기 위해 정보검색 분야에서는 사용자의 반응에 대한 많은 연구가 이루어져 왔다. 본 논문에서는 사용자의 묵시적 반응을 웹 페이지의 가중치에 적용하였다.

웹 페이지의 검색 순위를 결정하는데 있어서 일반적으로 사용되는 방법은 단어빈도 방법, 벡터 유사도 방법, 기계학습 방법, 확률 기반 방법, 등이 있으며 최근에 많이 사용되어지는 방법은 확률을 기반으로 한 베이지안 방법이다. 그러나 어휘와 문서 종속적인 방법은 검색 대상이 너무 많기 때문에 결과 웹 페이지에서 유용하고 중요한 웹 페이지를 추출하기 어렵다. 이러한 문제를 해결하기 위하여 사용할 수 있는 좋은 방법은 범주화된 사용자의 반응과 하이퍼링크 가중치이다.

사용자의 문서 선택 행위는 해당 웹 페이지에 대한 묵시적인 평가라고 할 수 있으며, 이러한 사용자의 정보 선별 행동들을 누적하여 웹 페이지의 가중치에 활용함으로써 웹 페이지에 인기도를 부여할 수 있다[19, 24]. 특정 질의어에 대해서 다수 사용자에게 많은 참조가 이루어진 웹 페이지는 중요도가 높은 웹 페이지이라 할 수 있을 것이다. 이러한 사용자 반응을 웹 페이지의 가중치 결정 방법에 도입하면 웹 검색엔진의 성능을 향상시킬 수 있다. 다이렉트히트는 웹 페이지 가중치 결정에 사용자의 반응을 이용한 검색엔진이다.

웹 페이지의 순위 결정의 기반 연구는 문서에 대한 검색에서 출발하였다. 과거에 사용한 순위 결정 방법은 어휘 종속적인 방법을 채택하였다. 그러나 어휘를 기반으로 한 적용 방법은 수량에 근거하게 되므로 한계성을 가지고 있다. 이러한 문제점을 해결하기 위해 기계학습 기법을 응용하기도 하며, 근래에는 확률 기법을 응용한 방법들을 채택하기도 한다. 대표적인 확률 기법으로는 베이지안 확률 기법을 예로 들 수 있다. 또한 결과 웹 페이지의 순위 결정에 질의어 이외에 사용자의 내력을 이용하기도 한다[21].

웹 페이지의 순위 결정 목적은 사용자의 요구에 맞는 웹 페이지를 쉽게 찾을 수 있도록 하는 것이다[8]. 순위는 정보 가치에 따라서 결정되는 것이며, 웹 페이지의 가치는 사용자의 요구에 따라 달라진다. 정보 추출과 사용자 편리성 측면에서 검색엔진에 의해 추출된 검색 결과는 사용자의 요구에 맞게 순위 정책을 적용하여 사용자에게 제시되어야 한다. 메타 검색엔진 분야에서는 웹 페이지의 순위 결정에 사용자가 요구하는 정보에 정확한 반응을 하기 위해 질의어에 종속적인 검색엔진을 선택함으로 웹 페이지 순위 결정 대상 집합의 오염도를 낮추어 웹 페이지 순위 결정에 도움을 주도록 유도하기도 한다[9]. 또한 웹 페이지의 순위 결정에 웹 페이지의 생성 날짜, 웹 페이지의 크기, 웹 페이지의 구조, 등을 이용하기도 한다[8].

웹 페이지의 순위 결정에 사용하는 또 다른 방법으로는 명성 평가 기법이 있다. 명성 평가 기법은 주제에 관한 웹 페이지의 명성을 평가하는 방법으로 웹 페이지의 하이퍼링크 결합과 내용 분석에 의해 특정 주제에 대하여 해당 웹 페이지가 얼마나 명성이 있는가를 계산하는 기법이다[1]. 명성 평가 기법은 검색 질의에 대한 페이지의 순위를 측정하기 위해 penetration과 focus라는 두 가지 비율을 사용한다. 주제를 t라 하고 페이지를 p라고 가정할 때, 주제 t에 대한 페이지 p의 penetration은 페이지 p를 가리키며 주제 t인 페이지 수를 주제 t에 관한 전체 페이지의 수로 나눔으로 측정된다. 주제 t에 관한 페이지 p의 focus는 페이지 p를 가리키며 주제 t인 페이지의 수를 페이지 p를 가리키는

페이지의 수(in-degree)로 나누어 측정할 수 있다[1]. penetration은 주제 t인 임의의 페이지가 페이지 p를 가리킬 확률을 말하며, focus는 페이지 p를 가리키는 임의의 페이지가 주제 t일 확률을 말한다.

이와 같은 하이퍼링크를 활용하여 웹 페이지의 순위 부여에 이용하는 연구는 Kleinberg에 의해 초기 연구가 이루어 졌으며, Kleinberg는 하이퍼링크의 in-link와 out-link를 활용하여 authority 페이지와 hub 페이지를 정의하여 웹 페이지의 가중치에 적용함으로 웹 페이지에 대표성(Representativeness)을 나타내었다[13]. 이러한 페이지의 구분 중, authority 페이지는 중요 정보를 많이 내포한 페이지라 할 수 있으며, hub 페이지는 중요 정보에 대한 링크를 많이 소유한 페이지라 할 수 있다. Authority와 hub에 대한 가중치 공식은 $a(p) := \sum_{q \to p} h(q)$, $h(p) := \sum_{p \to q} a(q)$이다. $p \to q$는 p 페이지가 q 페이지로의 하이퍼링크가 존재한다는 것을 뜻한다. 즉, a(p)는 페이지 p를 가리키는 모든 페이지 q의 h(q)의 합이고, h(p)는 페이지 p가 q로의 하이퍼링크로 모든 페이지 q의 a(q)의 합이 된다. a(p)와 h(p)의 가중치로 페이지 p의 중요도를 결정하는 것이 HITS 알고리즘이다. 이러한 하이퍼링크 구조 이용에 대해서는 Kleinberg의 HITS 알고리즘에 잘 소개되어 있다. 이와 같은 웹 페이지의 하이퍼링크 구조를 웹 페이지의 가중치에 적용한 검색엔진은 구글이다.

본 논문에서는 사용자의 웹 페이지 선택 행위를 모니터하여 웹 페이지의 협동적 평가 방법과 하이퍼링크 가중치를 웹 페이지 의미 카테고리 가중치에 저장하였으며, 의미 카테고리 저장 방식은 질의어의 의미를 이용하여 가중치를 카테고리별로 각각 저장함으로써 웹 페이지에 인기도와 대표성을 부여하고 변별력을 증가시킨다.

2.3 워드넷(WordNet)

사람들이 사용하고 있는 언어는 어느 국가의 언어이든 간에 다의어를 포함하게 된다. 다의어라 함은 정형화된 특정 형태의 단일 어휘가 여러 가지의 의미를 갖는 경우를 말한다. 이러한 어휘의 다의성 문제로 인하여 특정 어휘의 형태적 표현으로는 정확한 뜻을 알아 낼 수가 없다. 그리고 일반적으로 특정 단어나 단문의 표현 정도로는 어휘의 모호성을 해결할 수 없는 경우가 많다. 이러한 어휘의 다의성 문제는 사람들의 대화 내용뿐만 아니라, 검색엔진에서도 검색 질의어에 모호성 문제를 발생시킨다. 웹 검색엔진에서 다의성을 내포한 질의어를 사용하였을 경우, 질의어의 모호성 문제로 원하지 않은 결과 웹 페이지가 추출되는 경우가 발생하게 된다. 본 논문에서는 검색 질의어의 모호성을 해결하는 방안으로 워드넷(WordNet)을 이용하여 검색 질의어의 모호성 해결을 시도하였다.

일반적으로 검색엔진들은 어휘 사전이 필요하며, 어휘의 방대한 규모로 인해 저장 방법이나 어휘 추출 방법이 검색 속도에 많은 영향을 미친다. 워드넷은 어휘의 저장 형태가 어휘 의미를 이용한 계층적 구조로 이루어져 있으므로 어휘의 형태를 활용한 일반적인 어휘 사전에 비해 검색 효율이 좋다고 할 수 있으며, 연관성이 높은 언어를 검색할 경우에는 의미 계층을 활용하게 되므로 연관어휘 추출 효율이 일반 어휘 사전에 비해 탁월하다 할 수 있다.

워드넷은 1985년 프린스톤(Princeton)대학 인지과학연구실을 주축으로 연구되어 지금은 워드넷 버전 1.7.1까지 발표되었다. 워드넷을 개발하게 된 동기는 어휘의 형태만을 고려하여 어휘 사전을 구성하였을 경우에 관련 어휘를 검색하고 어휘들의 연관성을 나타내기엔 부적합한 면을 많기 때문이다. 이러한 문제점 해결을 위해 동의, 반의, 상위, 하위, 등과 같은 어휘의 연관 관계를 어휘 사전에 도입하게 되었다. 워드넷은 어휘의 의미에 대한 카테고리 분류가 잘 정의되어 있으며, 어휘들의 계층 구조와 연관 관계가 잘 표현되어져 있다.

워드넷의 원리는 개념 행렬을 기초로 정의되었다. 개념 행렬 모델을 <표 1>에 나타내었다. <표 1>은 워드넷의 어휘 의미 관계를 나타내는 기본 원리이다[13]. <표 1>에서, Fn은 어휘의 형태를 나타내고 Mn은 의미를 나타낸다. 그러므로 E1,1은 F1의 어휘 형태를 갖고서 M1의 의미를 나타낸다. F1과 F2는 M1의 의미에 있어서 동일한 의미를 가지고 있으면서 서로 다른 어휘의 형태를 가지고 있으므로 F1과 F2는 동의(synonym) 관계를 나타낸다. F2는 하나의 어휘 형태를 가지면서 어휘의 의미는 M1과 M2를 가지므로 다의(polysemy)어가 된다. {F1, F2}는 M1의 의미에 있어서 동의어 집합(set of synonyms, synset)이 된다[18].

〈표 1〉 어휘의 개념 행렬

Word / Meanings	F1	F2	F3	F4	....	Fn
M1	E1, 1	E1, 2				
M2		E2, 2				
M3			E3, 3			
:					....	
Mn						Em, n

예를 들어 {tabby, cat, pussy}와 같은 어휘 집합은 어휘의 형태는 다르지만 같은 의미를 소유한 단어로 구성되어 있으며, 이와 같은 어휘 집합을 동의어 집합(synset)이라고 한다. 반의(antonymy) 관계는 서로 상반되는 단어의 의미를 가지고 있는 어휘 사이의 관계를 나타내는 것이다. 예를 들면, 집합 A = {rise, ascend}와 집합 B = {fall, descend}는 각각 동의어 집합을 이루고 있으며, 집합 A와 집합 B는 반의 관계를 유지하게 된다. 하위(hyponymy) 관계와 상위(hypernymy) 관계는 단어들 사이의 의미 관계(semantic rela-

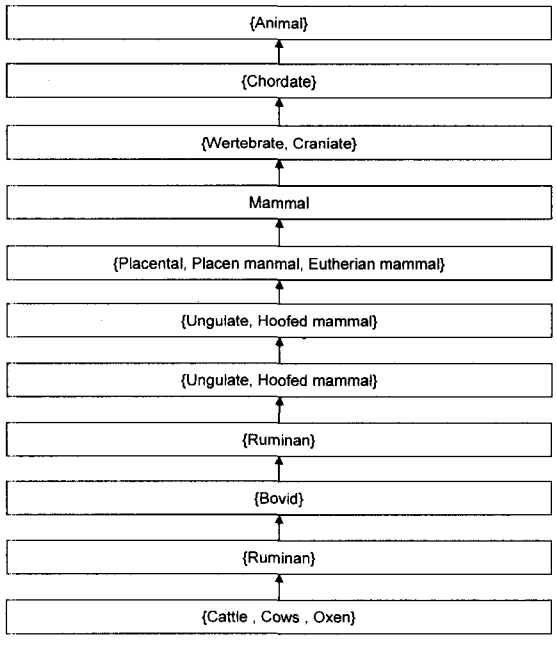
tion)로 단어의 의미 계층(semantic category)을 중추적으로 나타낸다. 예를 들어 "maple"은 "tree"의 하위어가 되며, "tree"는 "maple"의 상위어가 된다. 부분(meronymy) 관계와 전체(holonymy) 관계는 단어들 사이에서의 포함 관계를 나타낸다. 예를 들어 "leaf"는 "tree"의 부분 관계이며 "tree"는 "leaf"의 전체 관계가 된다. 이러한 어휘들의 연관 관계가 워드넷에서는 잘 정의되어 있다[18].

워드넷의 구조를 보면 위에서 언급한 기본 원리를 이용하여 어휘 데이터를 도식화 구조로 사용할 수 있게 하였다. 이러한 어휘의 도식화가 가능할 수 있었던 이유는 워드넷에서 사용한 어휘들은 의미를 이용하여 구조화되었기 때문이다. 워드넷에서는 명사 사전의 최상위 카테고리가 26개로 이루어져 있으며, 모든 어휘들은 최상위 카테고리에 하나 이상 포함되게 된다. 그러므로 특정 어휘의 형태적 모습은 워드넷의 구조에서 중요한 부분을 차지하지 않는다. <표 2>에서 명사 사전에서의 26개 최상위 카테고리를 나타내었다.

〈표 2〉 최상위 카테고리

명사 사전에서의 최상위 카테고리
Action, Animal, Artifact, Property, Body, Food, Location, Event, Cognition, Communication, Feeling, Motive, Object, Natural, Person, Plant, Possession, Process, Quantity, Relation, Shape, State, Substance, Time, Group, Top

어휘의 계층적 구조를 도식화하면 (그림 1)과 같이 나타낼 수 있다. (그림 1)은 소라는 의미의 동의어 집합인 {oxen, cattle, cows}에 대한 의미 계층 구조 관계를 나타낸 것이다. {oxen, cattle, cows}는 소과의 동물이므로 {bovine}은 {oxen, cattle, cows}의 상위어가 된다. 이와 같이 의미의 계층을



(그림 1) 동의어 집합의 계층 구조

따라가면 최상위 의미 계층은 {animal}이 된다. 이와 같이 어휘의 의미 계층 구조를 정보검색에 이용하게 되면 검색에 많은 도움을 줄 수 있다. 본 논문에서는 이러한 워드넷의 의미 계층 구조를 검색엔진에 활용하는 방안을 연구하였다. 본 논문에서 제시한 실험용 검색엔진에서는 질의어의 모호성 해결과 웹 페이지의 가중치 저장에 워드넷을 활용하여 검색엔진의 성능 향상을 이룰 수 있었다.

유럽이나 일본 등 선진국가에서는 어휘 의미를 이용한 사전의 중요성을 인식하여 자국어 워드넷을 소유하고 있다. 아직은 국내에서 워드넷에 관한 연구 활동은 미비한 상태이나, 앞으로 한국어 정보검색의 발전을 위해서는 필요한 연구 분야라 할 수 있다.

3. 실험용 검색엔진

전통적 방법인 내용기반 방법을 사용한 검색엔진들의 단점을 보완한 차대대 검색엔진과 워드넷의 분석을 통하여 실험용 검색엔진을 완성하게 되었다. 본 장에서는 실험용 검색엔진의 각 모듈들과 웹 페이지 가중치 부여 방법에 대하여 설명한다.

3.1 질의어의 모호성 해결과 재조합 질의어

사용자가 검색엔진을 통하여 질의를 할 경우, 검색엔진은 사용자가 던진 질의어의 형태만을 고려하고 웹 페이지를 추출하게 된다. 이러한 내용기반 방법은 사용자가 어떤 의미로 질의어를 사용한 것인지 정확한 판단을 내릴 수 없기 때문에 검색엔진에서 올바른 정보를 추출할 수 없게 된다. 간단한 예로 "자바(의미 : 섬)"에 관해서 조사하고 싶을 경우 사용자는 "자바"를 검색 질의어로 사용하게 될 것이다. 그러나 웹 검색엔진에서는 "자바(의미 : 섬)"에 대해 정확히 검색되지 않는다. 이러한 이유는 현재 프로그래밍언어로 각광을 받고 있는 "자바(의미 : 프로그래밍언어)"의 웹 페이지들이 높은 가중치를 소유하고 있기 때문이다. 이와 같은 원인으로 "자바(의미 : 섬)"에 대해 검색한 경우, "자바(의미 : 프로그래밍언어)"에 관한 웹 페이지가 더 많이 결과로 추출된다. 본 논문에서는 이러한 문제 해결을 위해 워드넷을 이용한 질의어의 모호성 해결 방안을 제시한다.

본 논문에서는 질의어의 모호성 해결을 위하여 워드넷 기반 사용자 인터페이스를 설계하였다. 사용자 인터페이스에서는 워드넷의 명사 사전을 활용하여 동의어, 상위어, 주석을 추출한다. 추출한 동의어, 상위어, 주석을 사용자에게 나타냄으로 질의어의 의미 선택에 편리성을 제공하였다. 사용자는 자신의 질의어 의미에 부합하는 의미를 선택하게 되고, 선택된 의미를 활용하여 검색에 이용될 재조합 질의어를 생성하게 된다. 이렇게 완성된 재조합 질의어는 검색엔진에서 정확한 검색을 위한 질의어로 사용된다. <표 3>은

질의어의 재조합을 위해 워드넷에서 추출해 낸 속성들의 예제이다.

〈표 3〉 워드넷을 이용한 java의 동의어, 상위어, 주석 추출

동의어	상 위 어	주 석
Java	island	an island in Indonesia S of Borneo ; one of the world's most densely populated regions
Coffee Java	beverage drink	a beverage consisting of an infusion of ground coffee beans
Java	object-oriented_ programming_ language	a simple platform-independent object-oriented programming language used for writing applets that are downloaded from the World Wide Web by a client and run on the client's machine

<표 3>에서 나타난 동의어, 상위어, 주석을 이용하여 질의어를 재조합할 경우 동의어와 상위어는 변형을 가하지 않고서 활용하게 된다. 그러나 주석을 활용하여 질의어를 재조합할 경우에는 주석의 형태가 문장으로 이루어져 있으므로 주석에서 불용어를 제거시킨 후 사용하게 된다. 재조합 질의어에 사용될 명사 집합들의 구성 예제를 <표 4>에 나타내었다. <표 4>의 예제는 질의어가 "java(의미 : 커피)"일 경우 질의어의 재조합에 사용되는 명사 집합을 나타낸 것이다.

본 논문에서는 동의어, 상위어, 주석을 이용한 검색 실험에서 주석이 가장 좋은 검색 성능을 나타냄을 확인하여 주석을 활용한 재조합 질의어를 이용하여 본 실험에 임하였다.

〈표 4〉 질의어 재조합을 위한 명사 집합의 구성 예제

	재조합에 활용될 원시 어휘들	재구성된 명사 집합
동의어	java, coffee	java+coffee
상위어	beverage, drink	java+beverage+drink
주 석	a beverage consisting of an infusion of ground coffee beans; he ordered a cup of coffee	java+beverage+infusion+coffee+bean

3.2 의미 카테고리 기반 협동적 평가 방법과 하이퍼링크 가중치

실험용 검색엔진은 중요한 웹 페이지에 높은 가중치를 주기 위해 의미 카테고리 기반 협동적 평가 방법과 하이퍼링크 가중치 방식을 적용하였다. 가중치에 협동적 평가 방법을 적용하기 위해 사용자들의 웹 페이지에 대한 반응을 이용하였다. 사용자 반응 정보의 축적을 위해 사용자 모니터를 설계하여 사용자의 웹 페이지에 대한 반응(웹 페이지의 클릭)을 가중치 데이터베이스에 저장하였다. 이때 사용자의 클릭 가중치는 원시 질의어의 의미에 해당하는 카테고리에만 가중치를 적용함으로써 질의어 모호성이 해결된 상태로 가중치가 저장되게 된다. 본 논문에서 활용한 의미 카테고리 방식은 워드넷 명사 사전의 최상위 카테고리 26개를 웹 페이지의 가중치 필드로 활용하였다. 이러한 방법을 채택함으로 질의어의 의미에 따라 웹 페이지의 가중치를 세분화하여 저장할 수 있다.

의미 카테고리를 사용한 협동적 가중치 방식의 장점을 간단한 예로 설명하여 보겠다. <표 5>에 나타낸 것은 두 개의 웹 페이지에 대한 협동적 가중치의 예이다. <표 5>에서 URL1은 섬 관련 가중치로 50, 프로그램언어 관련 가중치로 10, 단체 관련 가중치로 10을 가지고 있고 URL2는 섬 관련 가중치로 10, 프로그램언어 관련 가중치로 10, 단체관련 가중치로 100을 가지고 있다. 이러한 URL의 가중치가 저장되어 있을 때 검색 질의어로 "java(의미 : 섬)"가 사용되어진다면 검색엔진에서는 URL2의 가중치 총합이 가장 높게 나타나 있으므로 URL2가 보다 중요한 페이지로 인식되어 상위에 검색되게 된다. 그러나 질의어의 의미로 판단하게 되면 "java(의미 : 섬)"의 최상위 카테고리인 location에 대하여는 URL1이 URL2보다 높은 가중치를 가지고 있어 더 중요한 페이지임을 알 수 있다.

〈표 5〉 웹 페이지의 협동적 가중치 결정 방식에 대한 카테고리 분류 예제

검 색 어	URL1의 가중치	URL2의 가중치
Island(섬과 관련된 질의어)	50	10
Program Language(프로그램언어에 관련된 질의어)	10	10
Organization(단체에 관련된 질의어)	10	100
웹페이지 가중치의 총합	70	120

하이퍼링크에는 해당 페이지를 가리키는 인링크(inlink)와 다른 웹 페이지를 참조하는 아웃링크(outlink)가 있다. 본 논문에서는 하이퍼링크 가중치인 링크 가중치를 웹 페이지 가중치에 적용하였다. 본 논문에서는 웹 페이지의 정보가 풍부할 때 발생 빈도가 높게 나타나는 인링크를 링크 가중치로 활용하였다. 웹 페이지 가중치 중 클릭 가중치는 해당 웹 페이지의 카테고리별 클릭 가중치를 최대 클릭 가중치로 나눈 후에 log를 취하여 일반화한 후에 카테고리별 웹 페이지 총수의 반으로 나눔으로써 가중치에 부분적 작용을 하도록 하였다. 링크 가중치는 해당 웹 페이지를 가리키는 웹 페이지의 총수를 최대 링크의 총수로 나눈 후에 log를 취하여 일반화한 후에 카테고리별 웹 페이지의 총수의 반으로 나눔으로써 가중치에 부분적 작용을 하도록 하였다. 이러한 카테고리별 클릭 가중치와 링크 가중치를 합하여 웹 페이지의 가중치로 저장하였다. 웹 페이지의 가중치 적용 방법에 대한 수식은 <표 6>에 나타내었다.

본 논문에서는 웹 페이지 가중치 방식에 질의어 의미를 활용한 카테고리 기반 협동적 평가 방법과 링크 가중치 방법을 적용함으로써 검색의 정확도를 높일 수 있었다.

〈표 6〉 웹 페이지의 가중치 적용 방식

어휘의 의미	T_sen_j	카테고리별 웹 페이지들의 총수	$\sum(P\_cat_i)$
임의의 웹 페이지	P_k	웹 페이지 가중치	$Weighting(P_k)$
클릭 가중치	W_click	카테고리별 클릭 가중치	$U\_cat_i\_clk\_w$
링크 가중치	W_inlink	클릭 가중치의 최대값	$Max(U\_cat_i\_clk\_w)$
인링크 총수	$Inlink\_tot$	인링크의 총수 중 최대 값	$Max(Inlink\_tot)$
카테고리별 웹 페이지들의 집합		$P\_cat_i = \{ i \in I \mid P\_cat_1 \le P\_cat_i \le P\_cat_n \}$	

$$W\_click = \log_2\left(\frac{U\_cat_i\_clk\_w}{Max(U\_cat_i\_clk\_w)}+1\right)\cdot\frac{\sum(P\_cat_i)}{2}$$

$$W\_inlink = \log_2\left(\frac{Inlink\_tot}{Max(Inlink\_tot)}+1\right)\cdot\frac{\sum(P\_cat_i)}{2}$$

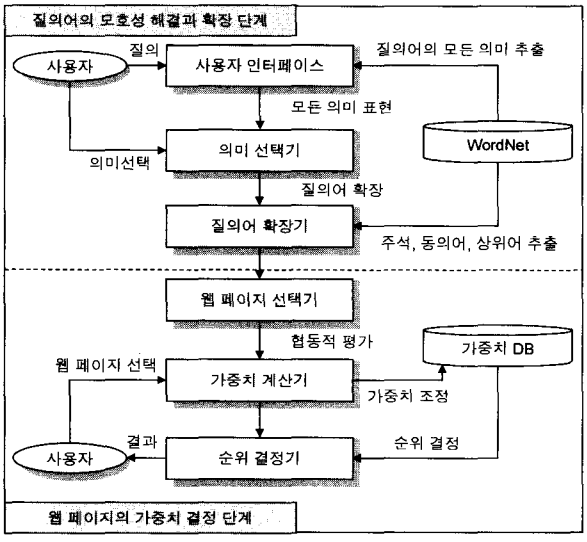
$$Weighting(P_k) = W\_click + W\_inlink = \left(\log_2\left(\frac{U\_cat_i\_clk\_w}{Max(U\_cat_i\_clk\_w)}+1\right)\cdot\left(\frac{\sum(P\_cat_i)}{2}\right)\right) + \left(\log_2\left(\frac{Inlink\_tot}{Max(Inlink\_tot)}+1\right)\cdot\left(\frac{\sum(P\_cat_i)}{2}\right)\right)$$

3.3 시스템 구현

본 연구에서는 실험을 위하여 펜티엄-III 1GHz, 메모리 512MB인 시스템을 사용하였으며, 시스템의 운영체제로는 레드햇(redhat) 리눅스 6.0 버전을 사용하였다. 웹 서버는 아파치(apache) 웹 서버를 사용하였으며, 웹 서버에서 자바 서블릿 프로그램을 실행하기 위해서 JServ 1.1 버전을 사용하였다. JServ는 아파치 자바 프로젝트(http://java.apache.org)에서 개발한 프로그램으로써 자바 서블릿 프로그램을 아파치 웹 서버에서 연동할 수 있도록 지원하는 프로그램이다. 실험을 위한 검색엔진은 자바(java)와 자바 서블릿(java servlet)을 이용하여 구현되었다. 구현에 사용된 자바 개발 도구로는 JDK(java development kit) 1.3 버전을 이용하였으며 자바 서블릿은 JSDK(java servlet development kit) 2.0을 사용하여 구현하였다. 또한 워드넷이 보다 효과적인 검색 및 정보 처리를 위해 자체적으로 개발한 파일 형식을 사용한 것처럼, 실험용 검색엔진의 데이터 처리도 자체적으로 규격화한 파일 데이터베이스를 사용하였다.

본 논문에서 제안한 협동적 평가와 하이퍼링크를 이용한 가중치 결정 방식의 실험을 위해 세부적인 모듈을 설계하여 실험용 검색엔진을 구현하였다. 본 연구에서 개발한 실험용 검색엔진의 구성도를 (그림 2)에 나타내었다.

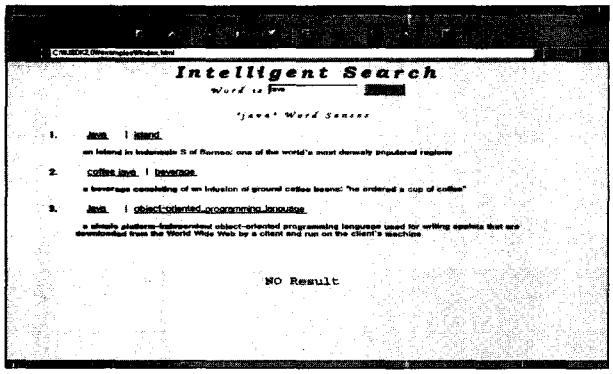
실험용 검색엔진에서 사용한 사용자 인터페이스는 워드넷과 연동되어 질의어에 해당하는 모든 의미를 사용자에게 나타냄으로써 사용자의 의미 선택을 요구하게 된다. 사용자가 선택한 질의어의 의미에 대해 질의어 확장기는 워드넷에서 추출한 명사 집합을 활용하여 재조합 질의어를 생성한다. 웹 페이지 선택기는 재조합 질의어를 이용하여 질의어의 의미에 적합한 웹 페이지들을 데이터베이스에서 추출하는 기능을 수행한다. 순위 결정기는 웹 페이지의 중요도를 판단하는 모듈로써 특정 질의어의 의미에 대해 웹 페이지의 순위를 결정하는 것으로 순위를 결정하기 위해 질의어의 의미에 해당하는 카테고리별 클릭 가중치와 링크 가중치를 활용한다. 순위 결정 과정을 거친 후에 선택된 결과 웹 페이지들은 중요도 순으로 사용자 인터페이스에 출력된다. 출력된 결과 웹 페이지들이 사용자의 검색(클릭)을 받게 될 경우, 가중치 계산기는 해당 웹 페이지의 카테고리별 협동적 가중치를 갱신한다. 이때 갱신되는 가중치는 질의어의 의미에 부합하는 가중치 필드만을 조정한 후에 해당 웹 페이지의 가중치 데이터베이스에 저장된다. 본 논문의 실험에서 사용된 사용자 인터페이스 화면은 (그림 3)과 같다.



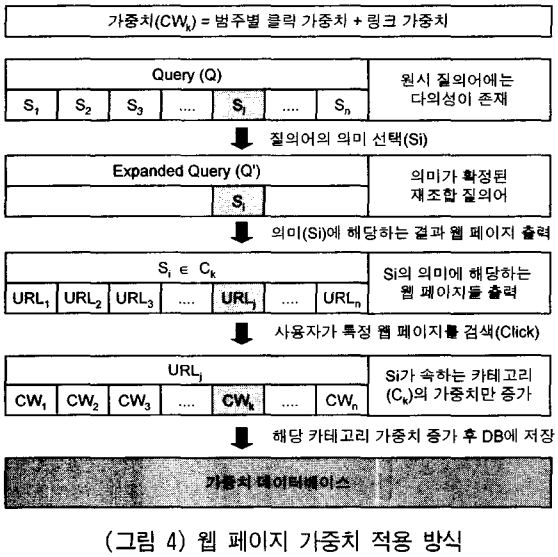
(그림 2) 시스템 구성도

질의 처리 과정과 협동적 평가에 의한 가중치 조정 과정을 (그림 4)에 나타내었고, (그림 4)는 의미 카테고리 기반의 가중치 적용 방법에 대한 예이다. (그림 4)를 보면, 원시 질의어(Q)가 입력되었을 경우 의미 선택기를 통하여 사용자가 사용한 질의어의 의미(S_i)를 결정한 후 질의어 확장기

를 통하여 새로운 재조합 질의어 (Q')를 완성한다. 재조합 질의어를 활용하여 결과 웹 페이지를 추출한 후에 순위 결정기에 의해 웹 페이지들의 순위가 결정되고 결과 웹 페이지들이 순위를 기반으로 사용자에게 보내지면 사용자는 자신이 원하는 정보에 합당한 웹 페이지를 검색하게 된다.



(그림 3) 사용자 인터페이스



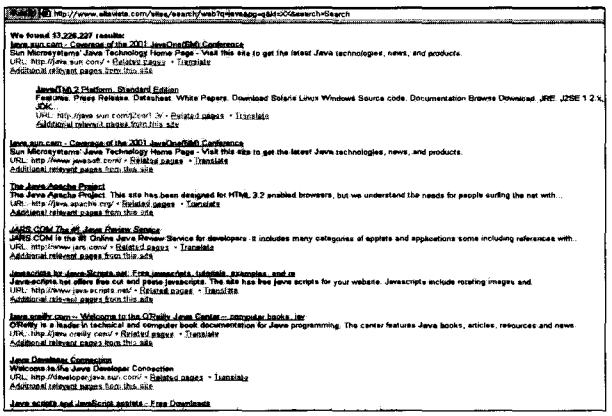
(그림 4) 웹 페이지 가중치 적용 방식

이때 검색된 웹 페이지는 질의어의 의미에 해당하는 가중치 필드의 가중치 값을 증가시킨다. (그림 4)에서 Si는 질의어에 해당하는 의미이며 Ck는 Si가 포함되는 워드넷의 최상위 카테고리이다. CWk는 웹 페이지의 카테고리별 가중치로 카테고리별 협동적 평가 가중치와 링크 가중치로 구성되어 있다.

결과 웹 페이지들은 사용자의 클릭을 받게 될 경우에 가중치 계산기에 의하여 가중치가 조정된다. 이때 조정되는 가중치는 질의어의 의미에 부합하는 가중치 필드만을 조정한 후에 가중치 데이터베이스에 저장된다.

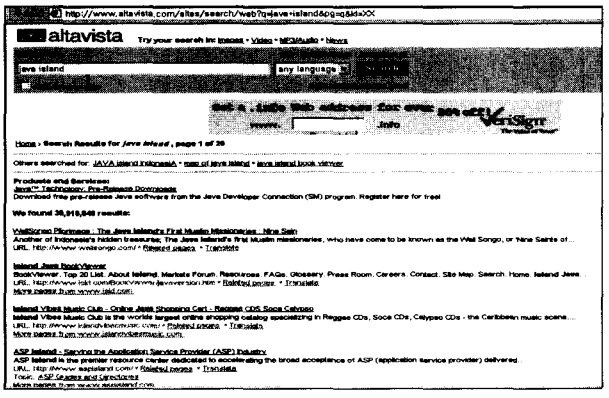
3.4 실험용 검색엔진과 일반 검색엔진의 비교

사용자가 "java(의미 : 섬)"에 관한 정보를 검색하고자 할 경우, 대다수의 일반 사용자들은 단일 키워드를 사용하여 검색하게 된다. 그러나 정보검색 숙련자들은 정확한 정보 추출을 위해 복합 키워드를 사용한다.

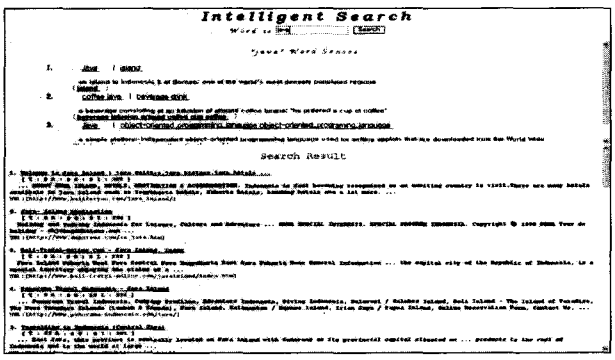


(그림 5) "java(의미 : 섬)"를 질의어로 사용하여 알타비스타에서 검색한 결과

(그림 5)는 단일 키워드를 활용하여 알타비스타에서 "java (의미 : 섬)"로 검색하였을 경우의 결과 화면이다. (그림 5)에서 알 수 있듯이 단일 키워드의 사용은 질의어의 모호성 문제로 인하여 일반 검색엔진에서는 사용자가 원하는 결과 웹 페이지를 정확히 추출할 수 없었다.



(그림 6) "java island(의미 : 섬)"를 질의어로 사용하여 알타비스타에서 검색한 결과



(그림 7) "java(의미 : 섬)"를 질의어로 사용하여 실험용 검색엔진에서 검색한 결과

(그림 6)은 복합 질의어를 활용하여 알타비스타에서 검색한 결과 화면으로 질의어는 모호성 해결을 위해 "java island(의미 : 섬)"를 사용하였다. (그림 6)을 보더라도 자바 섬에 대한 중요한 웹 페이지들은 검색되지 않음을 알 수 있다. 복합 질의어를 사용함으로 질의어의 모호성 문제가 완화될 수는 있으나 완전히 해결되지는 않으며, 내용기반 방법만을 사용할 경우는 중요한 웹 페이지를 상위에 위치시키지 못한다.

(그림 7)은 실험을 위해 설계된 실험용 검색엔진의 결과 화면이다. (그림 7)은 "java(의미 : 섬)"라는 질의어를 사용하여 질의를 하였을 경우의 결과 화면으로 일반 검색엔진에서의 결과보다 높은 정확도를 나타내고 있다. (그림 7)의 결과 화면을 보면 질의어의 모호성 해결이 검색엔진에 중요한 작용을 한다는 것을 알 수 있으며, 의미 카테고리 기반의 협동적 가중치와 링크 가중치를 웹 페이지 가중치에 적용함으로써 질의어의 의미에 적합한 중요한 웹 페이지를 상위 결과에 위치시킬 수 있다.

4. 실험 및 실험 결과

실험용 검색엔진의 실험을 위해 가중치 데이터베이스를 구성하였으며, 본 논문에서 제시한 웹 페이지의 가중치 방식에 대한 실험을 위해 사용자 반응이 필요하였다. 이러한 사용자 반응의 필요성으로 특정 사용자들에게 웹 페이지에 대한 반응을 보이도록 하였다.

4.1 실험 데이터 및 실험 방법

본 실험에 사용되어진 실험 데이터는 <표 7>에 나타나 있다. <표 7>을 보면 한 가지 형태를 가진 어휘가 여러 개의 의미를 소유하는 것을 확인할 수 있다. <표 7>에 나타난 어휘의 의미를 가지고 일반 상용화 검색엔진인 알타비스타를 이용하여 웹 페이지를 수집하였으며, 수집한 웹 페이지는 실험에 사용된 질의어의 의미 당 400개의 웹 페이지들을 수집하였다. 비교 실험에 사용할 검색엔진은 알타비스타로 선정하였다. 알타비스타의 검색엔진을 비교 대상으로 선정한 이유는 전통적 방식의 검색엔진의 모습을 가장 잘 나타내고 있기 때문이며, 실험용 검색엔진의 데이터베이스에 저장된 웹 페이지들이 알타비스타에 의해 수집되었기 때문이다.

실험용 검색엔진은 의미 카테고리 기반 검색엔진과 의미 카테고리를 사용하지 않는 검색엔진으로 나누어 실험에 임하였다. 두 시스템 모두 워드넷을 활용하여 질의어의 모호성은 해결한 상태에서 실험하였다. 본 연구에서는 웹 페이지의 협동적 가중치를 적용하기 위하여 동국대학교 컴퓨터공학과 학생들 300명으로부터 사용자 반응을 받아 가중치 데이터베이스를 구성하였다.

<표 7> 실험에 사용된 질의어

실험용 질의어					
질의어 형태	질의어 의미	질의어 형태	질의어 의미	질의어 형태	질의어 의미
Java	커피	Coach	코치	Mass	미사
Java	섬	Coach	객차	Mass	질량
Java	언어	Bill	법안	Capital	자본
Custom	통관	Bill	계산서	Capital	수도
Custom	관습	Sentence	문장	Court	코트
Horse	마약	Sentence	판결	Court	법정
Horse	말	Balance	잔액	Culture	문화
Character	배역	Balance	저울	Culture	재배
Character	문자	Ball	공	Bank	은행
		Ball	무도회	Bank	둑

검색엔진의 성능 평가를 위하여 결과 웹 페이지 중 상위 30개 페이지와 상위 10개 페이지에서 사용자에게 중요한 웹 페이지가 몇 개가 존재하는가를 비교하였다. 이때 웹 페이지의 정확도를 판단 내리기 위해 동국대학교 컴퓨터공학과 대학원생 중 정보검색 전공자 14명을 선정하였다. 웹 페이지의 중요도 평가에는 1부터 10까지를 활용하였으며, 평가 점수에서 상위 점수와 하위 점수는 삭제하고 나머지 점수만을 산술 평균하여 5이상의 값을 소유한 경우에만 중요한 웹 페이지라 판정하였다.

4.2 실험 결과

실험에 사용한 방식을 기술하면, 실험용 검색엔진에서는 사용자가 질의어의 모호성을 나타내고 있는 하나의 단어를 이용하여 검색하게 된다. 사용자 인터페이스는 워드넷을 활용하여 질의어의 모호성을 해결한 후 재조합 질의어를 완성시키고, 재조합 질의어를 활용하여 웹 페이지를 검색하는 방식을 따른다.

실험용 검색엔진은 두 가지 유형으로 실험에 임하게 되는데, 첫 번째 실험 유형은 질의어의 모호성이 해결된 워드넷 기반 검색엔진으로 의미 카테고리 가중치를 적용하지 않는 방법이고, 두 번째 실험 유형은 질의어의 모호성이 해결된 워드넷 기반 검색엔진으로 의미 카테고리 가중치를 적용한 방법이다.

상용화 검색엔진에서도 서로 다른 두 가지 유형으로 실험하였다. 실험에 사용한 첫 번째 실험 유형은 하나의 키워드를 활용하여 검색하게 함으로 질의어의 모호성이 존재되어 있는 방식이고, 두 번째 실험 유형은 정보검색 숙련자들이 많이 사용하는 복수 질의어의 조합에 의한 검색 방식이다. 두 번째 실험 유형에서는 사용자에게 질의어의 수량에 제약을 주지 않고 사용자가 원하는 질의어들을 자율적으로 선택하여 사용하게 하였다.

이러한 실험 조건을 선정하게 된 이유는 검색엔진 사용

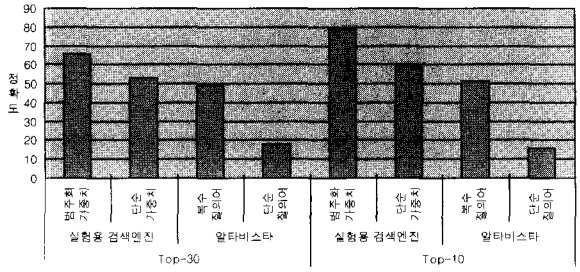
자들은 질의어의 조합을 이용하여 질의어 모호성 해결과 정보 추출의 정확도를 높이기 때문이다. 이와 같은 실험 방법을 통해 일반 사용자들은 하나의 질의어를 많이 활용하며, 정보검색 숙련자들은 2~3개 정도의 단어로 이루어진 질의어를 가장 많이 사용함을 알 수 있었다.

<표 8>은 상위 30개의 페이지와 상위 10개의 페이지에서 실험용 검색엔진과 상용화 검색엔진(알타비스타)의 비교실험에 대한 결과이다. <표 8>의 결과 중 상위 30개의 결과 웹 페이지를 보면 알타비스타에서 단일 키워드를 사용하였을 경우 평균 정확도는 18%로 낮은 성능을 나타낸다. 이와 같은 결과는 단일 질의어의 사용은 질의어의 모호성 문제를 해결할 수 없기 때문에 발생된다. 복수 질의어 사용에 의한 평균 정확도는 49%로 단일 질의어 사용보다 31%의 향상을 보였으나, 중요 웹 페이지를 상위에 위치시키지는 못하였다. 복수 질의어의 사용으로 질의어의 모호성 문제는 다소 해결되어 올바른 검색에 이용될 수는 있으나, 내

〈표 8〉 상위 30페이지에서의 성능 비교

질 의 어	의 미	실험용 검색엔진				Altavista			
		워드넷 기반				복수 질의어		단일 질의어	
		카티고리별 가중치		단순 가중치					
단어	의미	상위 30 페이지	상위 10 페이지	상위 30 페이지	상위 10 페이지	상위 30 페이지	상위 10 페이지	상위 30 페이지	상위 10 페이지
JAVA	섬	25	8	12	5	6	4	1	0
	언 어	30	9	25	8	27	10	26	5
	커 피	21	8	10	4	10	6	0	0
HORSE	말	29	9	22	5	21	6	18	3
	마 약	19	8	10	4	8	5	0	0
CUSTOM	관 습	12	9	7	2	8	3	4	2
	통 관	15	7	6	3	4	1	1	0
CHARACTER	문 자	13	8	13	7	12	5	8	3
	배 역	13	6	9	5	7	4	2	2
BANK	은 행	26	9	22	7	22	6	18	8
	둑	16	7	15	8	6	4	0	0
COACH	코 치	21	8	19	8	19	5	16	5
	객 차	16	7	15	7	15	6	6	2
BILL	법 안	23	8	18	6	15	4	4	3
	계산서	21	8	16	5	12	4	2	0
BALANCE	잔 액	18	9	16	7	13	2	1	0
	저 울	18	8	14	7	10	3	0	0
BALL	공	25	9	19	8	20	7	1	0
	무도회	22	9	16	6	18	6	3	1
CAPITAL	자 본	24	7	20	7	23	6	1	0
	수 도	23	8	25	5	19	2	1	0
COURT	경기장	20	7	21	6	17	7	5	1
	법 정	19	8	18	8	24	8	17	7
CULTURE	문 화	20	9	21	6	20	6	9	3
	재 배	14	7	13	8	14	7	4	1
MASS	미 사	18	6	15	3	13	8	1	0
	질 량	12	7	6	2	11	5	2	0
SENTENCE	문 장	25	9	23	8	18	6	6	3
	판 결	21	9	22	7	16	2	3	0
평균 정확도		**66%**	**79%**	**53%**	**59%**	**49%**	**51%**	**18%**	**16%**
상위 30 페이지에 대한 상위10 페이지의 정확도 증가율		**△13%**		**△6%**		**△2%**		**△2%**	
복수 질의 어휘에 대한 증가율(vs. Altavista)		**△17%**	**△28%**	**△4%**	**△8%**				
단일 질의 어휘에 대한 증가율(vs. Altavista)		**△48%**	**△62%**	**△35%**	**△43%**	**△31%**	**△35%**		

용기반 검색으로 인해 뛰어난 성능은 나타낼 수는 없었다.



(그림 8) 실험용 검색엔진과 알타비스타의 정확도

실험용 검색엔진에서는 클릭 가중치와 링크 가중치를 카테고리 활용 방식과 카테고리를 활용하지 않은 방식을 나누어 실험하였다. 카테고리를 활용하지 않은 가중치 적용 방식에 의한 상위 30개의 결과 웹 페이지의 평균 정확도는 53%의 성능을 나타내었다. 이러한 실험 결과는 알타비스타에서 복수 질의어를 사용한 평균 정확도보다 4%의 향상을 나타낸 것이며, 알타비스타에서 단일 키워드를 사용한 평균 정확도보다는 35%의 정확도 향상을 나타낸 것이다. 카테고리 기반 가중치를 적용한 실험에서는 상위 30개에 대한 평균 정확도가 66%를 나타내었다. 상위 30개에 대한 결과 중 카테고리 기반 가중치에 의한 평균 정확도는 알타비스타의 복수 질의어를 사용한 방식보다 17%의 성능 향상을 나타냈으며, 단일 질의어의 사용보다는 48%의 성능 향상을 나타내었다.

실험용 검색엔진의 두 가지 실험을 비교하면, 상위 30개에서 카테고리 기반 가중치를 사용한 검색엔진이 카테고리를 사용하지 않고 가중치를 적용한 검색엔진보다 13%의 정확도 향상을 나타내었다.

상위 10개의 결과 웹 페이지에 대한 정확도 평가에 대한 실험 결과를 보면 알타비스타에서 단일 키워드와 복수 키워드의 활용에 의한 평균 정확도는 16%와 51%이고, 실험용 검색엔진에서 의미 카테고리 활용 방식과 의미 카테고리를 활용하지 않은 방식은 79%와 59%의 평균 정확도를 나타내었다.

상위 30개와 상위 10개의 결과 웹 페이지의 평균 정확도를 활용하여 관련도 높은 웹 페이지를 상위에 위치시키는 것에 대한 비교에서 알타비스타는 단일 질의어를 활용한 경우 2%의 증가가 발생하였고 복수 질의어를 활용한 경우에는 2%의 증가가 발생되었다. 실험용 검색엔진은 카테고리 방식을 활용하지 않은 경우 6%의 증가가 발생되었고 카테고리 방식을 활용한 경우 13%의 증가가 발생되었다.

실험에 의해 검색엔진에서 질의어의 모호성이 검색 성능에 많은 영향을 미친다는 것을 알 수 있으며, 의미 카테고리 기반의 가중치 활용이 검색 성능에 많은 향상을 가져다 줄 수 있다는 것을 확인할 수 있었다. 질의어의 모호성 해결과 협동적 가중치 및 링크 가중치가 검색엔진의 성능에 중요한 작용을 하지만, 가중치 저장 방식에서 의미 카테고리 방식을 적용함으로 웹 페이지의 의미 세분화를 이룰 수 있게 되어 웹 페이지 변별력을 높이는 작용을 한다. 웹 페이지의 변별력 증대는 검색 정확도 향상에 많은 작용을 한다는 것을 실험으로 확인할 수 있었다. 또한, 의미 카테고리 방식을 활용하는 것이 웹 페이지를 상위에 위치시키는 것에 중요한 작용을 한다는 것을 실험으로 확인할 수 있었다.

5. 결 론

검색엔진에서 질의어의 모호성은 검색 성능 저하를 유발하고, 전통적 방식의 웹 페이지 가중치 부여 방식인 내용기반 방법은 대량의 정보가 내재된 웹에서는 정확한 정보 추출에 단점을 나타내고 있다. 이러한 검색엔진의 문제점을 해결하기 위해 본 논문에서는 질의어의 모호성 해결 방안 및 의미 카테고리를 이용한 협동적 가중치 부여 방식과 링크 가중치 부여 방식을 제안한다.

본 논문에서는 질의어의 모호성 해결을 위해 워드넷 기반 사용자 인터페이스를 설계하여 질의어의 모호성을 해결하였고, 전통적 방식의 가중치 부여 방식인 내용기반 방법의 단점을 보완하기 위해 협동적 가중치 방법과 링크 가중치 방법을 제안한다. 가중치 부여 방식은 의미 카테고리 구조를 기반으로 웹 페이지의 가중치를 결정한다. 이와 같은 의미 카테고리 기반 가중치 결정 방법은 웹 페이지의 가중치를 세분화하여 검색 성능을 향상시키고, 정보 가치가 높은 웹 페이지를 상위 결과에 위치시킨다는 것을 실험을 통해 확인하였다.

향후 연구 과제는 여러 개의 질의어가 주어지는 경우 사용자 반응을 요구하지 않은 상태에서 워드넷의 의미 카테고리를 활용하여 질의어의 모호성을 자동으로 해결하는 방법에 대한 연구이다.

참 고 문 헌

[1] E. Agichtein, S. Lawrence, and L. Gravano, "Learning search engine specific query transformations for question answering," In Tenth International World Wide Web Conference, Hong Kong, 2001.

[2] P. Adriaans, D. Zantinge, Data Mining, Addison-Wesley, 1996.

[3] J. M. Bradshaw, Software Agents, AAAI press, 1997.

[4] D. Dreilinger and A. E. Howe, "An information gathering agent for querying web search engines," Computer Science Technical report, CS-96-111, Colorado State University, 1996.

[5] D. Dreilinger and A. E. Howe, "Experiences with selecting

search engines using metasearch," ACM Transactions on Information Systems, Vol.15, 1997.

[6] W. Frakes, and R. Yates, Information Retrieval : Data Structures & Algorithm, Prentice-Hall, 1992.

[7] W. Frakes and R. Yates, Information Retrieval and Hypertext, Kluwer Academic Publishers, 1996.

[8] E. J. Glover and W. P. Birmingham, "Using decision theory to order documents," In Digital Libraries 98, Pittsburgh, PA, 1998.

[9] E. J. Glover, S. Lawrence, William P. Birmingham and C. Lee Giles, "Architecture of a Metasearch Engine That Supports User Information Needs," CIKM, pp.210-216, 1999.

[10] E. J. Glover, S. Lawrence, M. D. Gordon, W. P. Birmingham, and C. L. Giles, "Web Search - Your Way," Communications of the ACM, Vol.44, No.12, 2001.

[11] R. Hoch, "Using IR Techniques for text classification in document analysis," Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994.

[12] C. Junghoo, G. Hector and L. Page, "Efficient Crawling Through URL Ordering," 7th World Wide Web Conference, 1998.

[13] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," The Journal of the ACM, Vol.46, Issue 5, 1999.

[14] B. Krishna, and R. Monika, "Improved Algorithms for Topic distillation in a Hyperlinked Environment,"Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998.

[15] X. Li, S. Szpakowicz and S. Matwin, "A WordNet-based Algorithm for Word Sense Disambiguation," The 1995 International Joint Conferences on Artificial Intelligence, 1995.

[16] S. Lawrence and C. Giles, "Inquirus, the NECI meta search engine,"7th International World Wide Web conference, 1998.

[17] G. A. Miller, "WordNet : An On-Line Lexical Database," International Journal of Lexicography, 1990.

[18] G. A. Miller "WordNet : A Lexical Database for English," Communications of the ACM, Vol.38, Issue 11, 1995.

[19] D. Moldovan and R. Mihalcea, "A WordNet-Based Interface to Internet Search Engines," Proceedings of FLAIRS-98, 1998.

[20] S. Scott and S. Matwin, "Text Classification Using WordNet Hypernyms," Coling-ACL '98 Workshop, 1998.

[21] X. Shen and C. X. Zhai, "Exploiting query history for document ranking in interactive information retrieval," SIGIR 2003, pp.377-378, 2003.

[22] E. Siegel, "Disambiguating Verbs with the WordNet Category of the Direct Object," Coling-ACL '98 workshop, 1998.

[23] E. Voohees, "Query Expansion Using Lexical-Semantic Relations," Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994.

[24] http://none.cs.umass.edu/~schapira/thesis/report/.

[25] http://www.directhit.com.

[26] http://www.google.com.

김 형 일

e-mail : hikim@dongguk.edu

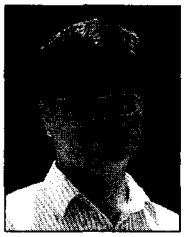
1996년 목원대학교 수학과(이학사).

1996년~1998년 (주)경기은행.

2001년 동국대학교 대학원 컴퓨터공학과 (공학석사)

2001년~현재 동국대학교 대학원 컴퓨터공학과(박사과정)

관심분야 : 지능형 에이전트, 정보검색, 기계학습, 전자상거래



김 준 태

e-mail : jkim@dongguk.edu

1986년 서울대학교 제어계측공학과 (공학사)

1990년 미국 Univ. of Southern California, Electrical Engineering-Systems(M.S.)

1993년 미국 Univ. of Southern California, Computer Engineering (Ph.D.)

1994년~1995년 미국 Southern Methodist University(Postdoc)

1995년~현재 동국대학교 컴퓨터공학과 부교수

관심분야 : 지능형 에이전트, 정보검색, 기계학습, 자연어처리, 데이터마이닝