질의어 의미별 사용자 선호도를 이용한 웹 검색의 성능 향상

(Improving Performance of Web Search using The User Preference in Query Word Senses)

김 형 일 † 김 준 태 ††
(Hyungil Kim) (Juntae Kim)

요 약 본 논문에서는 웹 검색의 성능 향상을 위해 질의어 의미별 사용자 선호도를 이용한 웹 페이지의 가중치 부여 방식을 제안한다. 일반적으로 검색엔진들은 검색 질의어와 웹 페이지의 어휘 비교에 의한 관련도 측정만을 사용하여 웹 페이지의 가중치를 부여한다. 웹과 같이 방대한 자료를 대상으로 검색을 할 경우 유사한 관련도를 가진 검색 결과가 매우 많으므로 어휘 비교만으로는 중요한 웹 페이지를 선별하기 어렵다. 본 논문에서는 질의어의 의미를 구분하도록 워드넷(WordNet)을 이용한 사용자 인터페이스를 구축하고, 사용자의 클릭 수를 각 웹 페이지의 가중치에 누적함으로써 다수 사용자의 검색 행위에 의한 묵시적 평가가 웹 페이지의 검색 순위에 반영되는 검색 시스템을 구현하였다. 클릭수의 누적에 있어서 질의어 의미별로 가중치를 구분하여 저장함으로써 일반적인 검색엔진보다 정확한 검색이 되었으며, 웹 페이지의 범주별 가중치와 질의어의 의미별 사용자 선호도를 이용함으로써 검색 시스템의 성능을 향상시킬 수 있다는 것을 20개의 어휘에 관련된 41개의 의미들을 대상으로 실험한 결과로 확인하였다.

키워드 : 정보검색, 검색엔진, 사용자 선호도, 워드넷

Abstract In this paper, we propose a Web page weighting scheme using the user preference in each sense of query word to improve the performance of Web search. Generally search engines assign weights to a web page by using relevancy only, which is obtained by comparing the query word and the words in a web page. 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. The experimental results with several keywords show that the precision of proposed system is improved compared to conventional search engines.

Key words : information retrieval, search engine, user preference, WordNet

1. 서 론

초기의 웹 검색 시스템은 정보 공유를 중요시하여 사용자가 원하는 정보를 웹에서 대량으로 추출하여 주는 것만을 고려하였으나, 현재와 같이 방대한 정보가 내재되어 있는 웹에서는 사용자가 원하는 정보를 얼마나 정확히 추출할 수 있는가가 보다 중요하다고 할 수 있다[1]. 대다수의 검색 시스템들은 검색 결과의 순위를 계산하는데 있어서 사용자가 사용한 질의어가 특정 웹 페이지에서 많은 분포를 이룰 경우 높은 가중치를 부여하는 방법을 활용하고 있다[2]. 이와 같은 방식은 사용된 질의어의 의미는 배제되고 해당 질의어 어휘가 많이 포함되어 있는 웹 페이지에 높은 가중치를 부여함으로써 모호성을 갖는 질의어에 대하여 올바른 가중치를 부여할 수 없으며, 또한 방대한 검색 결과로부터 중요한 웹 페이지를 선별해내기 어렵다는 문제를 가지고 있다. 이러한 웹 검색 시스템의 단점을 보완하기 위해 구글

· 본 논문은 정보통신부 대학기초연구지원 사업의 연구 결과임
† 학생회원 : 동국대학교 컴퓨터공학과
hikim@dongguk.edu
†† 종신회원 : 동국대학교 컴퓨터공학과 교수
jkim@dongguk.edu
논문접수 : 2002년 12월 27일
심사완료 : 2004년 5월 29일

(Google)나 다이렉트히트(DirectHit)와 같은 검색 시스템에서는 질의어의 비교 이외에 다양한 가중치 결정 방식을 사용하고 있으며, 일반 검색 시스템을 활용하여 웹 페이지 순위 조정을 통해 사용자에게 질 높은 결과를 제공하는 메타 검색엔진들은 질의에 따라 검색 시스템을 선택하는 방법 등을 사용하기도 한다[3].

본 논문에서는 웹 검색의 정확도 향상을 위하여 질의어 의미별 사용자 선호도를 이용한 웹 페이지 가중치 부여 방식을 제안한다. 본 논문에서 제안하는 방법은 사용자가 검색 결과 리스트 중 특정 페이지를 검색(클릭)하는 행위를 모니터하여 다수 사용자의 클릭 수를 각 웹 페이지의 가중치에 누적하는 것이다. 이와 같은 방법을 적용함으로써 다수 사용자의 묵시적 행위가 웹 페이지 검색 순위에 반영되는 효과를 얻을 수 있다. 특히, 클릭수의 누적에 있어서 질의어의 의미 범주별로 가중치를 구분하여 저장함으로써 웹 페이지 가중치를 세분화하여 정확한 검색이 되도록 하였다. 사용자 인터페이스는 질의어에 해당하는 모든 의미를 워드넷으로부터 추출하여 사용자에게 제시하고 그중 하나를 선택하게 함으로써 질의어의 모호성을 해결하고 질의어 의미 범주를 결정하도록 하였다.

본 논문에서는 웹 페이지의 변별력을 증가시키기 위해 단어빈도(TF)와 역문헌빈도(IDF)를 응용한 어휘 의미빈도(SF)와 범주 역문헌빈도(IDFC)을 적용하여 웹 페이지의 가중치에 범주별 클릭 가중치와 함께 적용하였다. 웹 페이지들은 여러 범주에 속할 경우가 있기 때문에 하나의 주제에 관련한 어휘로 문서 종속적인 가중치의 적용은 웹 페이지에 내재된 정보 이용량을 축소시키는 문제로 작용될 수 있다. 이러한 문제 해결을 위해 본 논문에서는 의미빈도와 범주별 역문헌빈도를 적용함으로써 웹 페이지의 가중치를 세분화하고 범주별 소속도를 나타냄으로써 웹 페이지의 변별력을 증가시켰다.

본 논문에서는 제안한 웹 페이지의 가중치 결정 방식을 실험하기 위해 실험용 웹 검색 시스템을 구축하였으며, 워드넷의 상위 범주와 20개의 검색어를 이용한 실험을 통해 제안한 방법이 웹 검색의 성능을 향상시킬 수 있음을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 질의어 활용과 모호성, 사용자 반응과 웹 페이지 순위, 워드넷을 이용한 정보 검색 등의 관련 연구에 대해 기술하고, 3장에서는 단일 질의어의 모호성 해결과 확장, 사용자 선호도를 이용한 범주별 가중치, 범주별 의미 가중치, 유사도 비교와 순위 결정 그리고 실험용 웹 검색 시스템에 대하여 설명한다. 4장에서는 실험용 검색 시스템의 성능에 관한 실험 방법과 실험 결과에 대하여 설명하고, 5장에서 결론 및 향후 연구를 제시한다.

2. 관련 연구

일반 검색 시스템에서 사용되는 검색 질의어는 적은 수가 사용되는 것이 일반적이며, 적은 수량의 질의어는 질의 해석에 어려움을 발생시킨다. 정확한 질의 해석을 위해 질의어의 길이를 증가하도록 유도하는 방법을 이용하기도 하지만, 명시적인 요구는 사용자에게 거부감을 발생시킬 수 있다[4]. 차세대 검색 시스템들은 사용자의 반응을 이용하여 검색의 정확도를 높이는 시도를 하고 있으며, 이러한 연구 중에는 사용자가 사용한 질의어의 내력(history)을 분석하여 질의어를 확장하는 방법들도 있다[5-6].

적은 수량의 어휘를 사용한 질의의 문제보다 더 중요한 문제는 단일 어휘를 사용한 질의이다. 단일 어휘를 사용한 질의는 사용자의 질의에 대해 명확한 판단을 내릴 수 없는 경우가 빈번히 발생되기 때문에 올바른 검색에 부작용을 발생시킨다[7-8]. 이러한 단일 어휘의 부작용이 발생되는 이유는 어휘의 의미 중의성 때문이다[9].

워드넷은 어휘에 대하여 동의, 반의, 상위, 하위 등과 같은 어휘의 연관성을 정의한 어휘 사전으로써 1985년 프린스톤(Princeton)대학에서 개발되어 버전 1.7.1까지 발표되었다. 워드넷에는 어휘의 의미에 대한 범주 분류가 잘 정의되어 있으며, 단어들 사이의 계층 구조와 연관 관계가 여러 형태로 표현되어 있다[10-12]. 워드넷에서 어휘 X와 어휘 Y의 의미가 같은 경우에는 동의어(synonym) 관계가 정의되고, 의미가 반대인 경우에는 반의어(antonym)관계가 정의된다. 의미상의 포함 관계는 상위어(hypernym)와 하위어(hyponym)관계로 정의되며, 부분(meronym)과 전체(holonym)를 나타내는 관계도 잘 정의되어 있다[13]. 예를 들면, 'rise'는 'ascend'와 동의어 관계이고 'tree'는 'maple'의 상위어이다. 동의어들은 하나의 동의어 집합(set of synonyms, synset)을 형성하며, 이들은 상·하위 관계에 의해 계층적인 표현도 가능하다.

워드넷에서 명사 사전의 상위 범주는 26개로 이루어져 있으며, 모든 어휘들은 상위 범주에 하나 이상 포함되게 된다. 워드넷의 어휘들은 의미로 구조화 되고 연관 관계도 잘 정의되어 있어서 질의어 모호성 해결 및 질의어 확장에 많은 연구가 되어지고 있다. 질의가 문장일 경우 질의어의 명확한 의미를 판단하기 위해 워드넷의 의미 관계를 이용하게 되면 문장 분석에 많은 도움을 얻을 수 있으며, 워드넷에 나타난 동의어나 상·하위어들을 사용하여 질의를 확장함으로써 효과적인 정보 검색을 수행할 수 있다[14-15].

웹 페이지의 순위 결정 목적은 사용자의 요구에 맞는 웹 페이지를 쉽게 찾을 수 있도록 하는 것이다[16-17].

순위는 정보 가치에 따라서 결정되는 것이며, 웹 페이지의 가치는 사용자의 요구에 따라 달라진다. 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 주는 방법을 활용하기도 한다[18]. 웹 페이지의 순위 결정에 웹 페이지의 생성 날짜, 웹 페이지의 크기, 웹 페이지의 구조, 등을 이용하기도 하며[19], 웹 페이지의 구조를 이용한 방법으로는 명성 평가 방법[20]과 Kleinberg의 HITS 알고리즘[21]이 있다.

메타 검색엔진에서 연구가 활성화된 질의 확장과 웹 페이지의 순위 결정 방법은 정보 추출 측면에서 높은 성능을 나타낼 수 있게 하는 중요한 방법들이다. 대표적인 메타 검색엔진은 MetaCrawler, SavvySearch, Dogpile, Inqurirus2, 등을 들 수 있으며, 질의어의 확장과 웹 페이지의 가중치 조정에 관한 연구가 활발히 진행된 메타 검색엔진으로 대표적인 것은 Inquirus2이다. Inquirus2는 웹 페이지에 대한 정확도를 높이기 위해 사용자에게 명시적인 반응을 요구한다.

어휘와 문서 종속적인 방법은 웹 정보의 방대함 때문에 검색 대상을 증가시키며, 중요한 웹 페이지를 선출하기 어렵다는 것이 문제가 된다. 이러한 문제를 해결하기 위하여 사용할 수 있는 하나의 방법은 사용자의 반응과 가중치의 범주화이다. 사용자의 문서 선택 행위는 해당 웹 페이지에 대한 묵시적인 평가라고 할 수 있으며, 이러한 정보 선별 행동들을 누적하여 웹 페이지의 가중치에 활용함으로써 웹 페이지에 인기도를 부여할 수 있다. 이러한 사용자 반응을 웹 페이지의 가중치 결정 방법에 도입하면, 웹 검색 시스템의 성능을 향상시킬 수 있다.

본 논문에서는 사용자의 웹 페이지 선택 행위를 모니터하여 사용자의 묵시적 평가를 웹 페이지의 범주별 가중치에 누적하는 가중치 방식을 사용하고, 사용자의 질의어 의미 선택을 이용하여 원시 질의어를 확장한다. 웹 페이지에 가중치를 부여할 경우 범주별 의미 가중치를 적용하였으며, 적용된 방법은 TF·IDF를 응용한 SF·IDFC를 적용하여 웹 페이지의 변별력을 증가시켰다.

3. 사용자 선호도와 범주별 가중치를 이용한 검색 시스템

이 장에서는 본 논문에서 제안한 질의어 모호성 해결 인터페이스 및 질의어 의미 기반 웹 페이지 가중치 부여 방법과 웹 페이지 순위 결정에 대해 설명한다.

3.1 단일 질의어의 모호성 해결과 확장

웹 검색에 있어서 질의어가 단일 어휘로 이루어졌을 경우, 일반적인 검색 시스템은 질의어의 어휘만을 고려하여 웹 페이지를 추출하므로 질의어 의미에 중의성이 존재하면 검색 결과의 정확도는 떨어지게 된다. 예를 들어, 사용자가 "자바(의미: 섬)"에 관해서 조사하고 싶을 경우, "자바"를 검색 질의어로 사용하게 되면 자바 섬에 대한 검색 결과를 얻기 힘들다. 이러한 결과 발생은 현재 프로그래밍 언어로 각광을 받고 있는 "자바(의미: 프로그래밍언어)"의 웹 페이지들이 높은 가중치를 가지고 있기 때문이다. 본 논문에서는 이러한 단일 질의어의 모호성을 해결하고, 질의어의 의미와 범주를 결정하기 위해 워드넷 기반 사용자 인터페이스를 설계하였다.

본 논문에서 설계한 사용자 인터페이스에서는 사용자가 질의어를 입력하면 의미 선택기가 워드넷의 명사 사전에서 해당 질의어에 내포된 모든 의미들과 해당 질의어의 의미들에 연관된 동의어, 상위어, 주석을 추출한다. 추출된 동의어, 상위어, 주석은 사용자 인터페이스를 통하여 사용자에게 전달된다. 사용자는 질의어의 모든 의미들에 대해 정확히 알고 있는 경우가 드물며, 사용자는 검색 질의어로 활용한 단일 어휘가 올바른 의미로 사용되었는지 판단하기 힘든 경우도 빈번하게 발생된다. 일반 사용자들에게 동의어, 상위어, 주석을 동시에 확인하게 함으로 올바른 질의어 의미 선정에 도움을 준다. 주석은 질의어의 의미에 해당하는 설명문으로써 간단한 문장 형태로 이루어져 있으며, 질의어의 확장에 주석을 활용할 경우에는 불용어를 제거한 후 사용하는 방식을 따른다.

사용자가 검색을 위해 원시 질의어 입력에서부터 재조합 질의어를 완성하기까지의 간단한 과정을 기술하면 다음과 같다. 사용자가 인터페이스를 통하여 단일 원시 질의어(Q)를 입력하는 과정을 거친 후에 자신이 원하는 의미에 부합하는 질의어 의미(Q_sense)를 인터페이스에 나타난 동의어, 상위어, 주석을 확인하여 선택한다. 선택된 질의어의 의미는 질의어 확장기로 전달되고, 질의어 확장기에서는 선택된 검색 질의어의 의미에 해당하는

표 1 재조합 질의어의 구성

단일 원시 질의어가 Q로 입력되고 질의어의 의미가 Q_sense2로 결정된 경우	
단일 원시 질의어	Q
원시 질의어에 해당하는 의미	Q_sense1, Q_sense2, Q_sense3
Q_sense2의 주석에서 추출한 명사집합	{word1, word2, word3, word 4}
단일 원시 질의어의 확장(재조합 질의어)	{Q_sense2, Word1, Word2, Word3, Word 4}

주석을 워드넷에서 추출한다. 추출된 주석에서 명사만을 선택하여 주석 명사 집합을 구성한다. 선택된 어휘들은 사용자가 선택한 의미에 소속된 어휘들이 되고, 이렇게 얻어진 어휘들을 이용하여 원시 질의어를 확장한다. 확장된 재조합 질의어는 의미가 선택된 어휘와 주석에서 추출된 명사 집합을 합하여 완성한다.

원시 질의어 확장은 동의어나 상위어를 이용할 수도 있으나, 동의어, 상위어, 그리고 주석을 활용한 질의어 확장에 대하여 각각 실험한 결과 주석을 이용한 질의어 확장이 가장 높은 성능을 나타내어 실험용 검색 시스템에서는 원시 질의어 확장에 주석을 이용하였다.

3.2 사용자 선호도를 이용한 범주별 가중치

본 논문에서는 검색 질의어의 의미에 가중치를 적용하는 범주별 가중치를 제안한다. 본 논문에서 제안하는 범주별 가중치는 결과 웹 페이지에 대한 사용자의 반응(검색 결과 중 특정 웹 페이지에 대한 선택 행위)을 해당 웹 페이지의 가중치에 누적하며, 누적하는 방식은 질의어의 의미에 해당하는 웹 페이지의 범주에 가중치를 적용하는 방식을 취한다. 실험용 검색 시스템에서 채택한 의미에 대한 범주는 워드넷 명사 사전의 상위 범주들을 이용하였다.

사용자 선호도를 이용한 의미 범주 가중치 방식의 장점은 다음과 같은 간단한 예로 설명할 수 있다. 표 2에 나타낸 것은 두 개의 웹 페이지에 대한 협동적 가중치의 예이다. 표 2에서 URL1은 "location"범주 관련 가중치로 50, "communication" 범주 관련 가중치로 10, "food"범주 관련 가중치로 10을 가지고 있고 URL2는 각 범주에 대하여 10, 10, 100을 가중치로 가지고 있다. 즉, URL1은 "location" 범주에 해당하는 질의어들에 대하여 사용자들로부터 50회 선택되었다는 의미이다. 이러한 경우, 가중치의 총합은 URL1(50,10,10)이 70, URL2(10,10,100)이 120이 된다. 이렇게 URL의 가중치가 저장되어 있을 때, 범주별 가중치를 고려하지 않고 단일한 가중치의 총합으로 가중치를 비교하고 검색 질의어로 "java(의미: 섬)"가 사용되었을 경우에 검색 시스템에서는 URL2의 가중치의 총합이 가장 높게 나타나 있으므로 URL2가 URL1보다 중요한 페이지로 인식된다.

표 2 웹 페이지의 협동적 가중치 예

Location(island의 상위 범주)	50	10
Communication(language의 상위 범주)	10	10
Food(coffee의 상위 범주)	10	100
웹 페이지 가중치의 총합	70	120

그러나 질의어의 의미를 범주별로 구분하게 되면 "java(의미: 섬)"의 상위 범주인 "location"에 대하여는 URL1이 50이고 URL2가 10이므로 URL1이 URL2보다 높은 가중치를 소유하고 있기 때문에 더 중요한 페이지라고 판단할 수 있다. 본 논문에서는 웹 페이지 가중치를 질의어 의미 범주별로 저장함으로써 웹 페이지 가중치의 변별력을 높이도록 하였다.

그림 1은 사용자가 검색을 하였을 경우에 재조합 질

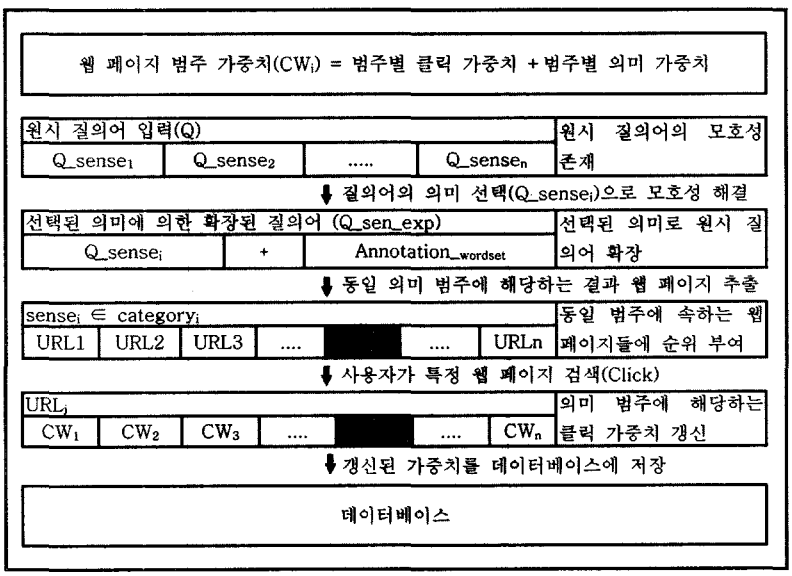

그림 1 범주별 가중치 조정

의어부터 웹 페이지 가중치 조정까지를 간단히 나타낸 그림이다. 웹 페이지의 범주별 가중치에는 범주별 클릭 가중치와 범주별 의미 가중치를 적용한다. 결과로 선택된 웹 페이지들에 대해 사용자가 검색을 하였을 경우에 질의에 사용된 의미 범주와 동일한 범주의 클릭 가중치를 증가시킨 후 데이타베이스를 갱신한다.

3.3 범주별 의미 가중치

웹 페이지들은 단일한 범주에만 속하지 않을 경우가 빈번하게 발생되므로 웹 페이지를 단일한 범주로 결정하는 것은 정보 활용 측면에서 위험한 문제를 야기 시킬 수 있다. 본 논문에서 이러한 문제점을 해결하기 위해 웹 페이지의 가중치에 범주를 이용하여 웹 페이지들의 변별력을 증가시켰으며, 범주 가중치에 확률 기법을 적용하여 웹 페이지의 타 범주 소속 가능도도 표현하였다.

웹 페이지는 여러 어휘들로 구성되어 있으며, 이러한 어휘들은 여러 가지의 의미를 소유하고 있게 된다. 웹 페이지에 N개의 단어들이 존재하고 각 단어들이 서로 다른 M개의 의미를 소유하고 있다면 해당 웹 페이지에는 N*M개의 서로 다른 의미들이 존재한다. 이때, 범주화가 가능할 수 있는 구조가 어휘에 존재한다면 N*M개의 서로 다른 의미들은 각각의 의미를 포함하는 범주로 소속될 수 있다.

본 논문에서는 웹 페이지의 가중치에 범주들을 이용하여 웹 페이지의 변별력을 높였고, 웹 페이지의 여러 범주에 대해서는 사용자의 선호도를 추가함으로써 웹 페이지의 인기도를 반영하였다. 사용자의 선호도는 사용자가 웹 페이지를 검색할 경우 사용자가 결정한 의미에 해당하는 범주의 선호도 가중치에 누적된다.

웹 페이지들의 범주별 의미 가중치를 결정하기 위해 본 논문에서는 TF·IDF를 변형하여 사용한다. 단어빈도(TF) 방식은 대상 문서에 특정 어휘들이 출현한 빈도가 다량으로 나타나는 어휘에 대해 높은 가중치를 부여하는 작용을 하고, 역문헌빈도(IDF) 방식은 변별력이 높은 어휘에 높은 가중치를 줄 수 있도록 하는 것으로 적은 수의 문서에 나타난 어휘에 높은 가중치를 부여하는 작용을 한다.

본 논문에서는 단어빈도 방식을 응용하여 어휘의 출현 빈도 가중치가 아닌 어휘의 의미를 이용한 어휘 의미빈도(SF)를 가중치에 사용함으로써 다의어의 문제를 해결하면서 어휘의 의미 세분화를 이루게 하였다. 역문헌빈도 방식은 정보검색 분야에서 높은 활용도를 보이며 오랜 기간동안 사용된 방식으로 어휘 중심적 방식이라 할 수 있다. 이와 같은 어휘 중심적 방식과 문헌 종속적 방식은 많은 문서를 검색해야한다는 단점을 내포한다. 또한 역문헌빈도 방식은 모든 문서를 대상으로 하지만, 범주별 역문헌빈도는 동일한 범주에서만 적용된다. 본 논문에서는 역문헌빈도 방식을 응용하여 범주별 역문헌빈도(IDFC) 방식을 적용함으로써 관련성 없는 웹 페이지와의 비교를 회피하여 비교 계산 시간을 단축하고 웹 페이지에 특성화된 어휘의 의미에 가중치를 부여한다. 본 논문에서 활용한 SF·IDFC 방식은 사용자 선호도 가중치와 함께 해당 웹 페이지의 범주 가중치에 적용된다. SF·IDFC 방식을 적용함으로 웹 페이지가 어떤 의미들을 내포하고 있는지 표현할 수 있으며, 어휘의 빈도를 사용하지 않고 어휘의 의미빈도를 활용함으로써 어휘의 활용도를 증가하게 하였다. 어휘 활용도의 증가는 웹 페이지의 가중치 부여에 높은 변별력을 부여하는 작용을 한다.

초기의 데이타베이스에 존재하는 웹 페이지들은 범주화가 이루어지지 않은 상태로 존재한다. 데이타베이스에 존재한 이러한 웹 페이지들에 대해 범주별 의미 가중치를 부여하기 위해서는 범주화가 필요하게 된다. 이런 범주화 작업을 위해 초기의 웹 페이지들에 대해서는 SF 방식만을 적용하여 해당 웹 페이지의 의미빈도 확률 값을 범주별 의미 가중치로 활용한다. SF의 적용으로 범주에 대한 웹 페이지들의 초기 의미 확률 값들을 부여받는다. 전체 데이타베이스에 존재한 웹 페이지에 SF 적용이 끝난 후, SF·IDFC 방식을 적용하여 각 웹 페

표 3 어휘의 범주별 의미 가중치 결정

어휘의 의미	T_sen_j	어휘의 의미빈도	S_jF	범주별 역문헌빈도	$IDFC_i$
범주별 웹 페이지들의 집합		P_cat_i		범주별 웹 페이지들의 총계	$\|P\_cat_i\|$
범주에서 동일한 어휘의 의미를 소유한 웹 페이지들의 집합					$P\_cat_i\_sen_j$
범주에서 동일한 어휘의 의미를 소유한 웹 페이지들의 총계					$\|P\_cat_i\_sen_j\|$
$S_jF = freq(T\_sen_j)$					
$IDFC_i = \log_2(\|P\_cat_i\|) - \log_2(\|P\_cat_i\_sen_j\|) + 1$					
$T\_sen_j\_cat_i\_w = S_jF \cdot IDFC_i$ $= freq(T\_sen_j) \cdot [\log_2(\|P\_cat_i\|) - \log_2(\|P\_cat_i\_sen_j\|) + 1]$					

이지들의 범주별 의미 확률 값을 재산정하여 초기 범주별 의미 가중치 값을 갱신한다. 이렇게 SF·IDFC 방식을 적용한 후에는 웹 페이지들은 범주별 의미 가중치 값을 소유하게 되고 범주별 사용자 선호도 가중치 값이 웹 페이지 가중치에 추가되어 웹 페이지의 범주별 가중치로 활용된다.

웹 페이지의 범주별 의미 가중치 값은 특정 웹 페이지(P_k)에 나타난 어휘의 의미들을 이용하여 결정한다. 웹 페이지들에 대해 범주별 의미 가중치 값을 부여하기 위해 단어빈도(TF) 방식을 응용한 의미빈도(S_iF)방식을 웹 페이지에 적용하여 범주별 의미 가중치 값을 부여한다. 의미빈도 방식은 웹 페이지에 존재하는 어휘들의 의미를 활용하여 어휘 의미의 범주별 빈도를 계산한다. 의미빈도를 적용하여 생성된 범주별 의미 가중치는 웹 페이지의 범주 소속 확률 값으로 활용될 수 있다. 의미빈도 적용이 모든 웹 페이지에 대해 수행되면 의미빈도·범주별 역문헌빈도($S_iF \cdot IDFC_i$)를 적용하여 웹 페이지의 범주별 의미 가중치를 재조정한다. 범주별 역문헌빈도($IDFC_i$) 방식은 범주 내에서 속하는 어휘 의미를 이용하여 웹 페이지의 변별력이 높은 어휘의 의미에 높은 가중치를 부여하는 방식이다. 의미빈도 방식과 범주별 역문헌빈도 방식을 적용함으로 웹 페이지들의 범주별 소속도를 확률적으로 나타낼 수 있게 되고, 이러한 범주별 소속도는 정보 활용도를 높이는 작용을 한다.

웹 페이지의 범주별 가중치는 질의어 의미에 해당하는 웹 페이지의 범주별 클릭 가중치와 범주별 의미 가중치를 결합하여 이용한다. 클릭 가중치에서는 해당 범주의 클릭 가중치를 최대 클릭 가중치($Max(U\_cat_i\_clk\_w)$)로 나누어 주고 해당 범주의 웹 페이지들의 총수($|P\_cat_i|$)의 1/2을 곱함으로써 가중치에 부분적인 작용만 할 수 있도록 하였다. 그리고 범주별 의미 가중치($T\_sen_j\_cat_i\_w$)는 특정 웹 페이지의 의미 범주 소속 확률이라 할 수 있으며, 범주별 의미 가중치는 최대 범주별 의미 가중치($Max(T\_sen_j\_cat_i\_w)$)로 나누고 해당 범주의 문서 총수($|P\_cat_i|$)의 1/2을 곱함으로써 가중치에 부분적인 작용만 하도록 하였다.

3.4 유사도 비교와 순위 결정

웹 페이지의 순위 결정은 질의어와 웹 페이지의 유사도 비교 값을 이용하여 순위 결정에 이용한다. 사용자가 인터페이스를 통하여 질의를 하면 실험용 검색 시스템은 질의어의 의미 범주에 해당하는 웹 페이지의 가중치만을 고려하여 웹 페이지 순위를 결정하며, 웹 페이지에 범주별 가중치는 범주별 의미 가중치와 범주별 사용자 선호도 가중치의 결합으로 이루어졌다. 순위 결정기는 결과 웹 페이지의 순위($Rank(P_k)$)를 결정하기 위해 질의어와 웹 페이지의 유사도($Sim(Q\_sen_j\_cat_i\_\exp, P_k)$)를 계산하며, 유사도 비교는 질의어 의미에 해당하는 범주의 웹 페이지 가중치 값을 대상으로 한다. 순위 결정은 결과 대상이 된 웹 페이지들의 범주별 가중치 값이 큰 순으로 순위를 부여하는 방식을 취한다. 웹 페이지의 순위 결정에 있어서 동일한 순위를 부여받게 될 경우는 원시 질의어의 의미빈도가 높게 발생되는 웹 페이지에 높은 순위를 부여한다. 원시 질의어의 의미빈도를 활용

표 4 웹 페이지의 순위 결정

의미를 적용한 확장 질의어	$Q\_sen_j\_cat_i\_\exp$	임의의 웹 페이지	P_k	어휘의 의미	T_sen_j
어휘 범주별 의미 가중치	$T\_sen_j\_cat_i\_w$	웹 페이지의 순위		$Rank(P_k)$	
범주별 웹 페이지들의 총계	$\|P\_cat_i\|$	어휘의 의미빈도 총계		$freq(T\_sen_j)$	
범주별 클릭 가중치	$U\_cat_i\_clk\_w$	클릭 가중치의 최대값		$Max(U\_cat_i\_clk\_w)$	
범주에서 동일한 어휘의 의미를 소유한 웹 페이지들의 집합				$P\_cat_i\_sen_j$	
범주에서 동일한 어휘의 의미를 소유한 웹 페이지들의 총계				$\|P\_cat_i\_sen_j\|$	
확장된 질의어와 웹 페이지 유사도				$Sim(Q\_sen_j\_cat_i\_\exp, P_k)$	

$$Sim(Q\_sen_j\_cat_i\_\exp, P_k) =$$
$$[(U\_cat_i\_clk\_w / Max(U\_cat_i\_clk\_w)) \cdot (|P\_cat_i| / 2)] +$$
$$[(T\_sen_j\_cat_i\_w / Max(T\_sen_j\_cat_i\_w)) \cdot (|P\_cat_i| / 2)] =$$
$$[(U\_cat_i\_clk\_w / Max(U\_cat_i\_clk\_w)) \cdot (|P\_cat_i| / 2)] +$$
$$[(freq(T\_sen_j) \cdot (\log_2(|P\_cat_i|) - \log_2(|P\_cat_i\_sen_j| + 1)) \cdot$$
$$((|P\_cat_i| / 2))]$$

$$Rank(P_k) = Sim(Q\_sen_j\_cat_i\_\exp, P_k)$$

한 경우에도 동일한 순위가 발생되면 웹 페이지의 타이틀에 나타나는 어휘의 의미빈도를 추가하여 결정하게 된다. 그러나 웹 페이지의 타이틀에는 많은 어휘가 나타나지 않게 되어 중요한 작용을 못할 경우가 있으며, 이러한 경우에는 웹 페이지의 어휘 총수를 이용하여 순위를 결정한다.

새로운 웹 페이지가 데이타베이스에 입력될 경우에는 범주별 의미 가중치 값과 사용자 선호도 가중치 값이 존재하지 않는다. 이러한 경우 가중치 계산기는 해당 웹 페이지의 의미빈도(S_iF) 가중치를 이용하여 해당 웹 페이지의 범주별 의미 가중치 값으로 활용한다. 이러한 임시적 범주별 의미 가중치 값은 새롭게 생성된 웹 페이지들이 일정량에 도달할 경우, 의미빈도와 범주별 역문헌빈도를 활용하여 정확한 범주별 의미 가중치로 갱신해야 한다. 새로 입력된 웹 페이지에는 사용자의 선호도가 발생되지 않았으므로 사용자 선호도 가중치는 존재하지 않는다. 이러한 경우 가중치 계산기는 새롭게 입력된 웹 페이지에 대해 범주별 사용자 선호도 가중치를 부여한다. 사용자 선호도 가중치를 적용하는 방법은 범주별 의미빈도 확률($\mathrm{Prob}(P\_T\_sen_i)$)을 이용하여 범주별 의미빈도 확률 값이 같은 웹 페이지들을 추출한 후 가장 낮은 사용자 선호도 가중치 값을 소유한 웹 페이지의 범주별 사용자 선호도 가중치를 새로운 웹 페이지의 범주별 사용자 선호도로 적용한다.

새롭게 생성된 어휘가 질의어로 입력될 경우에는 질의어의 의미가 결정될 수 없는 상태에 이른다. 이러한 경우가 발생될 경우에는 해당 질의어를 포함하고 있는 웹 페이지들을 단어빈도 방식을 적용하여 추출한다. 그리고 해당 질의어에 대해 단어빈도 방식으로 순위가 부여된 대상 웹 페이지들 중에 상위 30%에 해당하는 웹 페이지들의 의미 범주에 대한 확률 값을 더한 후 상위 30%에 해당하는 웹 페이지들의 총수로 나누어 의미 범주 평균을 산출하고 새롭게 생성된 어휘의 의미 범주로 채택하여 사전에 등록한다.

3.5 실험용 검색 시스템

본 논문에서 제안한 범주별 사용자 선호도 가중치와 범주별 의미 가중치를 실험하기 위해 실험용 검색 시스템을 구현하였다. 본 연구에서 구현한 실험용 검색 시스템의 구성도는 그림 3과 같다.

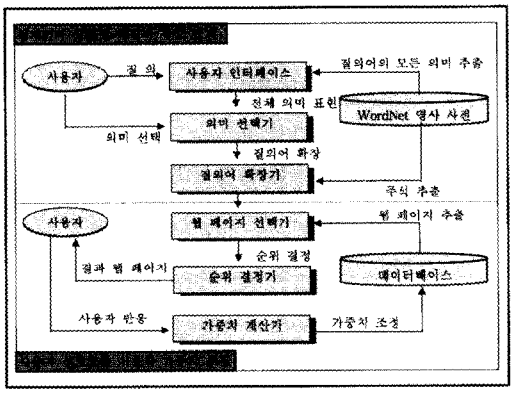

그림 3 시스템 구성도

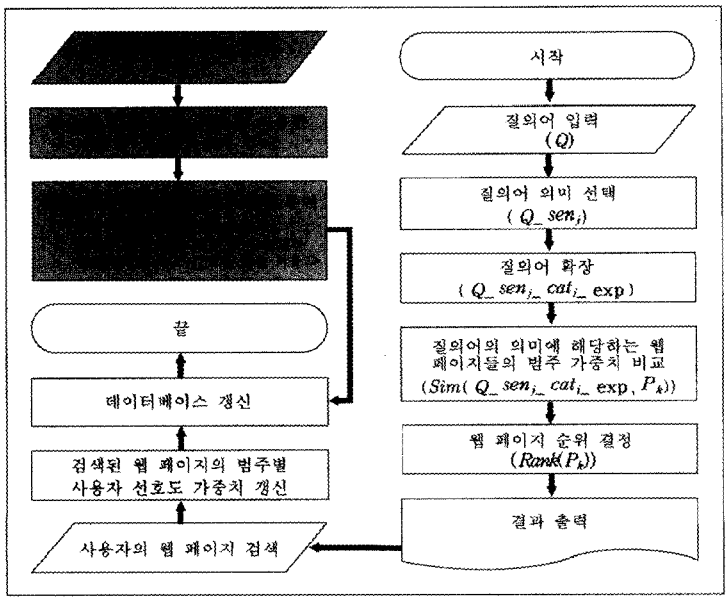

그림 2 결과 웹 페이지 순위 결정과 새로 입력된 웹 페이지의 가중치 결정

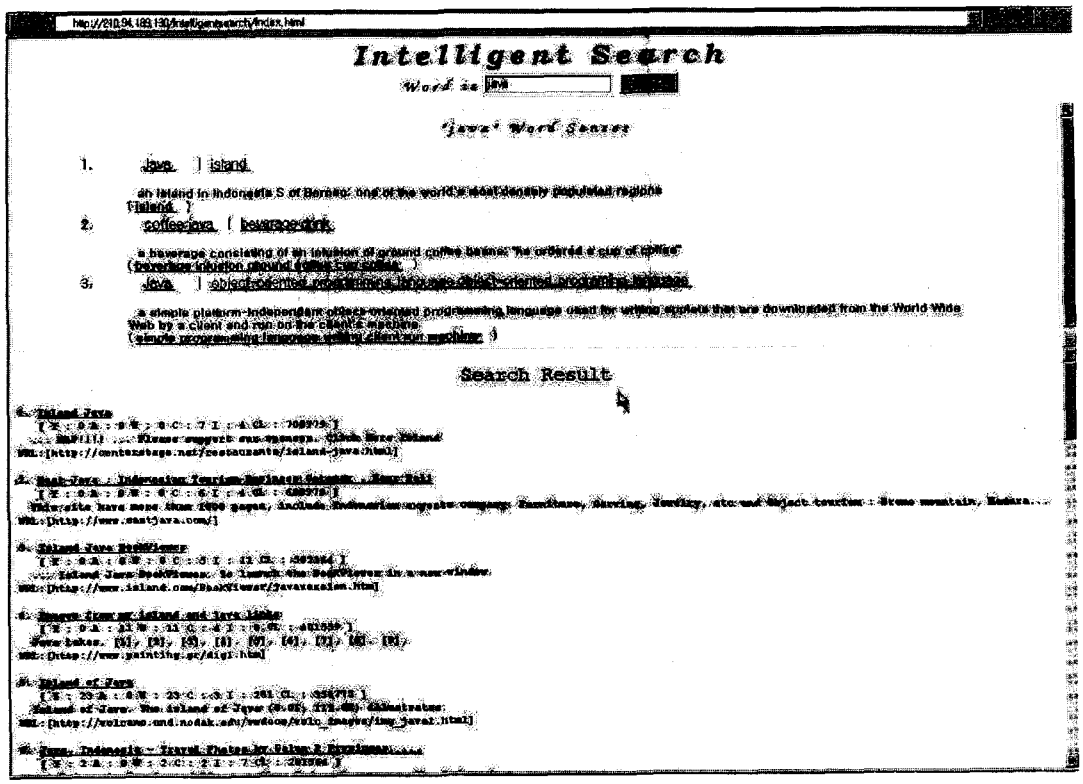

그림 4 사용자 인터페이스

사용자가 그림 4와 같은 인터페이스를 통하여 질의를 하면 의미 선택기는 질의어에 해당하는 모든 의미를 워드넷에서 검색하여 각 의미에 해당하는 동의어, 상위어, 주석을 사용자에게 제시한다. 질의어 확장기는 사용자가 선택한 질의어의 의미에 대해 주석을 이용하여 원시 질의어를 확장하고, 웹 페이지 선택기는 확장된 질의어를 활용하여 데이타베이스에서 질의어 의미와 동일한 범주에 있는 웹 페이지를 추출한다. 순위 결정기는 범주별 사용자 선호도 가중치와 범주별 의미 가중치를 이용하여 웹 페이지들의 순위를 결정한다. 사용자는 검색 결과 중에서 자신이 원하는 정보가 있다고 생각되는 웹 페이지들만을 선택적으로 검색하게 되며, 이러한 검색 행위는 사용자 모니터를 통하여 가중치 계산기에 전달된다. 가중치 계산기는 질의어의 의미 범주에 해당하는 웹 페이지의 가중치 범주에 사용자 반응 값을 누적한다.

4. 실험 및 결과

본 논문에서 실험용 검색 시스템의 성능을 측정하기 위해 실험 대상이 될 웹 페이지를 실험을 시작하기 전에 30일 동안 수집하여 실험용 데이타베이스를 구성하였으며, 실험을 위해 180일 동안 대학생 300명을 선정하여 실험용 검색 시스템을 활용하게 하여 사용자의 선호도 가중치가 적용된 데이타베이스를 구축하였다.

4.1 실험 데이타 및 실험 방법

본 논문에서는 웹 페이지 범주 가중치의 중요성과 질의어 의미 세분화의 중요성을 실험하기 위해 모호성이 있는 20개의 질의어들과 워드넷의 카테고리만을 활용하여 실험하였다. 사용되어진 질의어들은 표 5에 나타나 있으며, 표 5에 나타난 질의어들에 관련하여 데이테베이스 구축을 위해 30일 동안 알타비스타를 이용하여 웹 페이지들을 수집하였다. 실험 대상 질의어에 관련된 웹 페이지만을 수집하여 데이타베이스를 구축하였고, 질의어 의미별로 1,000개의 웹 페이지들을 수집하였다. 수집된 웹 페이지들은 범주와 가중치가 결정되지 않은 상태이기 때문에, 실험용 검색 시스템에서는 의미빈도와 범주별 역문헌빈도 방식을 이용하여 웹 페이지들의 범주별 의미 가중치를 생성하였다. 이와 같이 범주별 의미 가중치가 결정된 웹 페이지들은 실험용 데이타베이스에 저장되고, 웹 페이지에 사용자 선호도 가중치를 적용하기 위해 180일 동안 대학생들을 통해서 사용자 반응을 입력받았다.

웹 페이지의 수집을 위해 알타비스타를 사용한 것은

표 5 실험에 사용된 질의어

실험에 사용된 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	Engagement	전투
Bank	은행	Ball	무도회	Engagement	약혼
Bank	둑	Letter	편지	Plant	공장
Seal	봉인	Letter	임대인	Plant	식물
Seal	물개	Palm	손바닥	Pupil	학생
		Palm	야자	Pupil	동공

최근에 생성된 검색 시스템이나 메타 검색엔진을 활용하게 될 경우에 우수한 정보 검색 능력으로 인해 실험용 데이타베이스의 오염도를 낮추게 되는 원인으로 작용되는 것을 피하기 위해 초기의 검색 시스템의 형태를 가장 잘 나타내고 있는 알타비스타를 이용하여 데이타베이스에 저장될 초기 웹 페이지들을 수집하였다.

실험용 검색 시스템과 성능을 비교하기 위한 대상으로는 알타비스타와 구글을 선정하였다. 알타비스타를 비교 대상으로 선정한 이유는 알타비스타가 전통적 방식의 검색 시스템 모습을 가장 잘 나타내고 있으며, 실험 대상 웹 페이지들에 대한 영역이 동일함으로 비교 실험의 오류를 최소화 시킬 수 있기 때문이다. 구글을 비교 대상으로 선정한 이유는 구글은 최근 많은 각광을 받고 있는 차세대 검색 시스템으로 과거의 검색 시스템의 문제점들을 개선한 검색 시스템이기 때문이다. 알타비스타와 구글은 의미 중의성 해소 기능이 없는 상용화 검색 시스템이고, 상용화 검색 시스템들은 일반적으로 중의성 해소 기능이 없다. 본 논문에서 제안한 실험용 검색 시스템은 질의어 의미를 활용함으로 의미 중의성 해소 기능이 존재한다. 그래서 실험 방식에서 일반 상용화 검색 시스템에서는 복수 질의어 사용을 허락하고 실험용 검색 시스템에서는 단일 질의어만 사용하도록 했다. 이와 같은 실험 방식은 복수 질의어의 사용으로 의미 중의성을 해소할 수 있는지에 대한 중요한 비교 실험이 된다.

실험은 알타비스타와 구글에서는 단일 질의어만을 이용하여 검색하는 경우와 복수 질의어의 사용을 허용하여 검색하는 경우(예:"java code")에 대해 실험하였다. 실험용 검색 시스템에서는 범주별 의미 가중치를 기반으로 한 사용자 선호도 가중치 방법(범주화되지 않은 클릭의 총수를 가중치에 적용)을 적용한 실험과 범주별 의미 가중치와 범주별 사용자 선호도 가중치 방법(범주별 클릭 가중치를 가중치에 적용)을 적용한 실험으로 나누어 실험에 임하였다. 본 논문에서 제안한 실험용 검색 시스템을 단일 어휘로 제한하고 일반 검색 시스템에서는 복수 질의어 사용을 허용한 이유는 복수 질의어의 사용으로 질의어의 의미 중의성을 해결할 수 있는지 실험하기 위해서이며, 복수 질의어를 활용함으로써 관련성이 높은 웹 페이지를 검색 결과에서 상위에 위치시킬 수 있는지에 대한 실험을 하기 위해서이다. 실험용 검색 시스템에서 적용한 두 가지 실험 방식 모두 범주별 의미 가중치를 기반으로 하며, 하나의 방식은 사용자 선호도를 범주화시키지 않은 클릭의 총수를 가중치에 적용하는 방식이고 다른 하나의 방식은 사용자 선호도를 범주화시킨 범주별 사용자 선호도 가중치를 적용하는 방식이다.

결과 웹 페이지의 정확도 판단에는 상위 30개 및 상위 10개의 결과 페이지에서 질의에 적합한 중요한 웹 페이지가 몇 개 존재하는가를 기준으로 하였다. 검색 결과 웹 페이지에 대한 판단은 대학원생 중 정보 검색 전공자 10명에 의하여 수행되었다. 웹 페이지의 적합성 및 중요성에 대해 1부터 10까지의 점수를 부여하게 하고, 평가된 점수에서 최상위 점수와 최하위 점수를 빼고 나머지 점수를 산술 평균하여 웹 페이지의 평가 점수로 활용하였다. 평가 점수가 7이상의 값을 갖는 경우 정확한 검색 결과로 판단하였다.

4.2 실험 결과

표 6은 실험용 검색 시스템과 일반 검색 시스템을 비교한 실험 결과이다. 실험 결과는 정확도에 대해 질의어의 의미로 표현하지 않고 질의어별로 나타내었다.

알타비스타와 구글의 실험에서 상위 30개의 결과를

보면 단일 어휘를 사용하였을 경우에 평균 정확도는 12.6%와 21%의 낮은 성능을 나타내었으며, 이러한 결과는 단일 어휘를 사용하였을 경우에는 질의어의 모호성이 큰 영향을 미치기 때문이다. 복수 질의어를 사용하였을 경우에는 평균 정확도가 50%와 58%로써 단일 어휘를 사용한 경우보다 35%와 37%의 향상이 있었으며, 복수 어휘를 사용하는 경우 질의어의 모호성 문제가 다소 해소될 수 있기 때문에 검색의 정확도가 높아진다. 그러나 복수 질의어를 활용하여도 중요한 웹 페이지를 상위 결과로 위치시킬 수 있는 것은 아니다.

알타비스타와 구글에서 상위 페이지에 좋은 결과를 위치시키는 성능 비교에 상위 30개의 웹 페이지에 대한 상위 10개의 웹 페이지 정확도 증가율을 이용하였다. 단일 어휘를 질의어로 사용한 경우에 알타비스타는 1%의 향상이 있고 구글은 2%의 향상이 있었으며, 복수 어휘를 사용한 경우에는 알타비스타는 1%의 향상이 있고 구글은 5%의 향상이 있다. 단일 어휘의 사용에서는 알타비스타와 구글이 비슷한 성능을 보이지만 복수 질의어의 사용에서는 구글이 알타비스타보다 4%의 정확도 향상을 나타내었다.

알타비스타와 구글의 실험에서 어휘 종속적인 웹 페이지 가중치 결정은 좋은 결과를 상위에 위치시키기 어렵다는 것을 20개의 질의어 실험으로 알 수 있었으며, 우수한 정보를 내포한 웹 페이지를 검색 결과 상위에 위치시키기 위해서는 어휘 종속적이지 않은 가중치 방법(사용자 반응, 웹 페이지 구조, 등.)의 필요성을 20개의 질의어를 활용한 실험에 의해 확인하였다.

범주별 의미 가중치와 단순 클릭 가중치(범주화되지 않은 클릭 총수를 가중치에 적용)를 적용한 실험용 검색 시스템은 상위 30개와 상위 10개에서 56%와 62%의 평균 정확도를 나타냈으며, 범주별 의미 가중치와 범주별 클릭 가중치(범주화된 클릭 값을 가중치에 적용)를 적용한 실험용 검색 시스템은 상위 30개와 상위 10개에서 64%와 78%의 평균 정확도를 나타내었다.

표 6 실험용 검색 시스템과 일반 검색 시스템의 검색 성능 비교

질 의 어	의 미	실험용 검색 시스템				Google				Altavista			
		범주별 의미 가중치+범주별 클릭 가중치		범주별 의미 가중치+단순 클릭 가중치		복수 질의 어휘		단일 질의 어휘		복수 질의 어휘		단일 질의 어휘	
어휘	의미 개수	Top 30	Top 10	Top 30	Top 10	Top 30	Top 10	Top 30	Top 10	Top 30	Top 10	Top 30	Top 10
JAVA	3	61	24	37	10	44	17	29	8	38	20	26	6
HORSE	2	36	16	34	8	34	13	29	7	29	11	20	4
CUSTOM	2	21	14	13	5	21	6	7	1	11	3	4	2
CHARACTER	2	20	13	18	11	20	10	10	5	15	9	10	5
BANK	2	41	16	40	16	37	12	26	7	30	10	20	9
COACH	2	33	17	34	16	34	14	22	12	33	11	22	9
BILL	2	40	16	32	11	33	13	11	6	29	8	6	4
BALANCE	2	32	16	29	14	29	10	5	1	23	5	1	0
BALL	2	47	15	34	15	41	16	9	4	38	15	2	1
CAPITAL	2	47	17	45	12	45	15	4	1	44	8	2	0
COURT	2	42	15	39	14	41	14	21	8	42	14	23	7
CULTURE	2	35	14	33	13	39	13	21	9	36	13	14	7
ENGAGEMENT	2	36	14	36	13	39	12	11	5	33	9	6	2
LETTER	2	38	16	30	14	25	13	11	2	19	7	1	0
MASS	2	27	11	25	11	26	15	0	0	23	13	2	0
PALM	2	44	17	39	13	32	11	6	4	33	9	4	1
PLANT	2	52	17	45	17	46	14	13	6	42	13	11	5
SEAL	2	42	19	37	15	37	12	12	4	33	10	3	1
PUPIL	2	49	14	47	13	45	15	4	2	29	12	4	0
SENTENCE	2	49	19	47	15	46	14	11	3	34	8	8	4
총계	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식	잘못된 계산식
											52%	13%	18%
											4%	2%	

실험용 검색 시스템의 두 가지 실험에서 상위 페이지에 좋은 결과를 위치시키는 성능 비교에 상위 30개의 웹 페이지에 대한 상위 10개의 웹 페이지 정확도 증가율을 이용하였다. 범주별 의미 가중치와 단순 클릭 가중치를 적용한 실험용 검색 시스템에서는 6%의 정확도 향상을 나타내었고, 범주별 의미 가중치와 범주별 클릭 가중치를 적용한 실험용 검색 시스템에서는 14%의 정확도 향상을 나타내었다.

실험용 검색 시스템과 알타비스타와 구글을 비교하면, 실험용 검색 시스템이 두 가지 검색 시스템에 비해 20개의 질의어를 활용한 실험에서 좋은 결과를 나타내었다. 그러나 복수 질의어의 사용을 허락한 구글의 실험 결과와 단일 질의어를 활용한 실험용 검색 시스템 중에 범주별 의미 가중치와 단순 클릭 가중치를 기반으로 한 실험 결과는 복수 질의어를 활용한 구글에 비하여 상위 30개와 상위 10개에 대해 -2%와 -1%의 성능 차이를 보였다. 이러한 결과는 복수 질의어의 활용으로 질의어의 의미 중의성을 다소 해결할 수 있다는 것을 나타내는 결과이다. 범주별 의미 가중치와 범주별 클릭 가중치를 적용한 실험용 검색 시스템은 알타비스타와 구글에 비해 20개의 질의어를 활용한 실험에서 우수한 성능을 나타내었다.

질의어의 확장과 질의어의 의미 활용은 웹 페이지의 변별력을 증대하여 검색 성능을 향상 시킬 수는 있으나, 단일 질의어의 사용은 질의어의 의미 중의성으로 인해 좋은 검색 결과를 생성하는데 단점으로 작용한다는 것을 20개의 질의어를 활용한 실험으로 확인하였다. 단일 질의어를 활용한 경우에는 질의 확장을 통해 검색 성능을 향상할 수 있다는 것을 20개의 질의어를 활용한 실험을 통해 확인할 수 있었고, 웹 페이지의 범주별 가중치에 사용자의 반응을 이용함으로써 검색 시스템의 성능을 향상 시킬 수 있다는 것을 동일한 실험에서 확인하였다.

5. 결 론

본 논문에서는 웹 검색의 정확도 향상을 위하여 범주별 의미 가중치와 범주별 사용자 선호도 가중치를 웹 페이지의 가중치에 부여하는 방식을 제안한다. 본 논문에서 제안하는 방법에서는 사용자가 검색 결과 리스트 중 특정 웹 페이지를 선택하는 행위를 모니터하여 다수 사용자에게 적용할 수 있도록 범주별 사용자 선호도 가중치를 웹 페이지의 가중치에 적용하였으며, 웹 페이지에 범주별 의미 가중치를 의미빈도와 범주내 역문헌빈도 방식으로 적용함으로써 웹 페이지의 변별력 증대에 효과적인 작용을 할 수 있도록 하였다. 또한 질의어의 의미 중의성을 해결하고 질의어의 의미 범주를 결정하는 워드넷 기반 사용자 인터페이스를 구축하였다.

제안된 가중치 부여 방식을 실험하기 위해 실험용 웹 검색 시스템을 구축하고 의미 중의성을 갖는 20개의 질의어를 선정하여 실험을 수행한 결과 제안된 범주별 가중치 적용 방법이 알타비스타나 구글에 비해 실험 대상인 20개의 질의어에서는 우수한 검색 성능을 나타내었고 실험용 검색 시스템은 알타비스타나 구글에 비해 중요한 웹 페이지를 상위에 위치시킬 수 있다는 것을 동일한 실험 결과로 확인하였다.

향후 연구로는 질의가 문장으로 입력되었을 경우에 워드넷의 의미 구조를 활용하여 자동으로 질의 문장의 의미를 분석하여 정확한 웹 검색을 이루는 연구와 웹 페이지의 가중치에 웹 구조를 이용하는 연구이다.

참 고 문 헌

[1] 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.

[2] D. Dreilinger and A. E. Howe, "Experiences with selecting search engines using metasearch," *ACM Transactions on Information Systems*, Vol.15, 1997.

[3] 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.

[4] N. J. Belkin, D. Kelly, G. Kim, J. Y. Kim, H. J. Lee, G. Muresan, M. C. Tang, X. J. Yuan and C. Cool, "Query length in interactive information retrieval," *SIGIR*, pp. 205-212, 2003.

[5] S. Lawrence, "Context in Web Search," *IEEE Data Engineering Bulletin*, Vol.23, pp.25-32, 2000.

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

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

[8] E. Voorhees, "Using WordNet to disambiguate word senses for text retrieval," *Proceedings of the 16th ACM-SIGIR Conference*, 1993.

[9] M. Sanderson, "Word sense disambiguation and information retrieval," *Proceedings of SIGIR-94*, 1994.

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

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

[12] G. A. Miller, "Nouns in WordNet: A Lexical Inheritance System," *Communications of the ACM*, Volume 38, Issue 11, 1995.
[13] S. Scott, and S. Matwin, "Text Classification Using WordNet Hypernyms," *Coling-ACL '98Workshop*, 1998.
[14] A. S. Chakravarthy and K. B. Haase, "NetSerf : Using Semantic Knowledge to Find Internet Information Archives," *Proceeding of the ACM SIGIR Conference*, 1995.
[15] D. Moldovan and R. Mihalcea, "Using WordNet and Lexical Operators to improve Internet Searches," *IEEE Internet Computing*, Vol.4, No.1, 2000.
[16] M. Balabanovic, "An Adaptive Web Page Recommendation Service," *Proceedings of the First International Conference on Autonomous Agents*, 1997.
[17] L. Chen and K. Sycara, "WebMate : A Personal Agent for Browsing and Searching," *Proceedings of the 2nd International Conference on Autonomous Agents*, 1998.
[18] E. J. Glover, S. Lawrence, W. P. Birmingham and C. L. Giles, "Architecture of a Metasearch Engine That Supports User Information Needs," *CIKM*, pp. 210-216, 1999.
[19] E. J. Glover and W. P. Birmingham, "Using decision theory to order documents," *In Digital Libraries 98*, Pittsburgh, PA, 1998.
[20] 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.
[21] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," The Journal of the ACM, Volume 46, Issue 5, 1999.

김 형 일
1996년 목원대학교 수학과 졸업(이학사) 1996년~1998년 (주)경기은행. 2001년 동국대학교 대학원 컴퓨터공학과(공학석사) 2001년~현재 동국대학교 대학원 컴퓨터공학과(박사과정). 관심분야는 지능형 에이전트, 정보검색, 기계학습, 전자상거래

김 준 태
1986년 서울대학교 제어계측공학과 졸업(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, Computer Science and Engineering(Postdoc). 1995년~현재 동국대학교 컴퓨터공학과 부교수. 관심분야는 지능형 에이전트, 정보검색, 기계학습, 자연어처리, 데이타마이닝