http://dx.doi.org/10.5392/JKCA.2013.13.01.019

범주 기반 평가를 이용한 검색시스템의 성능 향상

Improving Performance of Search Engine Using Category based Evaluation

김형일*, 윤현님**
나사렛대학교 멀티미디어학과*, 한국폴리텍대학 안성여자캠퍼스 디지털정보과**

Hyung-Il Kim(hkim@kornu.ac.kr)*, Hyun-Nim Yoon(yhnim@kopo.ac.kr)**

요약

정보에 대한 공간 복잡도가 높은 현재의 인터넷 환경에서는 사용자가 원하는 정보를 정확히 제공하는 것이 검색엔진의 목표이다. 그러나 대다수 검색엔진이 활용하는 내용 기반 기법은 현재의 인터넷 환경에서는 효과적인 도구로 사용될 수 없다. 내용 기반 기법은 어휘의 형태적 특성을 이용하여 웹페이지 가중치를 결정하기 때문에 웹페이지에 대한 변별력이 우수하지 못하다는 단점이 있다. 이러한 문제점을 해결하여 사용자에게 효과적인 정보를 제공하기 위해, 본 논문에서는 범주 기반 평가 기법을 제안한다. 범주 기반 평가 기법은 질의어를 의미관계로 확장하여 웹페이지와 유사성을 측정한다. 웹페이지 가중치 적용에 있어서, 범주 기반 평가 기법은 웹페이지 검색에 대한 사용자 반응과 질의어 범주를 가중치에 활용함으로써 웹페이지에 대한 변별력을 증가시킨다. 본 논문에서 제안한 기법은 사용자가 원하는 정보를 검색엔진을 통해 효과적으로 제공할 수 있는 장점이 있으며, 다양한 실험을 통해 범주 기반 평가 기법의 활용성을 확인하였다.

■ 중심어 : | 검색시스템 | 정보검색 | 정보여과 | 정보범주 |

Abstract

In the current Internet environment where there is high space complexity of information, search engines aim to provide accurate information that users want. But content-based method adopted by most of search engines cannot be used as an effective tool in the current Internet environment. As content-based method gives different weights to each web page using morphological characteristics of vocabulary, the method has its drawbacks of not being effective in distinguishing each web page. To resolve this problem and provide useful information to the users, this paper proposes an evaluation method based on categories. Category-based evaluation method is to extend query to semantic relations and measure the similarity to web pages. In applying weighting to web pages, category-based evaluation method utilizes user response to web page retrieval and categories of query and thus better distinguish web pages. The method proposed in this paper has the advantage of being able to effectively provide the information users want through search engines and the utility of category-based evaluation technique has been confirmed through various experiments.

■ keyword : | Information System | Information Retrieval | Information Filtering | Information Category |

* 본 논문은 2012년도 나사렛대학교 학술연구비 지원에 의해서 연구되었음.
접수번호 : #121127-001
접수일자 : 2012년 11월 27일
심사완료일 : 2012년 12월 20일
교신저자 : 윤현님, e-mail : yhnim@kopo.ac.kr

I. 서론

초기 인터넷에는 많은 정보가 존재하지 않아, 검색엔진의 역할은 사용자가 원하는 정보를 다양하게 제공하는 것으로 충분하였다. 그러나 현재와 같이 정보에 대한 공간 복잡도가 높은 인터넷 환경에서는 사용자가 원하는 정보를 정확하게 추출하는 것이 중요하다[1].

인터넷을 기반으로 존재하는 웹페이지는 현대사회의 정보화로 인해 기하급수적으로 증가하였으며, 이러한 정보 대량화에서는 사용자가 원하는 정보를 정확히 추출하는 고급화된 검색 능력이 필요하다[2]. 그러나 대다수의 인터넷 사용자들은 고급화된 검색 능력을 소유하지 못한 경우가 많기 때문에, 원하는 정보를 인터넷에서 편리하게 획득한다는 것은 매우 어려운 일이다. 이러한 일반 사용자들의 정보검색 문제점을 해결해야 하는 주체는 검색엔진이다.

대다수의 검색엔진들은 사용자 질의와 대상 웹페이지의 유사성 측정을 수행할 때 내용 기반 기법을 활용하며, 내용 기반 기법은 질의어가 웹페이지에 출현하는 횟수를 기준으로 유사성을 측정한다[3]. 이러한 환경에서 정보의 대량화와 공간 복잡도 문제가 효과적인 정보검색 기능을 약화시키기도 하지만, 가장 큰 문제는 질의어의 모호성에 존재한다.

질의어의 모호성은 다의어로 발생하며, 다의어는 하나의 어휘 형태에 다양한 의미가 존재하는 어휘를 의미한다[4]. 이러한 다의어가 질의어로 활용될 경우, 검색엔진은 사용자 의도를 정확히 파악하지 못하여 잘못된 정보를 추출한다.

내용 기반 기법의 문제점을 해결하기 위해 구글 검색엔진에서는 하이퍼링크 정보를 웹페이지 가중치에 적용하여 웹페이지 순위 결정에 사용하였다. 하이퍼링크는 웹페이지를 연결하는 도구로 사용되며, 인터넷 환경에서 하이퍼링크를 이용하면 다양한 정보에 편리하게 접근할 수 있다[5].

구글 검색엔진에서 사용한 하이퍼링크 가중치는 다음과 같은 기본 개념에서 출발한다. 특정 정보에 대해 중요한 내용을 포함한 웹페이지는 다른 웹페이지로부터 많은 접근이 발생하고, 웹페이지 접근에 사용되는 도구는 하이퍼링크이다. 이러한 하이퍼링크 연결구조를 분석하면 특정 정보에 대한 중요 웹페이지를 쉽게 파악할 수 있다. 웹페이지 연결구조를 분석하여 웹페이지 가중치에 하이퍼링크 정보를 적용한 검색엔진인 구글은 현재 전 세계에서 가장 많이 사용되고 있으며, 구글 검색엔진에서 사용한 웹페이지 가중치 기법은 Kleinberg의 HITS 알고리즘에 소개되어 있다[6].

인터넷에 연결된 웹페이지의 구조를 분석하면 특정 정보에 대해 중요한 웹페이지를 쉽게 선별할 수 있는 장점이 있다. 또한, 웹페이지 가중치 기법에서 사용자들의 웹페이지에 대한 반응은 중요한 정보로 활용될 수 있다. 사용자들의 웹페이지 검색 행위를 분석하면 사용자가 원하는 정보에 적합한 웹페이지를 쉽게 추출할 수 있다는 장점이 있다[7][8].

검색엔진 사용자들은 인터페이스를 통해 질의어를 입력하고, 질의어와 관계성이 높은 웹페이지를 결과로 받는다. 결과로 주어진 웹페이지들을 대상으로 사용자는 자신이 원하는 웹페이지만을 검색하며, 이러한 묵시적인 사용자 검색활동을 사용자 반응이라 한다[9]. 이러한 사용자 반응을 이용하면 정보의 인기도를 측정할 수 있다[10]. 사용자가 원하는 특정 정보에 대해 많은 정보가 포함된 웹페이지는 다수의 사용자로부터 참조될 것이며, 이와 같이 특정 정보에 대해 많은 참조가 이루어진 웹페이지에 높은 가중치를 부여함으로써 내용 기반 검색엔진의 단점을 보완할 수 있다[11].

본 논문에서는 검색엔진의 성능 향상을 위해 검색 질의어의 의미 확장을 수행하고 질의 의미 범주를 활용하여 웹페이지 가중치를 결정하는 범주 기반 평가 기법을 제안한다. 본 논문에서 제안한 기법을 검색엔진에 활용하면 사용자가 원하는 정보를 쉽게 추출할 수 있는 장점이 있다.

논문의 2장에서는 정보검색을 중심으로 관련 연구를 소개하고, 3장에서는 시스템을 구성하는 모듈들에 대한 설명과 본 논문에서 제안한 범주 기반 평가 기법에 대해 기술한다. 4장에서는 제안한 기법의 실험결과를 분석하고, 5장에서는 결론과 향후 연구에 대해 기술한다.

II. 관련 연구

인터넷 환경에서 특정 정보를 표현하기 위해서는 웹페이지를 이용하며, 이러한 웹페이지는 텍스트를 중심으로 표현되는 것이 일반적이다. 웹페이지와 같은 텍스트 기반 환경에서는 단어들을 활용하여 정보검색을 수행한다.

내용 기반 기법은 특정 주제를 표현하기 위해 텍스트를 이용한 영역에서 주로 활용되는 정보검색 기법이다[12][13]. 내용 기반 기법이 적용된 시스템에서는 사용자가 원하는 정보를 취득하기 위해 텍스트 기반 질의어를 활용하며, 시스템은 텍스트 기반 질의어를 이용하여 사용자에게 적합한 정보물을 추출하는 방식을 취한다. 사용자에게 적합한 정보물을 추출할 때는 사용 질의어와 대상 정보물의 유사도를 이용하며, 질의어와 정보물의 유사도 측정에서는 사용 질의어가 대상 정보물에 나타난 출현빈도를 활용한다. 어휘의 형태적 정보를 활용하는 내용 기반 기법은 정보물의 내용정보를 활용하기 때문에, 서적이나 기사 등과 같이 내용정보가 풍부한 경우에 적합하다.

내용 기반 기법에서 가장 널리 사용되는 두 가지 방식은 단순빈도 방식과 상대빈도 방식이다. 단순빈도 방식에서 가장 대표적인 방식은 단어빈도 방식이고, 상대빈도 방식에서 가장 대표적인 방식은 역문헌빈도 방식이다. 단어빈도 방식은 특정 정보물의 유사성 측정에 단어의 출현빈도를 이용한다. 특정 정보물의 설명문에서 핵심 단어는 자주 출현하기 때문에, 설명문에 자주 출현하는 단어는 중요도가 높은 단어라 판단할 수 있다. 이러한 단어의 출현빈도를 이용하여 정보물의 유사성을 측정하는 방식을 단어빈도 방식이라 한다. 상대빈도 방식에서 가장 많이 활용되는 역문헌빈도 방식은 문헌빈도가 가장 낮은 단어에 높은 가중치를 부여하는 방식이다. 이와 같은 방식의 가중치 기법을 사용하는 이유는 다양한 문헌에 출현하는 단어는 문헌의 변별에 중요한 역할을 수행할 수 없기 때문이다.

내용 기반 기법은 단어를 기반으로 문서의 유사성을 측정하기 때문에, 가중치 부여 대상이 광범위하다는 문제점이 있다. 이와 같이 가중치 대상 범위가 광범위하면 공간 복잡도가 증가하여 사용자가 원하는 정보를 원활하게 취득할 수 없다[14][15]. 이와 같은 문제점을 완화하기 위한 방법은 정보 영역화이다[16][17].

정보 영역화를 이용하는 가장 대표적인 분야가 도서검색이며, 도서검색에서는 도서분류법을 활용하여 서적을 빠르고 쉽게 검색한다. 도서분류법은 정형화된 기호체계를 사용하여 서적을 분류하며, 이러한 도서분류에 적용되는 기호체계에 대한 규칙을 도서분류기호법이라 한다. 도서분류법의 종류에는 한국십진분류법, 듀이십진분류법, 국제십진분류법 등이 있다. 한국십진분류법은 듀이십진분류법을 기본으로 수정한 분류체계로, 활용하기 쉽고 배열이 편리하다는 장점이 있다. 한국십진분류법에서 주류는 총류, 철학, 종교, 사회과학, 자연과학, 기술과학, 예술, 어학, 문학, 역사로 이루어져 있다. 이와 같은 도서분류법을 웹페이지의 적합도 범주로 활용하면 사용자가 원하는 정보를 원활히 추출하는데 효과적이다.

검색엔진에서 사용하는 질의어에는 다의어가 포함되어 있으며, 이러한 다의어는 어휘 모호성이라는 문제를 발생시킨다[18]. 어휘 모호성 문제로 인해 검색엔진은 사용자가 원하지 않은 정보를 추출하여 사용자에게 제공하는 오류를 빈번하게 발생시킨다. 이와 같은 문제를 해결하기 위해서는 사용자가 사용한 질의어에 대한 모호성을 완화시켜야 하며, 어휘 모호성 완화에 효과적인 활용 요소는 어휘의 의미관계이다.

어휘의 의미관계가 잘 표현되어 있는 워드넷은 의미를 기반으로 어휘들을 계층적 구조로 연결하였기 때문에, 연관성 높은 어휘를 추출할 경우에 효과적으로 사용될 수 있다[19]. 워드넷에는 어휘의 의미에 대한 카테고리가 정의되어 있으며, 어휘들의 계층구조와 연관관계가 표현되어져 있다[20].

본 논문에서 제안한 범주 기반 평가 기법은 어휘 의미 범주와 사용자의 묵시적 반응을 활용하여 웹페이지 가중치를 생성함으로써 웹페이지 변별력을 높이는 장점이 있으며, 어휘 의미 범주를 활용함으로써 내용 기반 기법의 문제점을 완화할 수 있다. 질의 확장을 위해 워드넷에서 정의한 어휘 설명문을 활용하였으며, 질의 확장을 통해 어휘 모호성을 완화시킬 수 있다.

III. 시스템 구성과 범주 기반 평가 기법

본 연구를 위해 설계한 시스템은 크게 세 가지 모듈로 나뉘며, 세 가지 모듈은 의미생성모듈, 정보생성모듈, 순위결정모듈이다. 의미생성모듈에서는 질의색인집합, 질의범주집합, 질의확장집합을 생성하고, 정보생성모듈에서는 웹색인집합, 어휘가중집합, 어휘범주집합을 생성하며, 순위결정모듈에서는 질의가중치집합, 정보가중치집합, 사용자 프로파일을 생성한다.

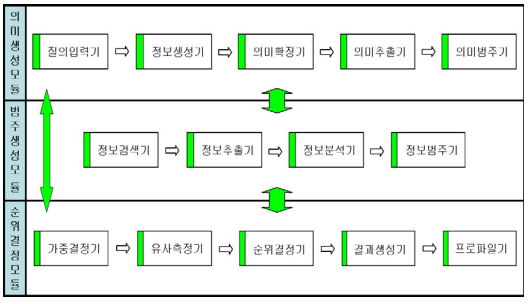

그림 1. 시스템 구성도

의미생성모듈은 질의입력기, 정보생성기, 의미확장기, 의미추출기, 의미범주기로 구성되며, 의미생성모듈에서의 정보 흐름은 다음과 같다. 질의입력기는 입력된 질의어에 대한 의미를 사용자에게 제공하고, 사용자의 의미 선택이 수행된 후에 질의입력기는 질의어와 선택된 의미에 대한 설명문을 정보생성기로 전송한다. 정보생성기는 어휘의 의미에 대한 설명문에서 중요 어휘만을 추출하여 의미확장기로 전송한다. 의미확장기는 질의어와 추출된 어휘집합을 이용하여 정보검색에 활용될 질의색인집합을 생성하고, 생성된 질의색인집합을 의미추출기로 전송한다. 의미추출기는 질의색인집합에 포함된 어휘들을 활용하여 각 어휘에 대한 범주를 추출하고, 추출된 범주와 질의색인집합을 의미범주기로 전송한다. 의미범주기는 질의색인집합에 포함된 단어와 질의범주집합에 포함된 범주를 이용하여 질의확장집합을 생성한다. 그러므로 질의확장집합에는 원시 질의어와 의미 설명문에서 추출한 색인어 및 해당 어휘들의 범주가 포함되어 있다. 최종적으로 생성된 질의확장집합은 정보생성모듈과 순위결정모듈로 전송된다.

정보생성모듈은 정보검색기, 정보추출기, 정보분석기, 정보범주기로 구성된다. 정보생성모듈은 사용자 질의를 기반으로 웹페이지를 추출하고, 웹페이지의 내용정보를 분석하는 모듈이다. 정보생성모듈로 입력된 질의확장집합은 정보검색기로 전송되며, 정보검색기는 입력된 질의확장집합을 이용하여 웹페이지를 검색한다. 검색된 대상 웹페이지는 정보추출기로 전송되고, 정보추출기는 웹페이지에서 중요 색인어를 추출하여 웹색인집합을 생성한다. 정보추출기에서 생성된 웹색인집합은 정보분석기로 전송되며, 정보분석기는 웹색인집합에 포함된 어휘들의 가중치를 계산하여 어휘가중집합을 생성한다. 어휘가중집합을 생성할 때는 웹색인집합에 포함된 어휘들의 웹페이지내 출현빈도를 활용한다. 정보분석기는 생성된 어휘가중집합과 웹색인집합을 정보범주기로 전송한다. 정보범주기는 입력된 웹색인집합을 활용하여 어휘범주집합을 생성한다. 정보생성모듈은 최종적으로 생성된 웹페이지에 대한 어휘가중집합과 어휘범주집합을 순위결정모듈로 전송한다.

순위결정모듈은 가중결정기, 유사측정기, 순위결정기, 결과생성기, 프로파일기로 구성되고, 순위결정모듈은 사용자에게 제공할 최종 결과 웹페이지를 결정하는 모듈이다. 가중결정기는 어휘가중집합과 어휘범주집합을 이용하여 웹페이지에 대한 정보가중치집합을 생성하고, 질의범주집합과 질의확장집합을 이용하여 질의가중치집합을 생성한다. 가중치결정기에서 생성된 질의가중치집합과 정보가중치집합을 이용하여 결과 대상 웹페이지와 질의어의 유사성을 측정하며, 유사성 측정에는 본 논문에서 제안한 유사성 평가함수를 이용한다. 유사측정기는 결과 대상 웹페이지의 유사성 평가값을 순위결정기에 전송하며, 순위결정기는 사용자 프로파일과 웹페이지와의 적합도를 평가한 후에 결과 웹페이지의 순위를 결정하여 결과생성기로 전송한다. 결과생성기는 최종 결과 웹페이지를 사용자에게 적합한 결과물로 변환하여 사용자에게 전송하며, 사용자 반응을 측정하여 프로파일기에 전송한다. 프로파일기는 전송된 사용자 반응을 이용하여 프로파일을 갱신하며, 프로파일에는 과거 사용자의 검색이력이 나타난다. 사용자 검색정보로 활용할 수 있는 속성은 사용 시간과 접근 횟

수 등이며, 본 연구에서는 접근 횟수만을 측정하였다.

본 연구에서 제안한 범주 기반 평가 기법을 식 1에 나타내었다. 식 1에서 DistFt(Q,R)은 검색 질의어 Q와 결과 대상 웹페이지 R의 거리를 나타내는 함수이고, DistFt(Q,R)은 S(Q,R)과 S(R,P)로 분해될 수 있다. S(Q,R)은 검색 질의어 Q와 결과 대상 웹페이지 R의 유사성 평가함수이고, S(R,P)는 결과 대상 웹페이지 R과 질의어 Q에 대한 검색이력 페이지 P의 유사성 평가함수이다. 검색이력은 사용자가 기존에 검색한 웹페이지를 기반으로 생성하며, 검색이력은 질의 범주를 기반으로 표현한 정보 페이지이다. 검색이력 페이지는 질의에 사용된 어휘와 범주 및 사용자 검색정보 등으로 구성된다. 본 연구에서 사용한 사용자 검색정보는 웹페이지 접근 횟수이다. 이러한 웹페이지 접근 횟수를 측정하면 사용자의 관심 분야를 쉽게 측정할 수 있는 장점이 있다.

유사성 평가함수에 사용되는 α와 β는 가중치 변수이고, 가중치 변수에 따라 유사성 평가함수의 적용도는 다양하게 표현된다. 일반적으로 질의어의 가중치 변수 α가 범주 가중치 변수 β보다는 변별력이 높기 때문에 본 연구에서는 α값이 2β값과 동일하도록 설정하였다. 특정 단어가 상위 의미로 이동되면 일반화되고, 하위 의미로 이동되면 특성화되는 것이 일반적이기 때문에 본 연구에서도 범주 가중치 변수 β의 최대값은 질의어 가중치 변수 α를 넘을 수 없도록 제한하였다. 결과 대상 웹페이지에 대한 검색 질의어의 빈도 평가값은 υ이고, 검색이력 페이지의 빈도 평가값은 $\jmath$이다.

식 1에서 유사성 평가함수 S(Q,R)을 검색 질의어와 결과 대상 웹페이지에 나타난 어휘를 중심으로 유사성 검사를 실시하기 위해 분해하면 식 2와 같다. 검색 질의어 Q는 질의어의 설명문을 이용하여 확장되며, 확장정보를 이용하여 질의벡터를 생성한다. 질의벡터는 $<q_1, q_2, \ldots, q_n>$으로 구성되며, 질의벡터에 사용되는 속성은 색인어이다. 결과 대상 웹페이지 R은 다양한 어휘들로 구성되어 있으나, 중요 정보를 포함한 색인어만을 이용하여 색인벡터를 생성하며, 색인벡터는 $<r_1, r_2, \ldots, r_n>$으로 구성된다. 각 벡터는 n차원으로 확장 가능하며, 질의벡터와 색인벡터를 이용하여 유사성 평가함수를 다시 표현하면 식 1은 식 2와 같다.

$$DistFt(Q,R) = \alpha \cdot (S(Q,R) \cdot \upsilon) + \beta \cdot (S(R,P) \cdot \jmath) \quad (1)$$

$$= \alpha \frac{(q_1 r_1 + q_2 r_2 + \ldots + q_n r_n)}{\sqrt{q_1^2 + q_2^2 + \ldots + q_n^2} \cdot \sqrt{r_1^2 + r_2^2 + \ldots + r_n^2}} \upsilon + \beta \cdot (S(R,P) \cdot \jmath) \quad (2)$$

식 1에서 유사성 평가함수 S(R,P)는 결과 대상 웹페이지와 검색이력 페이지의 유사성을 측정하며, S(R,P)를 페이지 요소를 이용하여 분해하면 식 3과 같다. 검색이력 페이지에서 사용한 요소는 확장 질의어에 대한 어휘, 의미, 범주, 접근 횟수 등이다. 검색이력 페이지에 나타난 범주는 워드넷에서 정의한 범주를 활용하였다. 검색이력 페이지에 나타난 범주를 활용하여 범주벡터를 생성할 수 있으며, 범주벡터는 $<p_1, p_2, \ldots, p_n>$으로 표현된다. 범주벡터와 결과 대상 웹페이지에 나타난 색인벡터 $<r_1, r_2, \ldots, r_n>$을 이용하여 검색이력 페이지와 결과 대상 웹페이지의 유사성 평가함수를 적용하면 식 3과 같다.

식 3에서 $\jmath$를 풀면 식 4와 같으며, 식 4에 나타난 C(R,P)는 결과 대상 웹페이지와 검색이력 페이지의 빈도 평가값이다. 빈도 평가값은 두 페이지의 유사 정보량을 의미한다. 빈도 평가값에 로그에 대한 절대값을 취함으로, 유사 정보 수량의 증가에 따른 결과 대상 웹페이지와 검색이력 페이지의 유사성 평가함수의 적용도가 증가된다.

$$\alpha \cdot (S(Q,R) \cdot \upsilon) + \beta \frac{(r_1 p_1 + r_2 p_2 + \ldots + r_n p_n)}{\sqrt{r_1^2 + r_2^2 + \ldots + r_n^2} \cdot \sqrt{p_1^2 + p_2^2 + \ldots + p_n^2}} \jmath \quad (3)$$

$$= \alpha \cdot (S(Q,R) \cdot \upsilon) + \beta \frac{(r_1 p_1 + r_2 p_2 + \ldots + r_n p_n)}{\sqrt{r_1^2 + r_2^2 + \ldots + r_n^2} \cdot \sqrt{p_1^2 + p_2^2 + \ldots + p_n^2}} \frac{1}{|\log C(R,P)|} \quad (4)$$

질의어와 결과 대상 웹페이지의 유사성 평가함수를 중심으로 υ를 풀면 식 5와 같으며, 식 5에 나타난 C(Q,R)은 질의어와 결과 대상 웹페이지의 빈도 평가값이다. 빈도 평가값 C(Q,R)은 질의어와 결과 대상 웹페이지의 유사 정보량을 의미하며, 해당 평가값을 질의어와 결과 대상 웹페이지의 유사성 평가함수에 적용함으로써 평가함수의 활용도를 증가시킨다.

질의벡터와 색인벡터를 이용한 유사성 평가함수와 범주벡터와 색인벡터를 이용한 유사성 평가함수를 활

용하여 벡터 기반 유사성 평가를 수행하면 식 6과 같이 표현되며, 식 6을 축약하면 식 7로 표현 가능하다.

$$\alpha\frac{(q_1r_1+q_2r_2+\ldots+q_nr_n)}{\sqrt{q_1^2+q_2^2+\ldots+q_n^2}\cdot\sqrt{r_1^2+r_2^2+\ldots+r_n^2}}\frac{1}{|\log C(Q,R)|}+\beta\cdot(S(R,P)\cdot\Im) \quad (5)$$

$$=\alpha\frac{\frac{(q_1r_1+q_2r_2+\ldots+q_nr_n)}{\sqrt{q_1^2+q_2^2+\ldots+q_n^2}\cdot\sqrt{r_1^2+r_2^2+\ldots+r_n^2}}}{|\log C(Q,R)|}+\beta\frac{\frac{(r_1p_1+r_2p_2+\ldots+r_np_n)}{\sqrt{r_1^2+r_2^2+\ldots+r_n^2}\cdot\sqrt{p_1^2+p_2^2+\ldots+p_n^2}}}{|\log C(R,P)|} \quad (6)$$

$$=\alpha\frac{\sum_{k=1}^{n}q_k\cdot r_k}{\sqrt{\sum_{k=1}^{n}q_k^2\cdot\sum_{k=1}^{n}r_k^2}}\frac{1}{|\log C(Q,R)|}+\beta\frac{\sum_{k=1}^{n}r_k\cdot p_k}{\sqrt{\sum_{k=1}^{n}r_k^2\cdot\sum_{k=1}^{n}p_k^2}}\frac{1}{|\log C(R,P)|} \quad (7)$$

본 논문에서 제안한 범주 기반 평가함수를 활용하면 질의어의 모호성을 완화할 수 있으며, 웹페이지의 가중치 변별력을 높일 수 있는 장점이 있다.

IV. 실험결과

본 논문에서 제안한 범주 기반 평가를 이용한 검색엔진의 실험을 수행하기 위해 실험 대상 검색엔진으로 구글을 선택하였다. 구글은 세계적으로 가장 널리 쓰이는 검색엔진으로 기존 검색엔진의 웹페이지 가중치 문제를 보완하기 위해 HIT알고리즘을 적용하였다.

실험용 검색엔진은 크게 네 가지로 나뉜다. 첫 번째 검색엔진은 내용 기반 검색엔진(tSE)으로, 내용 기반 검색엔진은 전통적 방식의 검색엔진이다. 내용 기반 검색엔진에서는 사용자 질의어와 웹페이지에 나타난 어휘와의 발생빈도를 측정하여 웹페이지 가중치를 결정한다. 두 번째 검색엔진은 웹페이지를 영역화한 영역 기반 검색엔진(ySE)으로, 영역 기반 검색엔진은 웹페이지의 정보에 따라 웹페이지를 영역화하여 웹페이지 가중치를 결정하는 검색엔진이다. 세 번째 검색엔진은 하이퍼링크 정보를 웹페이지 가중치에 적용한 구글 검색엔진(gSE)이고, 네 번째 검색엔진은 본 논문에서 제안한 범주 기반 평가를 적용한 검색엔진(cSE)이다.

범주 기반 평가를 적용한 검색엔진은 어휘 확장을 위해 어휘 설명문을 활용한다. 어휘 설명문 추출에 사용되는 워드넷을 [그림 2]에 나타내었다. [그림 2]에 나타난 결과를 보면 하나의 어휘에 다양한 의미가 존재한다는 것을 알 수 있다. 이러한 다양한 어휘 의미로 질의 모호성이 존재한다.

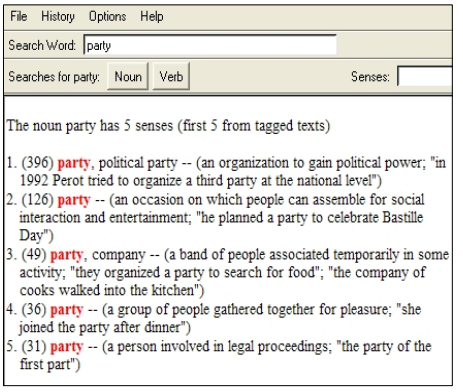

그림 2. 어휘 의미와 설명

실험에 사용한 어휘는 다양한 결과를 발생시키기 위해 중의성을 내포한 단어 20개를 선정하였다. 정보검색을 수행할 때는 의미를 고려하여 정보검색을 수행하기 때문에, 검색에 활용되는 의미 총수는 50개이다. 실험에 사용한 어휘는 note(메모, 지폐, 음계, 주석), party(정당, 파티), score(점수, 악보), school(학교, 학파), check(수표, 계산서, 점검), company(회사, 친구, 동료), draft(초안, 외풍), exercise(운동, 연습), race(경주, 인종, 경쟁), season(계절, 전성기), fortune(운, 재산), honor(명예, 체면, 영광), trade(무역, 거래, 직업), train(기차, 훈련, 행렬), ball(공, 무도회), bill(계산서, 지폐, 법안), book(책, 장부), capital(수도, 자본), credit(신용, 외상), delivery(배달, 강연, 분만)이다. 실험을 수행할 때는 각 단어에 나타난 의미를 활용하여 정보검색을 수행하며, 이와 같은 실험환경에 의해 정보검색 실험 분야는 총 50개이다.

영역 기반 검색엔진은 웹페이지를 영역화하기 때문에, 본 실험을 위해 질의어 의미별 2,000개의 웹페이지를 수집하여 웹페이지 영역화를 수행하였다. 웹페이지

수집에는 구글 검색엔진을 이용하였으며, 대학생과 대학원생으로 구성된 25명의 학생들이 웹페이지 수집과 검색엔진의 정확도 실험에 참여하였다.

결과 웹페이지의 적합도 평가를 위해, 실험 참여자들에게 0점부터 10점의 범위에서 웹페이지 적합도 점수를 부여하도록 하였다. 결과 웹페이지에 대한 최종 적합도 점수는 최상위와 최하위 점수를 제외한 전체 평균값을 활용하였다. 적합 판정에는 전체 평가를 종합한 평균값을 활용하였으며, 결과 웹페이지의 적합도 점수가 전체 평균값 이상일 경우는 적합으로 판정하였다.

웹페이지 적합도 평가 영역은 50개로 제한하였으며, 실험집합은 총 5개로 구성하였다. 5개 실험집합은 결과 웹페이지가 제공되는 수량을 기초로 나누었다. 첫 번째 실험집합은 Domain 1이고, Domain 1에서는 결과 웹페이지 20개가 사용자에게 제공된다. 실험집합 Domain 2, Domain 3, Domain 4, Domain 5에 나타난 결과 웹페이지는 각각 40개, 60개, 80개, 100개이다.

[그림 3]에 Domain 1의 실험결과를 나타내었으며, Domain 1에는 결과 대상 웹페이지가 20개 존재한다. Domain 1에서 tSE는 5%의 정확도를 나타냈고, ySE는 16%의 정확도를 나타내어 ySE는 tSE보다 11% 높은 정확도를 나타냈다. ySE는 웹페이지 영역화를 이용한 검색엔진이기 때문에, 관련성이 높은 결과 웹페이지가 tSE에 비해 우수하게 추출되었다. tSE는 단어 기반 검색이 수행되어, 관련성이 없는 결과 웹페이지가 자주 출현하였다. gSE는 23%의 정확도를 나타내어 tSE에 비해 18% 높은 정확도를 나타냈으며, ySE보다는 7% 높은 정확도를 나타냈다. 이와 같은 결과를 보더라도 웹페이지 연결구조는 웹페이지 가중치에 중요한 역할을 수행한다는 것을 알 수 있다. cSE는 32%의 정확도를 나타내어 tSE에 비해 27% 높은 정확도를 나타냈으며, ySE보다는 16% 높은 정확도를 나타냈다. 본 결과를 보더라도 범주 기반 평가가 단어 기반 평가나 영역 기반 평가에 비해 우수한 성능을 나타낸다는 것을 쉽게 확인할 수 있다. 본 실험에서 cSE는 gSE에 비해 9% 높은 정확도를 나타내어 범주 기반 평가가 웹페이지 연결구조 평가보다 우수하다는 것을 확인할 수 있었다. 범주 기반 평가는 질의어의 의미를 확장하여 가중치를 결정하기 때문에, 질의 모호성이 완화될 수 있는 장점이 있다.

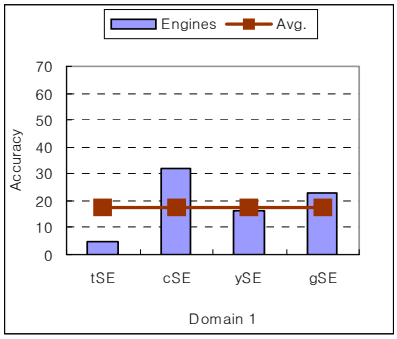

그림 3. Domain 1 실험결과

[그림 4]에 Domain 2의 실험결과를 나타내었으며, Domain 2에는 결과 대상 웹페이지가 40개 존재한다. Domain 2에서 tSE는 8%의 정확도를 나타내어 Domain 1에서의 성능보다 3% 향상되었다. 단어 기반 검색은 질의 모호성을 해소할 수 없기 때문에, 결과 대상 웹페이지 수량을 증가하여도 낮은 성능 향상을 나타냈다. 질의 모호성을 해결하지 않은 상태에서 결과 대상 정보만 증가시킨다는 것은 검색엔진의 활용성 측면에 치명적인 문제로 작용된다는 것을 본 실험으로 확인하였다. Domain 2에서 ySE는 18%의 정확도를 나타내어 Domain 1에서의 성능보다 2% 향상되었으며, ySE에서도 tSE에서와 같은 현상이 발생한 이유는 웹페이지 영역화에서도 여전히 질의 모호성은 해소되지 않았기 때문이다. ySE는 Domain 2에서 tSE보다 10% 우수한 성능을 나타냈다. gSE는 31%의 정확도를 나타내어 tSE에 비해 23% 우수한 성능을 나타냈으며, ySE보다는 13% 우수한 성능을 나타냈다. cSE는 Domain 2에서 45%의 정확도를 나타내어 tSE에 비해 37% 우수한 성능을 나타냈으며, ySE와 gSE보다는 각각 27%와 14% 우수한 성능을 나타냈다. 범주 기반 검색엔진은 정보의 수량이 증가할수록 더욱 우수한 성능을 나타낸다는 것을 본 실험결과를 통해 확인할 수 있었다.

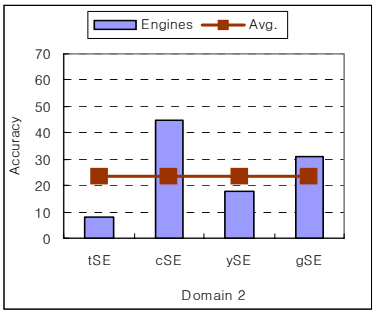

그림 4. Domain 2 실험결과

[그림 5]에 Domain 3의 실험결과를 나타내었으며, Domain 3에서는 결과 대상 웹페이지가 60개 존재한다. 본 실험에서 tSE는 15%의 정확도를 나타냈고, ySE는 24%의 정확도를 나타내어 ySE는 tSE보다 9% 우수한 성능을 나타냈다. gSE는 47%의 정확도를 나타내어 tSE에 비해 32% 우수한 성능을 나타냈으며, ySE보다는 23% 우수한 성능을 나타냈다. Domain 3에서 gSE 정확도는 Domain 1과 Domain 2에서의 정확도보다 각각 24%와 16% 높게 나타났다. 결과 대상 웹페이지가 증가함에 따라 gSE의 성능은 크게 높아진다는 것을 본 실험으로 확인할 수 있었다. 결과 대상 웹페이지가 증가한다는 것은 gSE 관점에서 웹페이지의 참고정보 활용도가 증가한다는 것을 의미하기 때문에, Domain 3에서의 gSE 정확도가 Domain 1과 Domain 2에서의 정확도보다 우수하게 나타난 것이다. cSE는 63%의 정확도를 나타내어 tSE에 비해 47% 우수한 성능을 나타냈고, ySE에 비해 38% 우수한 성능을 나타냈다. 본 실험에서 cSE는 gSE에 비해 15% 우수한 성능을 나타냈다.

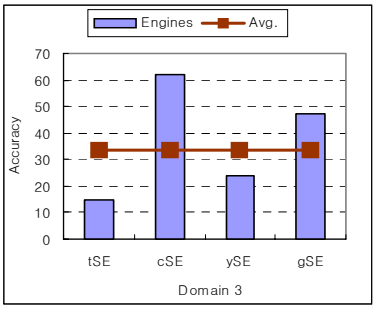

그림 5. Domain 3 실험결과

Domain 3에서 나타난 cSE의 성능이 Domain 1에서 나타난 성능보다 30% 높게 나타난 이유는 결과 대상 웹페이지가 증가됨으로써 범주 기반 평가가 사용자의 검색 의도 파악에 효과적으로 적응했기 때문이다.

[그림 6]에 Domain 4의 실험결과를 나타내었으며, Domain 4에서는 결과 대상 웹페이지가 80개 존재한다. Domain 4에서 tSE는 18%의 정확도를 나타내어 Domain 1에서의 성능보다는 13% 향상되었으며, Domain 2와 Domain 3에서의 성능보다는 각각 10%와 3%의 성능 향상을 나타내었다. 단어 기반 검색엔진에서도 결과 대상 웹페이지 수량이 증가함에 따라 성능 향상이 나타났으며, 이와 같은 성능 향상이 나타난 이유는 검색 대상 정보량이 증가하였기 때문이다.

Domain 4에서 ySE는 32%의 정확도를 나타내어 Domain 1에서의 성능보다 16% 향상되었으며, Domain 2와 Domain 3에서의 성능보다는 각각 14%와 8%의 성능 향상을 나타냈다. Domain 4에서 ySE는 tSE보다 14% 우수한 성능을 나타냈다. ySE에서 사용하는 영역화 평가는 대상 웹페이지가 증가할수록 내용 기반에 비해 우수한 성능을 나타냈다. 결과 대상 웹페이지를 영역화 함으로써 단어 기반 평가 기법의 단점을 보완할 수 있다는 것을 본 실험으로 확인하였다. gSE는 57%의 정확도를 나타내어 tSE에 비해 39% 우수한 성능을 나타냈고, ySE보다는 25% 우수한 성능을 나타냈다. cSE는 Domain 4에서 66%의 정확도를 나타내어 tSE와 ySE의 성능에 비해 각각 48%와 34% 우수한 성능을 나타냈고, gSE보다는 9% 우수한 성능을 나타냈다.

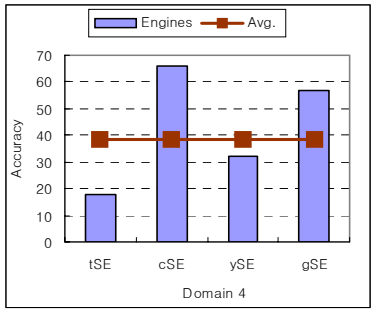

그림 6. Domain 4 실험결과

[그림 7]에 Domain 5의 실험결과를 나타내었으며, Domain 5에서는 결과 대상 웹페이지가 100개 존재한다. 본 실험에서 tSE는 25%의 정확도를 나타내어 모든 Domain에서 가장 우수한 성능을 나타내었다. ySE는 37%의 정확도를 나타내어 tSE의 정확도보다는 12% 높게 나타났다. gSE는 62%의 정확도를 나타내어 tSE와 ySE의 정확도보다 각각 37%와 25% 우수하게 나타났다. cSE는 69%의 정확도를 나타내어 tSE와 ySE의 정확도보다 각각 44%와 32% 높게 나타났다.

tSE는 14.2%의 평균 정확도를 나타내었고, ySE는 25.4%의 평균 정확도를 나타내어 tSE보다 평균 11.2% 우수한 성능을 나타냈다. gSE는 44%의 평균 정확도를 나타내어 tSE보다는 평균 29.8% 우수한 성능을 나타냈고, ySE보다는 평균 18.6% 우수한 성능을 나타냈다. cSE는 평균 54.8%의 정확도를 나타내어 gSE보다는 평균 10.8% 우수한 성능을 나타냈고, tSE와 ySE의 평균 정확도보다는 각각 40.6%와 29.4% 우수한 성능을 나타냈다.

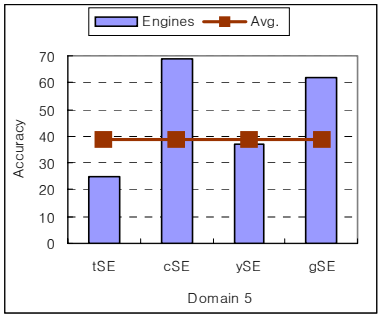

그림 7. Domain 5 실험결과

V. 결론 및 향후 연구

정보 대량화와 높은 공간 복잡도가 존재하는 현재의 인터넷 환경에서 내용 기반 정보검색을 활용하는 단어 빈도 방식은 비효율적인 도구이다. 대다수의 인터넷 검색엔진에서 활용하는 내용 기반 방법은 질의 모호성 문제로 사용자가 원하는 정보를 효과적으로 제공하지 못하는 단점이 있다.

이러한 문제를 해결하기 위해 본 논문에서는 범주 기반 평가 기법을 제안한다. 본 논문에서 제안한 기법은 질의 모호성을 완화하고, 웹페이지 가중치에 의미 범주와 사용자 반응을 적용함으로써 사용자가 원하는 정보를 쉽게 제공할 수 있는 장점이 있다.

여러 실험을 통해 내용 기반 정보검색의 문제점을 정보 영역화를 통해 보완할 수 있다는 것을 확인하였으며, 웹페이지 연결구조를 가중치로 활용하면 웹페이지 추출에 효과적이라는 것을 확인하였다. 질의 모호성을 완화시키기 위해서는 질의 의미로 정보검색을 수행하는 것이 효과적이며, 웹페이지 가중치 적용에 있어서 범주화와 사용자 반응이 중요한 요소로 작용할 수 있다는 것을 다양한 실험을 통해 확인하였다. 실험에서 tSE와 ySE는 14.2%와 25.4%의 평균 정확도를 나타내었고, gSE와 cSE는 44%와 54.8%의 평균 정확도를 나타냈다. 여러 실험결과를 통해 본 논문에서 제안한 기법이 웹페이지 정보검색에 효과적으로 적응할 수 있다는 것을 확인하였다.

향후 연구로는 다양한 사용자 반응을 분석하여 사용시간이나 이동 경로를 웹페이지 가중치에 적용하는 연구가 필요하다.

참 고 문 헌

[1] R. Kaul, Y. Yun, and S. Kim, “Ranking billions of web pages using diodes,” Communications of the ACM, Vol.52, No.8, pp.132-136, 2009.

[2] F. Liu, C. Yu, and W. Meng, “Personalized Web Search for Improving Retrieval Effectiveness,” IEEE Transactions on Knowledge and Data Engineering, Vol.16, No.1, pp.28-40, 2004.

[3] J. Zobel and A. Moffat, "Inverted files for text search engines," ACM Computing Surveys, Vol.38, No.2, 2006.

[4] K. W. Leung, W. NG, and D. L. Lee, “Personalized Concept-Based Clustering of Search Engine Queries,” IEEE Transactions on

Knowledge and Data Engineering, Vol.20, No.11, pp.1505-1518, 2008.

[5] A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, 2006.

[6] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment,"The Journal of the ACM, Vol.46, Issue.5, pp.604-632, 1999.

[7] T. Chen, W. Han, H. Wang, Y. Zhou, B. Xu, and B. Zang, "Content Recommendation System Based on Private Dynamic User Profile," International Conference on Machine Learning and Cybernetics, pp.2112-2118, 2007.

[8] M. N. Uddin, J. Shrestha, and G. Jo, "Enhanced Content-Based Filtering Using Diverse Collaborative Prediction for Movie Recommendation," 2009 First Asian Conference on Intelligent Information and Database Systems, pp.132-137, 2009.

[9] D. Billsus and M. Pazzani, "Learning Collaborative Information Filters," Proceedings of the 15th International Conference on Machine Learning, 1998.

[10] B. Krulwich, "Lifestyle Finder: Intelligent user profiling using large-scale demographic data," Artificial Intelligence Magazine, Vol.18, No.2, 1997.

[11] L. Deng, W. Ng, X. Chai, and D.L. Lee, "Spying Out Accurate User Preferences for Search Engine Adaptation," Advances in Web Mining and Web Usage Analysis, LNCS3932, pp.87-103, 2006.

[12] C. Jian, Y. Jian, and H. Jin, "Automatic content-based recommendation in e-commerce," The 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service, pp.748-753, 2005.

[13] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[14] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001.

[15] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.5, pp.657-668, 2005.

[16] S. H. Al-Harbi, "Adapting k-means for supervised clustering," Applied Intelligence, Vol.24, No.3, pp.219-226, 2006.

[17] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison Wesley, 2006.

[18] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," Information Processing & Management, Vol.42, No.1, pp.248-263, 2006.

[19] G.A. Miller and F. Hristea, "WordNet Nouns: Classes and Instances," Computational Linguistics, Vol.32, No.1, pp.1-3, 2006.

[20] S. K. Ray and S. Singh, "Blog content based recommendation framework using WordNet and multiple Ontologies," 2010 International Conference on Computer Information Systems and Industrial Management Applications, pp.432-437, 2010.

저 자 소 개

김 형 일(Hyung-Il Kim) 종신회원

- 1996년 ~ 1998년 : (주)경기은행
- 2004년 : 동국대학교 컴퓨터공학과(공학박사)
- 2005년 ~ 2006년 : 동국대학교 컴퓨터공학과 IT교수(정보통신부)
- 2007년 ~ 현재 : 나사렛대학교 멀티미디어학과 교수

<관심분야> : 정보검색, 추천시스템, 인공지능, 의료영상, 데이터마이닝, 임베디드시스템, 기계학습

윤 현 님(Hyun-Nim Yoon) 정회원

- 1996년 ~ 1999년 : (주)유니파이코리아
- 2009년 : 동국대학교 정보통신공학과(공학박사)
- 2001년 ~ 현재 : 한국폴리텍 대학 안성여자캠퍼스 디지털정보과 교수

<관심분야> : 그리드 컴퓨팅, 모바일 컴퓨팅, 멀티미디어 통신, 가상교육, 정보검색, 의료영상