웹 사용자 누적 사용정보 기반의 키워드 검색 모델

윤성희*

A Keyword Search Model based on the Collected Information of Web Users

Sung-Hee Yoon*

요 약

본 논문은 웹 검색 시스템의 사용자 질의에 대한 키워드 색인 기반의 검색 과정에서 적합 문서를 선별하기 위해 검색 키워드의 의미정보와 사용자의 누적 사용정보를 사용하여 검색 성능을 향상시키는 방법을 소개한다. 검색 키워드 의미 정보를 이용하는 검색 방법은 검색 결과로서 의미적으로 무관한 많은 문서들을 배제할 수 있고, 사용자의 누적된 사용정보는 관심사에 중심을 둔 검색문서들을 상위에 제시할 수 있다. 검색 키워드의 의미정보 지식베이스를 구축하고, 검색 문서들을 색인어와 해당 의미범주로 분류하며, 사용자의 정답 문서 참조 행위에 대한 누적 정보를 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

search engine, query word sense, ambiguity, user feedback
정보검색, 질의어, 검색 키워드 의미분류, 사용자 피드백

Ⅰ. 서 론

웹 정보검색 시스템에서 검색 대상 문서의 양이 급속하게 증가하고 있는 추세에 따라 검색 시스템의 성능에 대한 요구가 함께 높아지고 있으며, 사용자의 질의 의도에 맞는 문서들을 신뢰할 만한 순위로 제공하는 성능 향상에 초점을 두는 기술 개발이 요구되고 있다. 이 성능은 재현률과 정확률이라는 지표로 평가되며, 검색엔진에 대한 사용자의 만족도를 좌우하게 된다. 정보검색과 관련되는 기술의 발전을 위해서 미국 NIST(National Institute of Standard and Technology)의 주관으로 TREC(Text Retrieval Conference) 학술대회가 매년 열리고 있으며, 여러 관련 기술 분야별로 전문화된 트랙을 두어 기술 발전을 유도하고 있다.

웹 검색에서는 전문적이기 보다는 일반적이고 보편

* 상명대학교 컴퓨터소프트웨어공학과(shyoon@smu.ac.kr)
접수일자 : 2012. 05. 31 심사(수정)일자 : 2012. 07. 26 게재확정일자 : 2012. 08. 09

적인 용어들이 질의어로 많이 사용되고 있으며, 두 개 또는 세 개의 단어로 질의한다는 실험결과를 중심으로 볼 때, 검색 키워드의 의미를 정확히 판별하는 것이 검색 결과의 성능을 높이는데 매우 중요한 과정이다[1]. 현재 널리 사용되는 검색 방법은 질의어를 포함하는 문서들 중에서 대중적으로 많이 참조되는 문서를 중요한 문서라고 평가하여 상위에 랭크하여 보여주는 방식인데, 이러한 방법은 같은 질의어를 사용한 사용자들도 사용자마다 원하는 문서의 종류가 다르기 때문에 모두에게 가치있는 문서가 상위에 랭크되기 어렵다. 실험에 의하면 검색 시스템 사용자들은 일반적으로 상위에 제시되는 몇 개 문서만을 참조하는 경향이 있으므로 검색 시스템의 정확률(Precision)이 재현률에 비해서 더 중요하다고 평가되므로 개인의 관심사와 질의 의도를 반영한 검색결과에 대한 랭크 기법이 적용될 필요가 있다.

본 논문에서는 검색 대상 문서들의 의미정보 분류와 수집된 사용자의 검색 행위 정보에 기반한 검색 방법을 제안한다. 개별 의미범주에 따라 분류된 문서 집합으로부터 정답 문서를 선택하고, 정답 후보 문서 중에서 사용자가 참조하는 검색행위를 가중치로 누적하여 지속적으로 반영함으로써 검색 정확도를 높일 수 있다. 검색 과정에서 정답 후보 문서들 중 사용자가 보이는 적극적인 검색 행위를 수집하고 계량화하여 사용자의 반응 피드백으로 검색 문서의 순위 결정에 반영함으로서 순위 정확률을 높일 수 있다[1,2]. 사용자 웹 사용정보와 로그에 대한 분석은 웹페이지를 열람하는 사용자의 행위 및 웹 사이트 사용정보를 분석에 사용하고 웹페이지 추천을 위한 기반 기술로 유용하게 사용하는 방법이다[3,4]. 사용자의 누적 사용정보는 시간적 흐름에 따라 검색 성향이 바뀌거나, 관심사의 개념이 변경되는 것을 반영하여 최근의 검색 경향을 개인화정보로 적용하므로 최근의 관심에 적합한 문서들을 상위에 랭크할 수 있게 한다.

II. 관련 연구

2.1 사용자 검색 행위 분석

정보검색 시스템은 입력 질의(query)의 키워드와 웹 문서에 나타나는 색인어(index)의 형태적 일치를 검사함으로써 적합 문서 여부를 결정하며, 많은 웹 문서들을 주기적으로 방문하고 색인 데이터베이스를 갱신한다. 대부분의 검색 시스템 사용자들은 복잡한 검색식이나 연산자를 사용하지 않으며, 대신 매우 단순한 키워드를 검색 질의어로 입력하고 검색된 문서들에 대한 연관성 평가를 통해 더 관련이 많은 문서를 찾기 위해 재검색하는 것이 일반적이다.

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[8]. 이와 같이 사용자의 질의가 단일 어휘이거나 단순한 명사구의 형태를 갖는 많은 경우에 질의에 나타나는 키워드의 의미를 구분하고 초기 검색 결과에 대한 사용자 만족도의 척도가 되는 참조 행위를 반영하는 과정이 포함되어야 한다. 웹 검색에 대한 연구는 네트워크 사용자를 위한 문서 검색 뿐만 아니라 네트워크 내에서의 이미지를 비롯한 다양한 데이터 형식에 대한 내용기반 검색을 위해서도 널리 연구되는 중요한 주제이다.[5,6,7]

2.2 키워드 의미정보 추출

언어처리 분야에서 의미정보를 체계화하는 대표적인 방법 중 널리 사용되는 방법은 어휘에 대하여 동의어, 반의어, 상위의미, 하위의미 등과 같은 어휘의 연관성을 정의한 사전인 워드넷(WordNet)이다[9]. 워드넷에는 어휘의 의미에 대한 범주 분류가 잘 정의되어 있으며, 단어들 사이의 계층구조와 연관관계가 여러 형태로 표현되어 있다. 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.

실세계 지식을 보완하기 위해 시소러스로부터 검색어들을 추출하고, 사용자의 선택에 따라 검색질의를 확장하는 방법이 소개되기도 하였다[12]. 시소러스에서의 상하관계는 사전적인 의미가 아닌 필요에 따른 용례에 의한 관계가 반영되어서, 시소러스가 단어 간

의 관계에 대한 기준이 명확하지 않다고 평가되기도 한다. 한편 어휘 개념망 방법을 한국어 명사에 대한 의미 지식베이스 구축에 실험적으로 적용한 사례도 있다[12,13]. 사전의 뜻풀이를 중심으로 개념어들 간의 국어학적 의미관계를 연결하여, 단어들의 의미 포함관계를 명확하게 나타내고자 하였다.

TREC의 의미범주 계층을 활용한 실험에서는 질의문에 대한 정답을 웹 문서에서 찾아주는 질의응답 시스템에서 단어의 의미정보를 활용하여 질의문장의 유형을 분류하고 있는데[14]. 자연어처리 기술을 이용한 의미기반 색인과 검색 방법에 대한 연구도 활발히 소개되고 있다. 이들 관련 연구들은 어휘, 구절, 문장의 중의성을 해소하기 위해 형태소분석 및 태깅(tagging) 기술 뿐만 아니라 구문분석과 의미분석 등 자연어처리 기술을 이용하는 방법으로서, 마찬가지로 사용자 질의 의도를 추출하고 질의의 의미를 색인에 반영하여 검색성능을 높이는 것이 목적이다.

III. 사용자 정보 기반의 키워드 검색

본 논문에서 제안하는 검색 시스템의 검색 흐름은 다음의 그림 1에서와 같다. 의미기반 지식베이스로서 의미범주를 분류하는 지식베이스와 의미 범주별로 분류되고 지속적으로 가중치가 갱신되는 웹 문서 집합을 지식베이스로 갖는다. 검색 과정은 사용자의 질의입력, 키워드 의미분류 후 질의확장, 의미 선택된 색인으로 웹 문서 검색, 정답후보 문서에 대한 사용자 반응 수집과 가중치 갱신과 순위화의 순서로 이어진다.

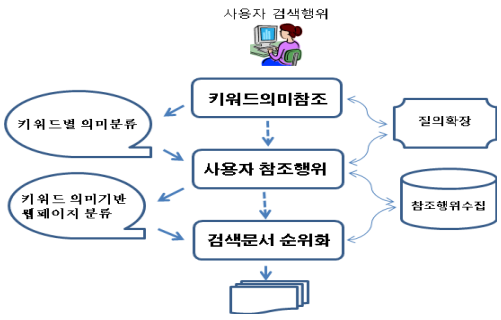

그림 1. 사용자정보 기반의 키워드 검색
Fig. 1 Keyword search based on user Information

3.1. 검색 키워드 의미 분류

앞에서 서술하였듯이, 정보검색에서 중의성 문제가 해결되면 정보검색 시스템의 정확도를 크게 향상시킬 수 있다. 이를 위하여 사용자 질의 키워드와 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 지식베이스을 구성하였다[10]. 의미적 모호성 해소를 위한 의미정보는 중의적 키워드의 의미범주 분류와 그 사전적 해석으로부터 추출할 수 있는 확장 정보들을 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 질의 확장과 유사도(similarity)에 기반한 순위결정에 사용될 수 있다. 중의성을 갖는 키워드가 의미 범주로 분류되는 체계의 예를 그림 2에서 보여준다.

사용자가 입력하는 질의어의 중의성을 해소하고, 질의어의 의미와 범주를 결정하기 위해 사용자 검색 행위 정보를 누적하는 인터페이스를 갖는다. 사용자가 입력한 질의어가 중의성을 갖는다면 의미정보 지식베이스에 복수의 의미범주 항목을 포함하고 있으며, 의미선택기는 각 질의어의 의미범주 단위로 동의어, 상위어, 사전적 설명 등을 추출한다. 추출된 복수의 의미범주는 사용자 인터페이스를 통해 사용자에게 전달되고, 사용자는 본인이 의도한 질의의 의미를 선택할 수 있다. 사용자 질의어와 선택된 의미범주을 색인으로 조합하여 문서 집합을 검색하게 된다. 개념이 선택된 검색 키워드는 의미 분류된 질의확장 과정을 거치며, 사용자의 원래 질의와 선택된 의미범주를 색인으로 조합하여 문서 집합을 검색한다.

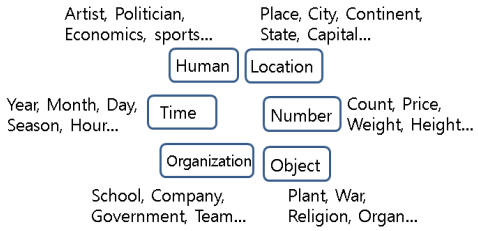

그림 2. 키워드 의미 분류 체계
Fig. 2 Keyword semantic categories

3.2 검색 문서 의미 색인과 사용자 반응 수집

검색 대상 문서의 검색 키(key)가 될 질의어가 중의적 성격을 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. 검색 문서들은 단일한 범주에만 속하지 않을 수 있으므로 검색 문서를 단일한 범주로 결정하는 것은 정보 활용 측면에서 위험한 문제를 일으킬 수 있다. 이를 해결하기 위해검색 문서를 색인어와 의미범주 단위로 다중 색인하여 검색의 변별력을 높인다. 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 부여하고 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 그림 3에서 보여주고 있다.

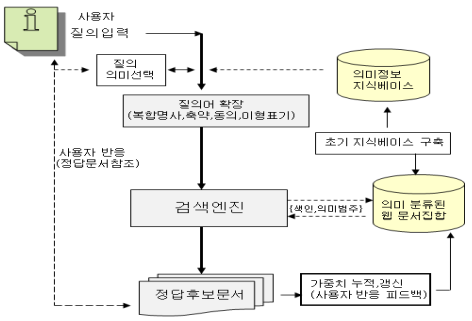

그림 3. 사용자 검색 행위와 검색 흐름
Fig. 3 User search action and search flow

Ⅳ. 실험 및 평가

본 연구의 실험에서는 개인, 학교, 기관 등의 홈페이지들, 학술 내용과 시사 뉴스 등의 내용을 담은 다수의 웹 문서들을 대상으로 하였다. 이 문서 집합은 선행 연구와 실험을 위해 구축된 바 있으며[12], 본 연구의 실험을 위해 시사 뉴스 분야의 웹 문서가 확장되고 검색 키워드에 대한 의미 범주별 색인 작업이 추가되었으며, 사용자의 최근 검색 행위에 대한 누적 정보를 개인화 검색을 위해 사용하는 방법으로 확장되었다. 실험 대상인 검색시스템의 사용자들에게 웹 문서에 대한 검색 질의를 입력하도록 하였으며, 실험에서 약 오천 회 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 1과

같이 선별하였으며, 사전적 의미 수를 보이고 있다. 실험에 사용된 검색 키워드의 80%이상이 한 번씩만 검색되었던 키워드들이므로 일정 회수 이상 검색된 키워드만을 실험 대상으로 삼았다. 중의적 검색 키워드는 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. 실제로 실험에서 상위 20개 페이지 중에서 19개, 상위 30개 웹 문서 중에서 24개의 문서들이 첫 번째 의미로 사용된 문서였다.

표 1. 실험에 사용된 검색 키워드 예
Table 1. Keyword samples for experiment

검색 키워드	의미 분류 수	검색 키워드	의미 분류 수
가사	5	인사	5
기도	5	은행	2
상용	3	조회	2
수도	5	정의	4
인도	5	지연	2

실험 결과로는 사용자의 질의 키워드에 대해서 중의성을 의미 분류하지 않은 채 검색한 경우와 사용자가 의미범주를 선택하고 의미 색인된 웹 문서들을 검색한 경우의 평균성능을 분석하였었다. 검색 결과 문서집합에서 상위 30개, 20개, 10개 문서에서 적합 문서를 찾은 경우를 분석해 보았을 때, 사용자 정보에 기반한 키워드의 의미를 분석과정을 거치지 않았을 경우에는 각각 42%, 56%, 68%였다. 반면, 본 논문에서 제안한 바와 같이 사용자의 검색 행위를 누적한 정보를 바탕으로 검색 키워드의 의미 분류 과정을 거친 검색결과 문서집합에서는 각각 58%, 77%, 87%의 비율로 적합문서를 참조할 수 있었다. 그림 4에서 가로축은 각각 상위랭킹 30개 문서, 20개 문서, 10개 문서의 경우를, 세로축은 적합문서가 포함된 비율을 나타낸다. 사용자의 검색 행위에 대한 개인화된 정보를 기반으로 의미 분류된 문서들을 검색한 경우에 상위 랭킹 문서에 상대적으로 높은 비율로 정답문서를 포함하고 있음을 볼 수 있다. 이 과정에서 사용자의 정답문서에 대한 참조 반응을 순위 갱신에 반영하기 위해 각 검색의 결과로 제시되는 정답 후보 문서들 중에서 사용자가 상위 정답문서로 평가하는 것들을 다섯 개 이상 선택적으로 참조하도록 하여 그 반응을 누적하였다. 실험이 누적된 정보는 검색결과의 문서들 간에 변경되고, 상위랭킹 문서들에 대한 만족도를 높이는 목적으로 적용된다. 실험의 결과에서 상위 10위 내의 문서들에 대한 정답 평가 정도가 크게 향상되어 87%에 이른다는 의미는 상위 10위 내의 문서들 중에 사용자가 만족하는 적합 문서들을 포함하는 것으로 평가할 수 있다.

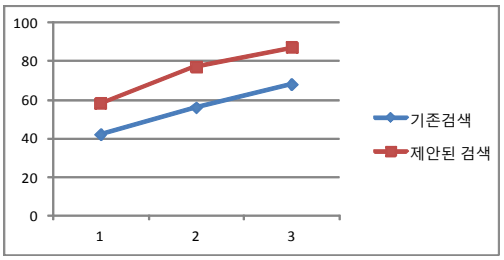

그림 4. 제안된 검색방법의 성능 향상
Fig. 4 Enhanced performance of proposed search model

V. 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 검색행위 정보를 누적 수집하여 순위에 반영하는 방법을 제안하였다. 본 연구를 위한 자원으로 검색의 색인어가 되는 명사들을 중심으로 의미정보 체계를 지식베이스로 구축하고, 웹 문서들을 색인어의 의미범주로 분류한다. 검색과정에서는 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 보여주었다.

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. 이와 더불어, 수시로 생성되고 소멸되며 내용이 변경되는 동적인 웹 환경을 위한 유연한 색인 데이터베이스 구축 시스템과 실험적 규모 이상의 의미 지식베이스를 구축할 필요가 있다.

감사의 글

본 논문은 2010년도 상명대학교 교내연구비 지원으로 수행된 연구의 결과임.

참고 문헌

[1] 김성진, “이용자 중심 웹 정보탐색 연구의 실체 이론 분석”, 정보관리학회지, 23권, 3호. pp. 127-146, 2006.

[2] 박건우, 이상훈, “질의어 패턴 자동분석을 통한 커뮤니티 기반 개인화 검색,” 한국정보과학회 논문지 D, 36권, 04호, pp. 321-326, 2009.

[3] 김태환, 전호철, 최중민, “페이지 랭크지수와 질의 확장을 이용한 재랭킹 방법”, 한국정보처리학회 논문지, 18-B권, 04호. pp. 231-240, 2011.

[4] 윤태복, 이승훈, 윤광호, 이지형, “웹 사용 정보에 기반한 다중 성향 키워드 모델의 설계와 응용,” 한국인터넷정보학회논문지, 10권, 05호, pp. 95-105, 2009.

[5] 김분희, “사용성 개선을 위한 P2P 그룹 검색 알고리즘”, 한국전자통신학회논문지, 5권, 2호, pp. 185-192, 2010.

[6] 김분희, “전처리 검색 기반의 P2P 그룹 검색 알고리즘”, 한국전자통신학회논문지, 5권, 5호, pp. 522-527, 2010.

[7] 김광백, 우영운, “HSI 컬러 공간과 신경망을 이용한 내용기반 이미지 검색”, 한국전자통신학회 논문지, 5권, 2호, pp. 152-157, 2010.

[8] 박상규, 이찬규, 윤경현, 김성희, 이준호, 2007, “검색엔진에서 질의어 분포의 정상성에 관한 연구”, 한국정보관리학회지, 24권, 4호. pp. 255-265, 2007.

[9] Moldova D. and Mihalcea R., "Using WordNet and Lexical Operators to improve Internet Searches," IEEE Internet Computing, Vol. 4, No. 1. pp. 36-43, 2000.

[10] 강현규, "개념 검색어 대체를 통해 질의 형식화를 도와주는 개념 마법사의 설계 및 구현" .정보처리학회논문지, 9-B권, 04호, pp. 437-444. 2002.

[11] Perez-Carballo Jose and Strazalkowski Tomek. "Natural Language Information Retrieval : progress report." Information Processing & Management, Vol. 36, No. 1, pp. 155-178, 2000.

[12] 윤성희, 장혜진,“검색엔진의 정확률 향상을 위한 질의어 의미와 사용자 반응 정보의 이용”, 정보관리학회지, 26권, 4호. pp. 81-92, 2009.

[13] 이용구, 정영미, 사전 정보를 이용한 단어 중의성 해소 모형에 관한 실험적 연구”, 한국정보관리학회지, 24권, 1호. pp. 321-342, 2007.

[14] TREC <http://trec.nist.gov/pubs.html>

저자 소개

윤성희(Sung-Hee Yoon)

1987년 서울대학교 컴퓨터공학과 졸업(공학사)

1989년 서울학교 대학원 컴퓨터공학과 졸업(공학석사)

1993년 서울대학교 대학원 컴퓨터공학과 졸업(공학박사)

1994년~현재 상명대학교 컴퓨터소프트웨어공학과 교수

1999~2000 University of Michigan 방문연구교수

2007~2008 University of Victoria 방문연구교수

※ 관심분야 : 인공지능, 자연어처리, 정보검색, 인터넷서비스