퍼지 및 지능시스템학회 논문지 2004, Vol. 14, No. 6, pp. 737-744

지식기반 의미 메타 검색엔진

Knowledge-based Semantic Meta-Search Engine

이인근 · 손세호 · 권순학*

In K. Lee, Seo H. Son, and Soon H. Kwon*

영남대학교 전자정보공학부

요 약

웹으로부터 사용자가 원하는 정보에 잘 부응하는 정보를 추출하는 것은 검색엔진이 갖추어야 할 기본적 요소라 할 수 있다. 그러나, 질의어와의 패턴 매칭 방식에 의존하는 기존의 대부분의 검색엔진은 질의어가 갖는 애매성으로 인하여 사용자의 요구에 부합하는 검색결과를 제공하기가 쉽지 않다는 단점을 지니고 있다. 이를 극복하기 위하여 본 논문에서는 다음과 같은 5가지 과정, 즉, (i) 질의어 형성, (ii) 질의어 확장, (iii) 검색, (iv) 순위 재생성 및 (v) 지식베이스로 구성되는 지식기반 의미 메타 검색엔진의 기본 구조를 제안한다. 영어로 구현된 웹 문서에 대한 모의실험을 통하여 본 논문에서 제안된 지식기반 의미 메타 검색엔진이 기존의 검색엔진(구글)을 사용하여 얻은 결과보다 좋은 결과를 보임을 확인할 수 있었다.

Abstract

Retrieving relevant information well corresponding to the user's request from web is a crucial task of search engines. However, most of conventional search engines based on pattern matching schemes to queries have a limitation that is not easy to provide results corresponding to the user's request due to the uncertainty of queries. To overcome the limitation, in this paper, we propose a framework for knowledge-based semantic meta-search engines with the following five processes: (i) Query formation, (ii) Query expansion, (iii) Searching, (iv) Ranking recreation, and (v) Knowledge base. From simulation results on english-based web documents, we can see that the proposed knowledge-based semantic meta-search engine provides more correct and better searching results than those obtained by using the Google.

Key words : 정보검색, 검색엔진, 메타 검색, 질의어 확장, PageRank 알고리즘, WordNet.

1. 서 론

매일 수백만 건의 검색이 이루어지고 있는 인터넷 검색엔진은 가장 비중이 큰 온라인 서비스로서 많은 유동적인 정보를 쉽게 검색할 수 있는 장점으로 인해 그 인기가 높아지고 있다. 일반적으로 검색엔진은 (i)검색자의 정보요구, (ii)인터넷이나 웹을 표현하기 위한 모델, 그리고 (iii)검색 방식과 검색 순위 결정 방법과 같이 크게 세 부분으로 요약될 수 있으며 [1], 현재 각 부분에 대해 많은 연구가 이루어지고 있다 [2-7].

인터넷과 웹에서의 가용 정보가 폭발적으로 팽창함으로써 많은 양의 정보 보다는 양질의 정보를 검색하는 것이 관건이 되었다. 하지만 현존하는 국내 · 외 검색엔진들의 검색방법으로는 양질의 정보를 효과적으로 검색하기가 어렵다. 이는 사용자가 사용하는 용어의 의미와 검색엔진이 인식하는 용어의 의미적 차이(Semantic gap) 때문이다 [8]. 용어의 의미적 차이가 발생하는 것은 인간이 사용하는 언어에 다의어와 동의어가 존재하기 때문이다. 다의어는 하나의 단어가 여러 의미를 나타내는 것이고, 동의어는 여러 단어가 하나의 의미를 나타내는 것이다. 이와 같이 정보 제공자와 검색자가 동일한 의미를 표현하는데 있어서 서로 다른 용어를 사용하고, 또한 서로 다른 의미를 표현하는데 있어 같은 용어를 사용하기 때문에 효과적인 정보 검색이 이루어지지 못한다. 게다가 대부분의 검색엔진들은 검색자가 질의어로 제시하는 검색어를 사용하여 단순한 키워드 매칭(Keyword matching)방식으로 검색을 하기 때문에, 검색자가 원하는 정보를 얻는데 상당한 시간과 노력이 요구된다. 따라서 기존 검색엔진에서의 키워드 매칭 검색방식과 달리 용어의 의미를 고려하는 새로운 개념의 검색방법이 필요하다. 새로운 개념의 검색방법중의 하나로 웹 문서로부터 추출하는 색인어(Index)에 의미를 부여하고 [9-12] 사용자의 검색의도를 파악하여 양질의 검색 결과를 제시하는 검색엔진에 대한 연구 [13-15]가 이루어지고 있다.

제한된 정보에서의 키워드 매칭에 의한 검색에서는 검색자 스스로가 정보여과(Information filter)의 역할을 수행하여 검색 결과로부터 필요한 정보를 얻을 수 있었다. 하지만 급변하는 시대의 방대한 자료 속에서는 검색자가 모든 검색 결과를 여과하는 것은 불가능하다. 이런 문제를 근본적으로 해결하기 위해서는 (i)정보 제공자가 의미 검색이 가능한 형태로 정보를 제공하고, (ii)검색엔진은 의미를 고려하여 정보를 수집해야 하며, (iii)검색자는 검색 의도를 분명히 밝혀 검색을 해야 한다. 그러기 위해서는 (i)정보 제공자에 대해 정보 전달 방법에 대한 교육이 필요하고, (ii)검색엔진은 웹 페이지를 다시 수집하여야 하며, (iii)검색자는 검색할 때마다 검색어에 대한 의미를 표현해야 하는 번거로움을 배제할 수 없다. 따라서 이미 구축된 정보베이스(Information Base)를 이

접수일자 : 2004년 3월 5일
완료일자 : 2004년 7월 26일
* Corresponding Author

용하고, 사용자는 단순한 검색어나 문장을 입력함으로써 양질의 정보를 검색할 수 있다면 검색에 소요되는 시간과 노력을 줄일 수 있다. 따라서 본 논문에서는 기존의 키워드 매칭 방식의 검색엔진에서 제공하는 정보를 그대로 사용하고, 검색자로 하여금 자신의 검색 의도에 맞는 양질의 검색 결과를 얻는데 효과적인 검색어를 제시함으로써 새로운 검색 시스템의 구축과 자료수집에 드는 비용과 수고를 줄이고자 한다.

본 논문에서는 검색자의 질의어에 사용된 검색어를 확장하고 인터넷 검색엔진으로부터의 검색 결과를 통합하여 새로운 검색 결과를 제공하는 지식기반 의미 메타 검색엔진(SMSE: Semantic Meta-Searching Engine)을 제안한다.

2. 인터넷 검색엔진

인터넷 검색엔진은 웹(Web)상에 존재하는 웹 문서(Web page)에 대한 검색을 가능하게 하는 정보검색 시스템으로서 "자료수집" 부분, "색인" 부분 그리고 "검색" 부분으로 크게 세 부분으로 나눌 수 있다. "자료수집" 부분에서는 스파이더(Spider), 크라울러(Crawler) 등으로 불리는 웹 문서 수집 프로그램이 링크(Link)정보를 바탕으로 네트워크로 연결되어 있는 전 세계의 컴퓨터에 저장되어 있는 웹 문서를 수집한다. "색인" 부분에서는 검색을 빠르게 하고 저장할 데이터의 용량을 줄이기 위해 수집한 웹 문서의 색인정보를 데이터베이스에 저장하게 된다. 그리고 "검색" 부분에서는 검색자가 원하는 정보가 입력될 때마다 데이터베이스에 저장된 색인정보를 검색하고 각각의 검색엔진에서 정한 순위 결정 시스템(Ranking system)에 따라 검색 결과에 대한 순위를 정하여 순위에 따른 검색 결과를 검색자에게 제공한다. 이러한 인터넷 검색엔진의 기본 구조는 그림 1과 같다.

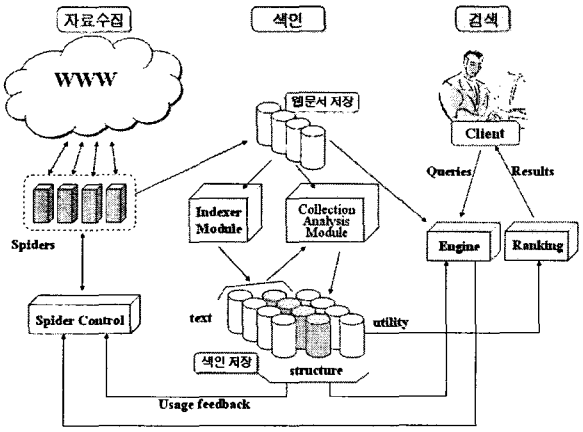

그림 1. 일반적인 검색엔진의 구조[2]
Fig. 1. The Structure of common searching engines

인터넷 검색엔진은 검색 방법에 따라 디렉토리 검색엔진, 키워드 검색엔진, 그리고 메타 검색엔진으로 구분된다. 디렉토리 검색엔진은 자료들을 주제어나 카테고리별로 구분하여 분류하고 설명 및 평가를 덧붙여 데이터베이스를 구축한 검색엔진을 말한다. 키워드 검색엔진은 웹 문서 수집 프로그램에 의해 웹 문서를 수집하고, 수집한 문서를 색인 과정을 거쳐 검색엔진의 데이터베이스에 저장해 놓고 사용자의 질의어에 대해 키워드 매칭 방식으로 원하는 정보를 검색해 준다. 메타 검색엔진은 다른 검색엔진으로부터 검색자의 질의어에 따른 검색 내용을 취합한 후 검색자에게 보여주기 때문에 검색자는 다양한 검색 결과를 얻을 수 있고, 기존의 검색엔진에서 질의어에 대한 결과를 종합하여 결과를 보여주기 때문에 내부적으로 데이터를 저장할 공간이 필요하지 않는 장점이 있다 [5, 9, 16].

2.1 PageRank 알고리즘

인터넷 검색엔진은 수십억 개의 웹 문서에 대한 검색 결과를 검색자에게 제공하기 때문에 검색 의도에 가장 잘 맞는 검색 결과를 우선적으로 제시할 필요가 있다. 따라서 검색엔진마다 각기 다른 순위 결정 방법을 사용하여 양질의 검색 결과를 검색자에게 제공하려 하고 있다. 검색 순위를 결정하는 방법중의 대표적인 것으로는 키워드의 출현 빈도(Frequency)를 이용하는 것과 웹 문서의의 인용정보(Back link)를 이용하는 방법 [3, 17]이 있다.

검색 순위를 결정하는 방법으로서 웹 문서의 인용정보를 이용하는 방법의 하나인 PageRank는 일반적인 사용자가 생각하는 어떤 페이지의 중요성과 그 페이지를 인용하는 다른 페이지를 바탕으로 한 인용 중요성의 객관적인 측정치다 [4]. 어떤 웹 문서가 얼마나 많이 인용되고 있는가를 측정함으로써 그 웹 문서의 중요성이나 품질을 추정해 볼 수 있는 것이다. PageRank는 단순히 모든 링크를 세는 것에서 한발 더 나아가 그 링크가 어떤 웹 문서로부터 왔는지를 차별화 했고, 링크를 해준 웹 문서가 링크하는 웹 문서의 개수를 이용하였다. 즉, 웹 문서 A를 가리키는 다른 페이지들이 T_1, T_2,$\cdots$, T_n와 같다고 할 때, 웹 문서 A의 PageRank $PR(A)$는 식 (1)과 같다.

$$PR(A) = q + (1-q)\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \quad (1)$$

$C(A)$는 웹 문서 A가 링크하는 다른 웹 문서의 개수이고, q는 웹 써퍼(Web surfer)가 써핑(Surfing)도중 무료함으로 인해 임의의 페이지로 넘어갈 확률을 뜻하며, $(1-q)$는 웹 써퍼가 현재 웹 문서에서 임의의 하이퍼링크(Hyper link)를 따라 다른 웹 문서로 이동할 확률을 의미한다. 따라서 PageRank는 '얼마나 유명하고 많은 웹 문서가 링크했느냐'와 '다른 웹 문서로부터 얼마나 중요하게 간주되느냐'에 따라 결정된다는 것을 알 수 있다.

3. 시스템 구성

본 논문에서는 검색자의 검색 의도를 파악하여 적절한 검색어를 생성하고, 기존의 키워드 매칭 방식 검색엔진에서 제공하는 결과를 분석하고 통합하여 검색자에게 양질의 검색 결과를 제공하는 지식기반 의미 메타 검색엔진(Knowledge-based semantic meta-searching engine)을 제안하고 이를 구축한다. 기존의 검색엔진에서 키워드 매칭 방법으로 검색을 할 경우 검색어의 의미가 반영되지 않으므로 검색자는 검색 의도에 맞는 정확한 검색 결과를 얻지 못한다. 경우에 따라서는 검색자가 적절한 검색어를 사용하여 질의어를 구성하면 보다 좋은 검색 결과를 얻을 수도 있다. 하지만 검색 초보자들이 검색 의도에 맞는 적절한 검색어를 생각해 내는 것은 쉬운 일이 아니다. 또한 적절한 검색어를 선택하여 질의어를 구성하고 검색을 하였다 하더라도 각 검색엔진에서 사용하는 순위 결정 방법에 따라 검색자의 검색 의도에 맞는 문서의 순위가

뒤로 밀려날 수 있다. 따라서 검색자의 검색 의도를 파악하고 적절한 검색어를 선택하여 질의어를 생성할 수 있게 하며, 기존 검색엔진에서 검색 결과로 제시된 문서의 순위를 파악한 후, 새로운 순위 결정 방법을 적용하여 사용자에게 검색 결과를 제공한다면 검색자가 정보 검색에 소요하는 시간과 노력을 줄일 수 있다.

본 논문에서 구축하고자 하는 지식기반 의미 메타 검색엔진을 SMSE라 칭하기로 한다. SMSE는 그림 2와 같이 크게 다섯 부분으로 구성된다. SMSE는 (i) 검색자의 검색의도를 검색엔진의 검색 방식에 맞게 질의어를 형성하는 "질의어 형성(Query formation)", (ii) 질의어를 분석하여 검색자의 검색의도를 파악하여 양질의 검색 결과를 얻기 위해 새로운 검색어를 질의어에 추가하는 "질의어 확장(Query expansion)", (iii) 검색엔진의 검색 API (Application Program Interface)를 사용하여 검색엔진으로부터 질의어에 대한 검색 결과를 가져오는 "검색(Searching)", (iv) 검색엔진으로부터 가져온 검색 결과를 분석하여 SMSE의 순위 결정 방법에 따라 순위를 재생성하는 "순위 재생성(Ranking recreation)", 그리고 (v) 질의어를 형성하고 질의어를 확장하며 순위를 재생성 하는데 있어 모든 자료와 규칙을 제공하는 "지식베이스(Knowledge Base)"로 구성된다.

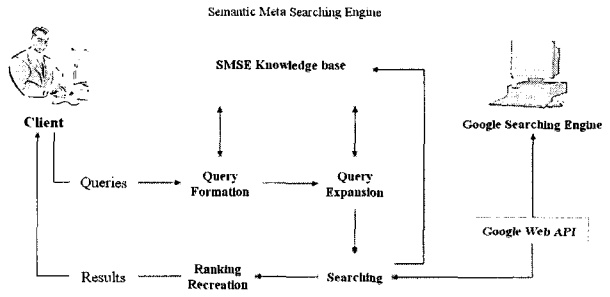

그림 2. 지식기반 의미 메타 검색엔진의 구조
Fig. 2. The structure of SMSE

일반적인 메타 검색엔진은 기존의 많은 검색엔진으로부터 검색결과를 얻는다. 하지만 본 논문에서는 검색결과를 얻는데 필요한 검색엔진을 하나로 제한함으로써 많은 검색 결과를 얻는 것 보다는 양질의 검색 결과를 얻는 것을 목적으로 한다. 따라서 현재 30억 이상의 웹 문서를 보유하고, 검색 결과에 대해 객관적인 순위를 제시하는 Google 검색엔진을 통해 웹 문서를 검색하고 검색 결과를 수정하여 검색자에게 제공한다.

3.1 지식베이스

지식베이스는 그림 3과 같이 지식을 적용함에 있어 자료를 저장하는 데이터 저장공간과 저장된 자료를 조합하여 새로운 정보를 생성할 수 있는 능력인 규칙부분으로 구성된다.

데이터 저장공간을 구성할 때는 이미 웹을 통해 공개되어 있는 Wordnet[18], Open Directory Web[19], 전자 사전[20], 불용어 사전, 원형 복원 사전을 일부 변경하여 사용하였고, 또한 SMSE 데이터 베이스와 질의어 관련 단어와 같이 본 논문에서 새로운 질의어를 구성하는데 필요한 정보는 직접 구축하여 사용하였다. 그리고 검색자의 검색 의도를 파악하여 새로운 질의어를 구성하는데 필요한 규칙을 구성할 때는 자연언어처리에서 흔히 사용되는 원형 복원 규칙, Soundex

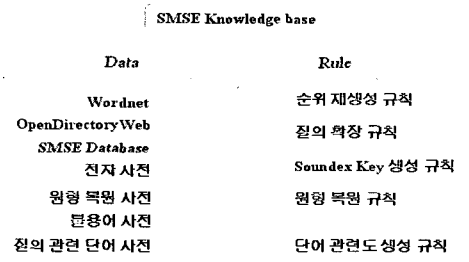

그림 3. SMSE 지식베이스 구성
Fig. 3. The component of SMSE knowledge base

Key 생성 규칙을 사용하였고, 또한 검색어 사이의 의미를파악하여 새로운 질의어를 확장하는 질의어 확장 규칙과 일반적으로 함께 사용되는 검색어들 사이의 연어(Collocation)관계를 미리 분석하여 관련 검색어를 추출하는 단어 관련도 생성 규칙을 만들어 사용하였다.

검색자가 자연언어로 질의를 할 경우, 질의어로부터 적절한 검색어를 추출하고 오타의 유무를 판단하며 유사단어를 제시하기 위해서 전자사전과 원형 복원 사전, 불용어 사전, Soundex Key 생성 규칙, 원형 복원 규칙이 사용된다. 그리고 검색자의 검색 의도를 파악하여 질의어를 확장할 때 WordNet, Open Directory Web, SMSE Database, 질의어 관련 단어 사전, 질의어 확장 규칙, 단어 관련도 생성 규칙이 사용된다. 그리고 검색결과에 대한 순위를 재생성하기 위해 순위 재생성 규칙이 사용된다.

3.2 질의어 형성

본 논문에서는 검색 초보자들이 검색 의도를 문장으로 표현하였을 경우 문장으로부터 검색어를 추출하여 검색에 이용한다. 그림 4는 문장으로 구성된 검색자의 질의어에서 검색어를 추출하는 검색어 추출 시스템의 구조를 보인다.

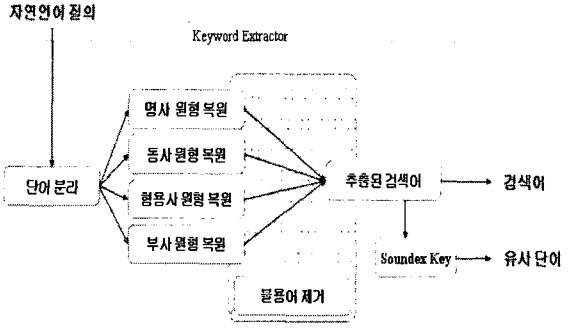

그림 4. 검색어 추출 시스템
Fig. 4. Keywords extraction system

자연언어에서 추출한 검색어로부터 의미 해석을 통해 의미에 맞는 확장 검색어를 제시하기 위해서 WordNet을 참조한다. WordNet은 단어의 관계를 계층적으로 표현한 어휘 데이터베이스로서 단어의 원형을 기준으로 구성되어 있으므로 변형된 단어를 원형으로 복원한다. 그리고 검색엔진이 수집한 웹 문서에서 색인정보를 추출할 때 색인어로 사용되지 않은 단어를 불용어라고 하는데, 불용어는 검색에 있어 큰 영향을 미치지 않으므로 본 논문에서도 검색어에서 불용어를 제거한다. 그리고 비슷한 발음이나 비슷한 철자를 가진 단어

에 동일한 코드를 붙여 두 단어 사이의 유사도를 나타내기 위한 방법으로 Soundex key를 이용하는데, 추출된 검색어의 Soundex key를 전자사전에 등록된 단어의 Soundex key와 비교하여 잘못된 검색어를 수정함으로써 올바른 검색 결과를 얻을 수 있도록 한다.

3.3 검색어 의미 해석

단어는 하나의 단어가 여러 의미를 내포하고 있는 "다의어"와 여러 단어가 하나의 의미를 나타내는 "동의어"로 구성된다. 그림 5와 같이 단어-0, 단어-1 그리고 단어-2가 나타내는 의미는 각각 여러 가지가 있을 수 있어 이 단어들은 다의어가 되며, 단어-0의 의미-0, 단어-1의 의미-0, 단어-2의 의미-0은 동일한 의미로서 이들은 동의어 관계가 된다.

WordNet에서 표현하는 의미는 전체적으로 명사 25개, 동사 15개의 최 상위 개념으로 분류되며, 그 하위 개념에 대응하는 단어들을 나열하는 방식이다. 각 단어의 상위개념을 나타내는 Synset을 노드라 할 때, 노드와 각 노드를 연결하는 아크가 단어의 의미가 되며, 두 단어 사이에 공통된 노드와 아크가 존재할 경우에는 두 단어가 의미적으로 서로 유사하다고 볼 수 있다. 따라서 공통된 상위노드가 존재하는 경우 두 단어를 공통의미의 단어로 판단하고 한 그룹으로 정한다. 그림 6에서 단어-1과 단어-2, 그리고 단어-2와 단어-3은 서로 공통된 상위노드가 존재하므로 세 단어는 {단어 1, 단어 2}, {단어 2, 단어 3}과 같이 의미적으로 두 부류로 분류된다. 의미적으로 분류된 단어들은 공통된 상위노드에 존재하는 단어를 통해 그 의미를 추측할 수 있다.

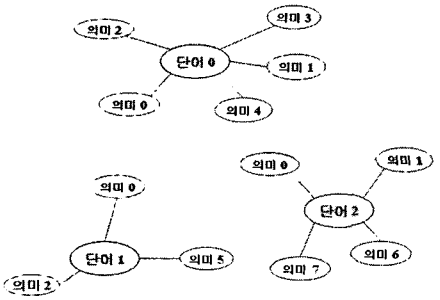

그림 5. 단어의 의미 관계
Fig. 5. Semantic relations among words

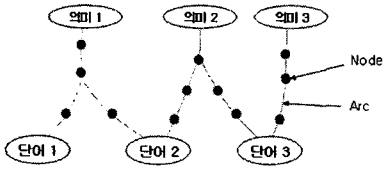

그림 6. 의미에 따른 단어 분류
Fig. 6. Word classes based on senses

ODW(Open Directory Web)에서 제공하는 카테고리도 WordNet의 의미네트워크에서의 공통된 의미를 찾는 것과 같이 공통된 카테고리를 찾음으로써 웹 상에서 사용되는 단어에 대한 의미를 찾을 수 있다. 그림 7은 "fuzzy"와 "zadeh"의 공통된 카테고리를 나타낸다.

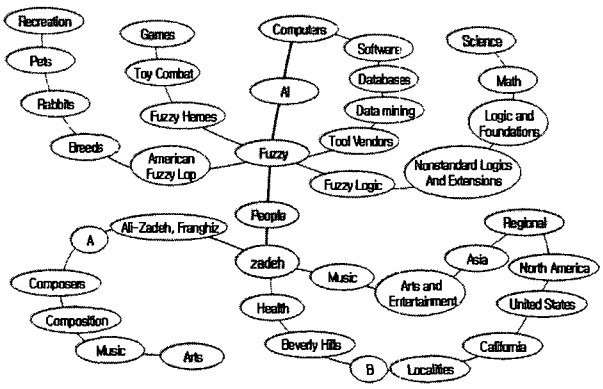

그림 7. ODW에서 "fuzzy"와 "Zadeh" 의 의미
Fig. 7. The senses of "fuzzy" and "Zadeh" in ODW

3.4 검색어 확장

SMSE는 검색자의 질의어를 구성하는 검색어의 의미 해석을 통해 사용자의 검색 의도를 파악하여 확장 검색어를 제시한다. WordNet에서의 확장 검색어는 각 단어의 의미로 표현되는 공통 상위어와, 공통 상위어를 설명하는 설명글(Metalanguage), 그리고 각 검색어의 동의어를 이용한다. Open Directory Web과 SMSE 데이터베이스에서의 확장 검색어는 각 단어가 속한 공통된 카테고리와, 각 카테고리에 대한 설명글(Anchor text)을 이용한다.

질의어를 확장함에 있어 확장된 단어 집합은 그 수가 다양하다. 확장규칙에 의해 확장되는 단어는 상위어, 동의어, 설명글에서 추출된 단어들로 구성되나 이것은 오히려 검색자들에게 혼란을 야기할 수도 있다. 따라서 지금까지 검색자들이 사용했던 질의어에서 검색어간의 연어(Collocation) 관계를 분석하여 새로운 검색자가 질의어로 사용한 특정 검색어와 가장 관련이 높은 단어를 우선적으로 제시함으로써 이런 문제를 해결하고자 한다. 예를 들면, 검색자가 검색어로 "computer"를 사용하였을 경우, "personal, apple, desktop, virus, science, network, technology, graphic, … "과 같이 지금까지 "computer"와 함께 검색어로 사용된 단어들이 그 사용 빈도순으로 검색자에게 제시된다.

3.5 검색

메타 검색엔진이 검색자에게 검색 결과를 제공하기 위해서는 기존의 검색엔진으로부터 검색 질의어에 대한 검색 결과를 요구해야 한다. 따라서 본 논문에서는 Google 검색엔진을 통해 40억개 이상의 웹 문서를 검색할 수 있는 Google Web API를 통하여 Google로부터 웹 문서의 제목, 주소, 설명글 등과 같은 검색 결과를 받는다.

3.6 순위 재생성

인터넷 검색엔진이 저장하고 있는 웹 문서에 대한 정보가 적은 경우에는 메타 검색엔진이 검색자에게 다양한 검색 결과를 제공할 수 있으나, 인터넷 검색엔진이 수십억 개의 웹 문서를 저장하고 있는 경우에는 검색된 문서의 양보다는 각 검색엔진에서 검색된 문서의 순위가 검색자에게는 더 중요할 수 있다. 따라서 본 논문에서는 기본 검색어와 확장 검색어를 차별화하여 검색에 이용하고, 검색결과의 순위를 재생성하여 검색자에게 검색 결과를 제공한다.

본 논문에서 제안하는 순위 재생성법은 다음과 같다. 기본

검색어와 확장 검색어를 동시에 검색에 사용하지 않고, 기본 검색어에 덧붙여 확장 검색어 하나를 추가한 후 검색을 시도한다. 만일 확장 검색어가 n개이면 총 $n+1$개의 검색 결과를 얻을 수 있다. 그림 8은 Google에서 질의어를 통해 검색한 웹 문서를 나타낸다. $Q=\{q_1,\cdots,q_m\}$는 기본 질의어 집합이고, $E=\{e_1,\cdots,e_n\}$는 확장 질의어 집합이며, $RD_{e_i}=\{d_1^i,\cdots,d_k^i\}$는 기본 질의어 집합 Q와 확장 질의어 e_i를 통해 검색된 웹 문서(d) 집합이다.

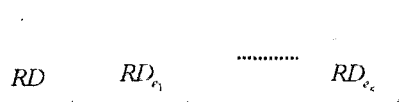

그림 8. 확장 질의어 검색 웹 문서 집합
Fig. 8. The searching results of web-document sets by using the expanded keywords

PageRank에 따른 각 검색 결과는 다음과 같이 순위가 다시 결정된다. 질의어 집합($Q\cup\{e_i\}$)을 통해 검색된 웹 문서(d)의 순위를 R_d^i라 하였을 때 생성 가능한 질의어 집합에 대한 전체 검색결과를 통합하였을 때의 문서(d)의 순위도 W_d는 식(2)로 얻어진다. 식(2)는 질의어 집합들의 검색을 통해 Google에서 제공하지 않는 불리언(Boolean) 검색기능인 OR 검색을 가능하게 하며, 웹문서가 다양한 검색어를 포함하고 있으면서도 PageRank가 높은 문서에 높은 순위도를 부여한다.

$$W_d=\sum_{i=0}^{n}\beta\left(\frac{\alpha+R_d^i}{R_d^i}\right) \qquad (2)$$

$$\alpha>0,\ \beta=\begin{cases}0, \text{if } d\notin RD_{e_i}\\ 1, \ else\end{cases}$$

예를 들어 기본 질의어 집합 Q와 확장 질의어 집합 E를 통해 표 2와 같은 검색 결과를 얻었다고 하자.

표 2. 질의어 부류별 검색 결과 예
Table 2. An example of searching result of each query class

질의 / 순위	Q	$Q+e_1$	$Q+e_2$
1	문서 A		
…			
30	문서 B		
…			
50	문서 C	문서 C	문서 B
…			
100	문서 D	문서 D	문서 D
…			

기본질의어집합 Q에 의해 검색된 결과에는 문서 A, 문서 B, 문서 C, 문서 D의 순위가 각각 1, 30, 50, 100 이다. 그리고 확장 질의어 집합 E에 의해 검색된 결과에는 문서 C, 문서 D의 순위가 각각 50, 100 이고, 문서 B, 문서 D의 순위가 각각 50, 100 이다. 검색 결과에서 보면 문서 D가 세 부류의 검색 결과에 공통으로 존재하며 문서 C와 문서 B는 두 부류의 검색결과에 공통으로 존재한다. 식 (3)~(6)은 문서 A, B, C, D 각각에 대해 $\alpha=1$ 일 때의 순위도를 계산한 것이다.

$$W_{Doc.A}=\left(\frac{1+1}{1}+0+0\right)=2 \qquad (3)$$

$$W_{Doc.B}=\left(\frac{1+30}{30}+\frac{1+50}{50}+0\right)=2.0533 \qquad (4)$$

$$W_{Doc.C}=\left(\frac{1+50}{50}+\frac{1+50}{50}+0\right)=2.04 \qquad (5)$$

$$W_{Doc.D}=\left(\frac{1+100}{100}+\frac{1+100}{100}+\frac{1+100}{100}\right)=3.03 \qquad (6)$$

계산 결과 세 부류의 결과에 공통으로 존재하는 문서 D의 순위도가 가장 높은 것을 알 수 있다. 그리고 두 부류의 결과에 공통으로 존재하는 문서 B와 문서 C 중에서 순위가 조금 높은 문서 B의 순위도가 더 높은 것을 알 수 있다. 따라서 검색 결과에서 가장 많이 출현한 결과에 우선순위를 주며, 같은 횟수로 출현했을 경우에 종합순위가 높은 결과에 우선순위를 준다.

4. 실험 및 검토

3절에서 제안된 SMSE의 성능 평가 실험을 위해 커피열매의 한 종류인 "java bean"과 과일의 한 종류인 "apple"로 만들 수 있는 것은 무엇인지를 알아보는 것을 검색 의도로 하여 다음과 같은 자연어 질의를 구성하였다.

Query: What can I make into jaba bean and apple?

검색자는 "java"의 정확한 철자를 몰라 비슷한 발음의 "jaba"를 질의에 사용하였다. SMSE는 먼저 자연언어의 질의를 각 어절별로 분리하여 원형복원을 한다. 그리고 복원된 단어로부터 검색어로서의 기능이 낮은 불용어를 제거하고 "jaba, bean, apple"을 검색어로로 추출한다. SMSE는 추출된 검색어의 Soundex key를 생성하여 전자사전과 비교한 후, "jaba"가 잘못된 단어라고 판단하고 동일한 Soundex Key를 가지는 "Joab, juba, jaap, jab, jape, java"를 유사어로서 검색자에게 제시한다. 검색자는 이 정보를 토대로 하여 "jaba"를 "java"로 수정한다.

그림 9는 검색자가 잘못된 단어가 포함된 자연어로서 질의를 하였을 경우 SMSE가 검색어를 추출하고 잘못된 단어에 대한 유사단어를 제시하는 것을 보인다.

자연언어 질의로부터 검색어를 추출하고 잘못된 검색어에 대한 수정이 이루어지면 검색어 사이의 의미 해석이 이루어진다.

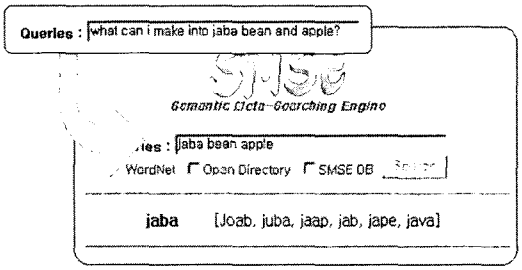

그림 9. 검색어의 추출과 수정
Fig. 9. Extraction and modification of keywords

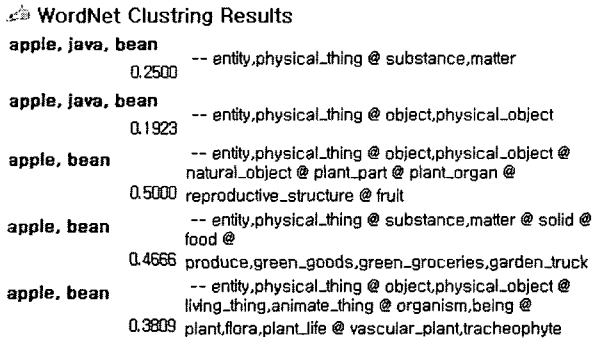

그림 10. 검색어 사이의 의미 관계

Fig. 10. Semantic relations between keywords

그림 10은 검색엔진으로부터 검색을 하기 전에 검색어를 확장하기 위해 검색어 "java, bean, apple"에 대한 의미 관계를 보여주는 화면이다.

그림 10에서 검색어 사이의 의미관계를 살펴보면 검색어 "java"와 "apple"이 의미 유사도가 0.5로서 가장 높고, 검색하려는 의도와 가장 잘 맞는 의미이다. 따라서 검색자는 두 단어의 공통된 의미로부터 검색어를 확장하기 위해 두 검색어의 세부 정보를 살펴본다. 검색자는 검색어의 세부 정보를 토대로 하여 질의어에 새롭게 추가할 확장 검색어를 선택한다. 그리고 검색자는 "bean, apple"의 의미 정보에서 두 단어를 대표하는 단어가 "fruit"이라는 것을 확인하고, 검색어 "bean"을 넓은 의미의 "fruit"으로서 대용한다. 따라서 기본 질의어였던 "java, bean, apple"에서 "bean"을 제거하고, SMSE가 제시한 의미 정보와 질의어 관련 단어로부터 "fruit"과 "cook"을 확장 질의어로 추가한다. 그리고 기본 질의어로는 "java, apple"을 사용하고 확장 질의어로는 "fruit, cook"을 사용하여 SMSE에서 검색을 하였다.

SMSE는 기본 질의어와 확장 질의어를 조합하여 "java, apple", "java, apple, fruit", "java, apple, cook", "java, apple, fruit, cook"과 같이 네 개의 검색 질의어를 구성하고, Google로부터 각 질의어마다 50개의 검색 결과를 받아 이들의 순위를 재생성 하였다. 순위를 재생성한 결과화면을 그림 11에서 볼 수 있다.

SMSE의 검색 결과에 대한 객관적인 평가를 위하여 문서수준 정확률(Document-level precision)을 계산하기로 한다. 문서수준 정확률은 문서수준 n에서의 정확률이 상위 n개의 문서들에 포함된 적합 문서들의 비율로서 정의되며 식(7)과 같이 표현할 수 있다.

$$\text{문서수준 정확률} = \frac{\text{상위 } n\text{개의 문서들에 포함된 적합문서 수}}{\text{문서수준 } n} \quad (7)$$

검색된 웹 문서가 분야별로 구분되어 있지 않으므로 검색된 결과에 대해서 평가자가 직접 문서를 분류하였다. "java"와 "apple"은 크게 "먹을 수 있는 것"과 "computer 분야"라는 공통된 의미를 내포하고 있다. 따라서 처음의 검색 의도에 따라 "java"를 커피에 관한 것, 그리고 "apple"을 과일에 관한 것으로 하고 웹 문서를 판단하기로 한다. 표 4와 그림 12는 순위 재평가 결과와 각 질의어에 대한 SMSE의 검색결과, SMSE에서 구성한 검색어 집합에 대한 Google 검색결과, 그리고 SMSE에서 사용한 모든 검색어에 대한 Google 검색 결과의 상위 50개에 대한 문서수준 정확률을 나타낸다.

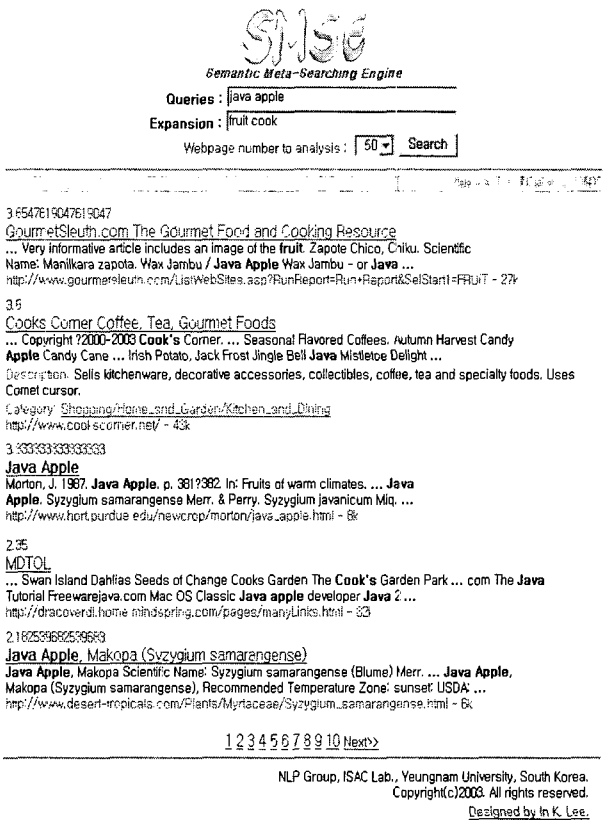

그림 11. 순위 재생성 결과

Fig. 11. The results of ranking reordering

표 4. 문서수준 정확률

Table 4. Document-level precision

검색 방법 / 순위	SMSE 검색	Google 검색			
	{java,apple} {fruit,cook}	java apple	java apple fruit	java apple cook	java apple fruit cook
10	0.7	0	0.3	0.2	0.7
20	0.6	0	0.4	0.15	0.8
30	0.5	0	0.4	0.17	0.83
40	0.5	0.025	0.475	0.2	0.85
50	0.46	0.02	0.48	0.2	0.88

그림 12의 문서수준 정확률 그래프에서 보듯이 검색어의 개수를 많이 사용할수록 정확률이 높아짐을 알 수 있다. 검색어의 수가 늘어날수록 정확률이 높아지는 것은 당연하나 상대적으로 재현율이 낮아지게 된다. 따라서 검색어를 하나라도 포함하지 않는 문서는 검색결과에서 배제되기 때문에 원하는 문서를 찾지 못할 수도 있다. 그리고 "java, apple", "java, apple, fruit", "java, apple, cook"을 검색어로 하였을 때의 검색결과를 살펴보면, Google의 PageRank 순위로는 문서수준이 높아질수록 정확률이 높아지는 것을 볼 수 있다. 이것은 다양한 의미로서 사용되는 "java"와 "apple"의 의미를 반영하지 못하는 검색 기법으로부터 기인한다고 볼 수 있다. 다시 말하면 검색의도에 맞는 웹 문서를 우선적으로 제시하지 못하고 검색어를 포함하는 문서 중에 PageRank가 높은 문서를 우선적으로 제시한다는 것이다. 하지만 SMSE에서 순위 재생성 결과를 보면, 문서수준이 10인 경우에 검

색어를 모두 사용하여 검색하였을 때와 동일한 최고 높은 정확률을 보인다. 하지만 Google의 검색 결과에서 문서수준이 높아질수록 정확률이 높아지는 것과 달리 SMSE의 검색 결과에서는 문서수준이 높아질수록 정확률이 낮아지는 것을 볼 수 있다. 이것은 전체 검색된 문서 중에서 사용자의 검색 의도에 맞는 문서를 우선적으로 제시하기 때문으로 볼 수 있다. 따라서 높은 재현율과 높은 문서수준 정확률의 결과를 얻을 수 있다고 해석할 수 있다.

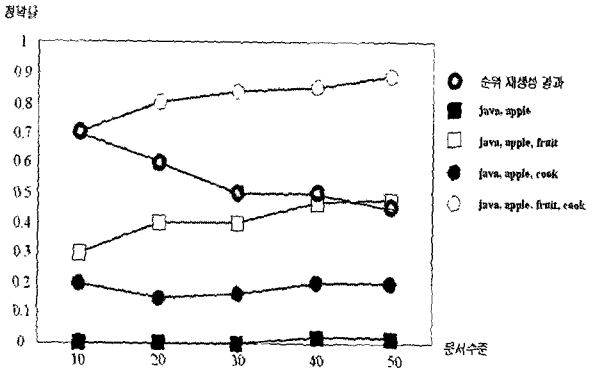

그림 12. 문서수준 정확률 그래프
Fig. 12. Document-level precision graph

이와 같이 SMSE는 Google의 PageRank에서 순위는 비록 낮으나 유용한 정보를 나타내는 웹 문서를 우선적으로 검색자에게 제공함으로써 검색자가 질의어를 수정하고 재검색하는 시간과 노력을 줄일 수 있었다. 또한 초기 질의어를 구성하는 검색어의 의미를 파악하여 질의어를 수정하고 검색어를 확장함으로써 높은 정확률의 결과를 얻을 수 있었다.

5. 결 론

본 논문에서는 검색자의 질의어를 확장하고 인터넷 검색엔진으로부터의 결과를 통합하여 새로운 검색 결과를 제공하는 지식기반 의미 메타 검색엔진의 구조를 제안하고 이를 구축하였다. SMSE (Semantic Meta-Searching Engine)는 검색자의 검색 질의어로부터 적절한 검색어를 추출하고, 편리한 인터페이스(Interface) 환경에서 검색어의 의미를 확장하여 새로운 질의어를 구성한다. 그리고 새롭게 구성한 검색 질의어를 사용하여 현존하는 인터넷 검색엔진 중에서 가장 많은 웹 문서를 보유하고 있고 객관적인 검색 순위를 제시하는 검색엔진인 Google을 통해 검색을 한다. 또한 검색된 결과를 통합하여 기존의 검색엔진보다 더 나은 검색 결과를 제공한다. 따라서 SMSE는 Google의 PageRank에서의 순위는 비록 낮으나 유용한 정보를 나타내는 웹 문서를 우선적으로 검색자에게 제공함으로써 검색자가 질의어를 수정하고 재검색 하는 시간과 노력을 줄일 수 있었다. 또한 초기 질의어를 구성하는 검색어의 의미를 파악하여 질의어를 수정하고 검색어를 확장함으로써 높은 정확률의 결과를 얻을 수 있었다.

그리고 본 논문에서 검색어 단위의 의미 해석에서 한발 더 나아가 문장 단위의 의미 해석을 통한 검색어 확장이 필요하며 이에 대한 연구는 계속 진행중이다.

참 고 문 헌

[1] M. Nikravesh, V. Loia, B. Azvine, "Fuzzy logic and the Internet (FLINT): Internet, World Wide Web, and search engines," Soft Computing - A Fusion of Foundations, Methodologies and Applications, Vol. 6, No.5, pp. 287-299, Aug. 2002.

[2] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke and A. Raghavan, "Searching the web," ACM Trans. Internet Technology, Vol.1, No.1, pp. 2-43, Aug. 2001.

[3] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," In Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, pp. 107-117, 1998.

[4] L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Proceedings of ASIS'98, Annual Meeting of the American Society for Information Science, 1998.

[5] E. Selberg and O. Etzioni, "The MetaCrawler architecture for resource aggregation on the Web," IEEE Expert, Vol.12, No.1, pp. 8-14, 1997.

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

[7] http://www.yahoo.com

[8] 박기선, 이덕남, 김우주, 이용석, "자연어 질의 문맥 구조 기반 개인형 메타 검색 에이전트," 한국정보과학회 가을 학술발표논문집, vol.29, No.2, pp. 688-690, 2002.

[9] S. Lawrence and C. L. Giles, "Inquirus, the NECI meta search engine," 7th International World Wide Web Conference, pp. 95-105, 1998.

[10] U. Hahn, "Making understanders out of parsers: semantically driven parsing as a key concept for realistic text understanding applications," International Journal of Intelligent Systems, Vol.4, No.3, pp. 345-393, 1989.

[11] D. D. Lewis and K. Sparck Jones, "Natural language processing for information retrieval," Communications of the ACM, Vol.39, No.1, pp. 92-101, 1996.

[12] J. Morris and G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," Computational Linguistics, Vol.17, No.1, pp. 21-43, 1991.

[13] 박현규, 오종훈, 김명호, 최기선, 이광형, "퍼지 추론에 의한 자연언어 정보 검색," 정보처리학회 논문지 B, 제8-B권, pp. 243-250, 2001.

[14] 강현규, "개념 검색어 확장을 통해 질의 형식화를 도와주는 '개념마법사'의 설계 및 구현," 정보처리학회 논문지 B, 제9-B권, pp. 437-444, 2002.

[15] 김형일, 김준태, "협동적 순위 평가와 워드넷을 이용한 검색 질의어의 모호성 해결," 한국정보과학회 가을 학술발표논문집, Vol.28, No.2, pp. 103-105, 2001.

[16] 박상위, 오정석, 이상호, "메타 검색엔진을 위한 패

이지 변경 탐지기 설계," 한국정보과학회 봄 학술발표논문집, Vol.28, No.1, pp. 205-207, 2001.
[17] J. Kleinberg, "Authoritative sources in a hyper-linked environment," In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, Vol.46, No.5, pp. 604-632, 1998.
[18] http://www.cogsci.princeton.edu/~wn/
[19] http://dmoz.org
[20] http://www.dictionary.com

저 자 소 개

이인근(In. K. Lee)
2001년 : 영남대학교 재료금속공학부 (공학사)
2004년 : 영남대학교 대학원 전기공학과 (공학석사)
현재 : (주)개선문 근무
관심분야 : 지능시스템 및 제어, 자연언어처리, 정보검색

손세호(Seo. H. Son)
2000년 : 영남대학교 전기전자공학부 (공학사)
2002년 : 영남대학교 대학원 전기공학과 (공학석사)
2004년 : 영남대학교 대학원 전기공학과 박사과정수료
현재 : 탑엔지니어링(주) 근무

관심분야 : 지능시스템 및 제어, 영상이해, 비젼시스템

권순학(Soon. H. Kwon)
1983년 : 서울대학교 제어계측공학과 (공학사)
1985년 : 서울대학교 대학원 제어계측공학과 (공학석사)
1995년 : 동경공업대학 시스템과학 (공학박사)
1996~ : 영남대학교 전자정보공학부 부교수

관심분야 : 지식 기반 지능시스템