어휘사전 워드넷을 활용한 의미기반 웹 정보필터링

변 영 태[†] · 황 상 규[††] · 오 경 묵[†††]

요 약

새로운 정보검색 환경인 인터넷에서의 정보필터링은 문헌정보 필터링뿐 아니라 기존의 뉴스그룹이나 전자메일을 대상으로 한 정보필터링과도 판이하게 다르다. 따라서 기존의 정보필터링 모형은 새로운 환경하에서는 정상적인 성능을 기대할 수 없다. 이러한 문제점을 해결하기 위해서 달라진 필터링 환경에 대해 조사하였으며, 이를 토대로 어휘사전 워드넷을 이용한 새로운 방식의 의미기반 정보필터링 모형을 제안한다. 의미기반 정보필터링 모형에서는 일반적인 필터링모델에서 이용하는 TF/IDF방식 대신에 새로운 방법인 SDCC알고리즘을 통해 웹 문서로부터 키워드를 추출하게되며, 이를 통해 인터넷 검색효율저하의 주요원인인 어휘의미중의성 발생을 예방하게 된다. 의미기반 정보필터링에서는 사용자 수준에 따른 선택적 필터링(selective filtering)이 가능하며, 이를 통해 사용자에게 보다 편리한 인터넷정보검색 환경을 제공하게 된다.

Semantic-Based Web Information Filtering Using WordNet

Young-Tae Byun[†] · Sang-Kyu Hwang[††] · Kyung-Mook Oh[†††]

Information filtering for internet search, in which new information retrieval environment is given, is different from traditional methods such as bibliography information filtering, news-group and E-mail filtering. Therefore, we cannot expect high performance from the traditional information filtering models when they are applied to the new environment. To solve this problem, we inspect the characteristics of the new filtering environment, and propose a semantic-based filtering model which includes a new filtering method using WordNet. For extracting keywords from documents, this model uses the SDCC(Semantic Distance for Common Category) algorithm instead of the TF/IDF method usually used by traditional methods. The word sense ambiguation problem, which is one of causes dropping efficiency of internet search, is solved by this method. The semantic-based filtering model can filter web pages selectively with considering a user level and we show in this paper that it is more convenient for users to search information in internet by the proposed method than by traditional filtering methods.

1. 서 론

인터넷은 수많은 정보들이 모인 정보의 바다이며, 지금 이 순간에도 수많은 새로운 종류의 정보가 끊임없이 생겨나고 소멸되고 있다. 인터넷에서 정보의 양은 계속 증가 추세에 있으며, 아무리 뛰어난 인터넷 정보검색시스템이라도 인터넷 전체 정보량의 대략 16% 정도밖에 찾아낼 수 없을 정도로 인터넷은 그 규모 면에서 이루 말할 수 없을 정도로 방대하다. 하지만 인터넷이라는 거대한 바다에서 정보를 찾고자 하는 대부분 사용자들은 인터넷과 정보검색의 도구인 인터넷 정보검색시스템에 대해 잘 알고 있지 못하다. 따라서 대부분의 사용자들은 자신이 필요로 하는 정보를 찾는데 어려움을 겪고 있다. 그들이 겪는 어려움의 대

※ 본 논문은 일부 STEPI의 소프트과학 기술 개발 사업 지원을 받아 연구되었음.
† 종신회원 : 홍익대학교 전자계산학과 교수
†† 준 회 원 : 홍익대학교 대학원 전자계산학과
††† 정 회 원 : 숙명여자대학교 정보과학부 교수
논문접수 : 1999년 10월 15일, 심사완료 : 1999년 11월 12일

부분은 찾고자하는 대상과 관련된 웹 문서를 인터넷 정보검색시스템을 통해 찾기 불가능하기 때문이기보다는 관련된 문서가 너무 많이 검색되기 때문에 원하는 정보를 찾기까지 너무 많은 시간이 소모된다는 점이다. 이러한 현상은 비단 인터넷만의 문제가 아니라 최근 대부분의 학술, 기업정보자료들이 전산화되어지고 있는 시점에서 점점 더 정보의 양이 폭발적으로 증가하고 있는 문헌데이터베이스검색에서도 동일한 현상이 벌어지고 있다. 이러한 관점에서 Belkin과 Croft는 정보의 필터링 과정이 정보검색시스템에서 중요한 부분을 차지할 것임을 예측하였다[1]. 문헌데이터베이스검색에 비하여 검색 정확률이 떨어지는 인터넷 정보검색의 특성을 고려해볼 때 정보필터링은 인터넷 정보검색에 있어 보다 필수적인 도구일 수 있으며, 앞으로 보다 다양한 활용적 가치가 기대된다.

인터넷 정보검색을 위한 정보필터링을 수행하기 위해서 먼저 달라진 필터링 환경과 그에 따른 문제점들에 대해 조사하였으며, 이를 토대로 새로운 정보필터링 방법인 의미기반 정보필터링 모형을 제시한다. 달라진 필터링 환경과 그에 따른 문제점들에 대해서는 2장에서 그 내용을 소개하고, 어휘 사전 워드넷을 활용한 새로운 정보필터링 방법인 의미기반 정보필터링 방법과 성능평가에 관해서 3장과 4장을 통해 다음과 같이 설명한다.

2. 관련연구

2.1 정보검색과 정보필터링

사용자 요구에 부합되는 웹 문서를 찾아낸다는 관점에서는 정보검색이나 정보필터링 모두 동일한 수행목적을 가지게 된다. 또한 실제 구현상에 있어서도 정보검색이나 정보필터링 모두 가장 널리 쓰이는 방법은 TF/IDF기법이며, 처리과정에 있어서도 상당부분 유사하다. 하지만 정보필터링은 그 목적상 정보검색과는 구별되어져야 하며, 몇 가지 면에서 차이점을 가지고 있다. Oard는 일반적인 정보검색과 정보필터링을 정보탐색의 측면에서 비교, 분석[2]하였으며 이를 기초로 인터넷 정보검색과 정보필터링, 일반적인 정보검색 등을 <표 1>를 통해 비교, 정리해 보았다.

먼저 인터넷 정보검색과 데이터베이스검색을 구분 지을 필요가 있는데, 인터넷 정보검색에서 대상이 되는 웹 문서들은 비정형화 되어 있는데 비하여, 데이터베이스검색에서는 대상이 되는 문서들이 정형화되어 있다는 특징을 가진다. 데이터베이스검색에는 보통 전문가의 손을 거쳐 저자, 서명, 출판사, 요약 정보 등이 정형화된 메타데이터(meta data)의 형태로 정리되어지며, 메타데이터를 이용한 보다 다양하고 정확한 검색이 이루어질 수 있다. 하지만 최근 문헌데이터베이스검색에 있어 전문(full-text) 검색이 차지하는 비중이 점점 높아져 가는 현 시점에서는 데이터베이스 검색 시에도 인터넷 정보검색과 동일한 문제점들이 대두되어지고 있다.

정보필터링이 정보검색과 구분되어질 수 있는 가장 큰 특징은 "특정 분야에 대한 장기간에 걸친 사용자의 관심사항"을 필터링 작업 수행의 근거로 한다는 점이다. 그에 반하여 정보검색에 있어서 "사용자의 관심사항은 거의 매번 질의마다 검색 목적이 달라진다"는 특징을 가지고 있다. 따라서 정보필터링 과정은 계속해서 생겨나는 웹 문서나 뉴스그룹, 전자메일 문서와 같은 정보원(information source)을 대상으로 하여 사용자의 관심사항에 부합되는 정보들을 골라내는 작업에 해당하며, 정보검색의 과정은 수집한 문서들을 대상으로 매번 달라지는 사용자 관심사항에 부합되는 웹 문서나 뉴스를 찾는 과정에 해당한다.

인터넷 정보검색 환경은 새로운 정보원이 계속해서 생성, 소멸되는 동적인 환경이며, 보다 최신의 정보를 사용자에게 제공하기 위해서는 정적인 정보수집방법(passive collection)보다는 능동적으로 새로운 정보를 찾는 동적인 정보수집방법(active collection)이 보다 효과적이다. 따라서 정보필터링에 관한 연구도 과거에는 정

〈표 1〉 Information seeking process

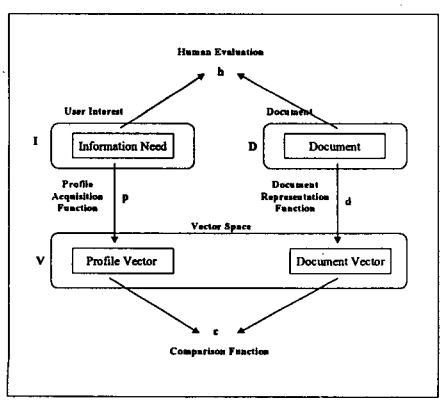
	정보필터링	인터넷 정보검색	정보검색	DB검색	브라우징
사용자 정보요구	Stable and Specific	Dynamic and Specific	Dynamic and Specific	Dynamic and Specific	Broad
정보원(Information Resource)	Dynamic and Unstructured	Dynamic and Unstructured	Stable and Unstructured	Stable and Structured	Unspecified
사용자 정보요구 변화	Slow	Fast	Fast		
정보원의 변화	Fast	Fast	Slow		

적인 정보수집방법 측면에서 정보필터링에 관한 연구[3]가 진행되어왔으나, 최근에는 동적인 정보수집방법 측면에서 정보필터링[4]으로 연구 방향이 전환되어지고 있다.

2.2 정보필터링 방법

정보필터링은 구현상 정보검색과 유사하며, 불리안모델, 확률모델, 벡터공간모델과 같은 정보검색 모델은 정보필터링을 위해서 이용되어질 수 있다[1]. Yan의 연구[5]에서는 불리안모델을 이용한 정보필터링의 자료구조 및 알고리즘을 소개하고 있으며, Callan은 확률모델 중 하나인 추론네트워크(inference network)를 이용하여 정보필터링시스템을 구현하였다[6]. 그밖에 벡터공간모델을 이용한 연구[7, 8]들이 있으며, 여러 정보검색 모델 중 벡터공간모델이 대체적으로 가장 나은 성능을 보이고 있다. 따라서 본 논문에서는 여러 문서필터링모델 가운데 대체적으로 가장 나은 성능을 보이고 있는 벡터공간모델에 대해 살펴보고자 한다. (그림 1)은 정보필터링 방법중 하나인 "content-based text selection techniques"를 기본 바탕으로 하여 일반적인 벡터공간모델의 정보필터링 과정을 모형화 하였다. "content-based text selection techniques"방식의 정보필터링 방법은 Belkin과 Croft의 연구[1]에서 대규모로 평가되어진바 있으며, 기본적으로 다음 4가지 요소로 구성되어 있다.

- Some technique for representing the documents
- Some technique for representing the information need
- Some way of comparing the profiles with the document representations
- Some way of using the results of that comparison



(그림 1) Filtering System Using the Vector Space Model

벡터공간모델에 가장 큰 특징은 사용자의 관심사항과 대상문서를 벡터화하는 과정을 거쳐 생성된 프로화일 벡터(profile vector)와 문서 벡터(document vector) 간의 유사도 계산을 통해 부적절한 문서를 제거하는 필터링 작업을 수행하게 된다. 먼저 대상문서에서 불용어를 제거하는 과정과 스테밍(stemming)과정의 전처리 단계를 거쳐 남아있는 단어가 문서 벡터 V_D를 구성하는 term이 된다. 각각의 term의 가중치는 보통 해당 문서에서 나타나는 term의 빈도수(TF)와 전체문서에서 나타나는 역문헌빈도수(IDF)를 고려하여 결정된다. 사용자의 관심사항으로부터 프로화일 벡터 V_P를 얻어내는 방법에는 사용자의 행동관찰을 통해 관심사항을 파악해내는 "implicit model"과 사용자가 직접 자신의 관심사항을 표명하는 "explicit model"이 있다[13]. 두 방법 가운데 "explicit model"의 방법이 보다 효과적이며 구현이 용이한 관계로 대부분의 시스템에서는 "explicit feedback"방식을 채택하고 있다. 프로화일 벡터와 문서 벡터가 생성되었을 때 두 벡터간의 유사도 계산방법은 다음과 같다.

$$V_D = ((d_1, w_1), (d_2, w_2), \ldots, (d_i, w_i), \ldots, (d_u, w_u))$$

$$V_P = ((p_1, u_1), (p_2, u_2), \ldots, (p_j, u_j), \ldots, (p_z, u_z))$$

$$\mathrm{sim}(V_D, V_P) = \frac{V_D \cdot V_P}{|V_D| \cdot |V_P|}$$

d_i : document term, w_i : weight of document term d_i

p_j : profile term, u_j : weight of profile term p_j

벡터공간모델을 이용한 필터링 시스템은 사용자가 자신의 관심사항과 각 문서들을 비교하여 불필요한 정보를 제거하는 인간의 정보취사선택 과정을 사용자의 관심사항과 대상문서를 벡터화하는 과정을 거쳐 생성된 프로화일 벡터와 문서 벡터간의 유사도 계산을 통해 부적절한 문서를 제거하는 과정으로 대신하게 된다.

2.3 사용자모형

대부분의 정보필터링 방법에서는 사용자의 관심사항을 조사하여 이를 토대로 각 개인별 사용자 모형(user model)을 구축한다. 구축된 사용자 모형은 대상 문서를 필터링하는 기준이 되며, 유저피드백을 통해 사용자모형을 갱신해나가게 된다. 하지만 대부분의 경우 사용자들은 자신의 관심사항을 표현하기 어려워하며, 사용자로부터 정보를 얻는 과정은 사용자에게 상당한

부담(information overload)으로 작용하게 된다[9]. 사용자의 관심사항은 개개인에 따라 각기 다르며, 일정시간이 지나면 관심대상은 변하게 된다. 사용자의 관심사항은 정보검색의 목적, 정보의 종류, 정보의 복잡도와 질과 같은 특성에 따라 달라지며 이러한 정보는 단순히 사용자가 입력한 키워드의 패턴만으로는 충분히 파악될 수 없다. 따라서 대부분의 사용자 모델에서는 관심대상과 연관성 있는 어휘들을 사용자가 직접 입력하는 방식을 채택하고 있다. 하지만 이러한 방법으로 사용자 모델을 구축하였다 하더라도 실제 필터링 작업을 수행하는 데에는 여전히 문제점이 남아 있다. 먼저 사용자가 입력한 연관성 있는 어휘들이 실제 웹 문서상에서 여러 가지 다양한 의미로 사용됨에 따라 필터링의 정확률을 떨어뜨리는 어휘의미중의성 문제를 유발할 수 있다. 이러한 현상은 정보검색과정에서도 동일하게 발생할 수 있으며, 어휘의미 중의성이 인터넷 정보검색 효율에 미치는 영향에 관한 연구[10]에서는 어휘의미중의성의 발생빈도가 과거의 정보검색 환경에 비해 인터넷 정보검색 환경에서 훨씬 더 높음을 확인한 바 있다. 또 다른 문제점으로 사용자가 자신의 관심사항을 표현하는 과정에서 사용한 어휘와 실제 웹 문서상에서 사용되어지는 어휘간의 불일치 역시 필터링의 효율을 떨어뜨리게 된다. 따라서 사용자로부터 얻은 연관성 정보는 그 자체로 유용한 정보임에도 불구하고, 정보 획득의 어려움과 운영상에 어려움으로 인하여 실제 시스템에서는 고의로 사용되어지지 않고 있다[11]. 인터넷 사용자의 검색 행동 성향에 관한 연구[12]에서도 유저 피드백 정보를 이용한 학습이나 이를 통한 검색식 재조정은 별로 실용적이 못하다고 결론짓고 있다. 사용자로부터 추가적인 정보를 획득할 수 있는 횟수가 극도로 제한되는 현실에서 인터넷 정보검색시스템은 제한된 정보를 효과적으로 활용하여 작업을 수행해나가야 한다. 이러한 상황은 정보필터링 수행 시에도 동일하게 적용되어질 수 있으며, 본 연구에는 사용자에게 추가적인 유저 피드백정보를 요구하지 않은 상황하에서 제한된 정보를 효과적으로 활용할 수 있는 새로운 정보필터링 방법을 연구하였다.

2.4. 인터넷 정보검색을 위한 정보필터링

현재까지 인터넷 환경 하에서 대부분의 정보필터링에 관한 연구는 뉴스그룹위주로 진행되어져 왔으며, 인터넷 정보검색을 위한 정보필터링에 관해서는 아직까지 충분한 연구가 이루어져 있지 못한 상황이다. 또한 뉴스그룹을 위한 정보필터링과 인터넷 정보검색을 위한 정보필터링의 차이점이 명확히 구분되지 않은 상황에서 연구가 진행되어져 왔다. 인터넷 정보검색을 위한 정보필터링은 몇 가지 고유한 특징을 지니고 있으며, 이러한 특징 때문에 뉴스그룹을 위한 정보필터링모델은 인터넷 정보검색에서는 원활히 수행되어질 수 없다. 본 연구에서는 벡터공간모델을 이용한 정보필터링모델을 중심으로 하여 인터넷 정보검색을 위한 정보필터링의 고유한 특징을 살펴보고, 발생하는 문제점과 이를 해결하기 위한 방안을 살펴보기로 한다.

먼저 인터넷 정보검색을 위한 정보필터링의 특징으로 사용자모델 구축에 어려움을 들 수 있다. 기존의 뉴스그룹을 위한 정보필터링시스템에서는 대부분 사용자의 관심사항을 조사하여 각 사용자별로 사용자모델을 구축하게 된다. 벡터공간모델을 이용한 정보필터링에서도 사용자의 관심사항으로부터 프로화일 정보를 얻어내는 가장 보편적인 방법으로 사용자가 직접 자신의 관심사항을 표명(explicit feedback)하고 이를 바탕으로 사용자모델(user model)을 구축하는 것이 일반적이다. 하지만 대부분의 인터넷 사용자가 초보자인 현 상황에서 자신의 관심사항을 구체적으로 표현하기 힘들어하는 관계로 사용자로부터 획득한 정보로부터 사용자모델을 구축하기까지에는 많은 어려움이 따른다. 또한 사용자들의 상당수가 장기간 계속해서 이용하는 기존의 뉴스그룹과는 달리 짧은 이용기간을 가지는 인터넷 정보검색시스템에서 사용자의 행동관찰을 통해 관심사항을 파악해내는 방식(implicit model)은 충분한 효과를 기대하기가 어려운 편이다. 따라서 인터넷 정보검색을 위한 정보필터링에서 사용자의 관심사항으로부터 얻어낼 수 있는 프로화일 정보는 사실상 사용자가 입력한 질의(query)만으로 한정되며 질의를 구성하는 키워드들이 직접 프로화일 벡터를 구성하는 용어가 된다. 이 경우 벡터공간모델을 이용한 필터링 작업이 원활히 이루어지기 위해서는 매번 사용자가 다수개 이상의 키워드를 입력해줄 필요가 있다. 하지만 대부분의 사용자가 한 개 내지 두 개의 키워드를 통해 검색을 시도하는 실제 상황하에서 키워드들만을 가지고 유사도계산을 통해 부적합 문서를 판별시 그 결과에 대해 신뢰하기 힘들어지며, 결국 만족할 만한 필터링 효과를 얻을 수 없게 된다.

그 다음으로 어휘의미중의성[10]에 의한 정보필터링

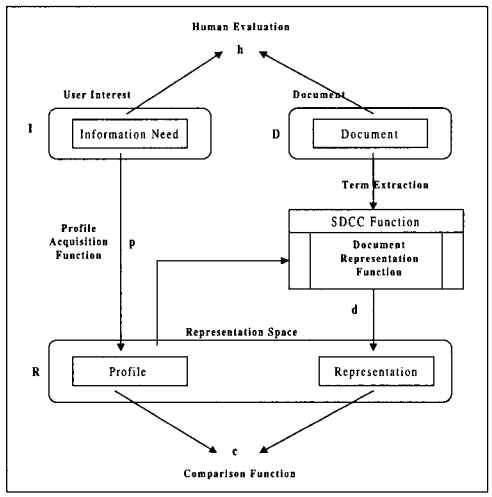
효율 저하를 생각해볼 수 있다. 인터넷은 특정 주제를 대상으로 한 뉴스그룹과는 달리 거의 모든 주제를 대상으로 하며 항목별로 정리되어 있지 않다. 이것은 사용자의 정보 요구에 부합되는 적합한 정보를 찾기가 매우 어려워짐을 의미한다. 실제로 어휘의미중의성문제는 과일인 'apple'과 관련된 문서만을 찾고자 단일질의 'apple'을 입력하였을 때 검색된 문서의 상당 부분이 'apple computer'와 같은 전혀 다른 내용을 담고 있는 상황을 유발하게 된다. 이러한 현상은 검색대상이 일상생활에서 자주 쓰이는 일반단어일 경우 보다 빈번하게 발생되어지며, 검색된 문서상에서 단어의 발생빈도수(TF)를 기반으로 하여 문서벡터의 가중치를 계산하는 벡터공간모델의 경우에는 필터링성능의 저하를 가져오게 된다.

3. 의미기반 정보필터링

기존의 벡터공간 정보필터링 모델이 가지는 문제점을 개선하기 위하여 인터넷 정보검색을 위한 새로운 방식의 의미기반 정보필터링 모형을 설계하였다. 인터넷 정보검색에서는 사용자 관심사항으로부터 사용자 정보를 추출하여 프로화일 벡터를 생성하는 과정에서 사용자 모델을 이용할 수 없다. 또한 단어의 발생빈도수(TF)를 기반으로 하여 문서 벡터의 가중치를 계산하는 방법 역시 어휘의미중의성에 의한 정보필터링 효율 저하로 충분한 성능을 기대할 수 없다. 이러한 제약조건하에서 효과적인 필터링을 수행하기 위해서는 기존과는 다른 새로운 문제해결 접근 방식을 필요로 하였으며, 이러한 내용을 바탕으로 하여 새로 구성된 정보필터링 모형은 (그림 2)와 같다.

정보필터링 모형의 전체 흐름은 사용자 관심사항으로부터 프로화일 벡터를 생성하는 과정과 각각의 웹 문서로부터 키워드를 추출하여 문서벡터를 생성하는 과정으로 나누어 살펴볼 수 있다.

먼저 사용자 정보를 추출하여 프로화일 벡터를 생성하는 경우, profile term들은 사용자 모델이 아닌 원 질의로부터 추출되어 진다. 이미 앞에서 언급한 바와 마찬가지로 현실적으로는 사용자 모델을 구축하기가 상당히 어려우며, 사용자 모델을 필터링에 활용하는 데에는 여러 가지 제약이 따른다. 또한 사용자의 대부분이 다양한 검색연산자를 사용하지 않고 한 개 내지 두 개의 검색어로 구성된 단순한 질의를 통해 검색이 이루어지는 상황에서 사용자가 입력한 질의를 그대로 프로화일 벡터로 이용하였으며, 가중치 값은 상수값(부록. 1의 수식[1])으로 설정하였다.



(그림 2) A Semantic-Based Filtering System

웹 문서로부터 키워드를 추출하는 과정은 일반적인 정보필터링모델에서 이용하는 TF/IDF방식 대신에 새로운 방법인 SDCC(Semantic Distance for Common Category)알고리즘을 통해 이루어지게 된다. SDCC알고리즘은 추출된 키워드들과 profile term들간의 연관성을 계산하여 최종적으로 적합한 키워드들만을 document term으로 결정하게 된다. 또한 profile term과 연관성을 계산하는 과정에서 각 document term의 가중치 값이 계산(부록 1)의 수식(2))되어 진다. 프로화일 벡터와 대상 웹 문서로부터 문서 벡터가 생성되었을 때 유사도 계산방법은 기존의 벡터공간모델과 동일하며 (부록 1)의 수식(3)과 같다. 의미기반 정보필터링의 또 다른 특징은 웹 문서 전체(full-text)가 아닌 일반 검색엔진에서 제공하는 요약정보(abstract information)만으로 필터링 작업이 이루어진다는 점이다.

인터넷 이용자성향을 조사한 선행연구결과[12]에서 사용자 평가성향 중 가장 두드러진 특징의 하나는 검색시스템으로 넘겨받은 타이틀(title)과 명세(description)정보로부터 평가자가 표시한 만족/불만족의 평가값이 실제 웹 문서를 방문 후 정확한 평가가 이루어진 결과와 큰 차이를 보이지 않는다는 점이다. 다시 말해 사용자가 검색시스템으로부터 넘겨받은 타이틀과 명세

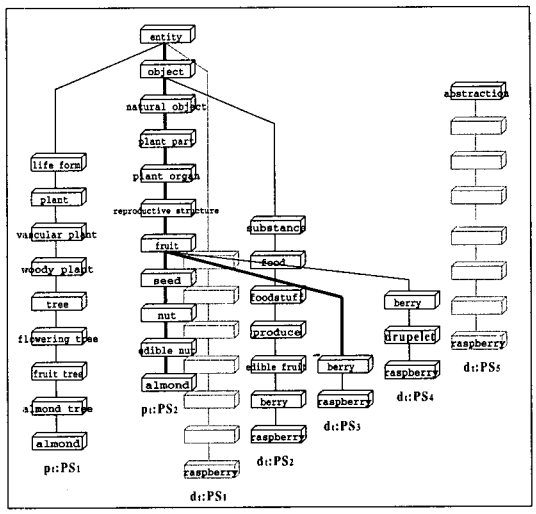
정보, URL만을 포함한 검색결과를 보고 별 가치 없으리라 여긴 웹 문서는 실제 그 문서를 방문한 후에도 가치 없다고 판단할 확률이 극히 높다는 점이다. 이는 정보필터링에도 동일하게 적용되어질 수 있으며, 필터링 평가 대상의 전체 크기가 줄어들게 되므로 응답 속도를 향상시킬 수 있다는 이점을 지닌다.

인터넷 정보검색시스템에서 정보필터링모듈은 정보검색과정과 별개이기보다는 실제 정보검색 시 사용자의 편의를 도모하고 검색 정확률을 높이는 보조적인 역할을 수행하게 된다. 대부분의 인터넷 정보검색이 실시간으로 이루어진다는 점에서 높은 필터링 정확률보다는 일정수준의 필터링효과와 빠른 응답속도가 보다 효과적일 수 있다. 의미기반 필터링 시스템은 사용자가 자신의 관심사항과 각 문서들을 비교하여 불필요한 정보를 제거하는 인간의 정보취사선택과정을 사용자가 입력한 원 질의와 SDCC알고리즘을 통해 웹 문서로부터 추출된 키워드들 간의 유사도 계산을 통해 부적절한 문서를 제거하는 과정으로 대신하게 된다.

3.1 워드넷

웹 문서로부터 추출된 document term은 profile term과 연관성 계산을 통하여 각각의 term의 가중치를 계산하며 최종적으로 적합한 키워드들만을 찾아내게 된다. 하지만 profile term과 연관성의 정도를 계산하기 위해서는 document term과 profile term간의 관련성을 알 수 있는 연관키워드 정보가 사전에 구축되어 있어야 한다. 하지만 광범위한 범위에 걸쳐 다양한 도메인에 대한 연관키워드 정보구축에는 상당한 노력을 필요로 하며, 그 과정 자체가 별도의 방대한 연구과정에 해당하므로 본 연구에서는 연관키워드 정보를 얻기 위한 별도의 지식베이스를 구축하는 대신 기존의 어휘사전의 일종인 워드넷(WordNet)으로부터 연관키워드 정보를 유도해낼 수 있는 새로운 SDCC알고리즘을 개발하였다.

Ontology의 일종으로 간주되고 있는 워드넷(WordNet)은 인간의 어휘지식에 대한 심리언어학 연구의 성과를 토대로 1985년부터 프린스톤대학 인지과학연구실이 구축해온 Lexical Database이다[14]. Ontology란 인공지능 분야에서 쓰이는 지식베이스의 일종으로, 인공지능 에이전트가 등장함에 따라서 Ontology는 지식표현의 호환성을 위한 도구로 널리 활용되어져 왔다. 현재 워드넷(WordNet)은 정보검색시스템, 자연언어처리와 정보검색 등의 여러 분야에서 널리 이용되고 있으며 영어판 이외에도 스페인어, 이탈리아어로의 다국어판의 구현도 시도되고 있다[15].



(그림 3) 워드넷 계층구조 예제

워드넷은 일상생활에서 많이 쓰이는 단어 약 12만개를 대상으로 하여 계층정보와 시소러스적 정보, 의미사전정보를 제공하는 일종의 어휘사전이다. 워드넷은 전체 12만개의 단어를 총 45개에 범주[16]로 분류하고 있으며, 본 연구에서는 명사에 해당하는 식물명(noun.plant)과 식물과 관련 있는 음식물명(noun.food)을 성능평가에 이용하였다. 평가단계 초기에는 순수 식물명으로 대상범위를 한정하였으나, 웹 문서에 특성상 식물에 관한 단일주제만을 담고 있는 경우가 드물었으며, 음식, 기업과 같은 여러 주제와 혼합되어 있는 경우가 대부분이었다. 필터링 성능평가를 위해 실험에 이용한 31개의 일반 식물명(부록 2)의 경우 검색목적과 부합되는 문서의 상당수가 식물(plant)과 음식(food) 두 가지 주제와 동시에 관련 있었다. 따라서 필터링 성능평가 시에도 검색된 문서가 식물과 관련 있는 음식을 주제로 한 웹 문서의 경우에도 적합한 문서로 판정하였다.

(그림 3)는 워드넷을 통해 얻을 수 있는 'almond'와 'raspberry'의 계층연관정보(hierarchy relation information)를 보여주고 있다. 워드넷은 일반적인 시소러스와는 구성방식이 다른데 word를 기본단위로 구성된 시소러스와는 달리 synset이라는 독특한 구조를 채택하고 있다. synset은 concept을 표현하기 위한 하나의

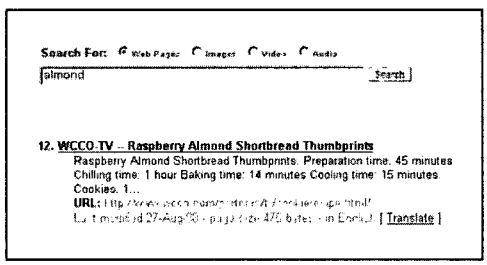
수단으로서 대상을 word 대신 "set of synonym"으로 표현한다[17]. 즉,

Set of synsets = {synset 1, synset 2, ..., synset n}

일례로 'raspberry'이라는 단어가 '나무딸기 열매'가 아닌 '나무딸기 나무'자체를 가리키는 의미로 사용될 때는 synset(raspberry, raspberry bush)이라고 표기한다. raspberry이라는 단어는 각각의 의미(sense)마다 synset이라는 형태로 표현되어지며, 그 단어가 가질 수 있는 의미의 개수 만큼 각 의미별로 여러 개의 계층링크(hierarchy link)가 존재하게 된다. 이러한 이유 때문에 (그림 3)의 워드넷의 계층구조는 일반적인 2차원적인 tree구조가 아닌 3차원적인 network topology형태를 취하게 된다((그림 3)에서는 공간적 제약에 의해 원래의 synset표기 대신 word로 표기를 대신함).

3.2 의미기반 정보필터링 과정

실제 검색엔진 Altavista에서 제공하는 요약정보(abstract information)만으로 필터링 작업이 진행되는 과정을 살펴보기로 한다. (그림 4)는 검색엔진 Altavista에 단일 질의 'almond'를 입력하였을 때 12번째로 검색된 웹문서의 요약정보이다. 요약정보인 Title, Description, URL내의 term들 가운데 "워드넷 어휘리스트에 존재하는 명사"를 제외한 나머지 term들을 제거하는 불용어처리 과정과 스태밍(stemming)과정을 거쳐 얻은 document D는 다음과 같다.

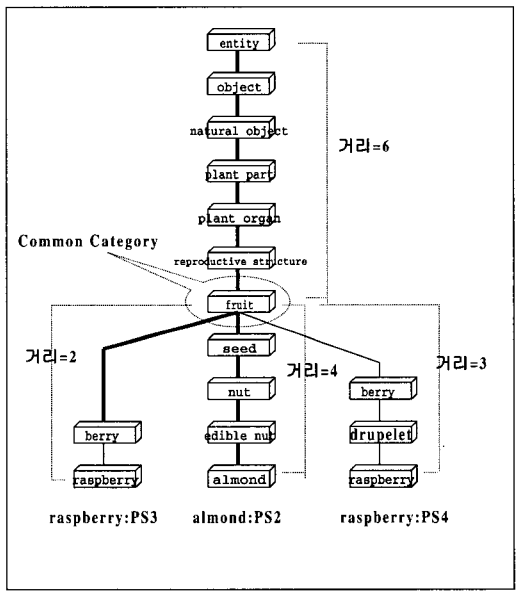


(그림 4) 검색엔진 Altavista의 검색결과

D = { (raspberry,dw1), (shortbread,dw2), (preparation,dw3), (time,dw4), (minute,dw5), (chill,dw6), (hour,dw7), (bake,dw8), (cool,dw9), (partner,dw10) }

profile P는 사용자가 입력한 원 질의로부터

P = { (almond, 0.5) } (부록 1의 수식 (1))



(그림 5) 'almond'와 'raspberry'의 계층연관정보

와 같이 구해진다.

먼저, 첫 번째 document term dt1 'raspberry'와 profile term pt1 'almond'의 워드넷 계층연관정보(그림 5)와 SDCC알고리즘을 이용하여 dt1의 가중치 dw1를 구하는 과정을 살펴보기로 한다. (그림 5)은 (그림 3)의 내용 중 필요한 부분만을 간략화시켜 2차원적인 트리구조로 정리하였다. (그림 3)에서,

Set of pt1(almond)'s synsets
= { $pt_1:PS_1$, $pt_1:PS_2$ }
Set of dt1(raspberry)'s synsets
={ $dt_1:PS_1$, $dt_1:PS_2$, $dt_1:PS_3$, $dt_1:PS_4$, $dt_1:PS_5$ }

이고, $pt_1:PS_2$와 $dt_1:PS_3$사이에 공통된 범주(Common Category)인 $C_1:PS_1$ 'fruit'이 존재한다. 또한 $pt_1:PS_2$와 $dt_1:PS_4$사이에도 공통된 범주 $C_2:PS_1$ 'fruit'이 존재한다.

먼저 $pt_1:PS_2$와 $dt_1:PS_3$를 통해 연관성의 정도(부록 1의 수식 (2))를 계산해보면,

$$SDCC(C_1:PS_1) = \frac{1}{2}\left(\frac{8-2}{8} + \frac{10-4}{10}\right) = \mathbf{0.675 > 0.5}$$

마찬가지로,

$$SDCC(C_2:PS_1) = \frac{1}{2}\left(\frac{9-3}{9} + \frac{10-4}{10}\right) = \mathbf{0.633 > 0.5}$$

따라서 d_{f1}의 가중치 dw1 는 다음과 같이 계산되어진다.

$$\begin{aligned} d_{w1} &= SDCCFunc(d_{f1}, p_{tj}) \\ &= \max\left(\sum_{no=1}^{k} SDCC(C_{no}:PS_u)\right) \\ &= \max(SDCC(C_1:PS_1), SDCC(C_2:PS_1)) \\ &= \max(0.675, 0.633) = 0.675 \end{aligned}$$

마찬가지로 나머지 document term들의 가중치를 구하면 다음과 같다.

D = { (raspberry,0.675), (shortbread,0), (preparation,0), (time,0), (minute,0), (chill,0), (hour0), (bake,0), (cool,0), (partner,0) }

따라서 최종적으로 profile P와 document D는 다음과 같이 정리되어질 수 있다.

P = { (almond, 0.5) }
D = { (raspberry, 0.675) }

최종적으로 검색된 웹 문서와 질의간의 유사도(부록 1)의 수식(3))를 계산하면,

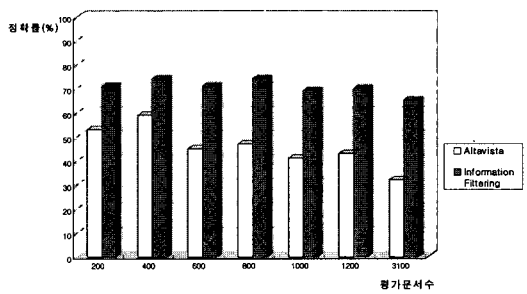
$$sim(D,P) = \frac{\sum_{wj=1}^{z}\sum_{wi=1}^{v}(d_{wi} * p_{wj})}{\sqrt{\sum_{wi=1}^{v} d_{wi}^2 + \sum_{wj=1}^{z} p_{wj}^2}} = \frac{0.675*0.5}{\sqrt{(0.675)^2 + (0.5)^2}} = 0.478$$

4. 성능평가

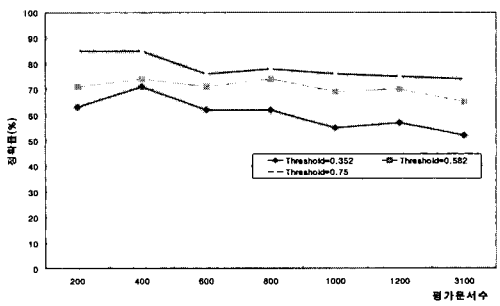
기존 정보필터링모델은 대부분 이용자모델을 기초로 필터링작업을 수행하게 되며, 이 경우 필터링의 성능은 필터링방법보다는 이용자모델구축에 좌우되어진다. 따라서 성능평가에 있어 기존의 정보필터링모델과 비교하는 대신 널리 사용되어지는 검색엔진 Altavista를 대상으로 하여 의미기반 정보필터링과정을 거치기 전과 의미기반 정보필터링과정 수행 후 정확률의 변화를 통해 필터링 효과를 확인해 보았다. (그림 6)은 필터링 임계치 0.582를 기준으로 할 때 의미기반 정보필터링 수행 후 정확률의 변화를 보여주고 있다. 성능 평가에 있어 문서 적합성 평가는 실무 경험을 갖춘 숙명여자대학교 문헌정보학과 대학원생 6명을 대상으로 각 문서마다 도메인 적합성을 확인하는 직접 평가를 실시하였으며, 총 3100개의 웹 문서를 평가하였다.

다음은 필터링의 임계치(threshold)를 변화시켜가면서, 필터링 효과를 조사해보았다. 문서수 증가에 따른 정확률(그림 7)과 누락율(snobbery ratio)의 변화는 (그림 8)과 같다.

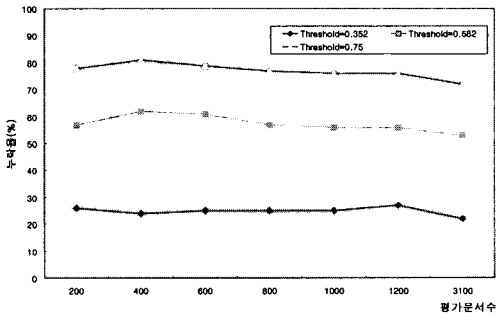
누락율은 전체 문서 중에서 실제 적합한 문서가 부적합 판정을 받은 비율을 보여 준다. (그림 7)과 (그림 8)에서 임계치 값이 높아짐에 따라 정확률의 증가에 비해 누락율의 증가비율이 상대적으로 높음을 살펴볼 수 있는데 이는 임계치 값이 높아짐에 따라 과도한 필터링이 발생할 수 있음을 의미한다. 이러한 현상은 인터넷 정보검색 시에도 비슷하게 발생하는데, 어휘의미 중의성이 인터넷 정보검색 효율에 미치는 영향에 관한 연구[10]에서도 키워드 수가 증가함에 따라 누락율의 급격한 증가를 살펴볼 수 있다. 이러한 현상은 여러 가지 복합적인 요인이 작용하여 발생되어지는 것으로 생각되어지며, 결과적으로 임계치값의 변화에 따른 필터링의 효과가 크게 달라질 수 있음을 확인할 수 있다.



(그림 6) 의미기반 정보필터링 효과비교



(그림 7) 임계치 변화에 따른 정확률의 변화



(그림 8) 임계치 변화에 따른 누락율의 변화

성능평가 실험에서 임계치 값을 변화시켜감에 따라 필터링의 강도를 조정할 수 있으며, 이는 사용자 수준에 따른 선택적 필터링(selective filtering)이 가능함을 의미한다. 충분한 검색능력을 갖춘 고급사용자에게는 별도의 필터링이 없겠으나, 초급이나 중급사용자들에게는 본 필터링기법이 유용하리라 여겨진다. 초급사용자들에게는 누락율이 상대적으로 높으나 대신 정확률이 높은 강한 필터링(strong filtering)이 중급사용자들에게는 정확률이 낮으나 상대적많은 관련 웹 문서들이 검색될 수 있는 약한 필터링(loose filtering)이 유용하리라 본다. 실험에서는 강한 필터링을 위한 임계치 값으로 0.582가 약한 필터링을 위한 임계치 값으로는 0.352가 적절하였음이 확인되었다. 앞의 예제에서 살펴본 (그림 4)의 12번째 웹 문서의 경우 약한 필터링을 적용할 때에는 적합문서로 강한 필터링을 적용할 때에는 부적합문서로 판정되어 사용자 수준에 따라 서로 다른 필터링 결과를 제공하게 된다.

5. 결론 및 향후계획

과거의 문헌데이터베이스 검색과는 다른 새로운 정보검색 환경인 인터넷에서 정보필터링에 관한 연구를 수행하였다. 연구 대상은 기존의 뉴스그룹이나 전자메일을 대상으로 한 정보필터링이 아닌 인터넷 정보검색을 위한 정보필터링이며, 이를 위해 달라진 필터링 환경에 대해 조사, 평가하였다. 새로운 정보검색 환경인 인터넷에서 정보필터링은 문헌데이터베이스 검색뿐 아니라 기존의 뉴스그룹이나 전자메일을 대상으로 한 정보필터링과도 판이하게 다르며, 기존의 정보필터링 모형은 새로운 환경 하에서 제대로 된 성능을 기대할 수 없다. 이를 확인하기 위해 대표적인 문서필터링 방식인 벡터공간모델 정보필터링 모형에 대해 자세히 살펴보았으며, 인터넷 정보검색을 위한 새로운 정보필터링 모형으로 의미기반 정보필터링 모형을 개발하였다. 의미기반 정보필터링 모형에서는 일반적인 필터링모델에서 이용하는 TF/IDF방식 대신의 새로운 방법인 SDDD알고리즘을 통해 웹 문서로부터 키워드를 추출하며, 이 과정에서 어휘의미사전인 워드넷을 활용하여 검색된 웹 문서와 사용자 관심사항간의 연관성의 정도를 계산하게 된다.

본 연구에서 개발한 의미기반 정보필터링 모형은 현재 개발중인 인터넷 정보검색에이전트를 위해 사용되어질 예정이다. 인터넷 정보검색에이전트는 검색엔진과 비슷한 역할을 수행하나, 그와 동시에 사용자의 수준을 고려하여 정보검색과정을 도울 수 있는 도우미역할을 수행하게 된다. 의미기반 정보필터링 과정 역시 획일적인 필터링의 수행이 이루어지는 것이 아니라 사용자의 수준에 맞추어 선택적 필터링을 수행하게 되며, 이를 통해 개별사용자에게 보다 쉽게 인터넷 정보검색이 가능하도록 편의를 제공하게 될 것이다.

부 록 1

의미기반 정보필터링 알고리즘

- 사용자 정보를 추출하여 프로화일 벡터를 생성하는 대신 사용자가 입력한 질의를 그대로 프로화일 벡터로 이용하였으며, document term을 결정하는 새로운 방법으로 SDCC알고리즘을 개발하였다. 최종적인 유사도 계산과정은 기존의 벡터공간모델과 동일하다.

document D가

$\mathbf{D} = \{ (d_{t1}, d_{w1}), (d_{t2}, d_{w2}), \ldots, (d_{ti}, d_{wi}), \ldots, (d_{tu}, d_{wu}) \}$이고,

profile P가

$\mathbf{P} = \{ (p_{t1}, p_{w1}), (p_{t2}, p_{w2}), \ldots, (p_{tj}, p_{wj}), \ldots, (p_{tz}, p_{wz}) \}$일 때,

모든 ProfileTerm들의 가중치 값이 항상

$$p_{w1} = p_{w2} = \ldots = p_{wj} = \ldots = p_{wu} = \lambda \qquad (1)$$

($\lambda = 0.5$)이라면, i번째 DocTerm d_{ti}의 가중치 값은 다음과 같은 과정을 통해 계산되어 질 수 있다.

원 질의를 통해 검색된 문서에서 추출된 DocTerm

d_{ti}와 ProfileTerm p_{tj}가
존재할 때($d_{ti} \neq p_{tj}$),

Set of d_{ti}'s synsets
= { $d_{ti}: PS_1$, $d_{ti}: PS_2$, ..., $d_{ti}: PS_a$, ..., $d_{ti}: PS_m$ }

Set of p_{tj}'s synsets
= { $p_{tj}: PS_1$, $p_{tj}: PS_2$, ..., $p_{tj}: PS_b$, ..., $p_{tj}: PS_n$ }

Synset은 원래 term에 의미태그(Possible Sense)정보가 추가 된 것임.

Synset $d_{ti}: PS_a$와 $p_{tj}: PS_b$의 공통된 범주(Common Category)인 $C_{no}: PS_u$가 존재하면,

$$SDCC(C_{no}: PS_u) = \frac{1}{q}\sum_{y=1}^{q}\left(\frac{Dy-dy}{Dy}\right), \quad n >= 2, \quad distance(root, C_{no}: PS_u) > 2$$

의미거리는 두 노드사이의 링크수의 합으로 계산되어 진다.

Dy = 루트로부터 각각의 synset $d_{ti}: PS_a$, $p_{tj}: PS_b$까지의 의미거리

dy = $C_{no}: PS_u$로부터 각각의 synset $d_{ti}: PS_a$, $p_{tj}: PS_b$까지의 의미거리 root로부터 $C_{no}: PS_u$까지의 링크거리(distance)는 반드시 2보다 커야 한다.

if ($SDCC(C_{no}: PS_u) >$ threshold θ)
then $SDCC(C_{no}: PS_u) = SDCC(C_{no}: PS_u)$
else $SDCCFunc(d_{ti}, p_{tj}) = 0$

threshold $\theta = 0.5$

i번째 DocTerm d_{ti}의 가중치 값은

$$d_{wi} = SDCCFunc(d_{ti}, p_{tj}) = \max\left(\sum_{no=1}^{k} SDCC(C_{no}: PS_u)\right) \quad (2)$$

이다.

최종적으로 유사도를 계산하는 sim(D,P)은 다음과 같이 계산될 수 있다.

$$sim(D, P) = \frac{\sum_{wj=1}^{z}\sum_{wi=1}^{y}(d_{wi} * p_{wj})}{\sqrt{\sum_{wi=1}^{y} d_{wi}^2 + \sum_{wj=1}^{z} p_{wj}^2}} \quad (3)$$

부 록 2

식물명 31개

apple almond aloe apricot cherry
camellia carnation coconut corn
chrysanthemum grape kiwi lily
mango mint medlar melon
olive peanut parsley pineapple
pepper potato pear pistachio
pumpkin rose sunflower tomato
tulip watermelon

참 고 문 헌

[1] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval : Two sides of the same coin?," Communication of the ACM 35(12), 1992.

[2] D. W. Oard, "A Conceptual Framework for text Filtering," the Medical Informatics project of the Pathology Department, a department of Education technology challenge grant, and the Logos Corporation, MDA9043C7217, 1996.

[3] P. J. Denning, "Electronic junk," Communication of the ACM,25(3), 1992.

[4] Mitchell F. Wyle, "Effective Dissemination of WAN Information," PhD thesis, LaSalle University, Mandeville, LA, 1995.

[5] T. W. Yan and H. Garcia-Molina, "Index structures for selective dissemination of information," Technical Report STAN-CS-92-1454, Stanford Univercity, 1992.

[6] J. Callan, "Document Filtering With Inference Networks," SIGIR '96 Zurich, Switzerland, 1996.

[7] T. W. Yan and H. Garcia-Molina, "Index Structures for Information Filtering Under the Vector Space Model," MDA972-92-J-1029 with the Corporation for National Research Initiatives(CNRI), 1994.

[8] A. Moukas and G. Zacharia, "Evolving a Multiagent Information Filtering Soluion in Amalthaea," Agents '97 Conference Proceedings, copyright 1997

ACM.

[9] S. Irene and K. Robert, "Modeling User's Interests in Information Filters," Communications of the ACM 35, 1992.

[10] 황상규, 오경묵, 변영태, "어휘의미 중의성이 인터넷 정보검색 효율에 미치는 영향에 관한 연구", 한국정보관리학회지 제16권 제3호, 1999.

[11] W. Patrick, "Unused Relevant Information in Research and Develop ment," Journal of the American of Information Science 46(1), 1995.

[12] 오경묵, 이용현, 황상규, "인터넷 이용자의 검색 행동 성향에 관한 연구", 한국문헌정보학회지 제33권 제3호, 1999.

[13] T. W. Yan and H. Garcia-Molina, "SIFT-A tool for wide-area information dissemination," Proceedings of the Third International Conference on Parallel and Distributed Information Systems, 1994.

[14] 이재윤, 김태수, "WordNet과 시소러스", 제11회 언어정보 연찬회 발표논문집, 연세대학교 상경관, 1998.

[15] EuroWordNet, "EuroWordNet : Building a Multilingual Database with WordNets for Several European Languages," University of Amsterdam Computer Centrum Letteren, <http://www.let.uva.nl/~ewn/>.

[16] Princeton University Cognitive Science Laboratory, "WordNet-a Lexical Database for English," Princeton University Cognitive Science Laboratory, <http://www.cogsci. princeton.edu/~wn/>.

[17] G. A. Miller, et al, "Introduction to WordNet : An On-line Lexical Database", International Journal of Lexicography 3(4), 1990.

변 영 태

e-mail : byun@cs.hongik.ac.kr

1977 서울대학교 전기공학과(학사)

1984 Indiana Univ. 미국 전산학 석사

1990 Univ. of Texas at Austin 미국 전산학 박사

1979~1982 D.E.C in Korea근무

1990~현재 홍익대학교 전자계산학과 교수, 홍익대학교 정보대학원장

관심분야 : Intelligent Agent, Knowledge Representation and Reasoning, Machine Learning임

황 상 규

e-mail : skhwang@cs.hongik.ac.kr

1998 홍익대학교 컴퓨터공학과 학사

1998~현재 홍익대학교 전자계산학과 석사과정 재학중

관심분야 : Intelligent Agent, Internet Search, Information Science임

오 경 묵

e-mail : kmoh@sookmyung.ac.kr

1986 연세대 문헌정보학과(학사)

1993 영국 Univ. of Sheffield, 정보학(석사)

1997 영국 Loughborough Univ. of Technology, 정보학(박사)

1997 첨단학술정보센터 팀장

1999~현재 숙명여자대학교 정보과학부 교수

관심분야 : 정보조직, 정보시스템 평가 및 방법론(Soft Systems Methodology)임