어휘의미 중의성이 인터넷 정보검색 효율에 미치는 영향에 관한 연구

A Study of Word Sense Ambiguation which Affects Efficiency of the Internet-based Information Retrieval

황상규(Sang-Kyu Hwang)*, 오경묵(Kyung-Mook Oh)**, 변영태(Young-Tae Byun)*

목 차

초 록

부적절한 검색어의 선정 및 검색식의 작성은 인터넷 정보검색 수행 시 검색 효율 저하의 주요 원인으로 작용하게 된다. 또한 정보검색 수행시 발생하는 어휘의미중의성(Word Sense Ambiguation) 역시 검색 효율 저하의 주요 원인으로 작용하는데, 어휘의미중의성에 의한 효율 저하 정도를 실험을 통해 확인하였다. 어휘의미중의성에 의한 검색 효율 저하란 검색어로 입력한 어휘가 문서에서 서로 다른 의미로 사용됨에 따라 의도하지 않은 다른 문서가 검색될 수 있음을 의미한다. 본 논문에서는 새로운 정보검색 환경인 인터넷기반 정보검색에 있어 어휘의미중의성이 검색 정확률에 미치는 영향을 살펴보고, 기존의 정보검색에 있어 어휘의미중의성에 관한 연구가 인터넷기반 정보검색에 있어서도 제대로 적용되는지를 조사분석하였다.

ABSTRACT

Internet users are often frustrated when they try to find "right" piece of information quickly. The reason is that the discovery of available and quality based-resources becomes more difficult to end users while the Internet continues to expand rapidly. Not only incorrect keywords and query expression but word sense ambiguation are the cause of dropping-off in efficiency on Internet search. In this paper, studies were conducted to analyze dropping off in efficiency for Internet search and discussed reducing user's frustration of the Internet and improving their search strategies.

키워드: word sense ambiguation, Internet search, WordNet, information retrieval

* 홍익대학교 전자계산학과
** 숙명여자대학교 정보과학부
■ 논문 접수일 : 1999년 9월 2일

1 서 론

대부분의 정보검색 시스템은 이용자가 입력한 질의를 근거로 질의어와 문서간의 키워드 매칭을 통해 적합 문서를 찾아내게 된다. 이 경우 보통 이용자가 입력한 질의어가 실제 문서상에서 한가지 의미로 사용되어지는 경우는 드물며, 대부분 여러 가지 다양한 의미로 사용되어진다. 하지만 키워드 매칭 방법에서는 사람과는 달리 검색시스템은 어휘의 의미를 구분할 수 없기 때문에, 과일로서의 'apple'과 컴퓨터 상호로 사용된 'apple'을 구분하지 못한다. 이러한 어휘의미중의성 때문에 실제 이용자가 의도한 바와는 전혀 다른 문서가 적합한 문서로 판정 지어질 수 있는 것이다. 정보검색에 있어 어휘의미중의성 문제(Word Sense Ambiguation Problem)란 사용자가 입력한 질의어가 실제 웹 문서상에서 여러 가지 다양한 의미로 사용됨에 따라 검색 정확률을 떨어뜨리는 요인으로 작용할 수 있다는 것을 의미한다.

이미 어휘의미중의성이 정보검색에 미치는 영향에 대해서는 많은 선행 연구자들에 의해 다양한 연구(Weiss 1973; Krovertz and Croft 1992; Voorhees 1993; Sanderson 1994)가 진행되어 왔으나, 아직까지도 다양한 실제상황에서 얼마나 심각한 영향을 미치는지 여부와 그에 대한 효과적인 해결방법이 뚜렷이 제시되지 못한 상황이다. 여러 가지 다양한 문제 해결 시도에도 불구하고 대부분 결과가 기대치에 미치지 못한 편이며, 실제 어휘의미중의성이 발생하는 상황을 살펴보면 매우 다양한 경우가 발생하여 해결책 제시를 어렵게 하고 있기 때문이다. 여러 연구가 계속 진행되어 가는 가운데 어휘의미중의성에 관한 연구를 비교 정리한 Sanderson의 연구 결과에서는, "어휘의미중의성 문제는 이용자가 극히 짧은 길이의 질의(very short queries)를 입력한 경우에 한해서만 정보 검색결과에 영향을 끼치게 된다"라고 밝히고 있다.

94년도 이후 인터넷이 정보량 및 이용자 수에 있어 폭발적인 증가를 가져옴에 따라 인터넷 정보검색은 중요한 정보검색의 수단으로 자리잡게 되었다. 하지만 인터넷기반 정보검색은 기존의 정보검색환경과 마찬가지로 텍스트 기반 정보검색을 기본으로 하나, 검색대상이 되는 문서나 서비스 이용자적 측면에서는 기존의 정보검색과는 여러 면에서 많은 차이가 있다. 일정한 주제를 대상으로 서로 비슷한 성격과 형식을 갖춘 정형화된 문서들의 모임인 서지데이터베이스와는 달리 인터넷은 각기 대상주제가 다른 비정형화된 문서들의 모임이며, 그 양에 있어서도 상대적으로 훨씬 더 방대하다. 또한 서비스를 이용하는 사용자 역시 소수의 검색 교육을 받은 전문가에서 다수의 평범한 일반 이용자들로 점점 더 확대되어 가고 있다. 매년 인터넷의 이용자수가 2배 이상씩 증가해 가는 현 상황에서 대부분의 이용자들은 전문적인 웹 정보검색 교육을 받아본 적이 없으며, 앞으로도 새로이 교육을 받을 수 있는 기회 또한 희박한 편이다.

이는 인터넷기반 정보검색을 보다 힘들게 하는 주요 요인으로 작용하게 되는데, 정확한 검색식을 작성한 능력을 갖추지 못한 이용자들은 부정확한 어휘를 검색어로 선정하기 쉬우며, 각기 다른 다양한 주제들을 담고 있는 웹 문서들을 대상으로 한 정보검색이기 때문에 어휘의미중의성에 의한 검색 정확률 저하의 가능성은 서지정보검색에 비해 훨씬 심각한 것으로 확인되었다. 누구나 실제 인터넷을 통해 정보검색을 해본 적이 있다면, 인터넷기반 정보검색의 결과가 서지정보검색

에 비해 훨씬 검색 정확률이 낮다는 사실을 쉽게 경험해 보았을 것이다. 하지만 지금까지 어휘의미중의성에 관한 선행 연구들은 서지정보검색 환경 하에서 연구가 진행되어져 왔으며, 새로운 정보검색 환경인 인터넷기반 정보검색환경 하에서는 어휘의미중의성이 인터넷기반 정보검색에 미치는 영향에 관해서는 아직까지 충분한 연구가 이루어지지 못한 상황이다.

어휘의미중의성이 인터넷기반 정보검색에 미치는 영향을 조사하고 기존의 어휘의미중의성에 관한 선행연구 결과가 인터넷기반 정보검색환경 하에서도 제대로 적용될 수 있는지를 확인하기 위하여 실제 현실에서 그 실태를 조사할 경우에는, 수많은 인터넷 이용자들의 성향 및 그들의 다양한 검색식을 검토해 볼 필요가 있다. 이 경우 조사 및 평가를 위해서는 많은 시간과 비용이 요구되며 충분한 조사가 이루어지기 힘들다는 문제점이 있게 된다. 따라서 본 연구에서는 설문 조사의 방법 대신 어휘의미중의성 문제가 실제 현실세계에서 발생할 수 있는 상황을 모델링하고, 시뮬레이션을 통해 검증해보기로 하였다. 먼저 Sanderson의 연구 결과에서 '극히 짧은 길이의 질의(very short queries)'를 해석하는데 있어서는 Allan과 Papka의 연구에서 조사된 결과(Allan and Papka 1998)를 바탕으로 Excite검색엔진에 이용자 평균 질의어의 개수 2.3개를 기준으로 삼았는데, 웹 정보검색 이용자의 대부분이 초보자이며 "그들은 대부분 단일질의를 통해 웹 정보검색을 시도한다(박창호, 박민규, 이정모 1998)"는 현실을 반영해 볼 때, 본 연구에서는 '극히 짧은 길이의 질의'를 검색 질의 길이의 평균치 2.3개보다 작은 1개 혹은 2개인 경우로 가정하였다. 이러한 가정을 전제로 실제로 웹 정보검색 이용자가 입력한 질의어의 수가 3개 이상인 경우에는 어휘의미 중의성에 의한 검색 정확률 저하가 더 이상 발생하지 않는지 여부를 웹 정보검색 초보이용자모형을 통한 시뮬레이션을 통해 검증해 보았다.

2 이용자 분석과 워드넷 활용

어휘의미중의성이 인터넷 정보검색시 미치는 영향에 대한 연구분석을 위해 사용한 인터넷 이용자들의 검색 방식을 분석한 방법과 결과 그리고 새로운 어휘 사전인 워드넷을 지식베이스로서 이용하는 방법은 다음과 같다.

2.1 인터넷 이용자 분석

DIALOG등과 같은 상용 데이터베이스와는 달리 인터넷 정보검색은 무료에 가까운 이용료를 지불함에 따라 더 이상 정보검색 시간에 제약을 받지 않는 상황하에서 정보검색을 수행하게 되며, 일반적으로 이용자들은 명확한 검색 목적을 가지고 검색에 임하지 않고 있다(Borgman, Hirsh and Hiller 1996). 또한 텍스트 문서, 비디오, 오디오 등 다양한 매체의 정보들을 접근할 수 있는 인터넷은 기존의 텍스트 위주의 상용 데이터베이스 검색과는 다른 다양한 서비스를 제공하게 되며, 그에 따라 이용자들의 서비스 이용 패턴 역시 과거와는 다른 양상을 보이고 있다(Abdulla, Fox and Abrams 1997). 이는 과거의 데이터베이스 위주 정보검색 이론들이 새로운 정보검색 환경인 인터넷에서도 제대로 적용되어질 수 있는지 여부를 확인하는 과정이 필요하며, 이미 다양한 연구들이 시도되어 지고 있다. 또한 인터넷 이용자들의 특성을 서비스를 제공하는 웹

서버에 기록된 이용자 정보(log file)분석을 통해 조사하는 방법을 시도해 보았다. 대학교, 고등학교, 관공서, 기업체 등 서로 다른 인터넷 이용자 그룹간의 이용 방식에 차이를 보이는지 여부를 조사해 보았는데, 분석 결과 대부분의 인터넷 이용자들은 그들이 서로 다른 환경에도 불구하고, 서로 유사한 이용패턴을 보이고 있음을 보고하고 있다(Abdulla, Liu and Fox 1998).

서로 다른 이용자들이 각기 개인적 성향 차를 지니고 있음에도 불구하고 서로 유사한 이용패턴을 보인다는 점에서 검색 결과에 대한 이용자의 만족도 측면에서도 인터넷 이용자들간의 어느 정도 유사한 패턴을 찾을 수 있지 않을까 생각할 수 있다. 이와 같은 가설이 실제로 맞는지를 확인하기 위하여 먼저 기존의 인터넷 정보검색 시스템에서 검색 결과에 대한 이용자의 만족도를 조사해보았다. 실험에서 성능 비교를 위해 서로 다른 특성을 지닌 4개의 웹 기반 정보검색 시스템을 선정하였다.

1) 지능형 정보검색 에이전트 - 에이전트란 사람의 일을 시스템이 대신하여 보다 편리한 작업을 진행할 수 있게 도와주는 프로그램의 일종이다. 정보검색 에이전트란 이용자의 정보검색을 보조해주는 도우미 역할을 하게 되며, 보통 특정 영역에 대한 전문지식베이스를 갖추고 있으며, 인공지능의 기법을 활용한 질의 확장으로 보다 나은 검색성능을 제공한다. 본 연구에서는 홍익대 전자계산학과에서 개발된 동물 영역을 대상으로 한 지능형 정보검색 에이전트 HIIA-1a를 실험에 이용하였다.

〔시스템〕 HIIA-1a

2) 로봇기반 검색 시스템 - 일반적인 자동화된 인터넷 검색엔진이 이에 해당되며, 보통키워드형 검색 방식을 제공한다.

〔시스템〕 Altavista, Excite

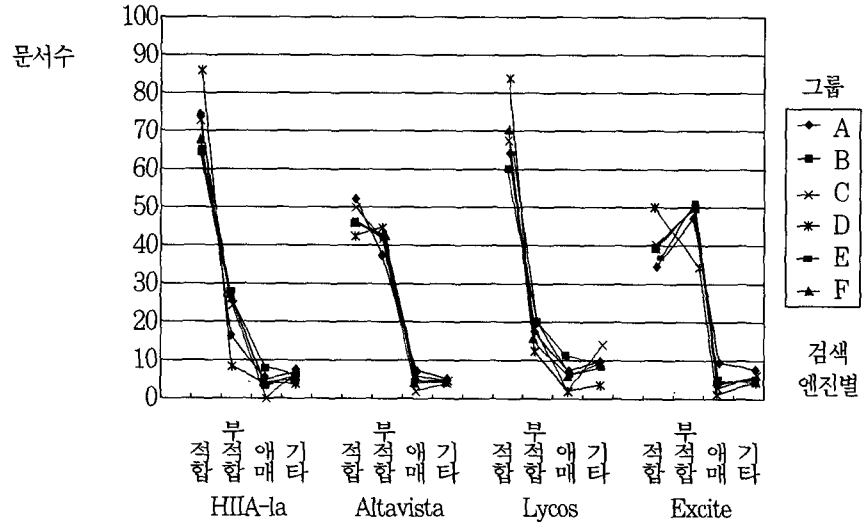
3) 혼합 검색 시스템 - 문서 수집을 위해서는 로봇을 이용하며, 수집된 문서 분류는 인간전문가에 의해 이루어진다. 문서 순위 결정은 시스템의 알고리즘에 의해 자동으로 이루어지나, 수집된 문서 분류정보 중에서 이용자의 질의를 통해 검색된 문서에 해당하는 분류정보가 존재하는 경우에는 분류정보가 문서 순위 결정에 반영되어 보다나은 성능을 기대할 수 있다.

〔시스템〕 Lycos

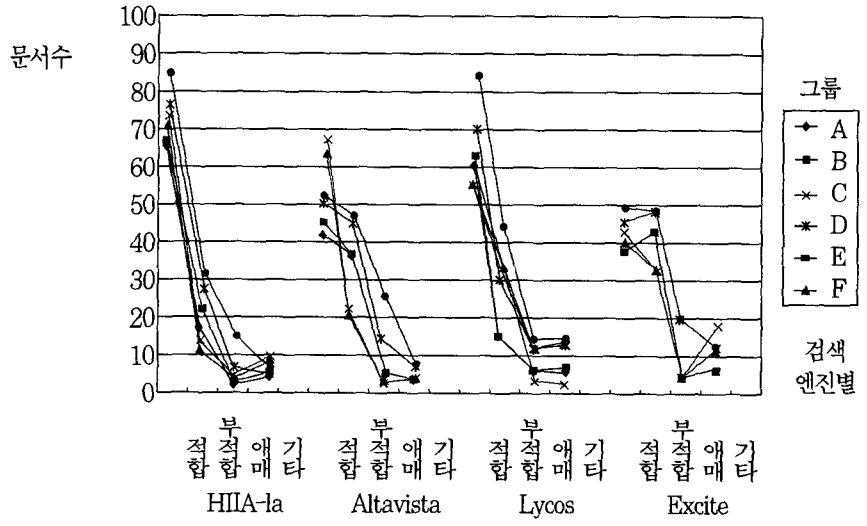
평가 인원은 홍익대 전자공학과 대학원생 19명을 대상으로 실시하였으며, 모든 응답자들은 기존에 웹 정보검색에 대한 교육을 별도로(공식적으로) 받은 적이 없으며, 웹 정보검색 방법을 친구나 그밖에 자료를 통해 스스로 습득한 일반 이용자들로 구성하였다. 동물에 대한 전문 지식베이스를 갖춘 지능형 정보 에이전트 HIIA-1a와 기존 검색엔진들과의 성능 비교를 위하여, 평가 참여인원은 동물에 관한 질의단어(전문단어인 canivora, cetacea, 일반단어인 gorilla 등을 질의어로 사용하였다)에 대해서 각각 4개의 웹 정보검색 시스템으로부터 100개까지의 문서를 평가하였다. 그리고 결과 분석을 위해 전체 인원을 3명씩 한 그룹으로 묶어 각 그룹간의 평가결과를 비교해 보았다.

이용자의 만족도 조사에서는 예상대로 평가 대상인 동물 영역에 대한 전문 지식베이스를 갖춘 지능형 정보 에이전트 HIIA-1a가 대체적으로 가장 높은 만족도를 보였으며, 인간의 수작업 분류 정보를 바탕으로 한 Lycos가 나머지 Altavista나 Excite보다 높은 만족도를 보였으나 그 차이가 미비하였다. 이는 질의 Set의 대상도메

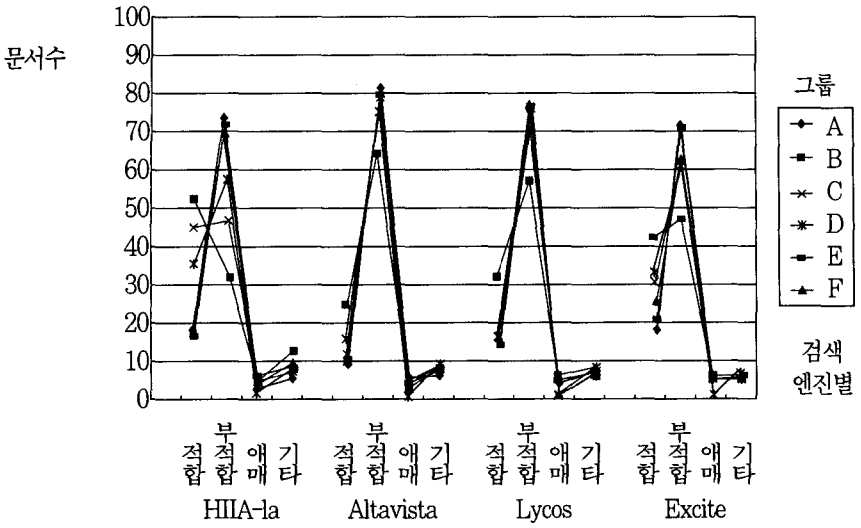
인이 동물로 한정되었으나, Lycos에서의 질의 Set에 동물 관련 전문 어휘와 관련된 인간의 수작업 분류 정보가 아직까지 갖추어지지 못한 경우가 대부분이었기 때문이다. 하지만 Lycos와 같은 대부분의 상용벤더에서는 교육 연구 분야와 같은 비상업적 분야는 상대적으로 소외되는 현 상황을 고려해 볼 때 단시일 내에 개선되기는 어렵다고 보여진다.



〈그림 1〉 질의가 전문단어 carnivora인 경우



〈그림 2〉 질의가 전문단어 cetacea인 경우

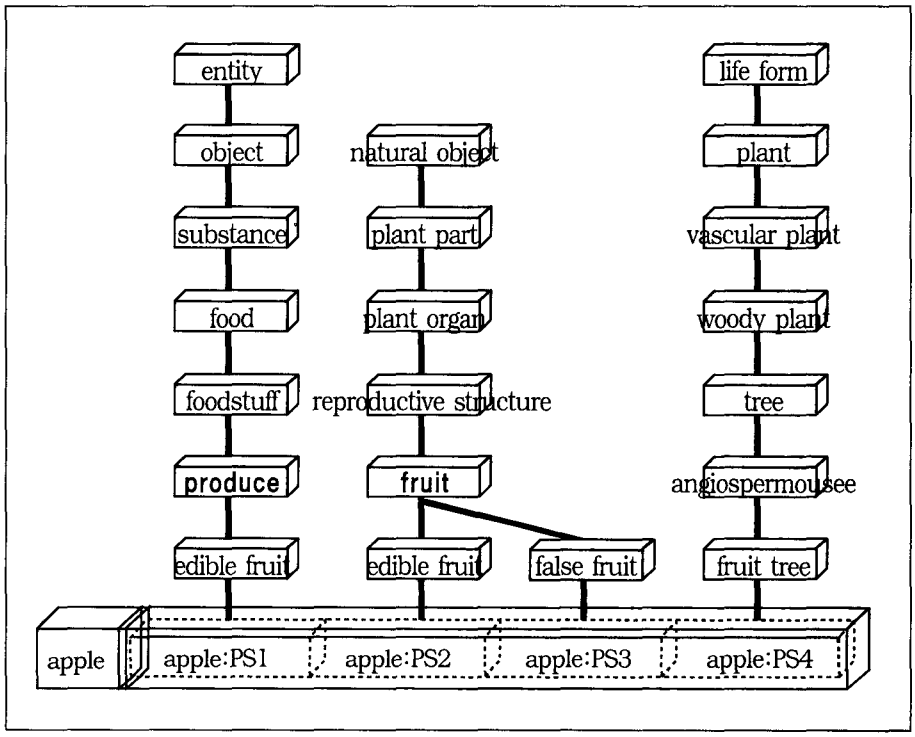


〈그림 3〉 질의가 일반단어 gorilla인 경우

분석작업을 진행해 가는 과정에서 흥미로운 상황이 발생하였는데, 질의 'gorilla'에서는 다른 결과를 확인할 수 있었다. 응답자 모두가 가장 나은 성능을 보인 지능형 정보검색 에이전트인 HIIA-1a를 포함한 모든 검색시스템으로부터 검색 결과에 대해 '불만족'을 가장 높게 표시한 것이다. 위의 그림들을 통해 이러한 차이를 비교해볼 수 있는데 각 그림들은 질의 'carnivora'와 'cetacea', 'gorilla'일 때 이용자 그룹별 검색 만족도를 각각 다르게 보여주고 있다. 이러한 원인이 발생한 이유로는 전문단어인 'carnivora', 'cetacea'에 비해 어휘의미중의성의 가능성이 큰 일반단어 'gorilla'가 단일 질의로 선택됨에 따라 어휘의미중의성에 의한 검색 정확률 저하가 발생한 것으로 확인되어 진 것이다.

서로 다른 이용자들이 각기 개인적 성향 차를 지니고 있음에도 불구하고 정보검색과 같은 인터넷 서비스를 이용하는 방법이나 검색 결과에 대한 이용자의 만족도 측면에서 그들간에는 어느 정도 공통된 성향이 존재한다는 점이 확인되었다. 이는 다시 말해 전체 인터넷 이용자 중 일부 이용자들의 행동 패턴 및 이용자 만족도를 분석, 조사해 얻은 이용자 정보는 대체적으로 전체 이용자 그룹의 특성을 파악하는 정보로 활용할 수 있음을 의미한다. 따라서 본 연구에서는 많은 수의 인터넷 이용자들을 대상으로 하여 실험을 수행하지는 않았지만, 실험의 결과에 대해서는 타당성을 보여준다.

이와 같이 인터넷 정보검색에 있어서 적절한 검색어의 선정이 이루어지더라도 일반단어와 전문단어의 선택에 따라 검색 정확률이 저하될 수 있음을 확인하였으며, 이런 상황을 발생시키는 한가지 요인으로 어휘의미중의성에 의한 검색 정확률 저하의 가능성을 확인하였다. 이러한 사실들을 토대로 본 연구에서는 기존의 선행 연구와는 다른 새로운 각도에서 인터넷 정보검색 효율저하에 미치는 원인에 대해 접근하였으며, 시뮬레이션을 통해 검색 효율 저하의 원인 파악 및 결과 분석을 시도하였다.



〈그림 4〉 워드넷의 계층구조

2.2 지식베이스로서의 워드넷(WordNet) 이용

Ontology의 일종으로 간주되고 있는 워드넷(WordNet)은 인간의 어휘지식에 대한 심리언어학 연구의 성과를 토대로 1985년부터 프린스톤 대학 인지과학연구실이 구축해온 Lexical Database이다(Miller et al. 1990; 이재윤, 김태수 1998). Ontology란 인공지능 분야에서 전문가 시스템에 쓰이는 지식베이스의 일종으로, 인공지능 에이전트가 등장함에 따라서 Ontology는 지식표현구조의 호환성을 위한 도구로 널리 활용되어져 왔다. 워드넷(WordNet)은 정보 검색시스템, 자연언어처리와 정보검색 등의 여러 분야에서 널리 이용되고 있으며 다국어판의 구현도 시도되고 있다(EuroWordNet 1998; Vossen 1997; Harabagiu 1998).

워드넷(WordNet)의 주된 특징은 단어형이 아닌 단어의 의미를 그 구성요소로 하였다는 점이다. 그에 따라 〈그림 4〉의 'apple'의 계층구조는 실제 구현상의 워드넷 계층구조와는 약간의 차이가 있다. 이는 워드넷이 단어(word)를 기본단위로 구성된 시소러스와는 달리 'synset'이라는 독특한 구조를 채택하고 있는데 기인한다. synset은 개념(concept)을 표현하기 위한 하나의 수단으로서 대상을 단어(word) 대신 동의어의 집합(set of synonym)으로 표현한다. 일례로 foodstuff이라는 단어가 "commodity sold by a grocer"의 의미로 사용될 때 실제 구현상으로 {grocery , foodstuff}이라는 동의어 집합으로 표현되어 진다. 이러한 synset의 개념에서 살펴보

면 〈그림 4〉에서 apple이란 단어는 "음식물로서의 사과(PS1)", "먹을 수 있는 과일로서의 사과(PS2)", "식용 가능한 과실류로서의 사과(PS3)", "사과나무로서의 사과(PS4)"의 서로 다른 4가지 의미로 사용되어질 수 있다는 점을 워드넷의 계층지식 정보로부터 유도해낼 수 있다.

본 연구에서는 어휘의미중의성이 검색 정확률 저하에 미치는 영향을 실제 이용자들이 질문에 응답하는 방법 대신 웹 정보검색 초보 이용자 모형을 통한 시뮬레이션을 통해 실험하였다. 실험을 수행하는데 있어서는 인간이 가지고 있는 지식 정보를 흉내내어 지식베이스를 구축하는 과정을 필요로 하게 되는데, 별도의 지식베이스를 구축하는 대신 워드넷(WordNet)이라는 어휘사전을 지식베이스로 활용하는 방법을 사용하였다.

3 실 험

3.1 기본전제

본 연구에서는 대상 도메인을 식물로 한정하였는데, 우리나라를 비롯하여 전세계적으로 가장 널리 사용되는 분류체계인 듀이십진분류표(DDC)를 사용하기로 하였다. DDC 21판의 580번대 예시주(example note)에서 일상 생활에 자주 쓰이는 식물명 31개를 선택하였으며, 각각에 대해 검색엔진 알타비스타를 통해 각각 100개씩의 문서를 수집하였다. 검색된 문서의 적합성 여부는 평가시 발생할 수 있는 개인적 성향 차를 감안하기 위하여, 숙대 문헌정보학과 대학원생 6명에 의해 평가되어졌다. 실험을 위해 먼저 다음과 같이 2가지 기본 전제를 설정하였다.

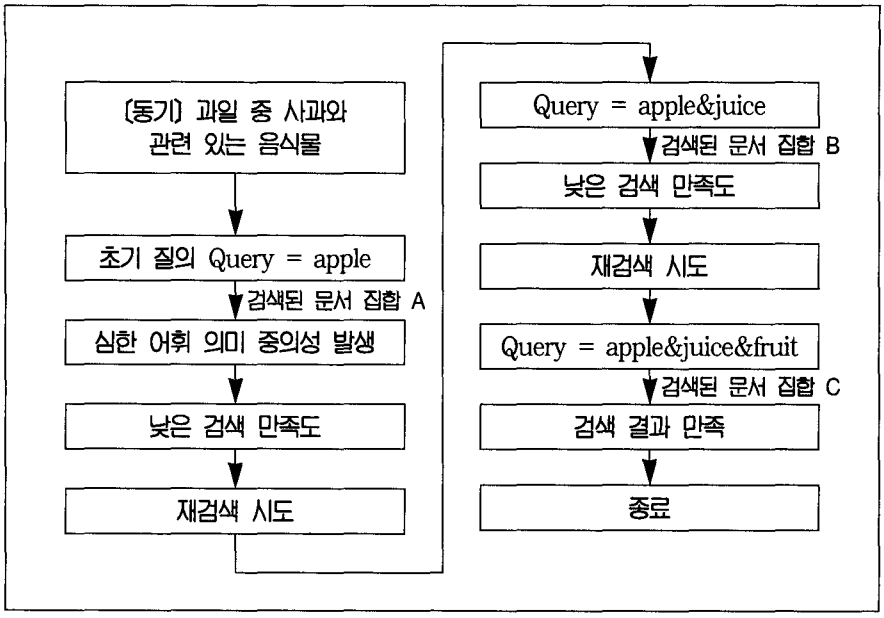
1) 일반적으로 초보 이용자가 입력하는 원 질의어의 수는 1개이며, 본 실험 역시 사용자가 입력한 검색어 1개 외에는 아무런 이용자 추가정보 없이 실험을 수행한다.

2) 일반적으로 초보 이용자가 입력하는 원 질의어는 특수한 식물명(전문단어)이기 보다는 일상생활에서 널리 쓰이는 식물명(일반단어)이라는 판단 하에 실험을 위한 검색어를 선정한다(선정된 검색어는 〔부록 1〕에 있음).

검색된 문서에 대해서는 적합성 여부를 판단하기 위하여 도메인 적합성 여부를 확인하였다.검색된 문서가 대상 도메인인 식물 관련 문서인지 여부를 확인하게 되는데, 만일 검색어가 'apple'인 경우, 검색된 문서에서 사용된 'apple'이란 단어가 과일로서의 'apple'인지 여부를 확인하여 문서의 적합성 여부를 판정 짓게 된다.

3.2 실험 모형 및 방법

Pollock과 Hockley는 인터넷에서 정보검색 수행 시에 예기치 못한 수많은 요인들이 검색 결과에 영향을 미치게 되며, 이러한 특성을 고려 할 때 일반이용자들의 인터넷 정보검색 수행시에는 보다 간단한 검색 방식을 취하라고 권고하고 있다(Pollock and Hockley 1996). 이러한 측면에서 인터넷 정보검색 수행시에는 여러 개의 검색어와 다양한 연산자로 구성된 복잡한 검색식을 구사하기보다는 1개 내지 2개의 검색어에서 시작하여 점차 검색어 수를 늘려가는 것이 보다 바람직한 검색 방식일 수 있다. 또한 대부분의 이용자들은 다양한 연산자를 이용하는데 익숙치 못하며, 일반적으로 AND나 OR정도의 연산자만을 이용하는 것이 보편적이다. 하지만 OR연산자의 경우 검색 문서 수를 크게 늘리는 역효과를 가져오기 쉬운 관계로 초보 이용자들의 경우 AND만

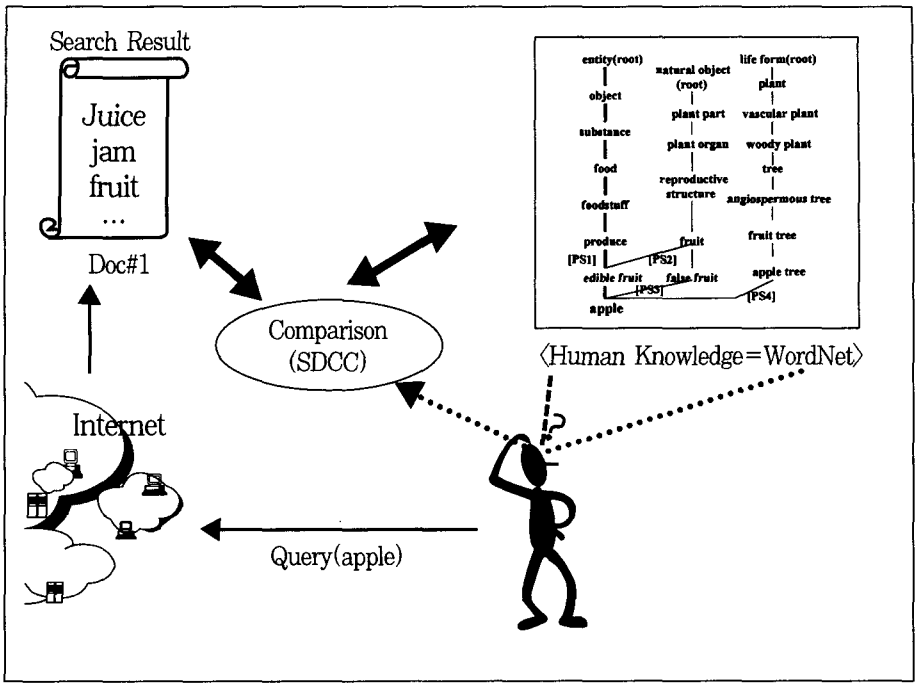


〈그림 5〉 검색 시나리오

으로 구성된 검색식을 작성하는 것이 보다 효과적인 검색방법일 수 있다. 지금까지 언급한 내용을 바탕으로 하여 실험을 수행하기 위한 하나의 검색시나리오를 〈그림 5〉와 같이 구성해보았다.

〈그림 5〉는 초보 이용자가 사과를 이용한 음식관련 정보를 찾고자 검색엔진 Altavista를 이용하여 실제 검색해 가는 일련의 과정을 단계별로 보여주고 있다. 정보검색시 이용자는 정보검색에 익숙치 못한 관계로 처음부터 적절한 검색식을 작성할 능력을 갖추고 있지 못하다. 따라서 맨 처음에는 단일 질의를 통해 검색을 시도하게 된다. 또한 이용자는 검색 결과에 대해 불만족 시 매번 새로운 검색어 하나씩을 AND연산자를 통해 기존의 검색식에 추가하게 된다. 이때 이용자가 매 검색단계마다 검색키워드로 선정하는 어휘는 전혀 검색의도와는 무관한 단어는 아니지만, 그렇다고 아주 적절하지도 않은 어휘라고 가정한다. 사과를 이용한 음식관련 정보를 찾고자 이용자가 맨 처음 단일 검색 키워드로 'apple'을 입력한 경우, 검색된 문서들에서는 상당히 심한 어휘의미 중의성에 의한 검색 정확률 저하를 예상할 수 있다. 이용자는 다음단계의 재 검색 과정으로 새로운 연관 검색키워드로서 'juice'를 기존의 검색식에 AND연산자로 추가하게 된다. 이때 이용자는 새로운 검색키워드 한 개씩을 추가하는 매 단계마다 검색된 모든 문서의 적합성여부를 확인하며, 전체 검색된 문서 중 상당수가 적합한 문서로 여겨질 경우 정보검색수행을 종료한다고 전제한다.

위와 같은 시나리오를 기본으로 하여 본 실험에서는 초보 이용자가 새로운 검색키워드를 생각해서 재 검색해나가는 과정을 인간의 지식을 대신하게되는 워드넷(WordNet)과 인간의 비교능력을 대신하는 SDCC(Semantic Distance for Common Category)알고리즘이라는 새로운 방법을 고안하여 자동화된 실험을 수행하였다([부록 2]참조). SDCC알고리즘은 워드넷을 이용하여



〈그림 6〉 초보 웹 정보검색 이용자 모형

이용자가 입력한 키워드와 연관성 있는 어휘를 자동으로 생성해주는 도구이다. SDCC알고리즘은 웹 문서 정보검색을 위한 필터링에 관한 연구(황상규 외 1999)에서 처음 개발되어 소개되어진 알고리즘이며, 본 연구에서는 SDCC알고리즘을 두 어휘간의 연관성을 계산해주는 도구로 활용하였다.

실험에서는 SDCC알고리즘을 이용하여 원 질의와 문서에서 추출한 키워드간에 연관성의 정도를 계산하게 된다. 만약 원 질의가 단일키워드 'apple'이고 이를 통해 검색된 문서 D에서 SDCC알고리즘([부록 2]의 예제참조)을 통해 문서의 대표 키워드로 찾아낸 어휘가 'juice'라면, 이는 앞의 시나리오에서 초보 이용자가 재 검색 과정으로 새로운 검색키워드로서 'juice'를 추가한 경우와 동일한 상태에 해당되게 된다. 이는 초보 이용자가 매번 주어진 상황을 고려하여 새로운 검색키워드를 생각해내는 과정을 시스템이 워드넷이라는 지식정보와 SDCC알고리즘을 이용한 일련의 처리 과정을 통해 매번 검색어 수 증가에 따른 검색결과를 얻게 되는 것과 동일한 효과를 얻게 되는 것이다. 지금까지 언급한 초보 웹 정보검색 이용자 모형을 그림을 통해 살펴보면 〈그림 6〉과 같다.

4 실험 결과

정보검색 환경 하에서 사용자가 입력한 질의어의 수가 3개 이상인 경우에는 질의어가 문맥상에서 발생하게 되는 의미중의성의 가능성이 극히 희박해진다는 기존의 연구 결과가 웹 정보검색

〈표 1〉 평가방법

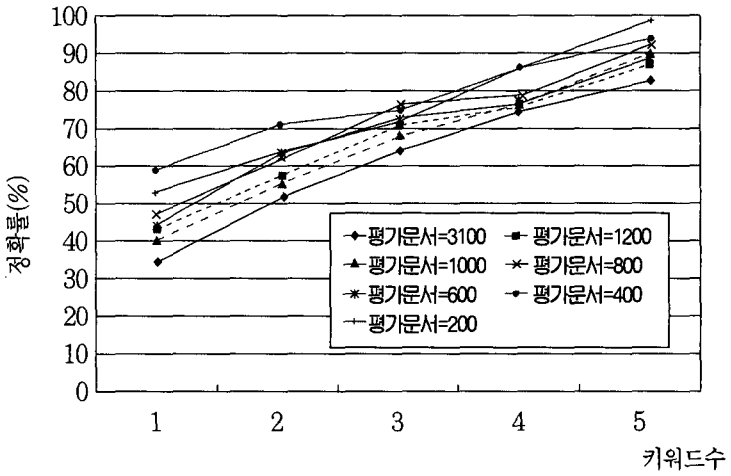
각각의 질의에 대해,
B : 평가자에 의해 적합판정을 받은 전체 문서 집합
C_i : SDCC 알고리즘을 통해 검색된 문서 집합(i= 키워드 수)
D_i : SDCC 알고리즘을 통해 검색되지 않은 문서 집합(i= 키워드 수)

정확률 $P_i = \frac{B \cap C_i}{C_i}$, 재현율 $R_i = \frac{B \cap C_i}{B}$

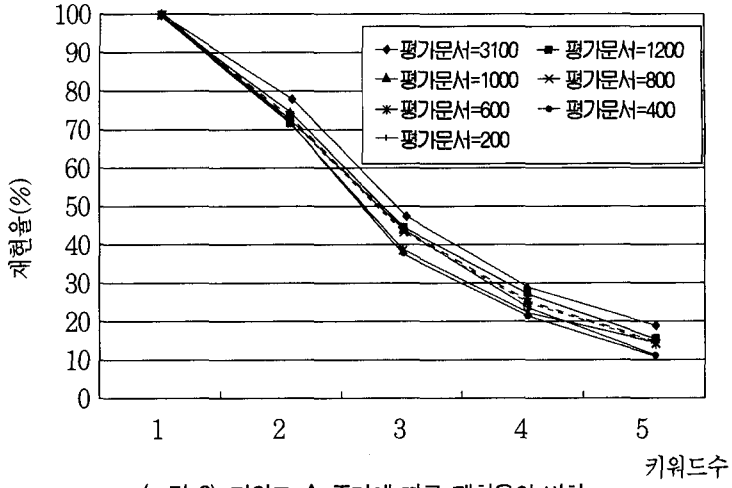
누락률 $F_i = \frac{B \cap D_i}{B}$

환경 하에서도 제대로 적용되는지 여부를 검증해 보기로 하였다. 본 실험에서는 총 31개의 질의 set을 대상으로, 각 질의마다 검색엔진 알타비스타를 통해 웹 문서 100개씩을 수집하였다. 전체 질의 set에서 무작위로 2, 4, 6, 8, 10, 12, 31개씩 7집단을 추출, 평가문서 수를 증가시켜 가면서 키워드 수 변화에 따른 정확률과 재현율, 누락률(snobbery ratio)을 조사하였는데 구체적인 계산 방법은 〈표 1〉과 같다.

초보 웹 정보검색 이용자 모형을 기반으로 한 실험 결과에서는 대체적으로 검색 키워드수가 3개에 이르게 되면, 문서의 정확률이 70%를 넘게 되어 어느 정도 검색결과가 만족할 수준에 이르며, 정보검색을 종료하게 되었다〈그림 7〉. 이는 기존의 어휘의미중의성에 관한 연구를 비교 정리한 Sanderson의 이론이 웹 환경 하에서도 제대



〈그림 7〉 키워드 수 증가에 따른 정확률의 변화

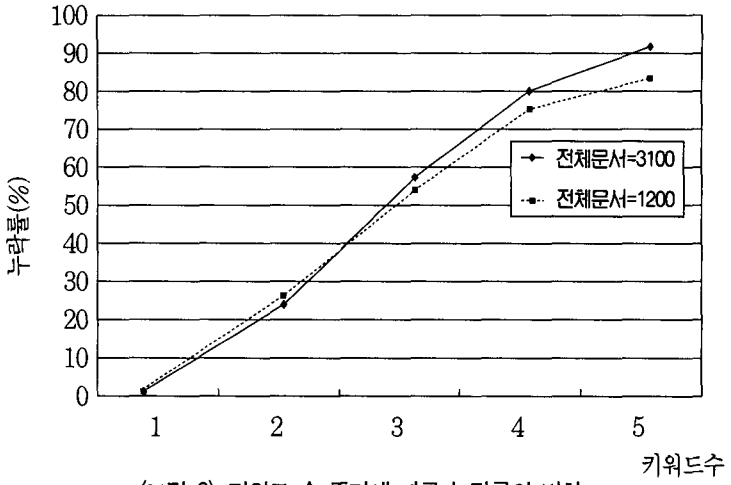


〈그림 8〉 키워드 수 증가에 따른 재현율의 변화

로 지켜진다는 것을 의미한다. 반면에 키워드수가 3개 이상 넘어서게 되면, 검색 재현율이 50% 이하로 낮아짐을 〈그림 8〉을 통해 살펴볼 수 있다. 또한 검색 키워드수가 5개 정도에 이르게 되면 검색 가능한 문서수가 원래의 10%정도 밖에 되지 않음으로서 무리한 검색이 될 수 있음을 확인할 수 있었다. 특히 검색 재현율은 평가 문서수나 평가자의 개인적 성향 차에 상관없이 거의 비슷한 수준의 감소율을 보였으며, 키워드 수 증가에 따른 정확률의 증가비율에 비해 상대적으로 높은 감소율을 보이고 있다. 이는 서지 데이터베이스에 비해 주제가 명확치 않은 대부분의 일반 웹 문서의 특성을 고려해 볼 때, 검색 키워드 수를 5개 이상 입력하는 방법은 과도한 필터링을 유발하는 비효과적인 검색 방법이 될 수 있음을 의미한다. 키워드 수 증가에 따른 정확률 증가에 비해 높은 감소율을 보이는 재현율의 저하는 전체적인 검색 만족도를 낮추는 역효과를 가져오기 때문이다.

검색 결과에 영향을 줄 수 있는 원인으로는 크게 세 가지를 고려해 볼 수 있다. 먼저 부적절한 키워드가 검색 결과에 영향을 줄 수 있으며, 검색식의 적절성 여부 역시 검색 결과에 영향을 준다. 마지막으로 검색어 수가 달라짐에 따라 검색된 문서수의 변화를 가져온다. 하지만 본 실험에서는 세 가지 검색 결과에 변화를 줄 수 있는 원인 중 마지막 원인만이 검색결과에 영향을 줄 수 있도록 조정하였다. 첫번째 원인인 키워드선정 과정은 SDCC알고리즘을 통해 유도되어지므로 부적절한 키워드가 선정되어지지 않으며, 검색식의 구성방법 역시 AND 연산자만을 이용하게 되므로 검색식의 구성방식이 검색 결과에 영향을 주지 않는 상황하에서 실험을 진행하였다.

검색어 수가 어휘의미중의성에 미치는 영향에 관해서는 이미 선행 연구자들에 의해 데이터베이스 환경 하에서 연구가 진행된 바 있으며, 새로운 정보검색 환경인 인터넷 환경 하에서도 검색어 수가 증가함에 따라 어휘의미중의성의 빈도가 감소함을 본 실험을 통해 확인할 수 있었다. 실험결과를 분석해 가는 과정에 있어서 주목할만한 새로운 사실을 확인할 수 있었는데, 키워드 개수 자체만으로도 인터넷검색 효율에 상당한 영향을



〈그림 9〉 키워드 수 증가에 따른 누락률의 변화

줄 수 있다는 것이다. 먼저 키워드 수가 너무 적은 경우에는 어휘의미중의성에 의해 검색 정확률이 낮아지고 이는 낮은 검색 만족도를 가져오게 된다. 대체적으로 키워드 수가 증가함에 따라 어휘 의미중의성에 의한 검색 정확률 저하의 가능성은 줄어들게 되는데, 이러한 상황을 고려해 볼 때 일반 이용자들이 인터넷기반 정보검색에 있어 적절한 키워드의 수는 3개 정도가 적당하다.

검색 키워드 수를 계속 늘릴수록 검색 정확률은 상당히 높은 수준까지 향상시킬 수 있지만, 이는 초보 웹 정보검색 이용자에게는 큰 부담으로 작용하며, 현실적으로 초보 이용자는 SDCC알고리즘과는 달리 키워드 수를 늘려 가는 과정에 있어 검색 목적과는 무관한 전혀 엉뚱한 단어를 키워드로 선정하여 쉽사리 '검색된 문서 없음'을 초래하기 쉽다. 또한 검색 키워드 수가 너무 많아지는 경우에는 실제 적합한 문서임에도 불구하고 부적합 문서로 판정되어 필터링 되어지는 누락률 〈그림 9〉의 가능성이 상당히 커지게 되며 결국 이 또한 검색 만족도를 낮추는 요인으로 작용하게 된다. 실험 결과 AND연산자로 묶을 수 있는 적절한 키워드 수의 범위가 2개에서 4개정도로 유효 범위 폭이 상당히 적으며, 키워드 수가 1개 증가 혹은 감소함에 따라 검색 결과가 상당히 달라질 수 있음을 확인하였다. 이는 키워드 수 증가에 따른 재현율의 급격한 감소에서 기인한다.

5 결 론

본 실험 결과에서는 인터넷기반 정보검색환경하에서도 기존의 Sanderson의 연구결과가 제대로 적용됨을 확인할 수 있었다. 또한 검색어 수가 증가함에 따라 재현율의 급격한 감소를 가져오는 인터넷기반 정보검색의 새로운 특성을 확인할 수 있었으며, 여러 가지 상황을 고려해 볼 때 일반 초보 이용자가 질의를 입력한 경우에는 검색어 수를 3개 정도로 유지하는 것이 검색 만족도 측면에서 가장 바람직한 검색 방식임을 확인할 수 있었다. 단일질의일 경우 어휘의미중의성에 의한 검색 효율의 저하를 가져오게 되며, 검색어 수가 5개 이상으로 너무 많아지게 되면 누락률의 급격

한 증가로 인하여 전체적인 검색 만족도가 낮아지게 된다.

새로운 정보검색 환경인 인터넷에서의 정보검색은 과거의 정보검색과는 여러 가지 면에서 상당히 달라진 모습을 보인다. 이는 과거의 정보검색 가이드라인이 새로운 인터넷 환경에서 동일하게 적용되어 질 수 없음을 의미하며, 보다 효과적인 인터넷 정보검색 수행을 위한 가이드라인을 제시하기 위해서는 무엇보다도 새로운 정보검색 환경인 인터넷의 특성에 대해 보다 많은 연구 및 조사가 선행되어 져야 한다. 이미 미국이나 유럽의 학계나 연구소, 기업체 등에서 인터넷이 활성화되기 시작한 1994년도 이후로 다양한 연구가 시도되었으며, 국내에서도 최근 수년 이래도 인터넷에 관한 여러 연구가 진행되어왔으나, 아직까지 대부분 초기 단계에 머물고 있는 실정이다. 인터넷에 관한 연구는 무엇보다도 인터넷이 하루가 다르게 발전해감에 따라 지속적인 연구가 필수적이며, 인터넷 정보 자원의 무한한 가능성을 고려해볼 때 보다 많은 연구 및 조사가 진행되어질 필요가 있는 것이다.

연구를 진행해가는 과정에서 많이 이용되어지는 인터넷 정보검색 관련 자료들을 참고(권군오, 유영기 1998; 홍승수 1997; 김윤식 1998)하여 실제 정보검색을 수행해본 결과 여러면에서 미흡한 점들이 발견되었으며, 인터넷 정보검색 수행시 발생하는 문제점들의 원인과 해결방안을 충분히 제시하지 못하는 것으로 확인되었다. 실제 검색 가이드라인에 맞추어 적절한 검색어의 선정, 검색식을 작성한 경우에도 그밖에 여러 가지 원인들에 의해 낮은 검색 만족 결과를 가져올 수 있다. 그런 원인으로 검색 엔진의 특성 같은 환경적 요인을 배제하더라도 검색 키워드 수 부족에 기인한 어휘의미중의성의 발생과 같은 개인적 요인이 검색결과에 상당한 영향을 미치게 되며, 인터넷 정보검색 수행 시에는 서지데이터베이스 검색에 비해 검색 키워드 수의 변화만으로도 검색 결과나 이용자의 검색만족도가 상당히 달라질 수 있다는 사실을 새롭게 확인할 수 있었다.

본 연구에서는 인터넷기반 정보검색 수행 시 발생할 수 있는 어휘의미중의성과 이에 의한 검색 효율 저하간의 상관관계를 조사하였다. 인터넷기반 정보검색에서는 웹 문서의 특성상 어휘의미중의성의 발생 빈도가 기존의 정보검색에 비해 상대적으로 높으며, 어떠한 방법으로든 어휘의미중의성 발생을 예방하는 것이 보다 나은 검색을 위한 방법임을 확인할 수 있었다. 이용자가 검색해 가는 과정에서 적절한 검색 키워드 수를 유지하는 것이 가장 효과적이나, 초보 이용자는 검색어 선정에 있어 실수하기 쉬우며 이 경우 검색엔진이 제공하는 유의어 확장기능과 같은 질의확장 도구를 활용하는 것이 보다 적절한 방법일 수 있다. 또한 웹 기반 정보 검색시스템에서 보다 나은 검색 효율성을 제공하기 위해서는 자동 질의 확장을 통해 검색키워드 수를 늘려 어휘의미중의성을 해소하는 것 역시 검색 정확률을 높이는 주요 방법일 수 있다. 또한 실험을 수행하는 과정에서 밝혀진 새로운 사실로 적절한 검색 키워드 수를 유지하는 것이 적절한 검색어의 선정이나 올바른 검색식 작성 못지 않게 효과적인 인터넷 정보검색을 수행하는 방법임을 확인할 수 있었다.

현재 실험에서는 실험 여건상 대상 키워드들을 DDC의 580번대인 식물 영역만으로 한정하여 테스트를 수행하였으며, 보다 정확한 평가 및 분석을 위하여 앞으로 대상 영역을 보다 확대하여 실험, 검토해야할 것이다. 본 연구에서는 실험대상을 검색시스템과 시스템을 이용하는 이용자로 그

범위를 한정하였으며, 주변 환경적 요소들에 대해서는 고려하지 않았다. 보다 전반적인 연구가 이루어지기 위해서는 여러 주변 환경적 요소들을 고려한 보다 광범위한 실험이 앞으로도 지속적으로 이루어져야할 것이다.

참 고 문 헌

권군오, 유영기. 1998. 『인터넷 정보검색사 한번에 끝내기』. 서울: 한글과 컴퓨터.

김윤식. 1998. 『인터넷 정보검색』. 서울: 21세기사.

박창호, 박민규, 이정모. 1998. 가이드라인이 인터넷 정보검색 수행에 미치는 영향. 『한국심리학회지: 실험 및 인지』, 10권 2호.

이재윤, 김태수. 1998. WordNet과 시소러스. 『제11회 언어정보 연찬회 발표논문집』. 1998년 2월 10일.

홍승수. 1997. 『인터넷 정보 검색사가 되자』. 서울: 삼양출판사.

황상규, 오경묵, 변영태, 1999. 어휘의미중의성이 인터넷기반 정보검색에 미치는 영향. 『제6회 한국정보관리학회 학술대회 논문집』

Abdulla, G., B. Liu, and E. A. Fox. 1998. "Searching the World-Wide Web: Implications from Studying Different User Behavior." WebNet98. 〈http://www.cs.vt.e d u/~nrg〉

Abdulla, G., Edward A. Fox and Marc Abrams. 1997. "Shared User Behavior on the World Wide Web." Association for the Advancement of Computing in Education (AACE). 〈http://www.cs.vt.edu/~nrg〉

Borgman, C. L., S.G. Hirsh, and J. Hiller. 1996. "Rethinking online monitoring methods for information retrieval systems." *Journal of the American Society for Information Science*, 47(7): 568-583.

EuroWordNet. 1998. "EuroWordNet: Building a Multilingual Database with WordNets for Several European Languages." University of Amsterdam Computer Centrum Letteren. 〈http://www.let.uva.nl/~ewn/〉.

Harabagiu, S. 1998. "Usage of WordNet in Natural Language Processing Systems." 〈http://www.ai.sri.com/~harabagi/coling-acl98/acl_work/acl_work.htm〉.

Krovertz, R, W. B. Croft. 1992. "Lexical Ambiguity and Information Retrieval." *ACM Tansactions on Information Systems*, 10(2): 115-141.

Miller, G. A., et al. 1990. "Introduction to WordNet: An On-line Lexical Database." *International Journal of Lexicography*, 3(4): 235-244.

Papka, R. and J. Allan. 1998. "Document Classification using Multiword Features." *Proceedings of the 1998 ACM 7th international conference on Information and knowledge management*: 124-131

Pollock, A. and Andrew Hockley. 1996 "What's

wrong with Internet searching?". In Designing for the Web: Empirical Studies. Seattle: Microsoft. ⟨www.microsoft.com/usability/webconf.htm⟩

Sanderson, M. 1994. "Word sense disambiguation and information retrieval." *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Devlopment in Information Retrieval. Dublin, Ireland*: 142-151.

Voorhees, E. M. 1993. "Using WordNet to disambiguate word sense for text retrieval." *Proceeings of ACM SIGIR Conference*, 16: 171-180.

Vossen, Piek. 1997. "EuroWordNet: a Multilingual Database for Information Retrieval." *Proceedings of DELOS Workshop on Cross-language Information Retrieval*: 715-728.

Weiss, S. F. 1973. "Learning to disambiguate." *Information Storage and Retrieval*, 9: 33-41.

Altavista. ⟨www.altavista.com⟩.

Lycos. ⟨www.lycos.com⟩.

Excite. ⟨www.excite.com⟩.

〔부록 1〕 식물명 31개

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

〔부록 2〕 SDCC알고리즘

〈Semantic Distance for Common Category〉

원 질의를 통해 검색된 문서에서 추출된 DocTerm d_{ti} 와 QueryTerm p_{tj}가
존재할 때 ($d_{ti} \neq p_{tj}$),

- Set of p_{tj}'s synsets
  $= \{p_{tj}:PS_1, p_{tj}:PS_2, \cdots, p_{tj}:PS_b, \cdots, p_{tj}:PS_n\}$
- Set of d_{ti}'s synsets
  $= \{d_{ti}:PS_1, d_{ti}:PS_2, \cdots, d_{ti}:PS_b, \cdots, d_{ti}:PS_m\}$

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

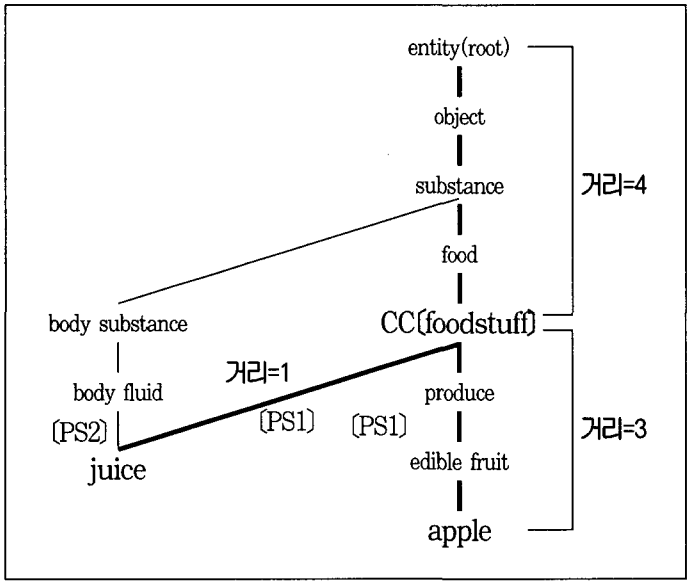
// 의미거리는 두 노드사이의 링크수의 합으로 계산되어 진다.
// Dy = 루트로부터 각각의 synset $d_{ti}: PS_a$, $p_{tj}: PS_b$까지의 의미거리
// dy = $C_{no}: PS_u$로부터 각각의 synset $d_{ti}: PS_a$, $p_{tj}: PS_b$까지의 의미거리
// root로부터 $C_{no}: PS_u$까지의 링크거리(distance)는 반드시 2보다 커야 한다.

- if (SDCC($C_{no}: PS_u$) > threshold θ)
  then SDCC($C_{no}: PS_u$) is valid
  else SDCC($C_{no}: PS_u$) is invalid

// threshold θ=0.5

예제) SDCC알고리즘

원 질의가 단일키워드 'apple'이고 검색된 문서 D에서 추출해낸 후보 키워드들 중 하나가 'juice'일 때 SDCC알고리즘을 이용하여 원 질의와 문서에서 추출한 후보 키워드간에 연관성의 정도를 계산하게 된다.



■ Set of pt1(apple)' s synsets
= { pt1:PS1, pt1:PS2, pt1:PS3, pt1:PS4 }
Set of dt1(juice)' s synsets
= { dt1:PS1, dt1:PS2}
이고, pt1:PS1와 dt1:PS1사이에 공통된 범주(Common Category)인 C1:PS1 'foodstuff'이 존재한다.

pt1:PS1와 dt1:PS1를 통해 연관성의 정도를 계산해보면,

■ $SDCC(C_1 : PS_1) = \frac{1}{2}\left(\frac{5-1}{5} + \frac{7-3}{7}\right) = 0.675 > 0.5$

따라서 두 어휘간의 연관성은 유효하며, 'juice'란 어휘는 대표키워드로 선정되어진다.