[UCI]I804:24011-200000232771

개념적 거리와 밀도를 기반으로 한 의미적 웹 문서 검색시스템

A Semantic Retrieval System for Web Documents based on Conceptual Distance and Density

2006 년 *8* 월

조선대학교 교육대학원

정보·컴퓨터교육전공

황 희 철

개념적 거리와 밀도를 기반으로 한 의미적 웹 문서 검색시스템

지도교수 김 판 구

이 논문을 교육학석사(정보·컴퓨터교육전공)학위 청구논문으로 제출합니다.

2006 년 *4* 월

조선대학교 교육대학원

정보·컴퓨터교육전공

황 희 철

황희철의 교육학 석사학위 논문을 인준합니다.

심사위원장 조선대학교 교수 인

심사위원 조선대학교 교수 인

심사위원 조선대학교 교수 인

2006 년 *6* 월

조선대학교 교육대학원

목 차

표 목차 ····· iii
그림 목차 ····· iv
ABSTRACT ····· v

Ⅰ. 서 론 ····· 1
A. 연구배경 ····· 1
B. 연구범위 및 내용 ····· 2

Ⅱ. 관련 연구 ····· 3
A. 문서검색 시스템을 위한 기반 기술 ····· 3
1. 문서 색인 ····· 3
2. 내용기반 문서 여과 ····· 4
3. 협동에 의한 정보 추천 ····· 6
4. 다중 에이전트에 의한 정보 검색 ····· 7
5. 사용자 관심 학습 ····· 8
B. 기존 인터넷 정보 검색 ····· 10
1. 검색엔진 ····· 10
2. 메타검색 엔진 ····· 11
3. 오프라인 브라우저 ····· 12
4. 추천 시스템 ····· 13
5. 뉴스 및 메일 여과 시스템 ····· 15
6. 푸시 서비스 ····· 16
C. 온톨로지 ····· 19

Ⅲ. 시스템의 제안 및 유사도 측정 ····· 21
A. 시스템의 구성 ····· 21
1. 교육정보 메타데이터 ····· 21
2. U-WIN ····· 22
B. 유사도 측정 ····· 26
1. 개념적 거리 측정 ····· 26
2. 개념적 밀도 측정 ····· 29

Ⅳ. 실험 및 결과 분석 ····· 32
A. 실험 내용 ····· 32
B. 결과 분석 ····· 35

Ⅴ. 결 론 ····· 38

참고문헌 ····· 39

표 목차

[표 1] 웹 문서간 개념적 거리를 통한 유사도 측정 결과 ····· 28
[표 2] 개념적 밀도에 의한 동음이의어 문제 해소 ····· 31
[표 3] '경제학' 관련 문서와 다른 문서간의 유사도 측정 결과(1) ····· 33
[표 4] '경제학' 관련 문서와 다른 문서간의 유사도 측정 결과(2) ····· 34

그림 목차

[그림 1] 다중 에이전트 시스템 ····· 7
[그림 2] 검색엔진 개념도 ····· 10
[그림 3] 메타검색 엔진 개념도 ····· 11
[그림 4] 내용기반 추천 시스템 ····· 13
[그림 5] 협동적 추천 시스템 ····· 14
[그림 6] 푸시 서비스 개념도 ····· 17
[그림 7] 온톨로지 관련 기술적 요소 및 응용 ····· 20
[그림 8] 제안한 웹 문서 검색 시스템의 개념도 ····· 21
[그림 9] U-WIN의 층위별 노드 수 ····· 23
[그림 10] U-WIN에서 단어별 의미 검색 ····· 24
[그림 11] U-WIN에서 계층 검색 ····· 24
[그림 12] U-WIN에서 '학문'에 대한 계층구조 ····· 25
[그림 13] '물리학'과 '운동학' 사이의 구조 ····· 27
[그림 14] 거리 및 유사도 측정 ····· 27
[그림 15] 단어 '수치'에 대한 의미 검색 ····· 29
[그림 16] 개념적 밀도 ····· 30
[그림 17] 동음이의어 단어의 예 ····· 30
[그림 18] '경제학' 관련 문서의 예 ····· 32
[그림 19] 메타데이터 추출 ····· 33
[그림 20] 검색엔진과의 실험 결과 비교 ····· 35
[그림 21] 검색엔진에서의 실험 환경 ····· 36
[그림 22] 검색엔진에서의 실험 예제 ····· 36
[그림 23] 정확률을 감소시키는 요인 ····· 37

ABSTRACT

A Semantic Retrieval System for Web Documents based on Conceptual Distance and Density

Heechul Hwang
Advisor : Prof. Pankoo Kim, Ph.D.
Major in Information and Computer Science Education
Graduate School of Education, Chosun University

Nowadays, there is so much quantity of lecture data acquired through the development and spread of internet techniques that we can easily get lecture data. We can obtain lots of data formed in various patterns ranging from simple text to multimedia data such as voice, movie and etc. But the data of custom templates can't be obtained easily using query in the web. So describing and organizing this vast amount of content is essential for realizing the full potential of the web as an information resource.

We should consider classification within larger context of subject-based metadata. Automatic classification is needed for at least two important reasons. The first is the sheer scale of resources available on the web and their ever-changing nature. And the second reason is that classification itself is a subjective activity. Different classification schemes are needed for different applications. Specific domain classification schemes, which can be quickly applied to large amounts of content using automatic methods, hold great promise for generating effective metadata.

If web content shifts from text-based pattern to multimedia oriented one, metadata will become even more important. Semantic interpretation of multimedia is incomplete without some mechanism in understanding semantic content which is not directly visible. For this reason, human assisted

content-annotation using natural language is one of the most common methods, particularly in multimedia retrieval applications, and it provides the means for exploiting syntactic, semantic as well as lexical information. A simple form of human-assisted semantic annotation is an attachment of textual description (such as keyword, or a simple sentence) to multimedia data. Problems in text (such as keywords) based image retrieval require exact remembrance of annotating words for retrieval and efficient management of semantic or concept clusters. More advanced technique like thesaurus-based term rewriting can be used to relieve the problem of traditional keyword-based information retrieval. However, keyword-based retrieval is still in the level of syntactic pattern matching. In other words, dissimilarity computation among terms is usually done by string matching not concept matching.

In this paper, we propose the way of classifying document by the subject using similarity measure or conceptual density based on subject classification of U-WIN from dispersed web data. Specially, we focus on similarity measure and conceptual density for automatic subject classification.

Ⅰ. 서 론

A. 연구배경

최근 인터넷 사용 증가와 관련 기술의 발달로 무수히 많은 정보들이 웹상에 분포하게 되었다. 이러한 웹 정보들을 사용자가 활용하기 위해 효율적이고 적합하게 검색하는 방법에 대한 필요성이 대두되면서 각 분야의 관련 연구가 진행되고 있고 관련된 시스템이나 검색 엔진들이 개발되고 있다.

그러나 현재의 검색기법은 텍스트 매칭(Text-matching)을 기반으로 동음이의어(homonym) 문제, 텍스트 자원이 구조화되지 않은 문제 외에 공통개념이 공유되지 않아 분산자료에서 정보통합의 어려움 등을 가지고 있어 정확한 검색 결과에 한계가 있다[1].

이런 문제들로 인해 차세대 지능형 웹인 시맨틱 웹이 등장하게 되었다. 시맨틱 웹 응용의 가장 중심적 개념 중 하나인 지식을 서술하는 온톨로지(Ontology)는 웹 기반의 지식 처리나 응용 프로그램 사이의 지식 공유, 재사용 등이 가능한 단어와 관계들로 구성된 일종의 사전으로서 생각할 수 있다[1, 2, 3]. 이러한 어휘 사전의 역할 이외에 지식을 효과적으로 표현하기 위해 정보에 의미를 부여하고 정보간의 관계를 설정하며 광범위한 도메인에 적용이 가능하도록 표준을 제시한다. 최근에는 의료나 전자, 기계 등의 특정 도메인에 관련된 단어들을 계층적으로 표현한 Domain Ontology와 WordNet 등 단어 간의 관계를 정의한 Linguistic Ontology 분야에 대한 연구가 활발히 진행되고 있다.

본 논문에서는 Linguistic Ontology의 일종인 울산대학교 어휘 지능망(UOU-Word Intelligent Network : U-WIN)을 기반으로 질의문서와 검색문서의 단어간 유사도를 측정하여 의미적으로 유사한 문서가 검색되도록 하는 시스템에 관하여 연구하고자 한다. 특히 검증을 위하여 분류 체계가 명확한 교육 분야의 웹 문서를 이용해 실험하였다.

B. 연구범위 및 내용

초기의 단순한 검색엔진에서 출발한 정보검색이 점차 각 사용자의 기호에 맞춰 정보를 찾고 제공하는 서비스로 변화하면서 검색엔진이나 푸시 서비스, 추천 시스템 등 상용화된 서비스나 프로그램들이 만들어지고 있다. 하지만 이러한 검색은 텍스트 매칭을 기반으로 하여 질의어와 문서 간의 키워드 출현 빈도 등에 의존하는 방식이므로 질의어와 의미적으로 유사한 문서에 대해서는 검색할 수 없다. 또 '한 언어로서 발음은 동일하나 의미가 다른 두 개 이상의 단어'인 동음이의어 문제 역시 해결되어야 할 과제로 남아 있다. 예컨대 질의어가 '스포츠'인 경우 검색될 문서의 내용에 '스포츠'라는 단어가 없더라도 '축구'나 '야구'와 같은 단어가 있다면 '스포츠'와 유사한 개념으로 보고 검색이 되어야 할 것이다. 또한 문장에서 '수학'이라는 단어가 존재할 경우 '물의 현상에 관하여 연구하는 학문', '학문을 닦음', '수량 및 공간의 성질에 관하여 연구하는 학문' 중 어느 것을 의미하는지의 문제가 발생된다. 이에 본 연구에서는 이러한 검색의 문제점들을 해결하여 사용자가 원하는 검색결과에 한층 더 다가가도록 개념적 접근 방안을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 지능형 정보검색을 위해 일반적으로 많이 사용되고 있는 기반 기술들과 기존의 인터넷 정보검색 및 온톨로지에 대해 고찰한다. 3장에서는 교육정보 메타데이터(MetaData), 울산대학교 어휘 지능망(U-WIN) 등을 이용하여 본 논문에서 제안하고자 하는 문서 검색 시스템의 전체적인 구성과 웹 문서에서 U-WIN을 기반으로 개념적 거리와 밀도를 이용하여 유사도 측정 및 동음이의어 문제를 해결하는 방법에 대해 기술한다. 4장에서는 의미적으로 유사한 웹 문서의 검색이 가능하도록 메타데이터 정보를 추출하여 개념간 유사도를 측정하는 실험을 실시하고, 마지막 5장은 본 연구에 대한 결론과 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

본 장에서는 기존의 문서검색 시스템의 기반 기술들을 고찰하고, 보다 편리하고 정확하게 정보를 수집하고 제공하기 위한 현재의 다양한 검색 시스템들에 대해 살펴보며, 특정하게 명시 되지 않은 데이터를 유추할 수 있는 온톨로지에 대해 기술한다.

A. 문서검색 시스템을 위한 기반 기술

지능적인 정보검색을 위해 많이 사용되고 있는 기반 기술들은 다음과 같다. 텍스트를 이루고 있는 단어들을 추출하여 통계적인 데이터를 산출하는 문서 색인 방법과 텍스트를 통계적으로 분석하는 내용기반 여과 방법, 유사한 취향의 다른 사용자가 선호하는 아이템을 추천하는 방식인 협동에 의한 정보 추천, 사용자를 대신해 작업을 수행하는 소프트웨어 에이전트, 그리고 사용자로부터 피드백을 얻는 방식과 사용자 프로파일을 변경하는 알고리즘에 따라 구분할 수 있는 사용자 기호에 대한 학습 방법 등이 있다.

1. 문서 색인

현재 정보검색의 대부분은 검색하려는 웹 페이지나 뉴스 메시지 등의 내용을 나타내는 텍스트를 비교 분석하여 이루어진다. 텍스트를 비교하기 위해서는 우선 텍스트를 이루고 있는 단어들을 추출하여 통계적인 데이터를 산출해야 하는데 이 과정을 색인이라고 한다.

문서에서 단어들을 추출하기 위해서는 사전을 사용하며 영문의 경우에는 간단한 stemming을, 국문의 경우에는 형태소 분석 과정을 거친 뒤 각 단어에 그 중요도에 따라 가중치를 부여한다[4, 25]. 단어에 가중치를 주는 방법으로는 각 단어의 문서

내 출현빈도(TF, term frequency)를 정규화 하여 사용하는 방법이 있다. 유한개의 키워드를 미리 선정하여 문서를 표현하는 경우에는 표본 문서 집단으로부터 역문헌빈도(IDF, inverted term frequency)를 계산하여 출현 빈도와 함께 사용하기도 한다. 이 경우 문서 j에 있는 단어 i의 가중치는 다음의 식으로 표현된다.

$$W_{ij} = freq_{ij} * (\log \frac{N}{DF_i} + 1)$$

여기서 $freq_{ij}$는 단어 i의 문서 j에서의 출현 빈도, N 은 총 표본 문서의 개수, DF_i는 단어 i를 포함하는 문서의 개수이다. 역문헌빈도의 의미는 모든 문서에 고르게 출현하는 단어는 검색을 위한 색인어로서의 가치가 낮으므로 낮은 가중치를 부여하는 것이다.

텍스트에서 단어들을 추출하는 작업 외에도 문서의 구조에 따라 각 문서를 특징지을 수 있는 요소들을 추출하는 과정이 필요하다. 예를 들면, 전자우편의 경우 Sender, Subject 등의 필드를 인식하거나, 웹 페이지의 경우에 HTML tag를 분리해 내는 작업을 수행하여야 한다[5].

2. 내용기반 문서 여과

사용자가 원하는 정보를 선택적으로 제공하기 위해 가장 일반적으로 쓰이는 것은 텍스트를 통계적으로 분석하는 내용기반 여과 방법이다[6]. 새로운 문서인 e-mail message, news article 또는 web page가 사용자의 관심인 user interest에 부합하는가를 판별하는 것은 사용자의 관심과 일치하는 문서의 카테고리를 'YES', 일치하지 않는 카테고리를 'NO'라고 했을 때 두 개의 카테고리를 갖게 된다.

내용기반 방법은 사용자의 관심과 일치하는 문서들과 일치하지 않는 문서들의 샘플로부터 단어들을 추출하여 사용자의 관심을 학습하고, 선택 대상 문서를 단어들과 각 단어의 가중치들로 나타낸 뒤 학습된 사용자 관심을 이용하여 'YES', 또는

'NO'로 분류하는 것이다.

내용기반 여과 방법으로 가장 대표적인 것은 벡터 유사도를 이용한 방법이다. 단어들을 추출하고 가중치를 계산하면 사용자의 관심과 선택 대상 문서를 모두 단어 가중치의 벡터로 표현할 수 있고, 이들 사이의 유사도를 두 벡터 사이의 각도를 나타내는 cosine similarity를 이용하여 계산한다. 문서의 표현에 총 n개의 단어가 사용되고, 사용자의 관심을 나타내는 단어 가중치 벡터를 P(profile), 선택대상 문서를 나타내는 단어 가중치 벡터를 D(document)라고 하면 이 두 벡터의 유사도는 다음과 같다.

$$Similarity\ (P,\ D) = \frac{P \cdot D}{|P|\ |D|}$$

사용자의 관심과 일치한다고 알려진 문서들의 집합을 S^+, 일치하지 않는다고 알려진 문서들의 집합을 S^-라고 가정한다. 벡터 유사도를 이용한 여과는 S^+의 평균 벡터로 사용자의 관심 P를 나타내고, 선택대상 문서 D는 P와의 유사도가 일정한 임계값(threshold) T를 넘으면 D를 사용자의 관심에 부합하는 것으로 판단하는 것이다.

내용기반으로 문서를 여과하는 또 다른 방법은 베이지안 확률을 이용하는 것이다. 이 방법은 S^+와 S^-로부터 확률 $P(Y)$, $P(N)$, $P(W_i|Y)$, $P(W_i|N)$ 등의 값을 추정한 뒤 단어 $W_1, W_2, \ldots, W_n$ 이 나타난 문서 D가 Y(YES) 또는 N(NO)으로 분류될 확률을 다음과 같이 비교하여 확률값이 높은 쪽으로 D를 분류한다.

$$P(Y|W_1,\ \cdots,\ W_n) = k * P(Y) * P(W_1|Y) * \cdots * P(W_n|Y)$$
$$P(N|W_1,\ \cdots,\ W_n) = k * P(N) * P(W_1|N) * \cdots * P(W_n|N)$$

유한개의 키워드와 기타 문서의 특성을 나타내는 요소들로 N개의 특징을 정의하여 문서를 표현하는 경우, 문서의 여과에 사례기반 추론(case-based reasoning 또

는 memory-based reasoning), 결정 트리(decision tree), 신경망(neural network) 등을 사용할 수도 있다.

사례기반 추론 방법은 D와 S^+, S^-의 모든 문서들과의 유사도를 계산하여 가장 유사한 문서가 S^+에 속한 것이면, D도 사용자의 기호에 맞는 것으로 판단한다. 결정 트리를 이용하는 방법은 S^+와 S^-의 문서들을 training set으로 하여 사용자의 관심을 판별하는 트리를 학습하고 학습된 결정 트리를 이용하여 새로운 선택 대상 문서를 'YES' 혹은 'NO'로 판별한다. 신경망을 이용하는 경우에는 N개의 특징이 input layer, 'YES'/'NO'가 output layer를 형성하는 multi-layered backpropagation network를 구성하고 S^+와 S^-의 문서들을 training set으로 하여 학습을 수행한다.

3. 협동에 의한 정보 추천

협동에 의한 정보 추천은 대상 문서들의 내용을 비교하는 대신에 사용자들 사이의 기호를 비교하는 것이다. 즉, 사용자가 이전에 선호한 정보를 토대로 새로운 아이템을 추천하는 것이 아니라 유사한 취향의 다른 사용자가 선호하는 아이템을 추천하는 방식이다[7].

예를 들면, 음반을 추천하는 경우 각 사용자는 몇 가지 음반에 대하여 "좋아한다, 싫어한다, 보통이다 등"의 등급을 지정한다. 추천 시스템은 다양한 음반에 대한 각 사용자의 응답을 이용하여 사용자들을 클러스터링 한다. 만일 사용자 A와 B가 같은 클러스터에 있고 A는 음반 1, 2, 3, 4에 대하여 좋다는 응답을 했으며, B는 1, 2, 3, 5에 대하여 좋다고 했다면, 사용자 A에게는 음반 5를, 사용자 B에게는 음반 4를 추천하는 것이다.

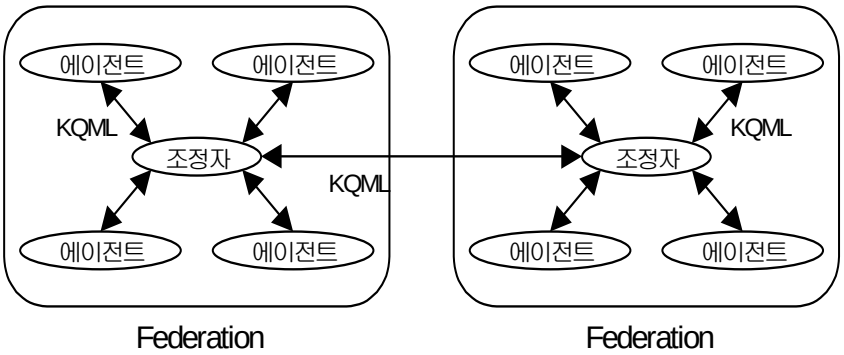
이를 위해 비슷한 취향의 사용자들을 그룹화 하는 작업이 필요하다. 기호가 같은 사용자를 찾는 방법으로는 각 사용자의 특성을 추천하려는 정보의 여러 속성으로 표현하고 유사한 문서를 계산하는 방법, 확률적으로 사용자를 클러스터링 하는 방법, 에이전트 사이의 정보 교환에 의해 같은 기호의 사용자 정보를 점진적으로 획득하는 방법 등이 사용되고 있다.

협동에 의한 정보 추천은 인간이 판단한 정보를 직접 이용하기 때문에 비문자 정보를 추천할 수 있다는 상대적 장점이 있지만, 그룹 내의 어떤 사용자도 보지 못한 완전히 새로운 아이템이 나타났을 때 추천이 불가능하며, 독특한 기호를 지닌 사용자를 수용할 적절한 방법이 없다는 단점이 있다.

4. 다중 에이전트에 의한 정보 검색

사용자를 대신하여 작업을 수행하는 소프트웨어 에이전트에 대한 연구는 최근 들어 많은 관심이 집중되는 분야이며, 인터넷에서의 정보검색에 에이전트 기술을 활용하기 위한 연구도 활발하게 이루어지고 있다[26].

에이전트의 속성, 에이전트의 구현에 이용되는 기술, 에이전트의 구조, 에이전트의 응용분야 등은 매우 다양하기 때문에 에이전트를 한마디로 정의하긴 어렵다. 많은 연구자들은 에이전트가 갖추어야 할 속성들에 대해 공통적인 성질들을 기술하고 있다. 에이전트의 대표적인 성질로는 사용자의 간섭을 받지 않고 사용자로부터 위임받은 일을 자율적으로 처리하는 자율성, 다른 에이전트나 사람과 상호작용을 하는 사교성, 추론과 학습 능력을 갖추어 새로운 지식을 습득하는 지능, 자신이 추구하는 목적을 달성하기 위한 능동적인 행동을 하는 선행위성, 사용자의 습관이나 작업 방식 및 취향에 따라 스스로를 변화시키는 적응성, 네트워크상에서 이동할 수 있는 능력인 이동성 등이 있다[8].



[그림 1] 다중 에이전트 시스템

인터넷에서와 같이 분산되어 있고 이질적인 정보 소스들로부터 다양한 매체의 자료를 검색하기 위해서 이러한 에이전트들이 각자의 목적에 따라 네트워크의 여러 곳을 돌아다니며 다른 에이전트와 협력하여 정보를 교환할 수 있는 다중 에이전트 환경이 연구되고 있다. [그림 1]은 이러한 다중 에이전트 시스템의 개념을 보여주고 있다.

다중 에이전트 환경에서 에이전트는 조정자라고 불리는 별도의 에이전트를 통해 다른 에이전트들과 통신한다. 하나의 조정자와 그에 등록된 에이전트들은 한 연방(federation)을 이루어 모든 에이전트들은 통신에 관한 자율성을 모두 그들의 조정자에게 위임하며 조정자는 그들을 관할한다. 다른 연방에 속한 에이전트와의 통신 또한 이들 조정자들을 통해서 이루어진다.

조정자는 에이전트 생성 시 에이전트의 식별자와 기능을 등록 받고 에이전트로부터 통신 메시지를 받으면 그 요구에 적합한 에이전트를 찾아 메시지를 전달한다. 필요에 따라 조정자는 메시지를 분할하여 여러 에이전트로 보내기도 하고 여러 메시지를 합성하기도 한다. 조정자라는 추가적인 에이전트의 이용은 에이전트끼리 직접 통신할 때 모든 에이전트가 다른 에이전트에 대한 정보를 갖고 있어야 하며 통신 협상도 직접 해야 하는 등의 부하를 감소시킬 수 있다.

이질적인 에이전트 사이의 통신을 위해서는 표준화된 에이전트 통신언어가 필요한데 ARPA의 KSE (Knowledge Sharing Efforts)에서 제안한 KQML (Knowledge Query and Manipulation Language)이 널리 이용되고 있다[27].

5. 사용자 관심 학습

정보검색 시스템에서의 학습이란 명시적, 혹은 묵시적으로 획득된 사용자의 피드백으로부터 사용자의 관심을 나타내는 프로파일을 갱신하여 사용자의 관심을 점차 정확하게 표현하는 것이다. 사용자의 관심을 학습하는 방식은 사용자로부터 피드백을 얻는 방식과 사용자 프로파일을 변경하는 알고리즘에 따라 구분할 수 있다.

사용자로부터 피드백을 얻는 방식은 크게 명시적(explicit) 방법과 암시적

(implicit) 방법으로 나누어 볼 수 있다. 명시적 피드백이란 사용자가 시스템이 제시한 특정 문서에 대해 직접 자신의 관심과 일치하는지 여부를 입력하는 방식으로, Yes/No의 두 단계 혹은 그 이상의 단계로 등급을 입력하는 것이다. 이에 반하여 암시적 피드백은 시스템이 사용자의 행위를 관찰하여 특정 문서에 대한 사용자의 관심 여부를 판별하는 것으로, 사용자의 파일 삭제 및 이동 행위, 사용자가 특정 문서를 읽는데 소비한 시간, 또는 웹 페이지에서 특정 링크를 클릭하는 행위 등을 바탕으로 사용자의 관심 여부를 추정한다.

사용자 프로파일을 변경하는 학습 알고리즘은 문서 여과 방식과 직접 관련이 있다. 일반적으로 많이 이용되는 벡터 유사도를 사용하는 경우 적합성(relevance) 피드백에 의한 단어 가중치 변경 방법이 사용된다. 적합성 피드백에 의한 사용자 프로파일 학습은 사용자가 관심과 일치한다고 반응한 문서에 있는 단어들의 가중치는 높이고, 일치하지 않는다고 반응한 문서에 있는 단어들의 가중치는 낮추어 프로파일의 내용을 변경하는 것이다. 현재의 사용자 프로파일 벡터를 P, 사용자의 기호와 일치하는 문서들의 벡터를 D^{+}, 사용자의 기호와 일치하지 않는 문서들의 벡터를 D^{-} 라고 하면, 학습에 의해 변경되는 새로운 프로파일 P'는 다음과 같이 계산된다.

$$P' = P + \frac{\alpha}{N} \sum D^{+} - \frac{\beta}{M} \sum D^{-}$$

위의 식에서 N과 M은 각각 D^{+}와 D^{-}의 개수이며, α와 β는 적합한 문서와 부적합한 문서에 대한 학습 가중치로, 이 값들의 조정함으로써 새로운 피드백에 대하여 학습하는 성향을 조정할 수 있다.

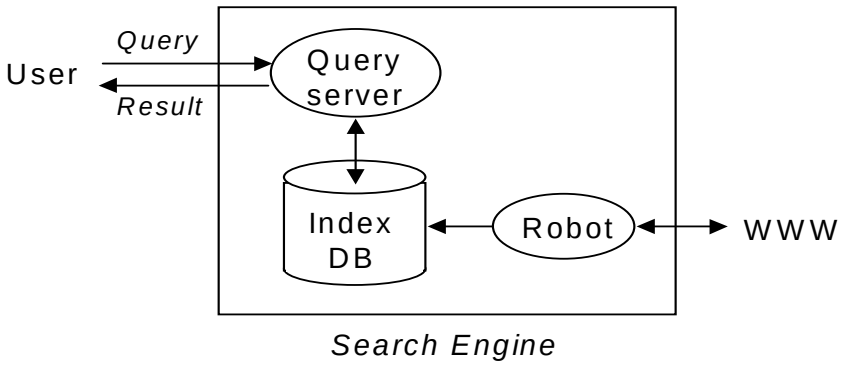
문서 여과에 결정 트리를 이용하는 경우 ID3 알고리즘, 신경망을 사용하는 경우 backpropagation 알고리즘 등이 학습에 사용되며, 사례 기반 추론이나 협동적 여과 방식을 사용하는 경우는 새로운 사례가 추가되는 것이 학습의 역할을 한다고 볼 수 있다.

B. 기존 인터넷 정보검색

기존의 인터넷 정보검색은 정보검색을 위해 개발된 시스템들의 형태와 용도에 따라 검색엔진, 메타검색엔진, 오프라인 브라우저, 추천 시스템, 메일 및 뉴스 필터링 시스템, 푸시 서비스 등으로 나누어 볼 수 있다.

1. 검색엔진

검색엔진은 웹에서의 정보검색을 위해 가장 먼저 시도되었으며 현재도 가장 보편적으로 사용되고 있는 시스템으로서 기본 구조는 [그림 2]와 같다. [그림 2]에서 로봇과 같이 spider, wanderer, crawler, warm 등은 검색엔진에서 웹 페이지들을 모아오는 역할을 하는 프로그램을 말한다.



[그림 2] 검색엔진 개념도

로봇은 인터넷에 있는 각각의 웹 서버에 직접 접근하여 웹 페이지들을 검색엔진으로 가져오며, 모아온 페이지들로부터 색인어들을 추출하여 인덱스 데이터베이스를 구축한다.

사용자는 웹 브라우저를 통해 검색엔진에 접속하고, 단어의 집합 형태로 질의를 던지면 검색엔진은 인덱스 데이터베이스를 참조하여 해당하는 URL들의 리스트를 보여준다. 로봇의 기본적인 형태는 특정한 웹 사이트들로부터 시작하여 하이퍼링크

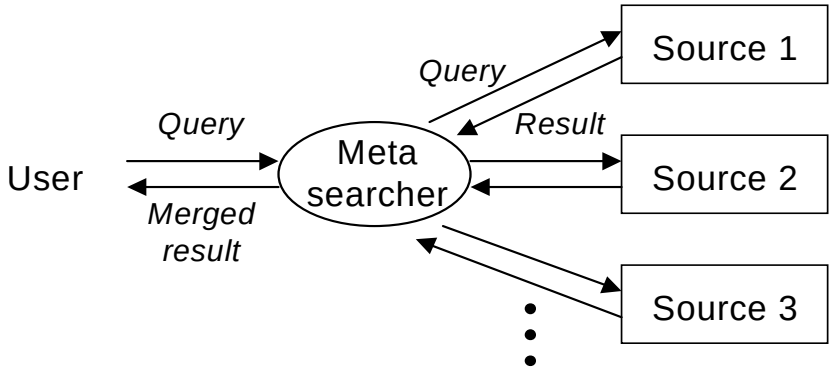
를 따라가면서 새로운 웹 페이지들을 가져오는 것으로, 하나의 웹 페이지를 가져오면 그 페이지 내에 있는 모든 URL들을 queue에 저장하고 다음 URL을 queue에서 선택하여 웹 페이지를 가져오는 작업을 반복한다[5].

색인 대상으로는 웹 페이지의 타이틀이나 처음의 일부 텍스트만을 이용하는 경우도 있고, 웹 페이지 안의 전체 텍스트를 색인하는 경우도 있으며, 색인어에 복잡한 통계적 가중치를 주는 시스템들도 있다.

검색엔진의 공통적인 문제점은 단순 질의에 대한 검색 결과가 지나치게 많고 정확도가 떨어진다는 것과, 로봇에 의한 색인은 텍스트에 국한되어 있고, 이미지와 같은 정보는 검색하지 못한다는 것이다.

2. 메타검색 엔진

메타검색 엔진은 검색엔진의 로봇과 달리 직접 웹 서버들에 접근해서 웹 페이지들을 모아오지 않고, 사용자가 질의를 주면 다수의 검색엔진에 질의를 던져 그 결과를 모아서 정리하여 사용자에게 제시하는 시스템이다[9]. [그림 3]은 메타검색 엔진의 개념을 나타낸다.



[그림 3] 메타검색 엔진 개념도

메타검색을 하기 위해서는 질의를 던질 대상 검색엔진의 URL, 각 검색엔진의 질

의 형태, 검색 결과의 양식 등을 알고 있어야 하며, 각 검색엔진의 검색 결과를 취합하여 중복을 제거하고 유사도에 따라 순위를 정하는 알고리즘이 정의되어야 한다.

웹 페이지를 검색하기 위해 검색엔진에 질의를 던지는 것과 비슷하게 쇼핑을 위한 상품 정보를 검색하기 위해 다수의 인터넷 쇼핑몰에 접속하여 정보를 찾아오는 시스템들도 있다. 또 사용자가 검색하고자 하는 상품의 종류와 사양을 선택하면 인터넷 쇼핑몰들로부터 관련 상품의 정보를 찾아 종합하여 보여준다.

3. 오프라인 브라우저

메타검색엔진과 비슷한 역할을 하는 독립적인 프로그램으로 사용자를 대신하여 백그라운드에서 사용자의 관심에 맞는 정보를 수집하여 저장하는 프로그램을 오프라인 브라우저 또는 정보 수집 에이전트라고 하며, WebZip, WebCompass, WebTamer, Agentware 등 다수의 상용화된 프로그램들이 있다.

WebZip은 사용자가 지정하는 웹 사이트에서 지정된 깊이까지 모든 웹 페이지를 수집하여 사용자의 하드디스크에 저장하는 역할을 한다. 사용자는 작업을 지시해 놓고 작업이 끝난 후 오프라인으로 빠르게 수집된 웹 페이지들을 검색할 수 있다.

WebCompass, WebTamer 등은 사용자의 질의를 받아 다수의 검색엔진과 ftp 사이트, Gopher 사이트 등에 접속하여 정보를 가져오고 결과를 취합하여 제공한다. 이들은 단순한 정보 수집 외에도 웹 사이트의 변화를 알려주는 등의 다양한 기능을 제공한다.

Agentware는 사용자의 자연언어 질의를 분석하여 패턴 매칭과 신경망에 의해 수집된 정보를 여과하며, 사용자의 행위를 관찰하여 학습을 수행한다.

Shopping Explorer는 오프라인으로 상품 정보를 찾아오는 쇼핑 에이전트로서 사용자가 검색하고자 하는 상품 정보에 대해 다수의 인터넷 쇼핑몰에 접속하여 정보를 수집해 제공하는 역할을 한다.

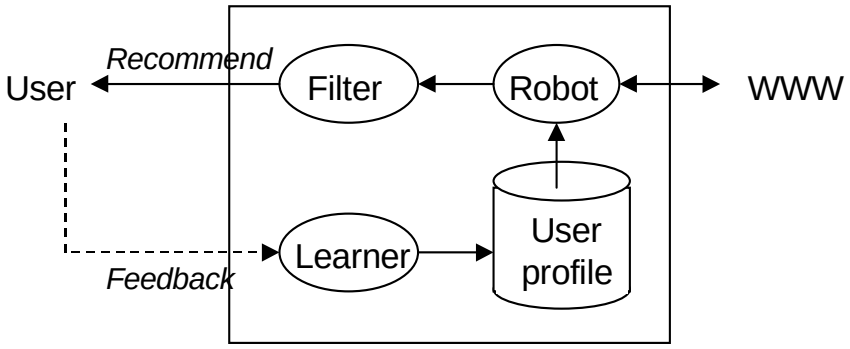
4. 추천 시스템

추천 시스템(recommendation system)은 지능형 에이전트로 사용자의 관심 분야를 스스로 파악하여 학습하고, 사용자의 관심 분야에 부합하는 정보들을 수집하여 제공한다. 대부분 학습을 위하여 사용자로부터 추천된 정보에 대한 관련 여부를 피드백으로 받는다. 정보 추천 에이전트는 추천 방식에 따라 브라우징 어시스턴트, 내용기반 추천 시스템, 협동적 추천 시스템 등이 있다[10, 11].

브라우징 어시스턴트는 사용자가 웹 페이지를 읽고 있는 동안 그 페이지 안의 링크들 중 사용자의 관심과 부합하는 링크를 추천하는 시스템으로, WebWatcher, Letizia 등이 있다.

WebWatcher는 카네기멜론 대학에서 만든 에이전트로, 사용자가 링크를 선택하는 행위를 관찰하여 사용자의 관심 분야를 학습한다. 사용자가 선택한 링크들을 학습 예제로 하여 단어들과 각 단어에 대한 가중치로 사용자의 관심 분야를 나타내고 새로운 링크들에 대하여 벡터 유사도에 의해 사용자의 관심과 부합 여부를 결정한다. Letizia는 웹 브라우저 위에서 동작하는 프로그램으로, 사용자가 보고 있는 페이지의 링크들을 미리 가져와서 평가하고, 사용자가 도움을 요청할 경우 링크를 추천한다.

내용기반 추천 시스템은 특정한 웹 페이지에 대한 사용자의 평가를 기초로 하여 사용자의 관심을 학습하고 새로운 웹 페이지들을 수집하여 제공하는 것으로 [그림 4]와 같다.



[그림 4] 내용기반 추천 시스템

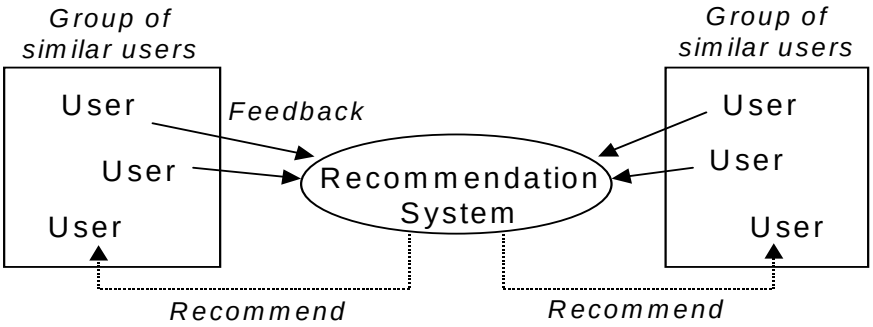
그밖에 Syskill&Webert, Fab, Amalthaea 등의 시스템이 있다. Syskill&Webert는 기본적으로 제공하는 인덱스 페이지의 일부 링크에 대한 사용자의 평가를 기초로 사용자의 관심을 학습한다. 학습을 위해 사용자가 평가한 페이지들로부터 특징 벡터를 추출하여 베이지안 분류, 유사한 문서, 결정 트리, neural network 등 다양한 방법으로 학습을 수행하고 그 결과에 따라 웹 페이지를 추천한다.

사용자 관심에 부합하는 링크를 추천하는 방식은 인덱스 페이지의 링크 중 나머지들을 추천하는 방식과, 사용자 관심을 질의로 표현하여 Lycos 검색엔진에 질의를 던져 새로운 웹 페이지를 수집하는 두 가지 방식으로 동작한다.

Fab은 다수의 사용자들을 관심 분야별로 그룹화 하여 내용기반 추천과 협동적 추천을 복합적으로 사용한다. Fab은 주제별로 웹 페이지를 모아오는 메타서치 형태의 collection agent와 사용자별로 웹 페이지를 제공하는 selection agent로 구성되어 있으며, 사용자별로 관심을 나타내는 프로파일을 구성하고 추천된 페이지에 대한 사용자의 평가에 따라 프로파일에 있는 단어들의 가중치를 수정함으로서 학습을 수행한다.

내용기반 추천은 벡터 유사도를 기반으로 하며, 협동적 추천은 한 사용자가 높은 평가를 한 웹 페이지를 같은 그룹에 있는 다른 사용자들에게도 추천함으로써 이루어진다.

Amathaea는 다수의 에이전트가 웹 페이지의 수집과 여과를 수행하며 유전자 알고리즘을 적용하여 추천 정확도가 높은 우수한 에이전트가 진화 과정을 거쳐 발전하도록 하였다.



[그림 5] 협동적 추천 시스템

협동적 추천 시스템은 웹 페이지나 뉴스 메시지 등의 텍스트를 분석하지 않고, 다수의 사용자를 그룹화 하여 일부 사용자들로부터 높은 평가를 받은 정보를 다른 사용자들에게도 추천하는 것으로, 내용기반 추천이 어려운 음악, 영화 등의 추천에 적합하며, [그림 5]와 같다.

Siteseer와 같은 시스템은 사용자들의 북마크에 있는 URL만을 이용하여 유사한 북마크 폴더들을 찾고, 유사한 북마크를 가진 사용자들 사이에 상호 URL을 추천한다.

Phoaks는 USENET의 각 뉴스그룹에 관련 URL을 추천하는 시스템으로, 웹 사이트를 추천하는 내용의 뉴스 메시지들을 인지하여 메시지 안에 언급된 URL들을 수집하고 언급된 횟수가 많을수록 관련 있는 웹 페이지로 간주하여 추천한다.

FireFly는 대표적인 협동적 추천 시스템으로, 영화나 음반에 대한 사용자들의 평가를 기초로 비슷한 취향의 사용자 그룹을 만들고, 같은 그룹 내의 사용자들로부터 높은 평가를 받은 영화나 음반을 새로운 사용자에게 추천한다.

WiseWire는 영화나 음반에 대한 정보뿐만 아니라 웹 페이지, 뉴스 메시지, ftp 사이트 등 다양한 소스로부터의 정보를 협동적 방식과 내용기반 방식을 복합적으로 사용하여 추천하는 시스템으로, 사용자의 평가를 기초로 한 신경망 학습을 사용한다.

5. 뉴스 및 메일 여과 시스템

메일 여과 시스템은 사용자가 전자우편을 읽고, 저장하고, 삭제하는 행위를 학습하여 사용자를 대신해 불필요한 메일을 자동으로 삭제하거나, 긴급한 메일의 도착을 알리거나, 또는 메일들을 분류하는 지능형 에이전트로서, 학습의 기능은 없으나 전자우편을 자동으로 여과하는 필터를 지정할 수 있다.

Information Lens나 Netscape mail은 사용자가 입력하는 명시적인 룰에 의해 메일을 여과하고 분류하는 시스템으로 학습의 기능은 없으며 메일의 각 필드의 텍스트에 스트링 매칭을 하여 룰을 적용시킨다.

뉴스 여과 시스템은 수만개에 달하는 USENET 뉴스 그룹과 매일 수백개 이상씩 올려지는 각 뉴스 그룹의 메시지들로부터 사용자의 관심에 부합되는 메시지들만을 선별하여 제공하는 시스템을 말한다.

GroupLens와 같은 시스템은 협동적 방법을 사용하여 각 메시지에 대한 뉴스 그룹 이용자들로부터의 평가를 기초로 높은 평가를 받은 메시지들만을 여과하는 시스템이다.

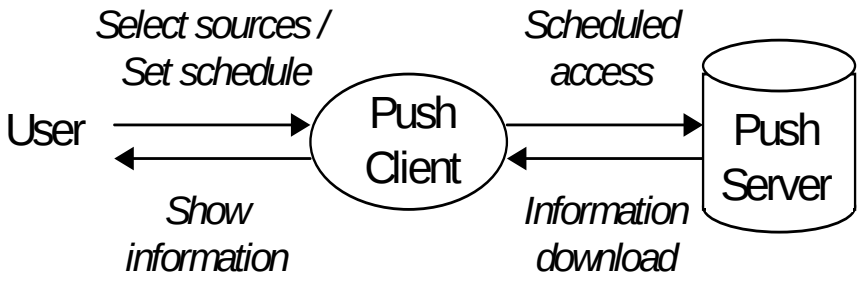
NewT는 내용기반 시스템으로, 사용자는 우선 적합한 뉴스 메시지와 부적합한 뉴스 메시지의 예들을 제공하여 시스템을 초기화 한다. 이후 NewT는 초기화된 정보를 바탕으로 계속적으로 뉴스 메시지를 여과하여 전달하고, 사용자는 NewT에 의해 여과된 메시지들에 대해 적합 또는 부적합 판정을 내린다. NewT에 의한 메시지의 여과는 단어들을 기반으로 한 벡터 유사도를 사용하며, 사용자의 반응에 따라 프로파일의 가중치를 변경함으로서 학습을 수행하여 사용자와의 상호작용이 계속됨에 따라 여과의 정확도가 향상된다.

6. 푸시 서비스

사용자가 브라우저를 통해 필요한 정보를 서버로부터 가져오는 기존의 방식을 정보를 끌어온다는 의미에서 풀(pull) 모델이라고 할 때, 이에 반해 정보를 갖고 있는 서버 측에서 정보를 밀어 주는 형식을 푸시(push) 모델이라고 한다[12].

푸시는 netcasting, webcasting, personal information delivery 등 다양한 용어로 표현되고 있다. 푸시 서비스는 1996년 PointCast가 뉴스 서비스를 무료로 전파하면서 본격적으로 소개되었지만 그 이전부터 사용하던 브라우저의 What's New나 What's Cool과 같은 기능은 서버가 사용자에게 능동적으로 정보를 제공했다는 점에서 푸시의 일종이라고 할 수 있다.

일반적으로 푸시 서비스를 이용하려면 우선 푸시 서비스를 제공하는 서버로 가서 사용자 등록을 한 후 거기서 제공하는 푸시 클라이언트 프로그램을 다운로드 받아 설치해야 한다.



[그림 6] 푸시 서비스 개념도

푸시 클라이언트 프로그램을 설치하고 나면 사용자는 자신이 원하는 정보의 종류와 그 정보들을 가져오는 주기 등의 환경설정을 할 수 있다. 푸시 클라이언트는 사용자의 환경 설정에 따라 정해진 시간에 서버에 접속하여 지정된 정보를 받아 놓으면 사용자는 자신이 편리한 시간에 저장된 정보를 볼 수 있다. 푸시 서비스의 구성도는 [그림 6]과 같다.

푸시 서비스를 이용하는 순서를 보면 정보 전달 방식이 클라이언트가 서버에 요청하여 클라이언트가 정보를 끌어오는 형태로서, 엄밀한 의미에서는 정해진 시간마다 자동으로 클라이언트가 서버로부터 정보를 끌어오는 자동화된 풀 혹은 폴링 푸시라고 할 수 있다. 이 과정은 사용자에게는 드러나지 않게 진행되어 사용자 입장에서는 서버가 자신에게 필요한 정보를 푸시해 준 것으로 인식되는 것이다.

현재 이용되고 있는 많은 푸시 제품들이 위와 같은 방식으로 동작하는데 서비스하는 정보의 종류나 정보의 전달 방식 등에서 차별화된다. 푸시 서비스의 선두 주자인 PCN의 PointCast는 CNN, 날씨, 스포츠 등의 각종 뉴스정보를 PointCast 서버에 통합하여 저장해 놓고 사용자가 선택한 뉴스들을 제공하며 클라이언트에서는 스크린 세이버로 갱신된 뉴스들을 보여준다.

반면에 Marimba의 Castanet은 뉴스보다는 소프트웨어나 응용프로그램 등 프로그램 단위의 정보를 다룬다. 사용자는 Castanet을 통해 소프트웨어 벤더(vendor) 등의 원하는 정보원에 등록을 하고 채널을 확보하면 이 채널을 통해 Castanet 클라이언트들은 정보원에 접근할 수 있다.

BackWeb은 사용자가 원하는 정보를 멀티미디어 클립 형태로 정기적으로 제공한다. 특히 플래쉬와 같은 기능은 사용자가 정상적인 컴퓨터 작업을 하고 있을 때 서버로부터 정보가 왔음을 알리는 작은 윈도우가 화면 하단에 나타나며 이를 마우스로 누르면 전달된 정보를 보여주고 그렇지 않으면 잠시 후 사라진다.

이와 같은 독자적인 푸시 제품들 이외에 마이크로소프트나 넷스케이프는 기존의 브라우저에 내장된 형태로 푸시 기능을 제공하고 있다. 마이크로소프트 인터넷 익스플로러 4.0은 Favorites 메뉴에서 원하는 정보원을 추가해 놓고 그 사이트에 가입만 하면 자동적으로 갱신된 정보를 받아와서 사용자가 오프라인으로 정보를 볼 수 있으며, 넷스케이프 커뮤니케이터 4.0에서는 커뮤니케이터의 한 구성 요소인 넷캐스터(NetCaster)를 통해 위에서 소개된 여러 푸시 제품들과 유사한 서비스를 제공한다.

푸시 서비스는 원하는 정보를 백그라운드 작업으로 혹은 오프라인으로 받아 볼 수 있어 사용자가 브라우저 앞에서 기다려야 하는 시간을 절약 해줄 뿐 아니라, 그 정보의 변화를 계속적으로 추적해야 하는 번거로움을 덜어 줄 수 있다[13, 14]. 그러나 푸시는 네트워크의 전송 부하를 가중시킨다는 문제점이 있으며, 누구나 능동적으로 원하는 정보원에 접근하도록 하는 개방된 인터넷이 일부의 푸시 서비스 제공자에 의해 과거의 TV와 같이 일방적으로 정보가 여과되어 전달되는 방식으로 후퇴하는 것이 아닌가 하는 비판도 있다. 이런 비판들에도 불구하고 푸시 서비스는 인터넷을 편리하게 사용하고자 하는 사용자들의 요구에 따라 계속 발전할 전망이다.

C. 온톨로지(Ontology)

온톨로지는 철학에서의 존재론이라고 부르며, 아리스토텔레스 이래 사물의 존재에 의미 부여를 논의하는 철학적인 연구영역이었다. 그러나 오늘날에는 지식공학이과 자연언어처리 등의 기술 분야에서 어휘, 개념 등과 같은 각각의 지식들이 지식 전체의 체계 안에서 어디에 위치하는가를 명확히 하는 연구 분야라고 할 수 있다. 특히 시맨틱 웹을 실현하기 위한 핵심 기술이 바로 온톨로지라고 할 수 있다[15].

온톨로지의 다양한 정의들 중에서 가장 본질적인 정의로 온톨로지를 설명하자면, "온톨로지는 공유하기(shared) 위한 개념들의 개념화를 형식적(formal)이고 명백하게(explicit) 설명해 놓은 명세서(specification)" 라고 할 수 있다[16]. 여기서 개념화는 어떤 현상에 대해서 관련이 있는 개념들을 식별하여 그 현상을 추상화한 모델로 설명하는 것이다. 명백하게 표현한다는 것은 개념의 사용 유형과 사용된 유형의 제약 조건들(constraints)을 명백하게 정의한다는 것을 의미한다. 형식적이란 것은 기계가 읽을 수 있어야 한다는 것을 말하며, 공유는 온톨로지가 표현하는 개념이 개별적이 아닌 해당 그룹 구성원 간의 합의된 지식에 바탕을 두고 있다는 것을 의미한다[1, 2].

온톨로지를 개발하려는 대표적인 목적 중 하나는 정보의 상호교통(interoperability)을 가능하게 하고 필요한 정보를 추출하기 위해서 정보의 구조를 사람들과 소프트웨어 에이전트들 간의 서로 공동 협의된 형태로 구성하는 것이다. 또 하나의 목적은 효율적인 재사용(reusability)이다[15, 17]. 데이터베이스 스키마 역시 개념화를 추구한 것이다. 하지만 이는 오직 하나의 애플리케이션에 관한 것이고, 온토롤지는 다양한 애플리케이션으로 적용이 가능하다는 것이 큰 차이점이다.

온톨로지를 이용한 지식 모델링이 가져다주는 기술적인 장점은 지식베이스의 일관성을 자동화하여 검사한다는 것, 특정하게 명시되지 않은 데이터를 유추할 수 있는 추론 메커니즘을 사용할 수 있다는 것, 사용자의 요구를 광범위하게 수용할 수 있고 검색을 용이하게 해준다는 것 등을 들 수 있다.

온톨로지의 종류는 다음과 같다.

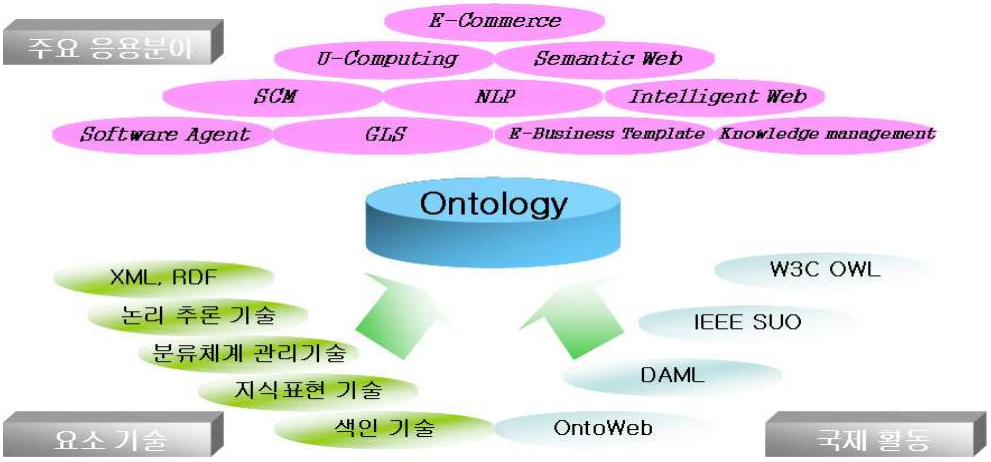
a. 도메인 온톨로지 : 특정분야에 한정되는 온톨로지(예 : 의료, 전자, 기계 등)

b. 메타데이터 온톨로지 : 온라인 정보자원의 내용을 기술하기 위한 어휘를 제공

c. 일상적 온톨로지 : 일상 세계에 대한 상식적 개념을 표현
(예 : 시간, 공간, 움직임 등)

d. 개념 표현을 위한 온톨로지 : 다른 개념을 온톨로지로 표현하는데 쓰이는 개념에 관한 것(예 : 프레임 온톨로지는 여러 개념을 객체 지향적으로 표현할 때 쓰이는 프레임)

[그림 7]은 온톨로지가 사용될 수 있는 응용분야와 필요로 하는 기술적 요소, 그리고 온톨로지와 관련된 국제활동 등을 보여주고 있다.

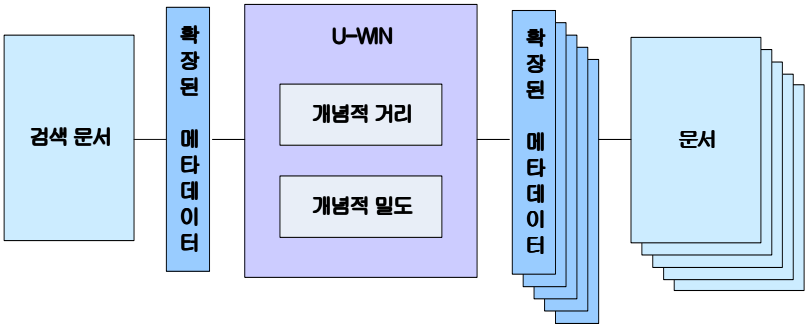


[그림 7] 온톨로지 관련 기술적 요소 및 응용

Ⅲ. 시스템의 제안 및 유사도 측정

A. 시스템의 구성

본 논문에서 제안한 웹 문서 검색시스템의 전체 구성은 [그림 8]과 같다. 사용자의 웹 문서에서 제목과 목차 부분의 명사를 추출하여 메타데이터 정보를 구축하고 또 검색의 대상이 되는 문서에서도 동일한 방법으로 메타데이터 정보를 구성한다. 이렇게 구성된 웹 문서들의 메타데이터들은 계층적 구조의 특징을 가지고 있는 U-WIN을 기반으로 개념적 거리를 이용해 단어간 유사도를 측정한다. 특히 동음이의어 단어에 대해서는 해당 단어와 문서 내의 다른 단어들과의 거리를 측정하는 개념적 밀도 측정법을 이용함으로써 의미적으로 유사한 문서의 검색이 가능하게 하는 시스템이다.



[그림 8] 제안한 웹 문서 검색 시스템의 개념도

1. 교육정보 메타데이터

일반적으로 메타데이터란 “데이터에 의한 데이터”를 의미 하지만 이러닝 시스템 표준에서의 메타데이터는 “학습 객체를 정의하고 설명하는 데이터”를 의미한다.

교육정보 메타데이터의 구현방법으로는 미국의 표준인 LOM 방식과 유럽의 표준

인 Dublin Core 방식, 기업 교육과 평생 교육에 초점이 맞춘 ERM 방식, 한국교육학술원(keris)에서 초·중등 교육을 위해 개발한 KEM 방식이 있다.

특히 Dublin Core 방식은 자원의 신속한 검색을 목적으로 합의된 형태이고, 가장 범용적인 메타데이터에 대한 표준을 제시하고 있으며 교육 분야에 특화된 한정어를 강화하는 추세이다. 또한 기본 요소로는 제작자(creator)나 발행자(publisher)가 부여한 표제인 title, 자원의 지적 내용에 주된 책임을 진 인물이나 단체인 Creator, 자료의 주제를 표현한 단어나 구(句)인 Subject, 초록이나 문장형식을 기술한 Description, 제작 기관인 Publisher, 자원의 유형을 기술한 Type, 현재 형태의 자원이 제작된 연도를 기술한 Date, 자원을 고유하게 식별하기 위한 문자열이나 부호인 Identifier등 총 15개의 요소로 구성되어 있다[18, 19, 20, 21].

본 논문은 Dublin Core 방식의 기본 요소에 착안해 웹 문서의 제목과 목차에서 명사만을 추출하여 메타데이터 정보를 구성한다. 이렇게 구성된 메타데이터들은 해당 문서의 성격을 좌우하게 하는 가장 기본이 되는 정보가 된다.

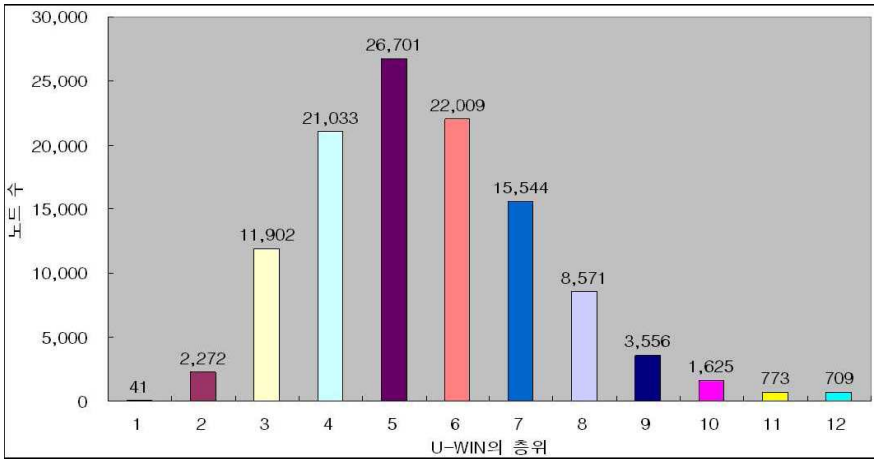
2. U-WIN

울산대학교 어휘 지능망 U-WIN(UOU-Word Intelligent Network)은 울산대학교의 한국어처리연구실에서 WordNet을 기반으로 한국어정보처리를 비롯한 정보검색, 기계번역, 시맨틱 웹 등 다양한 분야에 이용될 수 있는 어휘 데이터베이스이다. 특히 “인간이 가지고 있는 여러 관념 속에 공통적인 속성을 기반으로 인간의 보편적인 인지체계와 개념관계를 파악하여 이것을 표현한 언어를 대상으로 한 형식적이고 명세적인 어휘 네트워크”로써 우리말의 명사를 가지고 제작하여 현재 12만개의 어휘로 구성되어 있다[22].

U-WIN에서 단어와 뜻풀이의 관계는 1 : n이 아닌 1 : 1로 설정하여 관리함으로서 다의적 특성에 의한 의미적인 부담을 해소하고 의미적 계층구조의 세밀성 및 동의어, 유의어, 반의어에 대한 관계 설정에 용이한 계층적 구조로 구축하였다.

이러한 단어의 계층적 구조(hierarchical structure)는 의미론적 상·하 관계를 중

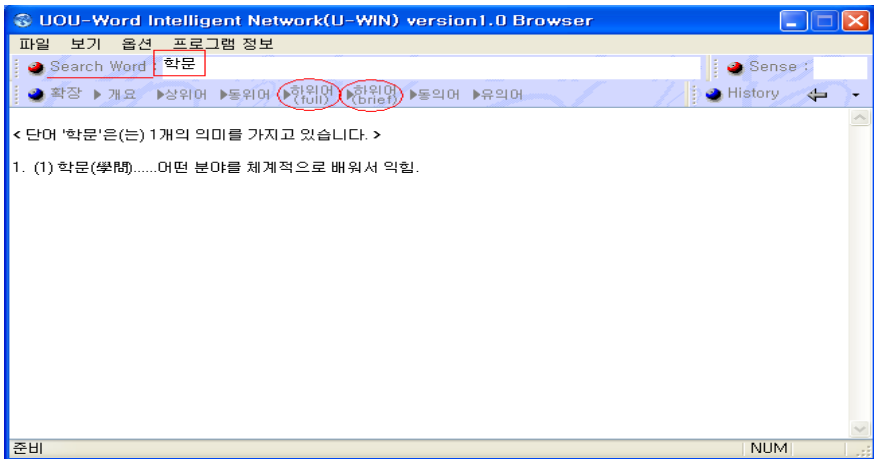
심으로 층위(level)를 이룬 집합체로 상·하위 층위(또는 노드)가 의미적으로 밀접한 연관성을 가짐과 동시에 엄밀한 의미에서의 상·하 관계가 형성되고, 보편적인 개념화(추상화) 과정으로 인식되는 구조로 U-WIN의 층위별 노드 수는 [그림 9]와 같다.



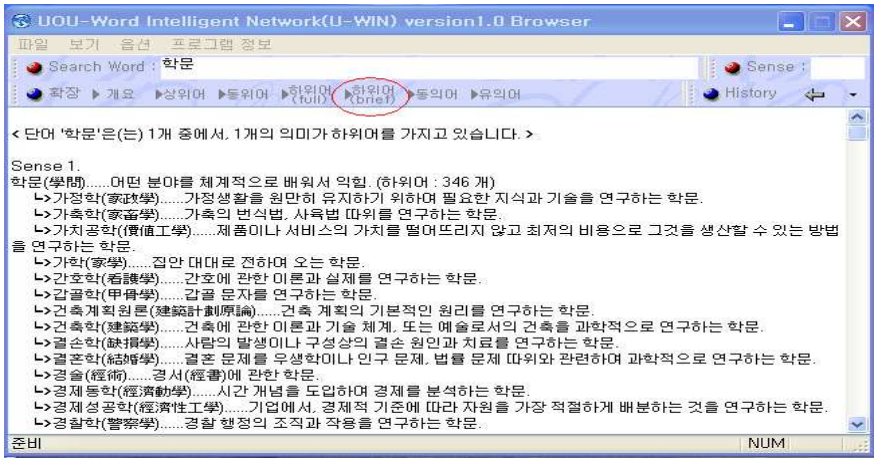
[그림 9] U-WIN의 층위별 노드 수

U-WIN은 최상위 노드 41개, 깊이(depth)는 12개로 계층이 구성되어 있으며[22], 특정 단어에 대한 계층 검색 방법은 다음과 같다.

U-WIN 브라우저 실행 후 [그림 10]과 같이 Search Word에 원하는 단어를 입력한 후, 하위어(brief)를 클릭하면, [그림 11]과 같이 검색하고자 하는 단어의 바로 아래 계층의 하위어들만 검색이 되고, 하위어(full)을 선택하면 검색하고자 하는 단어의 모든 하위어들이 검색이 된다. 단 하위어의 개수가 너무 많을 경우, 경고 메시지와 함께 검색이 불가한 경우도 있다.



[그림 10] U-WIN에서 단어별 의미 검색



[그림 11] U-WIN에서 계층 검색

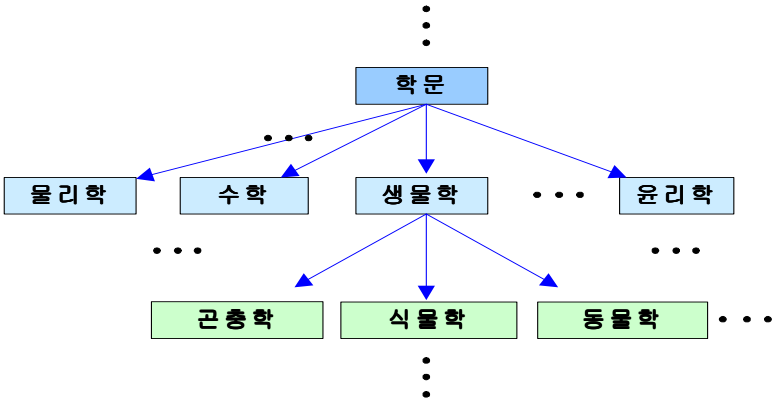
[그림 12]는 '학문'이라는 단어의 계층적인 구조를 검색한 결과이다.

'학문'은 '수학', '생물학', '역사학', '윤리학', '행정학' 등이 바로 아래 계층에 위치해 있는 반면,

'물리학'은 '학문'의 바로 아래가 아닌 "학문 --> 과학 --> 자연과학 --> 물리학"이라는 계층에 위치해 있고,

다시 '생물학'은 하위에 '식물생태학', '식물유전학', '식물생리학' 등이 위치해 계층

적 특성이 존재함을 보여주고 있다.



[그림 12] U-WIN에서 '학문'에 대한 계층구조

이러한 U-WIN의 계층적인 특징이 바탕이 되어 차후에 소개될 개념적 거리와 개념적 밀도를 이용한 유사도 측정이 가능하게 된다.

B. 유사도 측정

유사한 문서의 검색을 위해서는 확장된 메타데이터인 단어간 유사도 측정이 필요하다. U-WIN에서 단어의 개념적 접근을 위한 단어간 유사성 측정을 위해 고려될 수 있는 요소로는 다음과 같이 크게 다섯 가지로 분류될 수 있다[23].

a. 깊이(Depth) : U-WIN에서 노드의 위치로 깊이가 깊을수록 더 세부적이고 구체적인 개념이다. 측정하고자 하는 개념을 포함한 노드의 깊이가 깊을수록 유사성 측정시 더 큰 값을 반환한다.

b. 정보량(Information Content) : 노드 확률에 기반한 값으로 개념 c의 정보량은 $-\log P(c)$로 표현할 수 있다.

c. 밀도(Density) : 부모 노드의 정보량과 자식 노드의 정보량의 차로 차이가 작을수록 두 노드는 더 밀접하다.

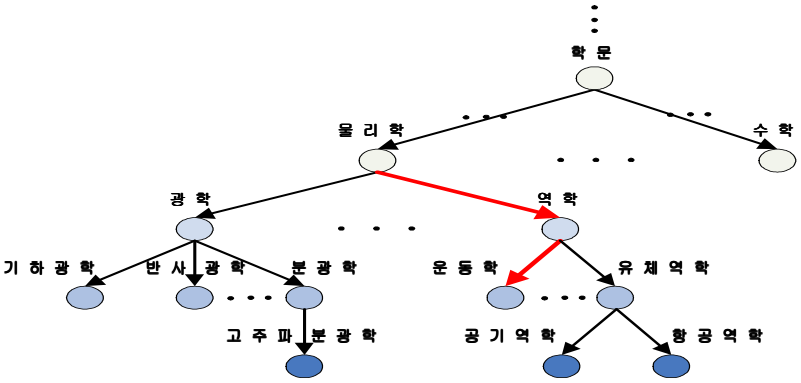
d. 거리(Distance) : 두 노드간의 거리로 노드간 에지 수를 의미하며, 거리가 짧을수록 두 개념은 더 유사하다.

e. 링크 타입(Link Type) : 인접한 두 노드 사이의 관계로 U-WIN에서는 상・하위어(IS_A) 관계를 비롯하여 전체, 부분, 반의 등과 같은 여러 관계들이 존재한다.

본 논문에서는 U-WIN의 계층적 특성으로 인해 단어의 깊이 정보가 상이함에 착안해 단어들 간의 거리와 동음이의어 문제를 해결하기 위한 개념적 밀도를 고려하여 유사도를 측정한다.

1. 개념적 거리 측정

개념간 유사도를 측정하는 방법으로는, 구성된 메타데이터 정보들의 개념적 거리를 이용하여 웹 문서 간의 유사도를 측정한다. [그림 13]은 U-WIN에서 '물리학'과 '운동학'의 개념간 구조 및 거리를 예로 나타내고 있다.



[그림 13] ‘물리학’과 ‘운동학’ 사이의 구조

두 개념간 거리를 이용한 유사도 측정 방법은 다음과 같다.

$$Distance = min[len(C_1, C_2)] \quad \text{----------} \quad (1)$$

$$Sim(C_1, C_2) = \frac{1}{Distance} \quad \text{----------} \quad (2)$$

여기서 $len(C_1, C_2)$는 비교 대상이 되는 C_1과 C_2의 개념간 거리로 최단거리 (min 값)를 구한다. 유사도는 거리에 반비례하므로 수식 (2)와 같이 구할 수 있다.

HyperTree:*Root*#n#1 인지#n#7 지식#n#2 학문#n#1
과학#n#1 자연과학#n#1 물리학#n#1 역학#n#2 **운동학#n#1**

HyperTree:*Root*#n#1 인지#n#7 지식#n#2 학문#n#1
과학#n#1 자연과학#n#1 **물리학#n#1**

최단거리 : 3, **유사도 측정값 = 0.3333333333**

[그림 14] 거리 및 유사도 측정

[그림 14]는 위 측정 방법을 이용하여 '운동학'과 '물리학' 간의 최단거리 및 유사도를 측정한 결과이다.

$Distance = min[len(C_1, C_2)]$ = 물리학-> 역학 -> 운동학 이므로 최단거리는 3이 되고,

$Sim(C_1, C_2) = \frac{1}{Distance} = \frac{1}{3}$ 이므로 유사도 측정값은 0.3333333333 이 된다.

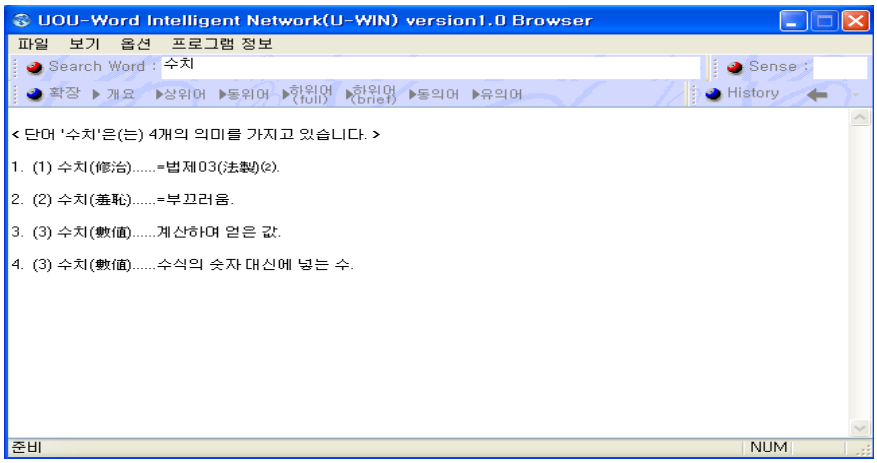
위와 같은 방법을 이용하여 '교육'과 관련된 웹 문서들의 유사도를 [표 1]과 같이 측정해 보았다. A문서는 유사도 측정시 기준이 되는 사용자의 문서로 5개의 메타데이터 정보(문명, 수학, 종류, 교과, 언어)를 가지고 있고, B문서는 검색 대상이 되는 문서중 하나로 7개의 메타데이터 정보(정수론, 소개, 수치, 연산, 개념, 군, 범위)로 구성되어 있다.

[표 1]에서 평균값 0.72460은 A문서와 B문서 간의 유사도를 측정한 결과이다. 이와 같은 방식으로 n개의 웹 문서들의 메타데이터 정보과 A의 메타데이터 정보 간의 유사도 측정을 통하여 평균값을 구한 후 값이 가장 높은 문서를 검색하게 된다.

[표 1] 웹 문서간 개념적 거리를 통한 유사도 측정 결과

A문서 / B문서		문명	수학	종류	교과	언어
		#n#2	#n#3	#n#2	#n#1	#n#1
정수론	#n#1	0.10000	0.50000	0.14285	0.10000	0.12500
소개	#n#1	0.10000	0.10000	0.14285	0.08333	0.12500
수치	**#n#3**	0.11111	0.11111	0.16666	0.09090	0.14285
연산	#n#2	•	•	•	•	•
개념	#n#1	0.12500	0.25000	0.20000	0.12500	0.16666
군	#n#16	0.12500	0.12500	0.20000	0.10000	0.16666
범위	#n#2	0.11111	0.11111	0.16666	0.09090	0.14285
평균,계	**0.72460**	0.67222	1.19722	1.01902	0.59013	0.86902

B문서에서 '연산'은 U-WIN에서 계층이 존재하지 않는 단어이므로 평균값 계산시 제외가 된다.



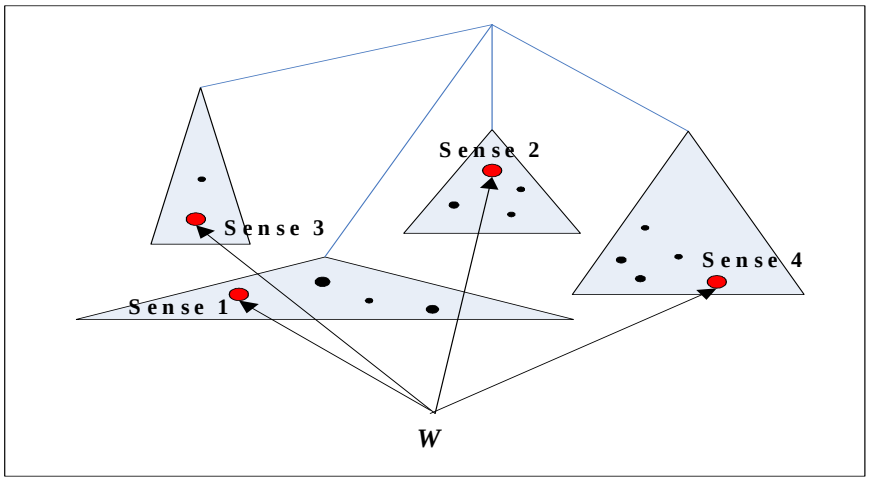
[그림 15] 단어 '수치'에 대한 의미 검색

'수치#n#3'에서의 'n'은 명사를 의미하고, '3'은 '수치'가 [그림 15]와 같이 U-WIN에서 갖는 4개의 의미 중 세 번째 의미인 "계산하여 얻은 값"이라는 것을 표현하고 있다.

2. 개념적 밀도 측정

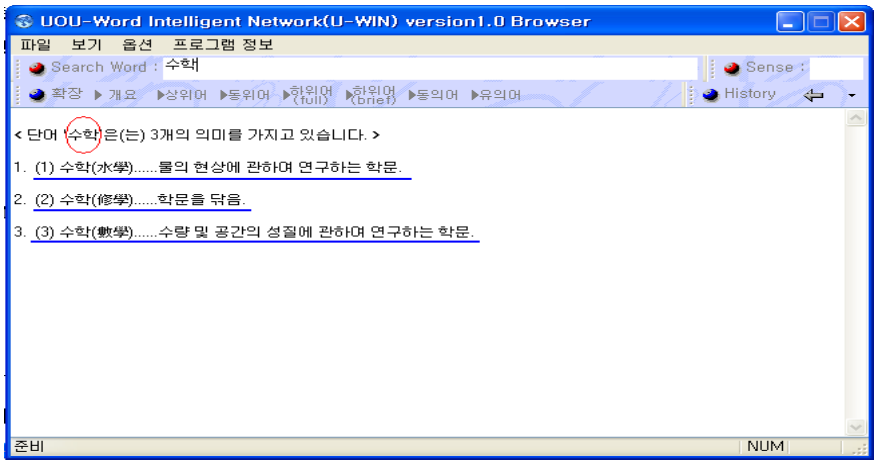
개념적 밀도란 한 단어가 여러 가지 의미를 가지고 있는 경우, 이를 해결하기 위한 것이다. 즉 동음이의어 문제를 가지고 있는 단어의 각 의미별 영역을 설정한 후 각 영역의 개념적 밀도를 구하기 위해 영역내 노드의 수를 세어 구할 수 있다. 이때 영역내 노드 수가 많을수록 높은 밀도값을 가지고 노드 수가 적을수록 상대적으로 낮은 밀도값을 가진다[24].

[그림 16]은 개념적 밀도를 통해 단어의 모호성을 해결하는 방법을 보여주고 있다. 예를 들어 단어 *W*는 4가지의 의미를 가지고 있다. 각각은 U-WIN의 subhierarchy에 포함되며, 이는 각 의미(Sense)별 영역이 된다. 영역내의 각 점은 중의성이 있는 단어 *W*와 *W*가 존재하는 문맥상의 다른 단어들 간의 유사도를 측정하여 그 결과에 따라 4가지 영역중 하나의 영역에 포함된다.



[그림 16] 개념적 밀도

각 영역 내에 존재하는 점인 노드의 수를 세는 방법을 이용하여 개념적 밀도를 구한다. 개념적 밀도가 클수록, 즉 영역이 클수록 각 문맥에서 정확한 의미를 표현할 가능성이 크다. [그림 16]의 경우 Sense 4의 밀도가 가장 크므로 찾고자 하는 문맥에 가장 알맞은 의미라고 볼 수 있다. 다음은 동음이의어 문제를 가진 단어를 이용한 예로 '수학'이라는 단어가 U-WIN에서 [그림 17]과 같이 3가지의 의미를 가지고 있다.



[그림 17] 동음이의어 단어의 예

이러한 동음이의어인 '수학'에 대해 개념적 밀도를 고려해 유사도를 측정한다. 개념적 밀도를 측정하는 방법으로는 #n#1, #n#2, #n#3의 각 노드의 위치와 비교 대상이 되는 통계학, 부호, 검정, 예제, 순위 등, 각 단어들 사이의 개념적 거리를 계산하는 방법으로 이루어진다. 이와 같은 방법으로 개념적 밀도를 측정한 결과 [표 2]와 같은 결과를 얻을 수 있었다.

[표 2] 개념적 밀도에 의한 동음이의어 문제 해소

단 어		통계학	부호	검정	예제	순위	부여	방식	비교	계
수학	#n#1	0.2500	0.1250	0.1250	0.0769	0.1111	0.1250	0.1428	0.0909	1.0467
	#n#2	0.0909	0.1111	0.0909	0.1000	0.1000	0.1111	0.1250	0.1666	0.8947
	#n#3	**0.5000**	0.1250	0.1250	0.0769	0.1111	0.1250	0.1428	0.0909	1.2967

[표 2]에서 #n#3과 '통계학'과의 결과 값이 0.5000으로 가장 높게 나옴으로써 이 문서가 #n#3의 '수학'이라는 것을 알 수 있어 동음이의어 문제가 해결될 수 있다.

Ⅳ. 실험 및 결과 분석

A. 실험 내용

개념간 유사도 측정을 통한 의미적 유사 문서를 검색하도록 하는 실험을 위하여 본 논문에서는 파워포인트로 된 웹 문서 90개를 5개 과목(생물학, 수학, 물리학, 경제학, 윤리학)으로 사전에 분류하였다. 또 각 문서에서의 제목과 목차를 구문분석하여 명사만을 추출한 후 메타데이터 정보를 구축하였고, U-WIN 1.0 버전을 이용해 개념적 거리와 밀도를 측정하였다. 실험결과에 대한 성능 평가의 척도로 정확률을 사용하였다.

실험 과정을 살펴보면 다음과 같다.

1. 구문분석을 통한 명사의 추출 : 문서에서 제목과 목차의 단어에서 명사를 추출한다.

[그림 18] '경제학' 관련 문서의 예

[그림 18]은 본 실험에서 사용한 경제학의 기준이 되는 문서로, [그림 19]와 같은 방법으로 명사를 추출하여 각 문서의 메타데이터 정보를 구성한다.

“경제학의 10가지 원리” ---> **경제학**(명사) / 의(조사) / 10 / **가지**(명사) / **원리**(명사)

“경제란, 단어는” ---> **경제**(명사) / 란(조사) / **단어**(명사) / 는(조사)

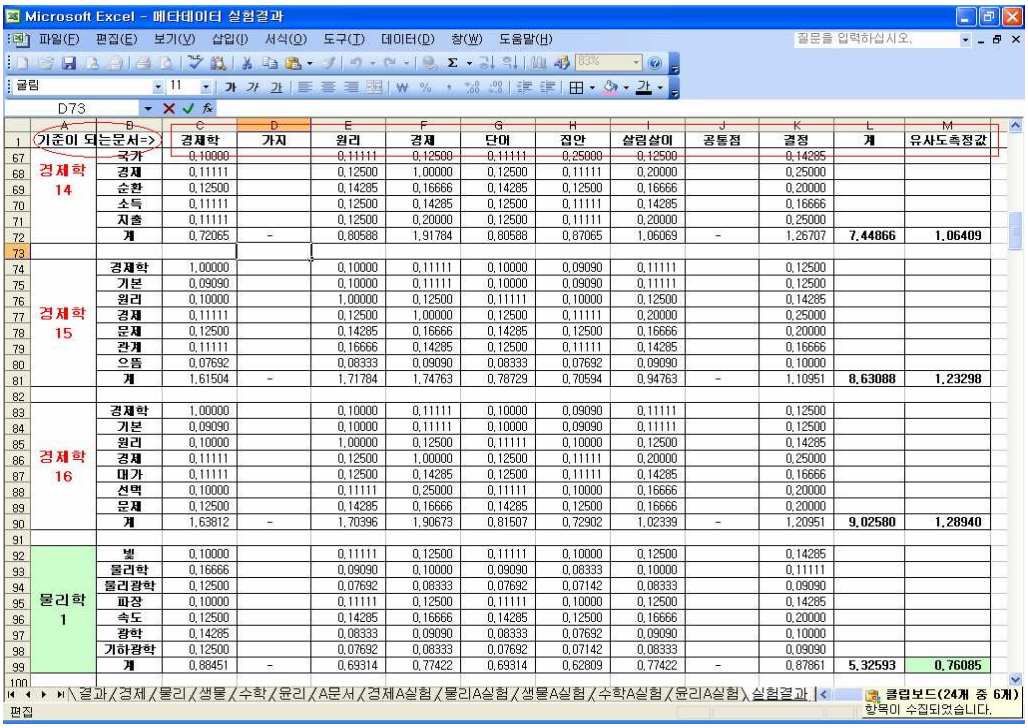
“집안 살림살이와 경제에는 공통점이 있다” ---> **집안**(명사) / **살림살이**(명사) / 와(조사) / **경제**(명사) / 에는(조사) / **공통점**(명사) / 이(조사) / 있다(동사)

[그림 19] 메타데이터 추출

위 방법으로 나머지 문서들에 대해서도 메타데이터 정보를 구성한다.

2. 구성된 메타데이터 정보들에 대해 각각 개념적 거리와 밀도를 이용하여 유사도를 측정한 결과 [표 3], [표 4]와 같이 각 문서의 유사도 측정결과를 알 수 있다.

[표 3] ‘경제학’ 관련 문서와 다른 문서간의 유사도 측정 결과(1)



기준이 되는문서=>		경제학	가지	원리	경제	단어	집안	살림살이	공통점	결정	계	유사도측정값
경제학 14	국가	0,10000		0,11111	0,12500	0,11111	0,25000	0,12500		0,14285		
	경제	0,11111		0,12500	1,00000	0,12500	0,11111	0,20000		0,25000		
	순환	0,12500		0,14285	0,16666	0,14285	0,12500	0,16666		0,20000		
	소득	0,11111		0,12500	0,14285	0,12500	0,11111	0,14285		0,16666		
	지출	0,11111		0,12500	0,20000	0,12500	0,11111	0,20000		0,25000		
	계	0,72065	-	0,80588	1,91784	0,80588	0,87065	1,06069	-	1,26707	7,44866	1,06409
경제학 15	경제학	1,00000		0,10000	0,11111	0,10000	0,09090	0,11111		0,12500		
	기본	0,09090		0,10000	0,11111	0,10000	0,09090	0,11111		0,12500		
	원리	0,10000		1,00000	0,12500	0,11111	0,10000	0,12500		0,14285		
	경제	0,11111		0,12500	1,00000	0,12500	0,11111	0,20000		0,25000		
	문제	0,12500		0,14285	0,16666	0,14285	0,12500	0,16666		0,20000		
	관계	0,11111		0,16666	0,14285	0,12500	0,11111	0,14285		0,16666		
	으뜸	0,07692		0,08333	0,09090	0,08333	0,07692	0,09090		0,10000		
	계	1,61504	-	1,71784	1,74763	0,78729	0,70594	0,94763	-	1,10951	8,63088	1,23298
경제학 16	경제학	1,00000		0,10000	0,11111	0,10000	0,09090	0,11111		0,12500		
	기본	0,09090		0,10000	0,11111	0,10000	0,09090	0,11111		0,12500		
	원리	0,10000		1,00000	0,12500	0,11111	0,10000	0,12500		0,14285		
	경제	0,11111		0,12500	1,00000	0,12500	0,11111	0,20000		0,25000		
	대가	0,11111		0,12500	0,14285	0,12500	0,11111	0,14285		0,16666		
	선택	0,10000		0,11111	0,25000	0,11111	0,10000	0,16666		0,20000		
	문제	0,12500		0,14285	0,16666	0,14285	0,12500	0,16666		0,20000		
	계	1,63812	-	1,70396	1,90673	0,81507	0,72902	1,02339	-	1,20951	9,02580	1,28940
물리학 1	빛	0,10000		0,11111	0,12500	0,11111	0,10000	0,12500		0,14285		
	물리학	0,16666		0,09090	0,10000	0,09090	0,08333	0,10000		0,11111		
	물리광학	0,12500		0,07692	0,08333	0,07692	0,07142	0,08333		0,09090		
	파장	0,10000		0,11111	0,12500	0,11111	0,10000	0,12500		0,14285		
	속도	0,12500		0,14285	0,16666	0,14285	0,12500	0,16666		0,20000		
	광학	0,14285		0,08333	0,09090	0,08333	0,07692	0,09090		0,10000		
	기하광학	0,12500		0,07692	0,08333	0,07692	0,07142	0,08333		0,09090		
	계	0,88451	-	0,69314	0,77422	0,69314	0,62809	0,77422	-	0,87861	5,32593	0,76085

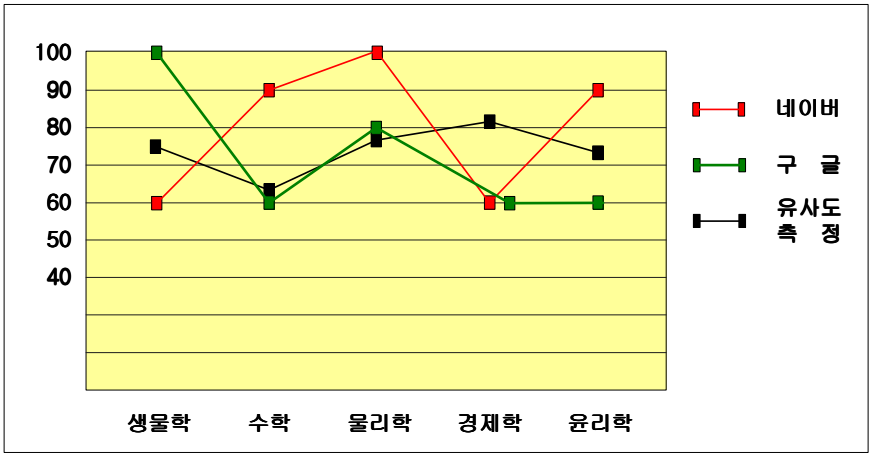
[표 4] ‘경제학’ 관련 문서와 다른 문서간의 유사도 측정 결과(2)

Microsoft Excel - 메타데이터 실험결과

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	기준이 되는문서=>		경제학	가지	원리	경제	단어	집안	살림살이	공통점	결정	계	유사도측정값
346	수학 15	도형	0,12500		0,14285	0,16666	0,20000	0,12500	0,16666		0,20000		
347		각	0,08333		0,09090	0,10000	0,09090	0,08333	0,10000		0,11111		
348		이등분선	0,08333		0,09090	0,10000	0,11111	0,08333	0,10000		0,11111		
349		성질	0,14285		0,16666	0,20000	0,16666	0,14285	0,20000		0,25000		
350		수직이등분선	0,08333		0,09090	0,10000	0,11111	0,08333	0,10000		0,11111		
351		계	0,68016	-	0,75913	0,87777	0,88170	0,68016	0,87777	-	1,01944	5,77613	0,82516
352													
353	수학 16	통계	0,09090		0,10000	0,16666	0,10000	0,09090	0,14285		0,16666		
354		논문	0,10000		0,11111	0,12500	0,11111	0,10000	0,12500		0,14285		
355		분석	0,09090		0,10000	0,16666	0,10000	0,09090	0,14285		0,16666		
356		통계학	0,20000		0,10000	0,11111	0,10000	0,09090	0,11111		0,12500		
357		소개	0,08333		0,09090	0,12500	0,09090	0,08333	0,12500		0,14285		
358		서비스	0,07692		0,08333	0,14285	0,08333	0,07692	0,11111		0,12500		
359		예제	0,07142		0,07692	0,12500	0,07692	0,07142	0,10000		0,11111		
360		계	0,71347	-	0,66226	0,96228	0,66226	0,60437	0,85792	-	0,98013	5,44269	0,77753
361													
362	윤리학 1	시민	0,06666		0,07142	0,07692	0,07142	0,06666	0,07692		0,08333		
363		윤리	0,10000		0,11111	0,12500	0,11111	0,10000	0,12500		0,14285		
364		필요성	0,12500		0,14285	0,16666	0,14285	0,12500	0,16666		0,20000		
365		도덕	0,11111		0,12500	0,14285	0,12500	0,11111	0,14285		0,16666		
366		문제	0,12500		0,14285	0,16666	0,14285	0,12500	0,16666		0,20000		
367		등장	0,12500		0,14285	0,16666	0,14285	0,12500	0,16666		0,20000		
368		특징	0,11111		0,12500	0,14285	0,12500	0,11111	0,14285		0,16666		
369		계	0,76388	-	0,86108	0,98760	0,86108	0,76388	0,98760	-	1,15950	6,38462	0,91209
370													
371	윤리학 2	경제	0,11111		0,12500	1,00000	0,12500	0,11111	0,20000		0,25000		
372		발전	0,11111		0,12500	0,14285	0,12500	0,11111	0,14285		0,16666		
373		윤리	0,10000		0,11111	0,12500	0,11111	0,10000	0,12500		0,14285		
374		관계	0,11111		0,16666	0,14285	0,12500	0,11111	0,14285		0,16666		
375		의미	0,09090		0,08333	0,09090	0,08333	0,07692	0,09090		0,10000		
376		조건	0,11111		0,10000	0,11111	0,10000	0,09090	0,11111		0,12500		
377		중요성	0,12500		0,14285	0,16666	0,14285	0,12500	0,16666		0,20000		
378		계	0,76034	-	0,85395	1,77937	0,81229	0,72615	0,97937	-	1,15117	7,06264	1,00895

B. 결과 분석

이러한 방법으로 나머지 '학문'(생물학, 수학, 물리학, 윤리학)에 대해서도 동일한 실험을 실시하여 그 결과를 정확률로 살펴본 결과 [그림 20]과 같다.



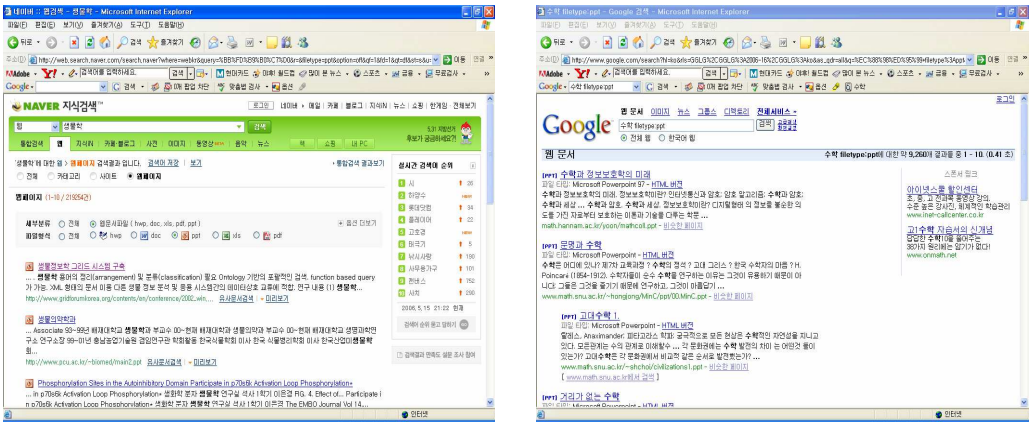
[그림 20] 검색엔진과의 실험 결과 비교

개념적 거리와 개념적 밀도를 이용한 유사도 측정으로 의미적인 유사문서 검색을 실시한 결과 정확률이 과목별로 적게는 64%에서 많은 과목은 82%로, 평균 76%의 정확률을 얻었다. 이 결과는 [그림 21]과 같은 실험 환경에서 [그림 22]와 같이 네이버(www.naver.com)와 구글(www.google.co.kr)에서 동일한 질의어(생물학, 수학, 물리학, 경제학, 윤리학)로 검색을 실시하였을 때 네이버의 평균 80% 보다 낮았지만 구글의 72% 보다는 정확률이 높음을 확인할 수 있었다.

기존 검색엔진의 실험

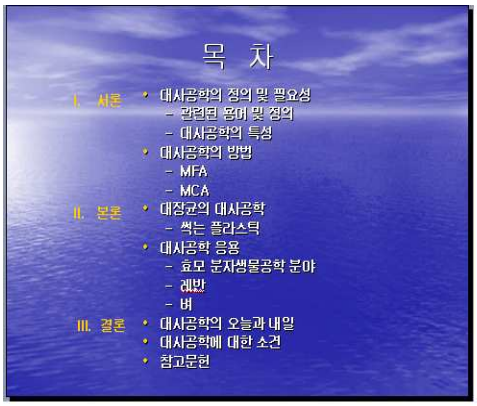
- 실험 일시 : 2006년 5월 15일 – 16일(2일)
- 실험 대상 : 네이버(www.naver.com)와 구글(www.google.co.kr)
- 실험 방법 : 1) 일반 검색 및 고급 검색 이용
  2) “학문” 관련 질의어 입력 후 검색(PPT 파일)
  3) 각 “학문”의 검색 결과 중 상위 10개의 문서만 추출
  4) 해당 “학문”의 내용과 유사성 관계 검토(수작업)
  5) 실행되지 않는 문서는 제외(차상위 문서로 대체)

[그림 21] 검색엔진에서의 실험 환경



[그림 22] 검색엔진에서의 실험 예제

이처럼 과목별로 검색 결과가 상이하거나 정확률을 다소 감소시킨 요인을 분석한 결과, 실험 문서의 폭을 파워포인트 문서로 제한한 점과, 정제된 메타데이터 정보를 사용하지 못했기 때문으로 판단된다. 또한 실험시 사용한 웹 문서가 규격화되어 있지 않아 메타데이터로 구성할 단어의 수가 적거나 외국어인 경우가 많았고, 중요 단어에 비해 비중이 낮은 단어들이 많이 존재해 정확률을 떨어뜨렸다.



[그림 23] 정확률을 감소시키는 요인

그 예로 [그림 23]과 같이 문서의 내용에 따른 사전분류에서는 생물학과 관련된 문서였으나 구문분석시 외국어와 외래어, 그리고 U-WIN에 존재하지 않는 '대사공학'과 같은 단어와 U-WIN에서 계층이 존재하지 않은 '효모', '분자생물공학' 단어들이 다수가 되어 유사도 측정시 비중을 차지하지 않는 단어 '정의', '필요성', '용어', '특성', '방법' 등에 의해 메타데이터 정보가 구성되어 정확률을 낮게 하는 요인으로 분석된다.

V. 결 론

본 논문에서는 온톨로지를 기반으로 개념적 거리와 밀도를 이용한 웹 문서의 의미적 검색을 제안하였다. 이를 위해 어휘의 계층적 특성을 지닌 Linguistic Ontology의 일종인 U-WIN을 이용하여 웹 문서를 대상으로 두 개념간 유사도 측정을 통해 의미적으로 유사한 문서를 검색하였다. 특히 개념적 밀도를 이용함으로서 복합적 의미를 지닌 단어의 의미적 판별이 가능하여 동음이의어 문제까지 해결할 수 있었다. 반면 정제된 메타데이터 정보를 사용하지 못함으로서 발생되는 문제점이나 새로운 단어 및 특정 분야에서 쓰이는 단어 등, U-WIN에 존재하지 않거나 계층이 없는 단어들에 의한 한계점은 본 연구의 성과를 저해하는 요인으로 파악되었다.

본 논문에서 제시한 개념간 유사도 측정을 통한 의미적 검색을 활용한다면 개념을 기반으로 하는 문서검색 분야나, 문서의 내용에 대한 의미적 분류가 가능하게 되어 사용자가 원하는 방식의 효율적인 카테고리별 분류가 가능함으로 문서분류 분야에 사용이 가능하고 추천시스템 분야에서도 응용될 수 있을 것이다. 또한 사용자가 보유중인 문서와 유사한 성격의 문서를 검색해 주는 유사문서 검색분야 등에 활용이 가능하게 되어 인간이 원하는 개념적인 사고방식에 더욱 접근할 수 있는 결과물을 얻을 수 있을 것으로 판단된다. 이러한 검색, 분류, 추천시스템의 활용 등은 문서에 국한되지 않고 이미지와 동영상 분야까지도 동일하게 활용이 가능할 것이다.

향후 연구과제로는 단순히 개념적 거리와 밀도만을 이용하기 보다는 개념적 인식률을 높여 보다 의미적인 문서의 검색이 가능하도록 유사도 측정 척도를 고안하고 이를 위해 개념의 깊이, 정보량, 또는 개념간 링크타입을 고려한 가중치를 부여하는 새로운 측정 방법에 관한 연구가 필요하겠다. 또한 본 논문의 정확률을 감소시켰던 메타데이터 선정·추출 등에 대한 연구가 필요하겠다. 더 나아가 좀 더 명확한 교육 분야 문서 분류를 위해서 '과목'이나 '학문'과 관련한 도메인 온톨로지를 구축하여 이를 활용한 유사문서 검색이 필요할 것이다.

참고문헌

[1] 정은경 외 3명, “온톨로지 기반의 정보검색”, 『한국정보과학회 2003년 추계학술대회』 VOL. 30, NO. 2-1. 2003

[2] 정민선 외 3명, “다양한 도메인의 효율적 표현을 위한 온톨로지 기반의 다계층 컨텍스트 모델링”, 『한국정보과학회 2005 추계학술대회』 VOL. 32, NO. 02. 2005.

[3] 최중민, “시맨틱 웹의 개요와 연구동향”, 『한국정보과학회 학회지』 VOL. 21, NO. 03, 2003.

[4] 최기선, “한국어 정보검색”, 『한국정보과학회 논문지』 제12권, 제8호.

[5] 김준태, 유견아, “인터넷 정보검색 시스템의 연구 동향”, 동국대학교 논문개제 연구비 지원, 1999.

[6] 정경용 외 3명, “내용 기반 여과와 협력적 여과의 병합을 통한 추천 시스템에서 조화 평균 가중치”, 『한국정보과학회 논문지 B』 VOL. 30, NO. 03, 2003.

[7] 이성진 외 2명, “추천 성능 향상을 위한 사용자별 가중치 자동 설정 기법”, 『한국정보과학회 2005 추계학술대회』 VOL. 32, NO. 02, 2005.

[8] 최중민, “에이전트의 개요와 연구방향”, 『정보과학회지』 pp.7-16, 1997.

[9] 박상위 외 2명 “메타 검색엔진을 위한 페이지 변경 탐지기 설계”, 『한국정보과학회 2001년 춘계학술대회』 VOL. 28, NO. 01, 2001.

[10] 김병만 외 4명, “추천시스템을 위한 내용기반 필터링과 협력필터링의 새로운 결합 기법”, 『한국정보과학회 논문지 B』 VOL. 31, NO. 03, 2004.

[11] 김병만, 이경, “항목 속성과 평가 정보를 이용한 혼합 추천 방법”, 『한국정보과학회 논문지 B』 VOL. 31, NO. 12, 2004.

[12] 최원태, 최미순, “푸시 기술의 개념 및 응용 사례에 관한 연구”, 국회도서관 관보, 1998.

[13] 이윤정, “Push 기술과 정보 배포 기술의 분석 및 동향 연구”, 『한국정보처리학회 추계학술발표논문집』 VOL. 7, NO. 2, 2000.

[14] 최은실 외 2명, “캐싱 기법을 적용한 푸시 에이전트 모델”, 『한국정보처리학

회 2003년 춘계학술대회』 VOL. 10, NO. 01, 2003.

[15] 공현장 외 3명, “특정 도메인에 대한 자동 온톨로지 구축 방법에 관한 연구”, 『한국정보과학회 2005 추계학술대회』 VOL. 32, NO. 02, 2005.

[23] 양정진 “시맨틱 웹에서의 온톨로지 공학”, 『한국정보과학회 학회지』 VOL. 21, NO. 03, 2003.

[17] 임수연, 송무희, 이상조, “전문용어의 처리에 의한 도메인 온톨로지의 구축”, 『한국정보과학회 논문지 B』 VOL. 31, NO. 03, 2004.

[18] 안정은, “멀티 플랫폼 환경에서의 e러닝 메타데이터 요소 개발”, 『한국정보과학회 한국컴퓨터종합학술대회』 VOL. 00, NO. 00, 2005.

[19] 최정화, 박영택 “메타데이터 기반 시맨틱 검색”, 『한국정보과학회 2005 추계학술대회』 VOL. 32, NO. 02, 2005.

[20] 한국소프트웨어진흥원, “동영상 콘텐츠의 온라인 유통 비즈니스 모델 및 메타데이터 연구”, 2002.

[21] 한국교육학술정보원, “교육정보 메타데이터 지침 해설서”, 2003.

[22] 김준수, 옥철영, “정제된 의미정보와 시소러스를 이용한 동형이의어 분별 시스템”, 『한국정보처리학회 논문지』 VOL. 12, NO. 07, 2005.

[23] 최준호 외 2명, “컬러 분포와 WordNet상의 유사도 측정을 이용한 의미적 이미지 검색”, 『한국정보처리학회 논문지 B』 VOL. 11, NO. 04, 2004.

[24] 조미영, 김판구, “정보량과 개념적 밀도를 이용한 단어 의미 중의성 해결”, 『한국정보처리학회』 VOL. 12, NO. 02, 2005.

[25] Gerald Salton, “Automatic Text Processing”, Addison Wesley Publishing Co., 1989.

[26] Michael R. Genesereth, “An agent-based framework for interoperability”, in Jeffrey M. Bradshaw(edited), Software Agent, The MIT Press, 1997.

[27] Tim Finin, Yannis Labrou, and James Mayfield, “KQML as an agent communication language”, in Jeffrey M. Bradshaw(edited), Software Agents, TheMIT Press, 1997.