지식관리시스템을 위한 FAH 기반 전문가 검색 방법론*

양근우** · 허순영***

FAH-Based Expert Search Framework for Knowledge Management Systems*

Kun-Woo Yang** · Soon-Young Huh***

Abstract

In Knowledge Management Systems (KMS), tacit knowledge which is usually possessed as forms like know-how, experiences, and etc. is hard to be systemized while managing explicit knowledge is comparatively easy using information technology such as databases. Recent researches in knowledge management have shown that it is more applicable in many ways to provide expert search mechanisms in KMS to pinpoint experts in the organizations with searched expertise so that users can contact them for help. In this paper, we propose an intelligent expert search framework to provide search capabilities for experts in similar or related fields according to the user's needs. In enabling intelligent expert searches, Fuzzy Abstraction Hierarchy (FAH) framework has been adopted, through which finding experts with similar or related expertise is possible according to the subject field hierarchy defined in the system. To test applicability and practicality of the proposed framework, the prototype system, Knowledge Portal for Researchers in Science and Technology, was developed.

Keyword : Expert Search, Text Categorization, Knowledge Management

1. 서 론

고도의 경쟁적인 비즈니스 환경하에서 조직의 경쟁력을 지속적으로 유지하기 위한 방법으로 조직 내부의 운영 지침이나 기준, 전문가의 노하우 등의 형태로 존재하는 조직의 암묵지(tacit knowl-

논문접수일 : 2004년 3월 8일 논문게재확정일 : 2005년 1월 17일
* 본 연구는 대학 IT연구센터 육성지원사업의 연구결과로 수행되었음.
** 계명대학교 사이버무역학과
*** 한국과학기술원 테크노경영대학원

edge)를 효과적으로 관리하는 것은 형식지의 관리만큼이나 혹은 그보다 더 중요한 것으로 인식되고 있다 [18]. 따라서, 이러한 암묵지의 효과적인 관리를 위한 기업의 요구는 지식관리시스템(KMS : Knowledge Management System) 연구의 많은 부분을 차지하고 있는데 그럼에도 불구하고 암묵지를 효과적으로 저장, 관리, 검색, 공유하기 위한 지금까지의 연구결과들은 특정 분야에 한정된 적용가능성과 비유연성 등의 이유로 다양한 분야에 적용되지는 못하고 있다.

최근의 연구들에서는 암묵지를 해당 지식 소유자로부터 의도적으로 분리하는 것 즉, 코드화하는 것은 불가피하게 해당 지식의 가치를 낮추는 결과를 초래하게 되므로 암묵지는 암묵지에 맞는 방법으로 관리되어야 함이 강조되고 있다[3, 12, 19]. 암묵지를 관리하기 위한 효과적인 방법의 하나로 조직내의 전문가 데이터베이스를 구축하고 필요한 전문 지식을 보유하고 있는 전문가를 효과적으로 검색할 수 있는 방법을 제공하는 것이 있다[3, 12, 14, 22]. 지속적으로 변화하는 비즈니스 경쟁 환경에서 지식관리 시스템의 지식베이스에 저장된 정적인 지식물들은 끊임없이 발생하는 모든 비즈니스 문제를 효과적으로 해결하는데 충분한 범용성을 제공하지 못할 가능성이 높다. 이는 지식 자체의 동적인 특성 및 지식의 상황 종속적인(context dependent) 특성에 기인하며[31] 따라서, 문제 해결에 필요한 지식을 보유하고 있는 전문가를 찾아 해당 지식을 제공 받는 것이 다양한 비즈니스 문제를 적시에 해결하기 위한 효과적이고 효율적인 방법이라고 볼 수 있다.

전문가 검색 기능을 위해서는 우선 조직 내에서 누가 어느 분야에 대한 전문 지식을 보유하고 있는지를 파악하는 전문가 식별(expert identification) 과정이 선행되어야 한다. 일단 이러한 과정을 통해 전문가 프로파일 데이터베이스가 구축되면 이를 기반으로 사용자는 필요한 전문가를 검색할 수 있게 된다. 이때, 각 시스템에서 제공하는 검색 방법을 이용하여 전문가 검색을 수행하게 되는데 기존의 질의 처리 방법을 이용할 경우 질의문에 표현된 모든 조건을 만족하는 정확한 값들만이 질의의 결과로 사용자에게 제공된다. 따라서, 만약 모든 질의 조건을 만족하는 값이 존재하지 않을 경우에는 질의 결과로 어떠한 정보도 제공할 수 없게 된다. 이러한 방식은 자신의 질의 결과로 최소한의 정보라도 제공 받고자 하는 사용자의 입장에서 본다면 만족스럽지 못한 질의 처리 방식이라고 볼 수 있다. 전문가 검색의 경우 특정 분야의 전문 지식을 보유한 전문가가 존재하지 않을 경우 시스템에서는 어떠한 유용한 정보도 제공할 수 없는데 이러한 문제를 해결하여 사용자의 전문가 검색 질의 만족도를 향상시키기 위해서는 사용자와의 상호 작용을 통해 유연한 질의 결과를 제공할 수 있는 방법이 제공되어야 한다.

본 연구에서는 주제 분야 분류 체계(subject field hierarchy) 상에서 주제 분야간 유사도를 퍼지 관계와 퍼지 연산을 활용하여 표현하고 계산하는 퍼지 추상화 계층(FAH : Fuzzy Abstraction Hierarchy) [25]을 도입하여 활용한다. FAH를 도입함으로써 KMS 사용자는 자신이 필요로하는 전문가를 검색할 때 퍼지 관계를 이용하여 계산된 유사도를 기반으로 검색 대상이 되는 분야는 물론 이와 유사한 혹은 관련이 있는 분야에 대한 전문가까지도 검색할 수 있게 된다. 본 연구에서는 기존 FAH 방법론이 가지는 초기 유사도 부여를 위한 개입을 없애고 이를 자동화할 수 있는 방법을 제안함으로써 해당 방법론을 개선하였다. 즉, 주제 분야간 초기 유사도 값을 도출하기 위하여 벡터 공간 모형(Vector Space Model)[4]이라는 문서 범주화 기법을 활용하게 되는데 이를 통해 미리 주제 분야가 결정된 학습문서를 이용하여 필요한 유사도 값이 자동으로 계산된다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 관련 연구를 간략히 정리하고 3장에서는 본 연구에서 제안하는 FAH를 위한 초기 주제 분야간 유사도 도출 과정을 포함하는 전문가 자동 분류기의 학습 과정을 설명한다. 이어서 4장에서는 사용자의 필요에 따라 유연한 검색 결과를 제공하는 지능형

전문가 검색 과정과 함께 본 연구에서 제안한 방법론을 적용하여 설계하고 구현한 원형시스템을 소개한다. 마지막으로 5장에서 향후 연구 방향과 함께 결론을 맺는다.

2. 관련 연구

2.1 지식관리시스템에서의 전문가 관리

암묵지와 관련한 기존의 연구에서는 암묵지를 획득, 저장, 검색하여 활용하는 것과 관련한 다양한 문제를 효과적으로 해결하지 못한 이유로 인해 여러 분야에서 널리 활용되지 못하고 있다[2, 3, 19, 32]. 이러한 연구의 예로 Design Rationale 시스템[7, 8, 15]을 들 수 있는데 이는 설계자의 설계 의도와 관련한 내용을 시스템적으로 저장하여 활용하고자 하는 시스템이다. 그러나, Design Rationale 시스템은 제한적인 적용 가능성과 특정 분야에 대한 의존성으로 인해 다양한 분야에서 널리 활용되지 못하고 있다. 암묵지 관리과 관련한 또 다른 연구 방향의 하나로 의도적으로 전문가가 보유한 전문 지식을 코드화하는 대신에 조직 내의 전문가를 효과적으로 검색하여 그들의 전문 지식을 활용할 수 있는 방법을 제안하는 연구들이 있다[2, 12]. 또한 이러한 기능이 실제로 몇몇 상업용 KMS 패키지에 탑재되어 활용되고 있기도 하다[13, 24, 30]. 이러한 시스템에서는 조직 내의 전문가에 대한 정보를 취합한 전문가 프로파일 데이터베이스를 구축한 후 일반적인 정보 검색 방법과 유사한 키워드 검색 등을 활용하여 필요한 전문가를 검색하게 된다.

전문가 검색 기능을 활용하기 위해서는 우선 조직 내의 각 전문가가 어떠한 분야에 어느 정도의 전문 지식을 보유하고 있는지에 대한 정보가 파악되어야 한다. 이와 같이 전문가의 전문지식에 대한 프로파일을 작성하기 위해서는 크게 두 가지 방법을 활용할 수 있다. 그 하나는 시스템 관리자나 전문가 자신이 시스템에 각 전문가의 전문 분야와 전문 지식의 정도를 수동으로 등록하는 방식이다. 또 다른 방법은 각 전문가에 의한 KMS에서의 지식 활동을 기반으로 프로파일 정보 수집을 자동화하는 것이다. 전문가 검색 기능을 제공하는 대부분의 상업용 KMS 패키지에서는 전자의 방식을 채택하고 있는데 이 방식이 구현의 용이함을 제공하는 반면 취합된 전문가 정보를 최신의 것으로 유지하기 위한 지속적인 관리 비용을 유발하는 단점을 가진다. 이 외에도 수동 전문가 프로파일 구축 방식은 다음과 같은 단점을 가진다. 첫째, 시스템 관리자나 전문가 자신이 프로파일 정보를 입력할 경우 해당 입력 정보의 객관성을 유지하기 어렵다. 자신이 자신의 전문분야나 전문 지식의 수준을 입력하게 되면 그 기준이 모호하고 서로 다른 전문 분야에 대해 서로 다른 기준이 적용될 가능성이 있다. 둘째, 사용되는 전문 분야 분류 체계가 수 많은 전문 분야를 포함하고 있고 한 사람의 시스템 관리자가 이렇게 다양한 분야에 대한 전문가를 할당하는 경우 모든 분야에 대한 분류를 수행할 수 있는 지식을 갖고 있기 힘들다는 한계가 있다. 셋째, 전문 분야별로 계속해서 새로운 개념이나 연구 경향이 나타나는 등의 변화를 가질 수 있는데 이러한 분야 자체의 변경을 반영하기가 쉽지 않다. 따라서, 이러한 수동 전문가 프로파일 구축 방식의 단점을 보완하기 위해 객관적인 기준에 의해 프로파일 정보를 자동으로 수집하고 전문가 프로파일 데이터베이스를 구축하는 효과적인 방법이 필요하다 할 수 있다.

프로파일 정보 수집을 위해 활용될 수 있는 KMS 사용자의 지식 생성 활동에는 지식베이스에 문서 파일을 올리거나, 전자 게시판에 글을 게시하는 것, 전자우편이나 메모를 주고 받는 행위 등이 포함된다. 이러한 지식 생성 활동의 결과물은 KMS 사용자의 전문 지식을 나타내는 것으로 볼 수 있는데 따라서 시스템의 관점에서 볼 때 사용자의 전문 지식을 자동으로 분석할 수 있는 유일한 방법은 사용자와 시스템과의 상호 작용에 의한 이와 같은 결과물의 내용을 분석하는 것이라고 할 수 있다. 이러한 관점에서 전문가 프로파일 정보 추출을 자동화하는 것은 시스템에 등록되는 지식물의 내용을 이해하고

이를 미리 정의된 주제 분야에 대해 분류해 내는 과정이라고 볼 수 있다.

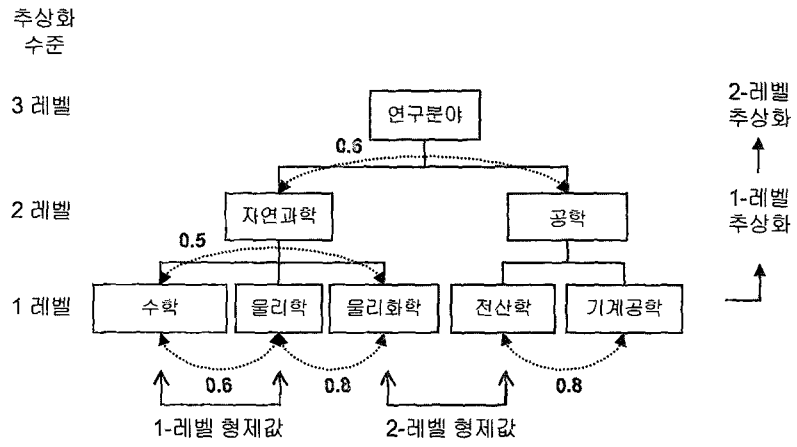
한편, KMS 사용자에 의한 지식 활동의 결과로 생성되는 지식물의 내용은 대부분 단순 텍스트 정보이거나 문서와 같이 텍스트 형태로 변경 가능한 것이다. 따라서 자동 문서 범주화 기법은 수동적인 관리자의 개입 없이도 시스템에 등록된 대부분의 지식물에 대한 분류가 가능하도록 하는 가장 이상적인 방법이다. 컴퓨터에 의해 텍스트로 이루어진 문서의 내용을 파악하여 미리 정의된 범주를 할당하는 자동 문서 범주화 기법에 관한 연구는 문서 범주화 기법의 효율을 높이기 위한 알고리즘 관련 연구와 이를 다양한 분야에서 활용하기 위한 연구 등을 포함하여 활발히 진행되고 있다[4, 20, 21, 32]. 본 연구에서도 관리자의 관리 노력을 최소화하면서 등록된 지식물을 자동으로 분류하여 해당 지식물을 등록한 사용자의 전문 분야를 구분해 내는데 이러한 자동 문서 범주화 기법을 활용하고자 한다.

2.2 퍼지 추상화 계층

전문가 프로파일 추출 과정을 자동화하고 이에 대한 검색 기능을 제공하는 궁극적인 목적은 KMS 사용자가 문제 해결을 위해 필요로 하는 전문 지식을 보유하고 있는 전문가를 효과적으로 찾을 수 있도록 돕는 것이다. 전문가 검색 결과에 대한 사용자의 만족도를 높이기 위해서는 사용자가 제시하는 검색 조건에 정확히 일치하는 결과는 물론 때로는 이와 유사한 특성을 가지는 질의 결과를 제공하는 것이 필요하다. 예를 들어, 사용자가 원하는 특정 분야의 전문가가 존재하지 않거나 검색 결과가 충분하지 않을 경우 시스템에서 해당 사용자가 찾고 있는 분야와 유사하거나 혹은 관련이 있는 분야의 전문가 정보를 제공해 줄 수 있다면 사용자의 질의 만족도를 훨씬 높일 수 있을 것이다. 즉, 이는 유사한 분야의 전문가일 경우 해당 사용자의 문제 해결을 위해 필요한 정보를 제공할 수 있는 가능성이 상대적으로 높으며 때로는 유사한 분야의 경험을 보유한 전문가가 특정 분야의 문제 해결을 위해 도움이 되는 경우가 많기 때문이다. 이것은 기존의 질의 처리 방법으로 사용자의 질의에 대해 아무런 결과도 제공해 주지 못하는 경우가 발생했을 때 질의 처리기의 개선을 통해 사용자에게 추가적인 유용한 정보를 제공해 줄 수 있는 방법이라 할 수 있다.

이와 같은 유사 분야 전문가 검색을 위해 본 연구에서는 사용자 질의에 대한 유사해(approximate answer) 제공이 가능한 질의 처리 방법론을 포함하는 퍼지 추상화 계층[25]을 도입하여 활용한다. FAH는 기존의 질의 처리 방법론과는 달리 사용자와의 상호 작용을 통해 사용자가 필요로 하는 정보를 좀 더 유연하게 제공하는 지능형 질의 처리 방법을 제공한다. 이를 통해 사용자의 의도에 맞추어 질의문의 조건을 완화하여 더 많은 질의 결과를 제공하거나 반대로 조건을 강화하여 더 적은 질의 결과를 제공하는 등의 질의어 변경을 수행한다. 또한 FAH는 이와 유사한 다른 데이터 추상화 방법론에 비해 전문가 검색 기능에 적용하는데 있어 다음과 같은 장점을 가지고 있다. 첫째, FAH는 데이터 값들간의 의미론적 관계를 데이터의 추상화 개념을 바탕으로 계층 (hierarchy)적으로 표현하는데 이는 다른 방법론에 비해 전문 지식 자체가 가지는 범주화 특성에 의한 체계적인 구조를 표현하여 전문가를 분류하는데 유리하다. 둘째, FAH에서 제공하는 방법을 이용하면 데이터 값들간의 정확한 유사도를 계산하고 표현하는 것이 가능한데 이를 통해 질의 결과값의 적합도 또는 유사도 등의 추가적인 정보 제공이 가능하게 된다.

이와 같이 분야간의 유사도를 부여하기 위한 방법의 하나로 웹사이트 혹은 문서의 내용으로부터 자동으로 온톨로지(ontology)를 구성하는 것과 관련된 연구들이 활발히 진행되고 있다[1, 29]. 이러한 연구들은 주로 웹마이닝(web mining)이나 텍스트마이닝(text mining) 분야에서 검색을 위한 방법 개선을 위해 웹페이지 혹은 문서간의 계층 구조 구성에 초점을 맞추고 있으며 도출된 유사도를 이용한 구체적인 유사 문서 검색 방안에 대한 고려는 미흡하다.



[그림 1] 퍼지 추상화 계층의 예

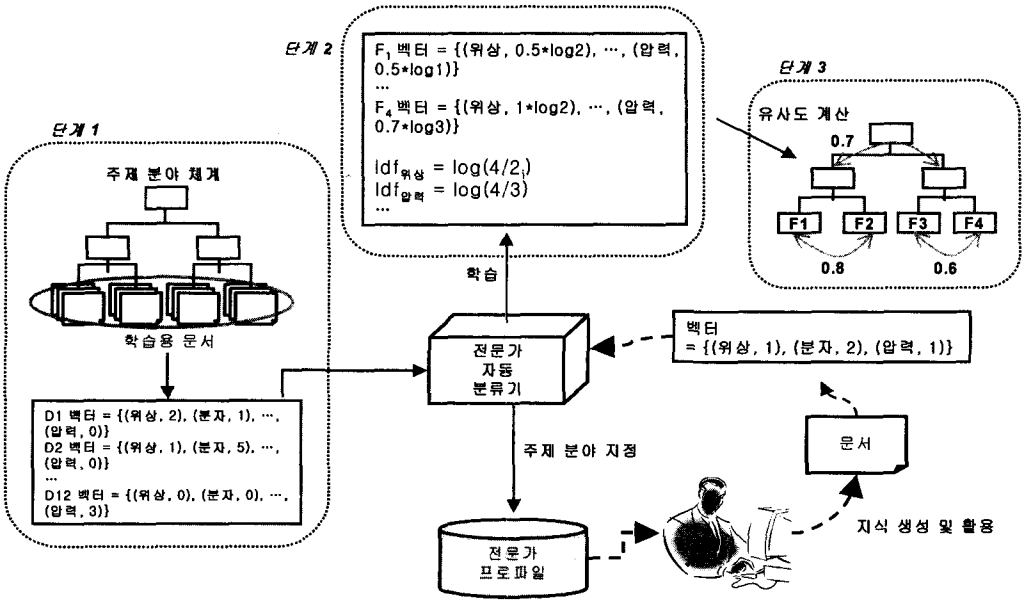
[그림 1]은 연구 분야에 대한 간단한 FAH의 예를 통해 FAH에서 각 개념간의 유사도와 관계를 어떻게 표현하는지를 개념적으로 보여주고 있다. 질의 조건을 완화하여 더 넓은 범위의 유사해를 제공하는 과정은 데이터 값들간의 유사도와 같은 대상 데이터베이스에 대한 추가적인 전문가의 지식을 필요로 한다. 이는 시스템 관리자나 이 분야의 경험이 있는 전문가가 [그림 1]에 나타난 바와 같이 부모값을 공유하는 형제값들간의 기본적인 유사도를 평가하여 할당해주어야 함을 의미한다. 또한, 이러한 데이터 값들간의 유사도에 대한 갱신 요구가 발생할 때마다 지속적인 관리 노력이 필요하게 된다. 이와 같은 유사도 할당 및 관리의 수동 작업 요구가 기존 FAH 방법론의 가장 큰 단점이며 수동으로 전문가 프로파일 정보를 관리하는 방식과 마찬가지로 (1) 주관적인 기준의 개입 가능성, (2) 관리자 한 사람에 의한 작업 수행의 어려움, (3) 각 주제 분야 자체의 변화 수용이 어려움 등과 같은 문제를 내포하고 있다.

이와 같은 문제점들을 고려하면 주제 분야 체계의 관리를 위한 보다 효율적이고 비용 효과적인 방법의 필요성이 대두되며 따라서 효과적인 주제 분야간 유사도 평가 및 할당을 위해 이러한 과정을 자동화하는 것이 필요하다. 본 연구에서는 이와 같은 주제 분야 분류 체계 관리 작업의 자동화를 위해서도 자동 문서 범주화 기법의 활용을 제안하는데 그 자세한 내용은 다음 장에서 설명하고자 한다.

3. 전문가 자동 분류기의 학습

이 장에서는 본 연구에서 제안하는 전문가 자동 분류기를 학습시키는 과정을 소개하고자 한다. 주제 분야간의 초기 유사도 값을 도출하여 할당하고 새로 등록된 지식물을 분류하여 전문가 프로파일 정보를 수집하기 위해서는 전문가 자동 분류기를 학습용 문서를 이용하여 학습시키는 과정이 필요하다. 이 장에서는 벡터 공간 모형(Vector Space Model)이라는 문서 범주화 기법을 기반으로 전문가 자동 분류기를 학습 시키는 과정을 설명한다.

많은 문서 범주화 기법 중, 본 연구에서는 다양한 관련 연구에 채택되어 좋은 문서 범주화 성능을 보인 벡터 공간 모델을 채택하여 전문가 분류기를 학습시키고 지식물을 분류하는데 이용한다[4, 28]. 본 연구에서는 벡터 공간 모델을 이용한 전문가 자동 분류기의 학습 과정을 세 단계로 보았는데 이러한 학습 과정을 통해 정의된 주제 분야 체계에서 부모값을 공유하는 주제 분야간의 초기 유사도 값이 계산되어 할당되며 이러한 초기 유사도 값은 주제 분야 체계 상의 모든 분야 쌍에 대한 유사도 계산을



[그림 2] 전문가 자동 분류기의 학습

퍼지 논리를 기반으로 수행하는데 활용된다. [그림 2]는 단계 1에서 단계 3을 거치는 전문가 자동 분류기의 학습 과정과 학습된 분류기를 이용하여 전문가 프로파일 정보를 생성하는 과정을 보여준다.

3.1 단계 1 : 학습용 문서 준비

전문가 자동 분류기를 학습시키기 위해서는 자신이 속한 주제 분야가 결정된 학습용 문서가 미리 준비되어야 한다. 학습용 문서를 이용하여 전문가 자동 분류기를 효과적으로 학습시키기 위해서는 다음과 같은 두 가지 기본 가정이 필요하다. 첫째, 학습에 사용되는 각 문서에는 정확한 주제 분야가 할당되어 있다. 둘째, 특정 주제 영역에 속하는 전체 학습 문서 집합은 해당 문서 집합에서 사용된 어휘들을 통해 해당 주제 분야를 완벽하게 표현한다. 이러한 가정이 성립되면, 학습용 문서들로부터 의미 있는 어휘들을 추출하여 각 주제 분야별 벡터 공간을 생성할 수 있는데 이렇게 생성된 벡터 공간은 새로 지식베이스에 등록된 지식물을 분류하는데 활용된다. 학습용 문서로부터 각 주제 분야에 대한 문서 벡터 공간을 생성하는 과정에서 필요한 개념들에 대한 정의는 다음과 같다.

[정의 1] 문서의 어휘 집합

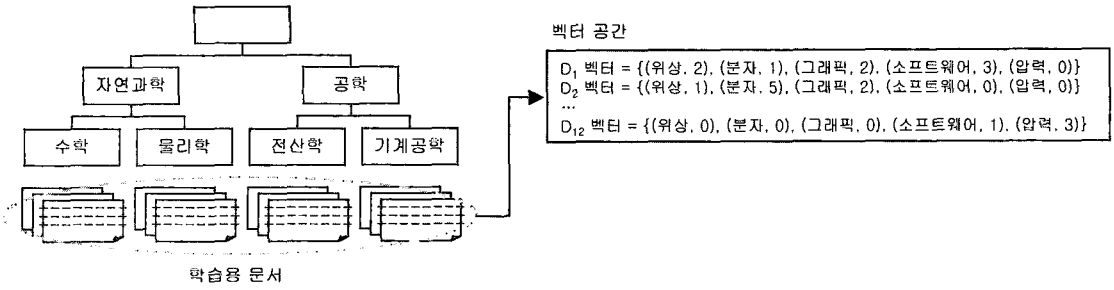
TS (Term Set)는 식 (1)과 같이 표현될 수 있는데 관사나 접속사와 같은 문법 요소들을 제외한 해당 문서의 실제 내용을 대표하는 의미 있는 어휘들의 집합을 뜻한다.

$$TS = \{t_1, \cdots, t_b\} \quad (1)$$

[정의 2] 분야의 통합 문서

AD_j (Aggregated Document)는 주제 분야 j에 속하는 문서 집합 $DS_j = \{D_{j1}, D_{j2}, \cdots, D_{jd}\}$로부터 생성되는 통합 문서를 의미한다. 여기에서 j = 1, ···, N이며 N은 정의된 전체 주제 분야의 수 그리고 j는 특정 주제 분야에 대한 색인을 의미한다. 또한 d는 주제 분야 j에 속하는 문서의 수를 나타낸다.

정의 1은 주제 분야를 구분하고 주제 분야와 지식물의 유사도를 계산하는데 사용되는 모든 어휘들이 각 문서로부터 추출되어야 함을 의미한다. [그림 3]의 예와 같이 5개의 의미 있는 어휘들로 구성된



[그림 3] 벡터 공간 생성

총 12개의 문서가 존재한다고 가정하자. 이 경우 전체 문서에 대한 어휘 집합 *TS*는 {*위상, 분자, 그래픽, 소프트웨어, 압력*}이 되며 각 어휘와 개별 문서에서의 출현 빈도의 쌍으로 이루어진 어휘 벡터 공간을 생성하여 활용할 수 있게 된다.

정의 2는 특정 주제 분야에 속하는 전체 문서에서 추출한 의미 있는 어휘와 그 출현 빈도의 쌍으로 구성된 통합 문서의 도출을 의미하는데 이렇게 구성된 통합 문서는 어휘와 그 출현 빈도를 기반으로 새로 등록된 지식물이 각 주제 분야에 얼마나 적합한지를 판단하는데 활용된다. [그림 3]의 예를 보면 수학, 물리학, 전산학, 기계공학의 네 분야가 존재하고 각 분야에는 세 개의 문서가 속하고 있다. 이러한 문서들이 각 분야별로 취합되어 분야 통합 문서를 생성하게 된다.

3.2 단계 2 : 분야별 통합 문서 벡터 생성

전문가 자동 분류기의 학습 과정 중 통합 문서 벡터 생성 단계에서는 각 주제 분야별로 취합된 분야 통합 문서(AD_j)로부터 분야 벡터(Field Vector)가 생성된다. 각 주제 분야에 대한 분야 벡터는 학습에 사용된 전체 문서 집합으로부터 추출된 색인 어휘 집합에 포함되어 있는 항목과 동일한 수의 요소를 포함하게 되는데 이 경우 분야에 따라 해당 어휘의 분야에 대한 관련성이 존재하지 않을 경우 출현 빈도가 0인 어휘들이 발생할 가능성이 있다. 특정 통합 문서 AD_j에 대한 분야 벡터인 FV_j는 전체 색인 어휘 집합에 속하는 어휘들과 그 어휘들의 특정 주제 분야에 대한 관련성 혹은 설명력 수치로 구성된다. FV_j에 대한 정의는 다음과 같다.

[정의 3] 주제 분야 벡터

FV_j는 전체 어휘 집합 *TS*에 속하는 어휘와 각 어휘의 주제 분야 j에 대한 설명력 혹은 관련성에 대한 가중치로 구성된다. FV_j를 계산하기 위한 수식은 다음과 같다.

$$FV_j = \{(t_1, w_{j1}), (t_2, w_{j2}), \cdots, (t_p, w_{jp})\} \quad (2)$$

여기에서 p는 전체 색인 어휘의 수를 나타내며 어휘 t_p의 통합문서 AD_j 혹은 주제 분야 j에 대한 설명력을 의미하는 w_{jp}는 다음과 같이 계산된다.

$$w_{jp} = tf_{jp} \times idf_p \quad (3)$$

$$tf_{jp} = freq_{jp} / \max_l(freq_{jl}) \quad (4)$$

$$idf_p = \log(N/n_p) \quad (5)$$

통합 문서 AD_j 즉, 주제 분야 j에서 어휘 t_p의 상대적 중요도는 어휘 t_p의 주제 분야 j에 대한 어휘 빈도 tf_{jp}를 이용하여 얻을 수 있으며 식 (5)의 n_p는 어휘 t_p가 출현하는 통합 문서의 수를 나타낸다. 어휘 빈도(term frequency)는 특정 문서에서 해당 어휘를 포함하는지의 여부를 고려하는 것으로 식 (4)의 $freq_{jp}$는 주제 분야 j에서의 어휘 t_p의 출현 횟수를 $max_l(freq_{jl})$는 주제 분야 j에서 가장 많이 출현하는 어휘의 출현 횟수를 의미한다. 또한, 역문서 빈도(idf : inverse document frequency)는 해당 어휘가 전체 문서 집합에서 출현하는지의 여부를 고려하는 것으로 전체 문서 집합

에서 드물게 출현하는 어휘가 그 어휘를 포함하고 있는 문서를 전체 문서 집합의 다른 문서들과 구분하는데 더 중요하다는 것을 반영한다. 따라서, 위 식 (3)에서 idf_p 는 어휘 t_p 의 역문서 빈도를 나타내는데 식 (5)에서와 같이 그 중요도는 t_p 를 포함하고 있는 문서 수 (n_p)에 역으로 비례한다. 그림 3의 예에서 네 개의 주제 분야에 대한 통합 문서 AD_j 가 가지는 각 색인 어휘의 빈도를 <표 1>과 같이 가정하자.

<표 1>의 정보를 이용하여 정의 3에서 설명한 분야 벡터를 서로 다른 네 주제 분야에 대해 생성할 수 있다. 전체 어휘 집합 TS 에 속하는 각 어휘의 설명력 즉, 해당 어휘가 특정 주제 분야를 얼마나 잘 설명하는지 혹은 해당 분야에 얼마나 잘 맞는지를 계산해 낼 수 있는 것이다. [그림 4]의 예는 '수학'과 '물리학'에 대한 설명력(가중치) 계산과 두 주제 분야에 대한 주제 분야 벡터 생성 과정을 보여주고 있다.

위 예에서는 분야 벡터 계산 과정의 결과로부터 '위상'이라는 어휘가 주제 분야 '수학'에 대해 가장 높은 설명력을 가지며 '소프트웨어'와 '압력'이라는 어휘는 '수학'이라는 주제 분야와 전혀 관련이 없는 것을 볼 수 있다. 동일한 계산 과정에 따라 '물리학,' '전산학' 그리고 '기계공학'과 같은 다른 분야에 대해서도 분야별로 5개의 벡터를 가지는 어휘 벡터 공간을 구성할 수 있다. <표 2>는 주제 분야별로 각 어휘가 가지는 상대 중요도 가중치를 보여주는 주제 분야 벡터 계산 결과를 요약하고 있다.

〈표 1〉 분야별 통합 문서에서의 단어 출현 빈도

	위상	분자	그래픽	소프트웨어	압력	*최다 출현 단어의 출현 빈도수*
수학	4	1	3	1	0	4
물리학	1	5	3	1	0	5
전산학	2	0	3	5	1	5
기계공학	0	0	0	1	3	3
단어 출현 분야수	3	2	3	4	2	

$w_{수학,\ 위상} = 4/4 \times \log(4/3) = 0.12$
$w_{수학,\ 분자} = 1/4 \times \log(4/2) = 0.08$
$w_{수학,\ 그래픽} = 3/4 \times \log(4/3) = 0.09$
$w_{수학,\ 소프트웨어} = 1/4 \times \log(4/4) = 0$
$w_{수학,\ 압력} = 0/4 \times \log(4/2) = 0$

$w_{물리학,\ 위상} = 1/5 \times \log(4/3) = 0.02$
$w_{물리학,\ 분자} = 5/5 \times \log(4/2) = 0.3$
$w_{물리학,\ 그래픽} = 3/5 \times \log(4/3) = 0.07$
$w_{물리학,\ 소프트웨어} = 1/5 \times \log(4/4) = 0$
$w_{물리학,\ 압력} = 0/5 \times \log(4/2) = 0$

$FV_{수학} = \{(위상, 0.12), (분자, 0.08), (그래픽, 0.09), (소프트웨어, 0), (압력, 0)\}$
$FV_{물리학} = \{(위상, 0.02), (분자, 0.3), (그래픽, 0.07), (소프트웨어, 0), (압력, 0)\}$

[그림 4] 분야 벡터 생성 예

〈표 2〉 벡터 공간

	위상	분자	그래픽	소프트웨어	압력
수학	0.12	0.08	0.09	0	0
물리학	0.02	0.30	0.07	0	0
전산학	0.05	0	0.07	0	0.06
기계공학	0	0	0	0	0.30

3.3 단계 3 : 주제 분야간 유사도 계산

FAH를 이용하여 유사한 혹은 관련 있는 분야의 전문가를 검색하려면 우선 주제 분야 체계에 존재하는 분야들 사이의 유사도 혹은 거리가 사전에 계산되어야 한다. 기존의 FAH 방법론에서는 분류 체계 내에서 동일한 상위 단계를 공유하는 분야간 즉 노드간의 초기 유사도 값을 측정하고 계산하는 작업이 전적으로 이러한 작업을 수행하는 개인의 경험 즉 해당 전문가의 주관적인 지식에 의존하고 있다. 하지만, 본 연구에서는 FAH를 도입하여 KMS 내의 주제 분야간 관계를 표현함에 있어서 문서 범주화 기법을 이용한 초기 유사도의 자동 추출 방법론을 제안한다. 이는 특정 두 주제 분야간의 유사도를 해당 분야에 속하는 일련의 학습용 문서들간의 유사도를 측정하는 방법을 통해 도출할 수 있기 때문이다. 다시 말해, 각 분야에 속하는 문서 파일이나 전자게시판 게시물 등이 포함된 지식물의 내용은 해당 주제 분야간의 유사도를 측정하는데 활용될 수 있는데 이는 특정 주제 분야로 분류된 지식물의 내용은 해당 주제 분야 자체를 설명하는 것으로 이해할 수 있기 때문이다.

각 주제 분야에 대한 통합 문서가 구성되고 주제 분야별 분야 벡터가 생성되면 이를 이용하여 전문가 자동 분류기 학습 과정의 3번째 단계인 주제 분야 쌍에 대한 초기 유사도 계산 과정이 수행된다. 이전 학습 단계를 통해 생성된 주제 분야별 벡터는 추출된 어휘 집합 내 각 어휘의 주제 분야별 설명력을 포함하고 있는데 이를 이용하여 주제 분야간의 유사도를 코사인 유사도 함수(Cosine Similarity Function)를 적용하여 구할 수 있다[4]. 일반적으로 두 벡터 사이의 유사도는 두 벡터로부터 계산되는 $\cos\theta$값으로 표현할 수 있는데 다음의 식 (6)는 이러한 코사인 유사도 함수의 수학적 표현을 나타낸다.

$$sim(FV_1, FV_2) = \cos\theta = \frac{FV_1 \cdot FV_2}{\|FV_1\| \times \|FV_2\|} \quad (6)$$

여기에서 FV_1FV_2는 두 벡터 FV_1과 FV_2의 내적(inner product)을 나타내며 $\|FV_1\|$은 벡터 FV_1의 절대값을 표시한다. [그림 4]의 예를 보면, '수학'과 '물리학'의 주제 분야 벡터는 각각 $FV_{수학}$ = {(위상, 0.12), (분자, 0.08), (그래픽, 0.09), (소프트웨어, 0), (압력, 0)}과 $FV_{물리학}$ = {(위상, 0.02), (분자, 0.3), (그래픽, 0.07), (소프트웨어, 0), (압력, 0)}이 되며 두 주제 분야의 유사도는 다음과 같이 계산된다.

$$\begin{aligned} &sim(\mathrm{FV}_{수학}, \mathrm{FV}_{물리학}) \\ &= (0.12\times0.02+0.08\times0.3+0.09\times0.07+0\times0+0\times0) / \\ &\quad (0.12^2+0.08^2+0.09^2+0^2+0^2)^{1/2} \\ &\quad \times(0.02^2+0.3^2+0.07^2+0^2+0^2)^{1/2} \\ &\approx 0.0327 \,/\, 0.17\times0.3087 \approx 0.6231 \end{aligned}$$

주제 분야간 유사도 값은 일반화(generalization)와 추상화(abstraction) 관계를 기반으로 하는 FAH를 구성하게 되는데 이렇게 분야간의 유사도로 계산된 값을 활용하기 위해서는 유사도 값이 FAH의 정의에 따라 0과 1사이의 값을 가져야만 한다. 두 분야 벡터 간의 유사도 값은 해당 벡터 사이의 $COS\theta$값으로부터 도출되므로 코사인 값이 항상 0과 1사이의 값을 가지는 특성에 따라 이러한 조건을 만족한다.

3.4 전문가 프로파일 데이터베이스의 구축

전문가 프로파일 추출 과정은 특정 전문가에 의해 등록된 지식물과 등록된 주제 분야간의 유사도를 계산하는 단계를 포함하는데 이러한 계산을 위해 학습 단계에서 도출된 분야별 문서 벡터를 활용하게 된다. 본 연구에서는 전문가 프로파일을 한 전문가가 보유하고 있는 전문 지식의 분야와 각 분야별 전문 지식의 수준을 나타내는 정보의 모음으로 정의한다. 전문가 검색 기능을 활용하기 위해서는 먼저 각 전문가가 보유하고 있는 전문 지식에 대한 상세한 정보를 가지는 전문가 프로파일 데이터베이스를 구축해야 한다. 특정 전문가는 하나 이상의 주제 분야에 대한 전문 지식을 보유할 수 있으며 각 전문가의 전문가 프로파일은 해당 전문가의 관심이

변하거나 또는 해당 전문 분야 자체가 시간의 흐름에 따라 변화하는 등의 이유로 인해 수시로 변경될 가능성이 있다. 따라서, 수집된 전문가 프로파일 정보를 항상 최신의 것으로 유지하고 그 유용성을 보존하기 위해서는 지속적이고 시스템화된 전문가 프로파일 갱신 방안이 필요하다.

전문가 프로파일 데이터베이스를 구축하기 위해서는 전문가가 보유하고 있는 전문 지식의 수준이 측정되어야 하는데 전문 지식은 (1) 활동성(activeness), (2) 우수성(excellence), (3) 평가(assessment)의 세가지 요소를 기준으로 측정될 수 있다. 본 연구에서 활동성은 전문가가 시스템에 지식물을 등록하는 지식 생성 활동을 얼마나 빈번히 수행하는지를 의미한다. 활동성은 또한 빈도(frequency), 최신도(recency), 분량(volume)의 요소로 구성되어 있다고 볼 수 있다. 빈도란 특정 전문가에 의해 시스템에 등록된 지식물의 숫자를 의미한다. 더 많은 지식을 등록한 전문가가 해당 분야에 대한 전문 지식을 상대적으로 더 많이 보유하고 있다고 볼 수 있는 것이다. 지식은 시간이 지남에 따라 그 가치가 변하게 되는데 최신도는 이러한 지식의 특성을 지식물의 평가에 반영하도록 한다. 따라서, 지식물은 그 가치를 평가함에 있어 시간 요소에 의해 조정되어야 할 필요가 있는 것이다. 마지막으로 분량 요소는 양이 방대한 지식을 등록한 전문가에 대한 평가를 위해 도입된다. 일반적으로 말해서, 분량이 큰 지식물이 더욱 많은 정보를 전달할 수 있다고 볼 수 있는데 따라서 시스템에서는 이러한 분량이 큰 지식물을 등록하는데 드는 시간과 노력을 인정해 줄 필요가 있는 것이다.

전문 지식을 구성하는 두 번째 요소인 우수성은 등록된 지식물이 얼마나 특정 주제 분야에 적합한가를 의미한다. 등록된 지식물이 속하는 주제 분야와 해당 분야에의 적합도를 계산하는 것이 가능하다면 우리는 이러한 요소를 이용하여 해당 지식물이 특정 주제 분야를 나타내는데 얼마나 우수한가를 결정할 수 있게 된다.

전문가 자동 분류기는 해당 지식물에 대해서 주제 분야 체계(subject field hierarchy)에 등록되어 있는 모든 분야와의 적합도를 각각 계산하여 제공한다. 즉, 모든 분야와의 적합도 중 가장 높은 점수를 보이는 분야가 해당 지식물이 속하는 분야이자 해당 적합도 점수가 그 분야에 대한 우수성 점수가 된다. 계산의 복잡도를 고려하여 본 연구에서는 최고 점수 하나만을 활용한다. 그러나 만약 최근 연구의 경향인 복합 분야 지식으로의 확장을 고려한다면 적합도 계산 목록에서 상위 두 개 혹은 세 개의 점수를 활용할 수 있겠다.

평가 역시 등록된 지식물의 유용성을 평가하는데 중요한 요소 중 하나로 볼 수 있다. 실제로 매우 가치 있는 지식이 시스템에 등록되었다 하더라도 시스템을 활용하는 다른 사용자가 해당 지식물을 이용하고 그로부터 해당 지식의 가치를 인정하기 전에는 그 지식이 유용하거나 의미 있는 것이라고 보기 힘들다. 따라서 등록된 지식에 대한 다른 사용자의 평가를 해당 지식물 등록자의 전문 지식을 측정하는 과정에 포함시키는 것이 필요하다.

4. FAH를 이용한 지능형 전문가 검색

이 장에서는 유사 분야 전문가 검색을 위해 도입한 FAH 방법론의 활용과 본 연구에서 제안하는 전문가 검색 방법론을 적용하여 개발된 원형시스템을 소개하고자 한다.

4.1 전문가 검색을 위한 FAH의 적용가능성

KMS에 등록된 전문가를 통해 그들이 보유하고 있는 암묵지를 효과적으로 활용하기 위해서는 각 전문가가 전문 지식을 보유하고 있는 주제 분야를 정확히 구분하는 것뿐만 아니라 사용자의 검색 조건에 맞는 특정 전문 지식을 보유하고 있는 전문가를 검색하기 위한 효과적인 검색 방법을 제공하는 것도 매우 중요하다. 전문가를 검색함에 있어서 시스템은 등록된 전문가의 부족, 특정 분야 전문가의

부재 등 다양한 원인에 의해 사용자의 검색 요구에 대해 만족스러운 결과를 제공할 수 없는 경우가 흔히 발생한다. 이처럼 사용자의 요구에 부합하는 전문가의 검색이 불가능할 경우 시스템이 이와 유사한 혹은 관련 있는 분야의 전문가를 대신 검색하여 제공한다면 사용자의 검색 만족도를 크게 향상시킬 수 있을 것이다. 이는 특정 분야의 전문가가 존재하지 않을 경우 이와 유사한 분야의 전문가가 활용 가능한 인력 중 사용자의 문제 해결을 위해 유용한 정보를 제공할 수 있는 가능성이 가장 높기 때문이다. 또한 다수의 특정 분야 전문 인력이 필요한 상황이라면 해당 분야 전문가의 부족 시 유사 분야의 전문가가 가장 최선의 대안이 될 수 있다.

유사 분야 전문가 검색을 위해 본 연구에서는 지식 표현 방법론인 FAH를 도입하여 활용한다. 유사 분야 전문가 검색 결과의 일부로 검색 대상 분야와의 적합도 혹은 검색 결과내의 순위 등을 제공하는 방법으로 검색된 전문가가 얼마나 검색 조건을 만족하는지를 사용자에게 알려줄 수 있다면 이 또한 사용자에게 매우 유용한 정보가 될 수 있다. FAH와 문서 범주화 기법을 접목하면 전문가 검색 기능의 향상을 위한 두 가지 추가 정보 제공이 가능해진다. 그 하나는 FAH의 퍼지 논리 계산을 이용한 주제 분야 체계에 정의된 주제 분야간의 유사도이며 이는 유사한 분야의 전문가를 검색하기 위해 활용된다. 또 다른 하나는 검색된 전문가가 특정 주제 분야에 대한 전문 지식을 어느 정도 보유하고 있는지를 판단하는데 활용할 수 있는 해당 분야와의 적합도 점수이다. 이와 같은 주제 분야간의 유사도와 각 전문가의 분야별 적합도 점수는 KMS 사용자가 필요로 하는 전문가를 보다 효과적으로 검색하는데 크게 도움이 된다.

4.2 Max-Min 연산을 이용한 유사도 계산

이 절에서는 퍼지 논리의 Max-Min 연산을 이용하여 주제 분야 체계에 정의된 주제 분야들 중 서로 상위 분야를 공유하지 않는 분야들간의 유사도를 계산하는 과정을 설명하고자 한다[33]. 이 계산을 위해서 3장에서 설명한 전문가 자동 분류기의 학습 과정에서 도출된 상위 분야를 공유하는 주제 분야간의 초기 유사도 값이 활용된다. 한편, FAH는 데이터 값들간의 유사 정도를 명시적으로 할당해야 하는 데이터 쌍의 수를 줄여주어 분류 체계의 구축과 관리 비용을 크게 낮추어 주는데 이는 상위 개념의 분야를 공유하는 주제 분야 쌍에 대해서만 초기 유사도 값을 명시적으로 할당하고 그 외의 분야간 유사도는 Max-Min 합성 연산을 이용하여 도출할 수 있기 때문이다. 분야 체계에 존재하는 주제 분야간의 유사도 계산을 위해서는 다음과 같은 추가적인 개념 정의가 필요하다.

[정의 4] n-레벨 추상값(abstract value)

특정 값으로부터 n이라는 추상화 수준의 차이를 가지는 값을 해당 값의 n-레벨 추상값이라 한다.

[정의 5] n-레벨 형제값 (sibling value)

동일한 추상화 수준에 위치한 여러 값들 중 자신과 동일한 n-레벨 추상값을 상위값으로 공유하는 값들을 자신의 n-레벨 형제값으로 정의한다.

또한, FAH의 다층 구조 추상화 수준 개념을 바탕으로 동일한 수준에 위치하는 형제값들간의 유사도에 대해서 형제값 수준의 차이에 따라 일정하게 감소하는 유사도를 가지도록 하기 위해 다음과 같은 정리를 만족할 필요가 있다.

[정리 1] 유사도의 단조 감소

(n+1)-레벨 형제값 사이의 유사도는 n-레벨 형제값 사이의 유사도보다 작다. 단, n>=1.

정리 1은 추상화 수준이 커짐에 따라 두 값 사이에 존재하는 유사도는 지속적으로 감소한다는 특징을 설명하고 있는데 다시 말해 추상화 수준 n이 증가함에 따라 n-레벨 형제값간의 유사도는 계속해서 작아지게 된다. 예를 들어, 2-레벨 형제값들간의 유사도는 1-레벨 형제값들 사이의 유사도보다 작으며 동시에 3-레벨 형제값들이 가지는 유사도보다는 반

드시 커야 한다.

한편, 1-레벨 형제값들 사이의 초기 유사도 값은 이전 절에서 설명한 바와 같이 벡터 공간 모형을 이용하여 자동으로 도출되며 n-레벨 형제값(단, n >= 2) 사이의 유사도는 퍼지 관계의 Max-Min 연산을 통해 계산된다. 이는 초기에 실제로 유사도 값을 부여해야 하는 형제값 쌍의 수를 크게 줄이는 장점이 있다. 예를 들어, [그림 1]에 정의된 2-레벨 형제값인 '수학'과 '전산학'의 유사도는 다른 1-레벨 형제값들 사이의 유사도가 부여되어 있다면 이를 이용하여 계산될 수 있는 것이다.

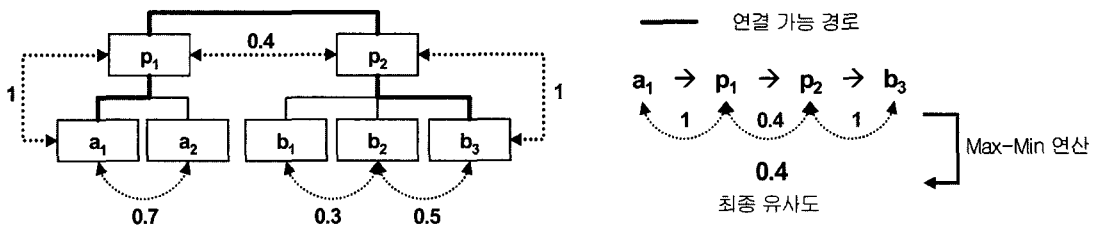
3장에서 설명한 바와 같이 1-레벨 형제값들 사이의 유사도는 전문가 자동 분류기의 학습 과정을 통해 자동으로 도출된다. 반면에 2-레벨 이상의 관계를 가지는 형제값 쌍에 대한 유사도는 퍼지 관계의 Max-Min 연산을 이용하여 계산하게 된다. 형제값들 사이의 유사도를 그 레벨 차이에 관계없이 모두 1-레벨 유사도 도출 방식과 동일하게 자동으로 부여하는 것이 가능하지만 이러한 방법은 다음과 같은 단점을 가지고 있다. 첫째, 조합 가능한 모든 쌍의 형제값에 대해 필요한 유사도를 도출하여 할당해야 하므로 고려하고 있는 주제 분야 체계의 크기에 따라 유사도 도출을 위한 계산의 복잡도가 크게 증가하게 된다. 이는 계산의 복잡도뿐만 아니라 상당한 시간과 비용을 수반하게 된다. 둘째, 모든 유사도 값을 자동으로 도출할 경우 각 레벨 별로 도출된 유사도 값이 항상 정리 1을 만족함을 보장할 수 없다. 즉, 특정 주제 분야에 속하는 예외적인 학습용 문서의 영향 등 다양한 원인에 의해 2-레벨 형제값의 유사도가 1-레벨 형제값의 유사도보다 더 크게 할당될 가능성이 존재한다. 이는 정리 1에 위배되는 것으로 1-레벨 형제값들의 유사도만을 학습 과정을 통해 도출하고 2-레벨 이상의 형제값들 사이의 유사도는 다른 방식으로 계산하여 도출하는 방법을 통해 정리 1을 만족시킬 수 있다.

[그림 5]는 간단한 FAH를 기반으로 Max-Min 연산을 이용하여 2-레벨 형제값 사이의 유사도를 계산하는 과정을 보여주고 있다. 그림에서 사각형은 주제 분야를 이들을 연결한 실선은 p_1이 a_1과 a_2의 부모값이 되는 것과 같은 주제 분야간의 계층 관계를 표시하고 있다. 또한, 굵은 실선은 FAH 상에서 계층 관계를 가지는 두 주제 분야를 연결하는 연결 가능 경로를 나타내며 점선 화살표와 함께 표시된 숫자는 해당 화살표로 연결된 두 주제 분야간의 유사도를 표시한다.

[그림 5]에서 a_1과 b_3의 유사도를 계산해보자. 예에서 a_1과 b_3은 각각 상위 수준 주제 분야 p_1과 p_2에 속하고 그들의 부모값인 p_1과 p_2는 그림에 표현되지는 않았으나 동일한 부모값을 공유하고 있다. 따라서 a_1과 b_3은 2-레벨 형제값이 된다. a_1과 b_3에 대한 Max-Min 연산을 이용한 퍼지 관계 합성의 결과는 다음과 같다.

$$(Derived)\ sim(a_1, b_3) = \text{Max}\,[\text{Min}\,[sim(a_1, p_1), sim(p_1, p_2), sim(p_2, b_3)]] \quad (7)$$

[그림 5]에서 보는 바와 같이 FAH는 하나의 자식값이 하나의 부모값만을 가지는 계층 구조를 가지고 있다. 따라서, a_1에서 b_3까지는 $a_1 \rightarrow p_1 \rightarrow p_2 \rightarrow b_3$라는 하나의 연결 가능 경로가 존재하며 만약 모



[그림 5] Max-Min 연산을 이용한 유사도 계산

든 부모값과 자식값 사이의 유사도를 일정하게 '1'로 가정한다면 식 (7)은 다음과 같이 변환될 수 있다.

$$\begin{aligned}(Derived)\ sim(a_1, b_3) &= \text{Min}\ [sim(a_1, p_1), \\ &\qquad sim(p_1, p_2), sim(p_2, b_3)] \quad (8)\\ &= \text{Min}\ [1, sim(p_1, p_2), 1]\\ &= sim(p_1, p_2)\\ &= 0.4\end{aligned}$$

이 결과로부터 부모값과 자식값의 유사도를 '1'로 고정할 경우 a_1과 b_3의 도출된 유사도는 그들의 부모값인 p_1과 p_2 사이의 유사도와 일치하는 것을 알 수 있다. 한편, a_1과 b_3사이에서 식 (8)로부터 도출된 유사도($sim(a_1, b_3)$) 0.4는 b_1과 b_2의 유사도($sim(b_1, b_2)$) 0.3보다 큰데 이는 정리 1에 위배된다. 이러한 모순을 해결하기 위해 본 연구에서는 분류체계상에서의 추상화 수준으로부터 얻을 수 있는 형제값 레벨의 개념을 포함한 식 (9)와 같은 확장된 유사도 개념을 정의하였다.

$$(Extended)sim(a_1, b_3) = \frac{1}{(Derived)sim(a_1, b_3) + \text{레벨차}} \quad (9)$$

단, $(Derived)\ sim(a_1, b_3) = sim(p_1, p_2)$이고 n-레벨 형제값 a_1과 b_3의 레벨차는 (n-1)이다.

식 (9)는 최종 유사도가 두 분야 사이에서 도출된 유사도와 레벨차로 구성되며 단조 감소 제약을 만족시키기 위해 역수를 취하고 있음을 보여주고 있다. 다시 말해, 두 주제 분야간의 차이가 커질수록 유사도는 그에 따라서 작아지게 되며 이와 마찬가지로 형제값 사이의 레벨차가 커지게 되면 최종 유사도 또한 감소해야 함을 의미한다. 따라서, a_1과 b_3의 최종 유사도는 이 경우 $sim(a_1, b_3) = 1/(0.4+1)$ 약 0.714가 된다.

4.3 유사 분야 전문가 검색 과정

일단 전문가 프로파일링 과정이 완료되어 각 주제 분야별 전문가가 그들이 보유하고 있는 전문 지식에 따라 할당되고 전문가 자동 분류기의 학습 과정을 통해 주제 분야간의 초기 유사도가 도출되면 사용자는 KMS에 등록된 전문가 중 그들이 필요로 하는 전문 지식을 보유한 전문가를 검색할 수 있으며 이와 함께 검색 대상이 되는 분야와 유사한 혹은 관련 있는 분야의 전문가까지도 검색할 수 있다.

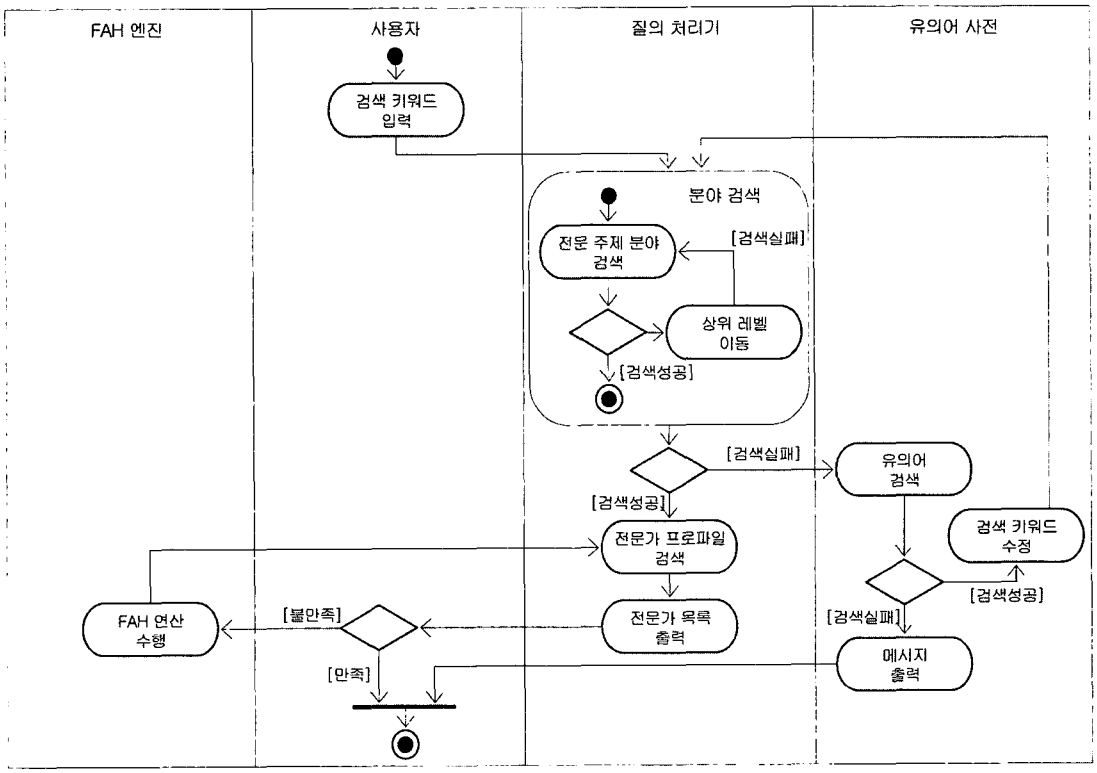
[그림 6]은 본 연구에서 제안하는 지능형 전문가 검색의 일반적인 과정을 UML(Unified Modeling Language)의 활동도를 이용하여 표현하고 있다[5]. 활동도는 시스템 내부에서 발생하는 일련의 행위 또는 순차적 사건 등을 모델링하는데 활용되는 도구로서 둥근 사각형은 각 행위를, 이러한 사각형을 연결하는 화살표는 행위간의 선행관계를 표현한다. 또한 마름모꼴은 조건 분기를 위해 활용되며 각 행위를 서로 다른 열에 표시함으로써 해당 행위를 수행하는 행위자 혹은 개체를 표현할 수도 있다.

전문가 검색은 사용자의 검색어 입력으로부터 시작되는데 시스템은 정의된 주제 분야 분류 체계에서 사용자가 입력한 검색어와 일치하는 주제 분야명이 존재하는지를 검색하게 된다. 이러한 주제 분야명 검색은 주제 분야 체계에서의 가장 하위 수준으로부터 시작하며 만약 일치하는 분야명을 찾는데 실패하면 다음 상위 수준 분야들을 대상으로 검색이 수행된다. 최하위 수준에서 최상위 수준까지의 검색 과정에서 만약 일치하는 주제 분야명이 존재할 경우 전문가 프로파일 데이터베이스로부터 해당 분야의 전문가 목록이 구성되어 사용자에게 제공된다. 이 과정에서 사용자에게 추가적인 정보를 제공하기 위해 유사 분야 전문가 검색을 수행할 수 있는데 이에 앞서 한 단계를 더 거치게 된다. 만약 주제 분야 검색을 주제 분야 체계의 가장 상위 수준까지 수행하였음에도 불구하고 일치하는 주제 분야명을 찾는데 실패하였다면 검색 유의어 사전 기능이 검색 과정의 일부로 동작한다. 이러한 유의어 사전 방식은 검색어 입력 과정에서 흔히 발생하기 쉬운 오타나 약자, 동의어로인한 문제들을 해결하기 위해 도입되었다. 예를 들어, 주제 분야 체계 상에

'전자상거래'라는 주제 분야가 존재한다고 가정하자. 이 경우 사용자가 해당 어휘 대신 'EC'라는 검색어를 입력했다면 시스템에서는 단어 대 단어 비교 방식 이외의 다른 방법을 활용하지 않고서는 어떠한 전문가 검색 결과도 제공하지 못할 것이다. 이때 유의어 사전은 입력된 검색어 'EC'를 '전자상거래'로 변경하는 역할을 하여 시스템이 이후의 검색 과정을 진행할 수 있도록 한다. 마찬가지로 검색 유의어 사전은 자주 발생하는 오타 입력의 경우 이를 자동으로 수정하기 위해서도 활용된다.

전문가 검색의 결과로 구성된 전문가 목록은 그 결과가 만족스럽지 못할 경우 사용자의 필요에 따라 질의 완화 혹은 질의 강화를 포함하는 FAH 연산을 수행하는데 활용된다. 만약 사용자의 검색 요구를 만족시키는 전문가의 수가 지나치게 많거나 혹은 적을 경우 각각 그 하위 혹은 상위 수준의 주제 분야에 대한 질의 처리를 통해 질의 결과를 줄이거나 늘릴 수 있다. 조금 더 구체적으로 살펴보면 만약 너무 많은 전문가가 검색될 경우 질의 강화를 적용하여 시스템에서 제공하는 더 구체적인 하위 주제 분야 목록으로부터 특정 주제 분야를 선택하여 검색 영역을 줄임으로써 검색 조건을 강화할 수 있다. 반대로 너무 적은 검색 결과가 제공될 경우 질의 완화를 통해 유사 분야 전문가를 포함하는 더 많은 전문가의 목록을 제공하는 것이 가능하다.

[그림 6]에 소개된 FAH 연산을 이용한 지능형 전문가 검색 과정을 더 자세히 예시하고자 'EXPERT {(id, name, field, title)}'의 테이블 스키마로 정의된 간단한 전문가 데이터베이스를 이용하기로 하자. EXPERT 테이블은 전문가의 전문 분야 정보를 제공하고 있는데 예를 들어 사용자가 '수학' 혹은 이와 유사한 분야의 전문가를 찾아라라는 질의를 수행하고자 한다면 이에 해당하는 질의문은 다음과 같이 작성될 것이다.



[그림 6] 전문가 검색 활동도

• 원 질의문

select name, field
from expert
where field =? '수학'

질의문 내에서 질의 완화 조건은 완화 연산자 '=?'를 이용하여 표시되었다. 사용자의 질의 완화 요청에 대해 '수학'의 1-레벨 부모값(추상값)을 검색하는 완화된 질의문은 다음과 같이 작성된다.

• 완화 질의문

select name, field
from expert
where field *is-a* Generalize('수학', 1)

위의 질의문에서 추가적인 요소인 *is-a*는 세분값과 추상값 즉, 자식값과 부모값 사이의 일반화 관계를 표현하고 있다. 완화된 질의문에서 'Generalize '(수학', 1)'은 '수학'의 1-레벨 추상값 즉 '자연과학'을 결과로 돌려주게 되며 따라서 질의 조건은 'where field *is-a* '자연과학"과 같이 완화된다. 또한, '수학'의 2-레벨 형제값들 중 특정 유사도 (예에서는 0.5) 이상을 가지는 주제 분야는 다음과 같은 근사 질의문을 통해 구할 수 있다.

• 근사 질의문

select name, field
from expert
where field *is-a* Approximate('수학', 2, 0.5)

[그림 1]을 보면 위의 근사 질의문을 통해 '수학'의 2-레벨 형제값 중 0.6의 유사도를 가지는 '전산학'과 '기계공학' 두 분야가 검색됨을 알 수 있다. 또한, 질의 강화는 'Specialize'를 이용하여 표현되는데 아래의 예에서는 '자연과학'의 1-레벨 자식값인 '수학,' '물리학,' '물리화학'이 검색된다. 이렇게 검색된 세분값 중에서 사용자는 원하는 주제 분야를 선택하여 질의 조건을 강화할 수 있다.

• 강화 질의문

select name, field
from expert
where field *is-a* Specialize('자연과학', 1)

〈표 3〉 유사 분야 전문가 검색을 위한 유사도 계산 예

전문가	전문분야	전문 지식 수준	'수학'에 대한 sim / (Derived) sim	'수학'에 대한 (Extended) sim	전문지식수준 × (Extended) sim
John	수학	23.5	1	$\frac{(2-1)+1}{2}=\frac{2}{2}=1$	23.5 × 1 = **23.5**
Sally	수학	12.6	1	$\frac{(2-1)+1}{2}=\frac{2}{2}=1$	12.6 × 1 = **12.6**
Sam	수학	7.8	1	$\frac{(2-1)+1}{2}=\frac{2}{2}=1$	7.8 × 1 = **7.8**
Peter	물리학	11.3	0.6	$\frac{(2-1)+0.6}{1.6}=\frac{1.6}{2}=0.8$	11.3 × 0.8 ≈ **9.0**
Smith	물리학	4.1	0.6	$\frac{(2-1)+0.6}{1.6}=\frac{1.6}{2}=0.8$	4.1 × 0.8 ≈ **3.3**
Brown	기계공학	22.1	0.6	$\frac{(2-2)+0.6}{0.6}=\frac{0.6}{2}=0.3$	22.1 × 0.3 ≈ **6.6**
Lucy	기계공학	9.8	0.6	$\frac{(2-2)+0.6}{0.6}=\frac{0.6}{2}=0.3$	9.8 × 0.3 ≈ **2.9**
Kevin	기계공학	1.6	0.6	$\frac{(2-2)+0.6}{0.6}=\frac{0.6}{2}=0.3$	1.6 × 0.3 ≈ **0.5**

이 절에서 설명한 FAH 연산을 위한 질의문들은 사용자의 추가적인 질의 수행 요청에 따라 시스템에서 자동으로 생성되며 질의처리기를 통해 일반적인 SQL(Structured Query Language)과 같이 처리된다. 추가적으로 '수학' 분야에 대해 검색된 전문가의 상대적인 전문 지식의 수준은 분야간의 유사도를 기준으로 계산되어 사용자에게 제공되는데 앞서 정의한 전문가 프로파일 데이터베이스를 기준으로 이러한 계산 과정의 예를 <표 3>에 정리하였다.

4.4 원형시스템 소개

본 연구에서 제안하는 지능형 전문가 검색 방법론의 적용가능성을 테스트하기 위해 전문가 자동 분류 및 검색 기능을 가진 KMS인 '과학기술 연구를 위한 지식포탈'을 원형시스템으로 개발하였다. 정보 포탈(information portal) 혹은 기업 정보 포탈(EIP : enterprise information portal)은 다양한 형태의 정보 시스템을 통합하고 이러한 정보시스템에 대한 효과적이고 효율적이며 동시에 개인화된 접근 경로를 제공하는데 적합한 시스템 아키텍처로 인식되고 있다[10, 11]. 이러한 포탈 시스템을 이용하면 사용자는 정보 원천의 종류에 상관없이 필요한 모든 정보에 친숙한 사용자 인터페이스인 웹 브라우저를 이용하여 접근하는 것이 가능하며 시스템은 개개인의 사용자가 적시에 중요한 비즈니스 결정을 내리는데 도움을 줄 수 있는 개인별 맞춤 정보를 제공하게 된다. 본 연구에서는 전문가 자동 분류 및 전문가 데이터베이스 구축을 통한 전문가 정보 검색을 사용자에게 친숙한 인터페이스를 통해 제공할 수 웹 기반 KMS를 개발하는데 포탈 기반 아키텍처를 기반 구조로 선택하였다.

본 연구에서 제안하는 지능형 전문가 검색 방법론은 현재 공개 서비스 중인 '과학기술 연구를 위한 지식포탈(http : //www.z4you.net)' 사이트에 적용되었다. 전문가 자동 분류기는 자바 언어를 이용하여 개발되었으며 이를 위한 문서 범주화 엔진으로는 Carnegie Mellon University에서 개발된 RAINBOW[23]를 이용하였다.

지식 포탈에서는 FAH를 위한 주제 분야 체계 구축을 위해 '한국과학재단(KSF)'에서 제정한 과학기술 분야 분류 체계를 이용하였다. KSF는 방대하고 복잡한 과학 기술 관련 연구 분야를 대, 중, 소의 3개 수준으로 자세히 분류하였는데 크게 '자연과학,' '생명과학,' '공학,' '복합학' 등 4개의 대분류 항목을 가지며 이러한 대분류는 총 69개의 중분류 항목으로 구분된다. 또한 각각의 중분류 항목은 평균적으로 7-8개의 소분류 항목을 가지게 되는데 총 523개의 소분류 항목이 존재한다. 지식 포탈에서는 이러한 주제 분야 체계가 (1) 사용자의 관심 분야 등록, (2) 사용자에 의해 등록된 지식물의 자동 분류, (3) 프로파일링 과정을 통한 각 전문가의 주제 분야 자동 할당, (4) 지능형 전문가 검색을 위한 FAH 연산 수행 등을 위해 활용된다.

전문가 자동 분류기의 초기 학습을 위해서는 KSF에서 제공한 1400개의 연구 과제 제안서를 이용하였다. 각 연구 과제 제안서는 해당 연구 과제가 다루고자 하는 특정 주제 분야가 미리 정해져 있으며 각 과제 제안서에 명시된 분야는 반드시 KSF 분야 체계에 정의된 주제 분야 중 하나가 된다. 따라서 본 연구에서 설계하고 구현한 전문가 자동 분류기를 학습시키고 그 성능을 평가하기 위해서 이러한 KSF 주제 분야 체계에 맞는 연구 과제 제안서가 매우 적합한 자료가 될 것으로 판단하였다.

전체 문서 중 약 60%를 학습에 이용하였으며 나머지 580개의 문서를 전문가 자동 분류기의 분류 정확도를 검증하기 위해 활용하였다. 사용 가능한 문서수의 제약으로 인해 중분류까지의 분류 정확도만을 계산하였으며 580개의 검증용 문서 중 37.24%의 문서가 정확한 주제 분야로 할당되었다. 이러한 다소 낮은 분류 정확도는 학습용 문서수의 부족에 기인한 것으로 판단되는데 이는 중분류까지의 분류 정확도만을 계산하였음에도 불구하고 많은 분야가 5개 미만의 학습용 문서를 이용하여 학습 과정을 거쳤다는 사실에 기인한다. 모델의 정확한 학습을 위해 통계적으로 충분한 수의 학습용 문서가 필요

할 것으로 판단되어 해당 분야에 속하는 문서수가 가장 많은 상위 10개의 주제 분야만을 고려하여 검증하였을 경우 약 73.25%의 분류 정확도를 보였다. 따라서 각 주제 분야별로 충분한 수의 학습용 문서를 확보할 수 있다면 만족할만한 수준의 분류 정확도를 유지하면서 동시에 소분류 항목까지를 포함한 더욱 정교한 전문가 자동 분류기의 학습이 가능할 것으로 예상된다.

5. 요약 및 결론

조직내의 암묵지를 효과적으로 관리하는 것은 비즈니스 세계에서 지속적으로 경쟁력을 유지하기 위해 중요하고 필수적인 요소로 인식되고 있다. 지금까지 다양한 연구를 통해 암묵지 관리를 위한 많은 기법들이 개발되고 제안되었으나 그 결과는 그다지 성공적이지 못하였다. 따라서 이 분야의 또 다른 연구 방향으로 조직내의 전문가에 대한 검색 기능을 제공하는 것이 암묵지를 관리하기 위한 실행 가능하면서 동시에 효과적인 방안으로 제안되고 있다. 이는 대부분의 경우 조직의 일상 업무나 운영 지침에 내재되어 있는 형태로 존재하는 무형의 전문가 지식을 코드화하거나 그 소유자인 전문가 자신으로부터 분리해 내는 것이 항상 가능하지도 또한 바람직하지도 않기 때문이다.

본 연구에서는 KMS를 위한 지능형 전문가 검색 방법론을 제안하였다. 제안된 방법론을 채택한 KMS를 통해 전문가 프로파일 수집을 자동화하고 지능적인 전문가 검색 기능을 활용할 수 있게 된다. 또한 본 연구를 통해 지식 표현 방법론인 FAH를 개선하였는데 자동 문서 범주화 기법을 도입하여 기존에 도메인 전문가에 의해 할당되어야 했던 분류 체계상의 데이터 값들간 초기 유사도를 자동으로 계산하여 부여하고 지속적인 갱신 처리를 할 수 있도록 하였다. 본 연구에서 제안하는 지능형 전문가 검색 방법론은 FAH를 도입함으로써 검색 조건에 부합하는 전문가뿐만 아니라 이와 유사하거나 관련 있는 분야의 전문가까지도 그 유사 정도에 따라 검색하는 것이 가능하다.

KMS에서 전문가 검색 기능을 제공하려면 미리 각 전문가가 보유하고 있는 전문지식에 대한 상세한 정보를 제공하는 전문가 프로파일 데이터베이스가 구축되어야 하는데 이를 위해 본 연구에서는 그 기준이 되는 전문가 프로파일의 구성 요소들을 제안하였으며 이것을 기준으로 자동 문서 범주화 기법을 이용하여 전문가의 전문지식을 측정하고 이러한 정보를 모아 전문가 프로파일 데이터베이스를 구성하게 된다.

현재 복합 분야 지식물을 처리할 수 있도록 전문가 관리 및 검색 방법론을 개선하는 것과 개별 전문가 검색 기능을 확장한 팀 구성 자동화 방안에 관한 연구가 진행 중이다. 또한, 전문가 자동 분류기의 성능 향상을 위한 알고리즘 개선 연구도 함께 수행하고 있다.

참 고 문 헌

[1] Alani, H. et al., "Automatic Ontology-Based Knowledge Extraction from Web Documents," *IEEE Intelligent Systems*, Vol.18, No.1(2003), pp.14-21.

[2] Alavi, M. and D.E. Leidner, "Review : Knowledge Management and Knowledge Management Systems : Conceptual Foundations and Research Issues," *MIS Quarterly*, Vol.25, No.1 (2001), pp.107-136.

[3] Augier, M. and M.T. Vendelo, "Networks, Cognition and Management of Tacit Knowledge," *Journal of Knowledge Management*, Vol.3, No.4(1999), pp.252-261.

[4] Baeza-Yates, R. and B. Riberio-Neto, *Modern Information Retrieval*, ACM Press, New York, 1999.

[5] Booch, G., J. Rumbaugh and I. Jacobson, *The Unified Modeling Language User Guide*, Addison Wesley, Boston, 1999.

[6] Braga, J.L., A.H.F. Laender and C.V. Ramos, "A Knowledge-Based Approach to Cooperative Relational Database Querying," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol.14(2000), pp.73-90.

[7] Buckingham-Shum, S.J. and N. Hammond, "Argumentation-based Design Rationale : What Use at What Cost?," *Human-Computer Studies*, Vol.40, No.4(1994), pp.603-652.

[8] Conklin, J.E. and K.B. Yakemovic, "A Process-oriented Approach to Design Rationale," *Human-Computer Interaction*, Vol.6, No.3-4 (1991), pp.357-391.

[9] Cai, Y., N. Cercone and J. Han, *Attribute-Oriented Induction in Relational Databases*, in *Knowledge Discovery in Databases*, AAAI Press/The MIT Press, 1993.

[10] Deltor, B., "The Corporate Portal as Information Infrastructure : Toward a Framework for Portal Design," *International Journal of Information Management*, Vol.20(2000), pp. 91-101.

[11] Dias, C., "Corporate Portals : a Literature Review of a New Concept in Information Management," *International Journal of Information Management*, Vol.21(2001), pp.269-287.

[12] Desouza, K.C., "Barriers to Effective Use of Knowledge Management Systems in Software Engineering," *Communications of the ACM*, Vol.46, No.1(2003), pp.99-101.

[13] Handysoft, BizFlow KMS, *http://corona.handysoft.co.kr/eng/*.

[14] Hansen, M.T., N. Nohria and T. Tierney, "What's Your Strategy for Managing Knowledge?," *Harvard Business Review*, Vol.77, No.2(1999), pp.106-116.

[15] Hu, X. et al., "A Survey on Design Rationale : Representation, Capture and Retrieval," *Proceedings of 2000 ASME Design Engineering Technical Conferences*, Baltimore, Maryland, September, 2000.

[16] Huh, S.Y. and J.W. Lee, "Providing Approximate Answers Using a Knowledge Abstraction Database," *Journal of Database Management*, Vol.2(2001), pp.14-24.

[17] Huh, S.Y. and K.H. Moon, "A Data Abstraction Approach for Query Relaxation," *Information and Software Technology*, Vol.42 (2000), pp.407-418.

[18] Kakabadse, N.L., A. Kouzmin and A. Kakabadse, "From Tacit Knowledge to Knowledge Management : Leveraging Invisible Assets," *Knowledge and Process Management*, Vol.8, No.3(2001), pp.137-154.

[19] Kreiner, K., "Tacit Knowledge Management : The Role of Artifacts," *Journal of Knowledge Management*, Vol.6, No.2(2002), pp.112-123.

[20] Lam, W., M. Ruiz and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," *IEEE Transactions on Knowledge and Data Engineering*, Vol.11, No.6(1999), pp.865-879.

[21] Lee, D.L., H. Chuang and K. Seamons, "Document Ranking and the Vector-Space Model," *IEEE Software*, Vol.14, No.2(1997), pp. 67-75.

[22] Liebowitz, J., "Knowledge Management and Its Link to Artificial Intelligence," *Expert Systems with Applications*, Vol.20(2001), pp. 1-6.

[23] McCallum, A.K., "Bow : A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering," *http://www.cs.cmu.edu/~mccallum/bow*, 1996.

[24] Microsoft, Microsoft SharePoint Products and Technologies, *http : //www.microsoft.com/sharepoint/*.

[25] Moon, K.H. and S.Y. Huh, "An Integrated Query Relaxation Approach Adopting Data Abstraction and Fuzzy Relation," *Submitted for publication to Information Systems Research.*

[26] Nonaka, I. and H. Takeuchi, *Knowledge Creating Company*, Oxford University Press, New York, 1995.

[27] Rus, I. and M. Lindvall, "Knowledge Management in Software Engineering," *IEEE Software*, Vol.19, No.3(2002), pp.26-38.

[28] Salton, G. and M.E. Lesk, "Computer Evaluation of Indexing and Text Processing," *Journal of the ACM*, Vol.15, No.1(1968), pp. 8-36.

[29] Vargas-Vera, M. et al., "Knowledge Extraction Using an Ontology-Based Annotation Tool," *Workshop on Knowledge Markup & Semantic Annotation*, ACM Press, New York, 2001.

[30] Verity, "Verity K2 Architecture : Unprecedented Performance, Scalability and Fault Tolerance," Verity White Paper, *http : //www.verity.com/products/k2_enterprise/*.

[31] Wenger, E., R. McDermott and W.M. Snyder, *A Guide to Managing Knowledge : Cultivating Communities of Practice*, Harvard Business School Press, Boston, Massachusetts, 2002.

[32] Zack, M., "Managing Codified Knowledge," *Sloan Management Review*, Vol.40, No.4 (1999), pp.45-58.

[33] Zadeh, L., "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," *IEEE Transactions on Systems Management and Cybernetics*, Vol. SMC-3, No.1(1973), pp.28-44.