한국경영과학회/대한산업공학회 2003 춘계공동학술대회
2003 년 5 월 16 일-17 일 한동대학교(포항)

기업지식포탈을 위한 지능형 지식추천 모델 비교

A Comparative Analysis of Knowledge Recommendation Model for Enterprise Knowledge Portal

임남구*, 김광래*, 이홍주, 변현진**, 김종우***, 박성주****

** 충남대학교 통계학과*
대전광역시 유성구 궁동 220번지
Tel: +82-42-821-5432, Fax: +82-42-822-0260, E-mail: {nglim, paladink}@stat.cnu.ac.kr

*** 한국과학기술원 테크노경영대학원*
서울특별시 동대문구 청량리동 207-43번지, 130-722
Tel: +82-2-958-3647, Fax: +82-2-958-3604, E-mail: {hjlee, hjbyun, sjpark}@kgsm.kaist.ac.kr

**** 한양대학교 경영대학 경영학부*
서울특별시 성동구 행당동 17번지, 133-791
Tel: +82-2-2990-1067, Fax: +82-2-2990-1169, E-mail: kjw@hanyang.ac.kr

Abstract

의사결정에 관련된 지식을 선별하고 이를 효과적으로 활용하기 위하여 많은 기업들이 지식관리시스템을 도입하여 활용하고 있다. 방대한 지식에서 사용자에게 적합한 지식을 제공하는 지식추천 기능은 지식관리시스템의 주요한 기능 중의 하나이다. 대부분의 시스템들이 사용자에게 직접 관심분야를 입력 받고 이 정보를 바탕으로 지식추천을 하고 있으나 사용자가 과거 지식관리시스템을 활용하면서 보인 관심표명 행동들을 활용한 지능적인 지식 추천 방안에 대한 연구는 미진한 편이다. 본 연구에서는 지식 카테고리 또는 문서 키워드를 활용하여 지식을 추천하는 방안과 사용자의 관심분야를 표현하는 프로파일 생성을 위한 다양한 방안을 설계하고 각 방안들의 지식추천 성과를 비교하였다.

Keywords:

기업지식포털, 지식 추천, 사용자 프로파일, 지능형 정보제공

1. 서론

지식기반 경제구조에서 지식 창출과 활용은 기업경쟁력의 근간이 되고 있으며, 이를 위한 지식 관리 프로세스가 매우 중요하다. 또한 기업들은 정보나 지식을 다양한 경로를 통해 흡수하기 때문에, 적합한 지식을 선별하고 활용하여 의사결정 문제를 해결할 수 있느냐에 많은 관심을 가지고 있다[1]. 다양한 원천의 정보들과 다양한 지식활동의 장에 접근하는 통합된 접근점으로서 지식포털(knowledge Portal), 기업지식포털(Enterprise knowledge Portal, EKP)이 활용되고 있다[5]. 효과적인 지식포탈이 되기 위해서는 다양한 원천에 흩어진 지식들의 효율적인 수집과 분류와 함께, 사용자에게 적합한 지식을 선별하여 제공하는 지식 추천 기술의 활용이 필요하다. 본 연구에서는 기업지식포탈 내에서 활용 가능한 지식 추천 방안들을 도출하고 이들의 성과를 비교하고자 한다. 정보검색 기술에 기반한 키워드 방식과 전자상거래 분야의 상품 추천 방식 중 하나인 카테고리 방식을 비교하며, 다양한 사용자 프로파일 구성 방안에 기반한 지식추천방안을 설계하고 신문기사를 활용한 실험을 통해 지식추천 방안의 성과를 비교한다.

2. 지식추천 모형

2.1 지능형 정보제공 방식

2.1.1 카테고리 방식

카테고리 방식은 지식 추천 모델 중 가장

간단하고 구현이 쉬운 방식이다. 사용자에게 미리 자신의 선호분야를 선택하도록 하고 해당분야에 분류된 문서를 추천하는 방식이다. 이 방안에서는 일단 사용자가 자신의 관심 분야를 입력하여야 하며 사용자가 관심분야를 수정하지 않는 한 같은 분야의 문서가 추천된다. 그러나 사용자의 관심분야가 고정되어 있지 않고 시간의 흐름에 따라 변화할 때는 사용자의 관심분야 변화를 반영하지 못할 수 있다. 자동 분류 기반 카테고리 방식은 이와 같은 단점을 해결하고자 사용자가 직접 관심분야를 입력하지 않고 사용자가 관심을 보인 문서들이 가장 많이 포함된 분야를 사용자의 선호분야라고 파악하고 정보를 추천하는 방식이다. 그러나 카테고리 방식은 작성된 사용자 프로파일의 일부분만을 가지고 지식을 추천한다는 측면에서 문서와 사용자에 관련된 많은 유용한 정보를 간과하고 있다는 단점이 있다.

2.1.2 키워드 방식

키워드 방식 중 대표적인 벡터공간(Vector space) 모델에서는 문서와 질의를 벡터공간에서 하나의 벡터로 취급한다. 즉, 문서집합에 나타나는 색인어들을 축으로 하여 벡터공간이 정의되고, 문서와 질의는 포함된 키워드의 빈도수에 의해서 각각 벡터로 표현된다. 벡터공간 모델에서 두 문서간의 유사도는 두 벡터 사이각의 코사인 값을 이용하고 있는데, 이는 각이 작을수록 유사도가 높다는 기본적인 아이디어에서 출발한다. 사용자 프로파일은 사용자가 작성한 문서 또는 사용자가 관심을 보인 문서들의 합으로 정의될 수 있으며, 사용자의 특정 문서에 대한 선호도는 앞의 문사가 유사도와 유사하게 정의될 수 있다. 즉 문서는 $D_i = (w_{i1}, \cdots w_{im})$ 로, 사용자 프로파일은 $P = (p_1, \cdots p_m)$ 으로 표현하면, 이 두 벡터의 유사도는 다음과 같이 계산된다.

$$sim(D_i, P) = \frac{\vec{d_i} \cdot \vec{p_i}}{|\vec{d_i}| \, |\vec{p_i}|} = \frac{\sum_{n=1}^{m} w_{nd} \quad w_{np}}{\sqrt{\sum_{n=1}^{m} w_{nd}^2} \sqrt{\sum_{j=1}^{t} w_{np}^2}}$$

2.2 사용자 프로파일 구성

2.2.1 개요

사용자 프로파일의 작성과정은 사용자의 흥미와 관심분야를 시스템이 활용할 수 있는 형태로 전환하는 과정이다. 지식추천 시스템을 위한 기존 연구에서는 사용자 프로파일이 갖추어야 할 조건을 다음과 같이 제시하고 있다[3][7].

1. **Specialization**: 무작위로 정보를 사용자에게 제공하는 것이 아니라 사용자의 관심분야에 맞는 정보를 제공할 수 있는 정보의 개인화.
2. **Adaptation**: 개인의 관심분야는 시간에 따라 변화하기 때문에 그러한 변화에 맞게 사용자 프로파일의 변화.
3. **Exploration**: 주어지는 정보만을 제공하는 것이 아니라 시스템이 개인의 관심분야에 맞는 정보를 직접 탐색하여 찾아낼 수 있는 능력.

사용자의 관심분야는 사용자가 직접 자신의 관심분야에 대해 입력할 수도 있지만 사용자의 편의성을 위해서는 사용자가 열람한 문서의 특성, 작성한 이메일이나 게시판 글이 사용자의 관심분야를 나타낸다고 간주하여 이러한 문서에서 추출된 벡터를 통해 사용자 프로파일을 작성하는 것이 가능하다. 또한 이렇게 함으로써 사용자의 관심분야변화를 반영할 수 있다. 본 연구에서는 **Average document vector, Rocchio algorithm, clustering** 방식을 통해 사용자 프로파일을 구성하였다. 개별 사용자 프로파일 구성방안에 대한 설명은 다음 절에서 이루어진다.

2.2.2 Average document vector

이 방법에서는 사용자가 관심을 보인 문서에 포함된 각 단어의 출현빈도를 합한 후 관심 있는 문서 전체의 수로 나눈 평균치를 사용자 프로파일로 사용한다. 사용자 프로파일 P는 개별 단어들의 출현빈도의 집합으로 이루어 지며, 개별단어들의 출현빈도 P_i 는 다음과 같이 계산된다.

$$P_i = \frac{\sum_{j=1}^{m} f_{ij}}{m}$$

f_{ij}는 i번째 단어가 j문서에서 나타나는 빈도수이며, m은 관심 있는 문서 전체 수를 나타낸다.

2.2.3 Rocchio algorithm

Rocchio algorithm은 사용자가 표현한 선호와 비선호 정보를 모두 고려하여 사용자 프로파일을 구축하는 방식이다. 사용자 프로파일 P는 다음과 같이 계산된다.

$$P = \alpha\mu_1 - \beta\mu_2$$

여기서 μ_1은 선호하는 문서들의 평균벡터, μ_2는 선호하지 않는 문서의 평균 벡터를 나타낸다. α, β는 선호문서와 비선호 문서에 대한 가중치로

실험에 따라 다르게 주어지지만 기존 문헌에 의하면 $\alpha = 0.75$, $\beta = 0.25$ 에서 가장 효과적인 결과가 나온다고 한다[4]. 이렇게 초기에 사용자 프로파일을 구축하게 되면 향후 사용자 프로파일 갱신 시에는 기존의 선호도 정보와 새로운 선호/비선호 문서에 따라 다음과 같이 갱신된다.

$$P^{new} = \gamma P^{old} + \alpha\mu_1 - \beta\mu_2$$

여기서 γ, α, β 는 각각 기존 프로파일, 신규 선호문서, 신규 비선호 문서에 대한 가중치이다[3][4].

2.2.4 클러스터링 방식

사용자의 관심 분야를 고려한 지식추천 과정에서 관심분야의 개수에 대한 논의가 필요하다. 즉, 사용자에게 하나의 관심분야만 할당하는 방안 이외에도 다수의 관심 분야를 지원하는 방법이 가능하다. 하나의 관심분야를 지원하는 경우에는 사용자 프로파일의 관리가 쉽고, 계산의 복잡함을 감소시킬 수 있다는 장점이 있다. 하지만 사용자의 관심 분야가 한 분야에 국한되지 않고, 정보를 얻는 분야가 여러 분야에 걸쳐서 종합된 정보를 가져오는 상황에서는 단지 하나의 관심분야만을 가지고는 사용자의 지식 요구를 제대로 반영하지 못한다. 이와 같은 문제의 해결을 위해 클러스터링 방식을 통한 지식추천 방식을 제시되었다. 이 방식은 휴리스틱 군집분류 알고리즘에 의해 적당한 수의 군집으로 사용자의 선호한 문서를 클러스터링하게 된다. 즉, 각각의 클러스터는 사용자의 다른 관심분야를 나타나게 된다. 즉, 사용자의 프로파일이 하나의 벡터로 표현되지 않고, 다수의 벡터(클러스터 별 하나씩)로 표현된다. 추천 대상 문서와 사용자 프로파일간의 유사도는 이들 클러스터 벡터들과 대상 문서의 벡터간의 유사도 중 가장 작은 값으로 결정된다.

3. 지식 추천 모델 비교 실험

3.1 지식추천 모델

본 연구에서 비교 실험된 지능형 정보제공 모델들을 구체적으로 정리하면 [표1]과 같다.

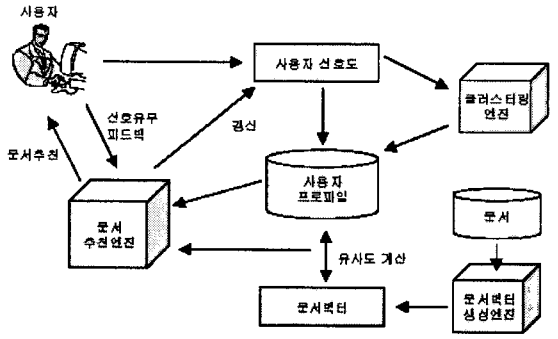
[표 1] 지식추천 모델

추천 방식	사용자 프로파일	구분	설명
카테고리방식 (Category based model)	사용자가 관심분야 직접 선택	CU	최초에 사용자가 자신의 관심분야를 직접 입력하고, 입력된 분야의 기사를 추천
	자동으로 사용자 관심분야 선택	CA	사용자가 선호한다고 선택한 기사 중 가장 많이 선택된 분야를 자동으로 선택하여, 해당분야 기사 추천
벡터공간모델 (Vector space model)	Average document vector	VA	Average document vector + 벡터공간모델
	Rocchio algorithm	VR	Rocchio algorithm + 벡터공간모델
	Clustered user profile	CS	Clustered user profile + 벡터공간모델
무작위추천		RA	전체 기사 중 무작위로 추출하여 추천

3.2 추천 모형의 구현

추천 모형의 구현은 Java 언어를 사용하여 이루어졌으며 전체적인 구조는 [그림 1]과 같다. 추천 대상이 되는 문서는 시스템이 활용할 수 있는 형태로 일차적으로 변환된다. 즉, 각 문서에서 단어와 그 빈도수를 추출하여 이를 벡터화한다. 이를 위해서 먼저 형태소 분석기를 거쳐 단어의 어근과 어미를 분리하고, 접속어, 조사 등을 제외하는 선작업 (Preprocessing) 과정을 거친다. 이렇게 추출된 단어는 문서의 길이로 인해 빈도수가 영향 받는 것을 없애기 위해 문서의 길이에 따라 빈도수를 조정하는 정규화 (Normalizing)과정을 거치게 된다. 정규화 과정을 거친 문서는 정보 필터링을 통해 단어의 상대적인 중요도(Weight)를 가리는 과정을 거치게 된다. 즉 하나의 문서에 단어가 얼마나 자주 나오는지(tf: term frequency)와 전체 문서 집합에서 그 단어가 출현하는 정도에 대한 역(idf: inverse document frequency)를 곱하여 계산한다. 이 과정에서 일정한 임계치(threshold)에 미치지 못하는 단어는 전체 추천에 미치는 정도가 미미하다고 간주하여 제외한다. 위에서 생성된 문서벡터와 최초에 생성된 사용자 프로파일과의 유사도 계산을 실시하여 유사도가 높은 문서를 추천하게 된다. 또한 처음에 사용자가 선택한 선호분야에 해당하는 정보를 사용자 프로파일에 포함시켜 그 분야에 해당하는 문서를 추천한다. 추천된 문서에 대해 사용자는

자신의 선호/비선호 여부를 표현하여 사용자 프로파일을 갱신하며 이를 다음 추천에 활용한다.



[그림 1] 지식추천의 구조

3.3 실험 환경 및 과정

실험은 웹 기반으로 실시되었으며, 사용자의 선호도를 표시할 수 있는 인터페이스 및 환경은 JSP에 의해 작성되었고 각각의 추천방식의 알고리즘은 JAVA로 구현되었다.

3.3.1 자료의 수집

실험자료는 신문사이트(중앙일보) 기사 중 정치, 경제, 사회, 국제, 문화생활, 정보과학, 스포츠 7개 분야에 해당하는 기사 중 실험실시 일주일전의 자료를 가져와서 기사 ID, 제목, 내용, 카테고리에 따라 DB에 저장하였다. 전체 기사의 수는 각 분야별 100개씩 총 700개의 기사를 수집하였다. 기사의 활용은 아래의 표와 같이 이루어 졌다.

[표 2] 실험 자료의 구성

Group	비율(개수)	설명
초기 사용자 프로파일 생성 대상	20%(140개)	최초에 주어지는 기사에 대해 사용자가 선호/비선호를 표시하는데 사용
1차 추천 대상	40%(280개)	1차 추천을 위한 데이터 집합
2차 추천 대상	40%(280개)	사용자 프로파일 피드백 후의 2차 추천을 위한 데이터 집합

3.3.2 실험 과정

1. 사용자가 실험사이트에 가입하며, 자신의 선호분야를 정한다.
2. 최초 사용자 프로파일 생성을 위해 140개의 기사에서 각 분야별로 3개씩 총 21개의 기사를 설문자에게 보여준다.
3. 사용자가 제시된 신문기사에 대해 자신의 선호/비선호 여부를 체크하게 되면, 그 결과에 의거 **Average document vector, Rocchio algorithm, Clustering algorithm**의 3가지 사용자 프로파일이 생성되고 저장된다.
4. 1차 추천 대상 280개 기사 중 사용자 프로파일과 VA, VR, CS의 3가지 추천 방식간의 유사도 계산에 의해서 상위 4개의 기사를 추천한다. 또한 CA, CU에 의해 4개, RA에 의해 4개가 추천되어 총 24개의 기사가 추천된다.
5. 추천된 기사에 대해 사용자는 5점 척도(매우 선호, 선호, 보통, 비선호, 매우 비선호)에 의해 자신의 선호여부를 선택하게 된다. 그러면 그 결과에 의해 각 3가지 사용자 프로파일 및 CA를 위한 관심분야 정보도 갱신된다.
6. 갱신된 사용자 프로파일과 2차 추천 대상을 비교하여 4번째 단계와 같은 실험이 반복되고 설문 참여자가 자신의 선호여부를 체크하면 설문은 종료된다.

실험을 통해 얻어진 자료의 분석을 위해서는 SPSS와 MS EXCEL이 사용되었으며, 실험의 샘플 수는 총 207명이었으나 무성의하게 작성된 설문응답은 제거되어 총 173명에 대해서만 분석되었다.

3.4 실험결과

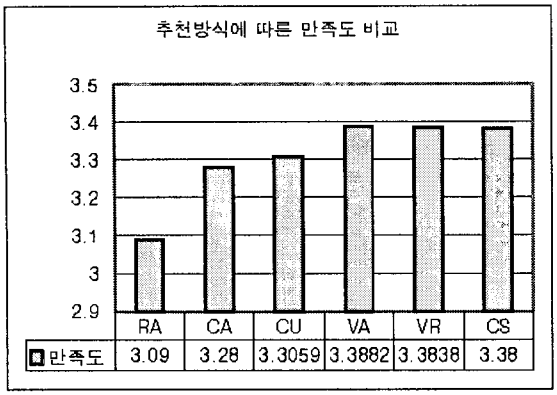
3.4.1 설문 응답자의 특성

실험에 참여한 응답자들은 주로 대학생과 대학원생들이며, 이들이 설문사이트를 통해 선호한다고 명시적으로 표현한 관심분야는 표 3과 같다.

[표 3] 카테고리 별 선호도

선호분야	빈도수(명)	퍼센트
정치	12	6.9%
경제	37	21.4%
사회	8	4.6%
국제	6	3.5%
문화생활	47	27.2%
정보과학	27	15.6
스포츠	36	20.8
합계	173	100%

3.4.2 각각의 추천 모델에 대한 만족도 비교



[그림 2] 추천 방식에 따른 만족도 비교

[그림 2]는 각 추천 방식에 의해 추천된 기사에 대한 설문자의 평균 만족도를 나타낸 것이다. 가장 높은 만족도를 나타내고 있는 것은 키워드 방식 중에 Average document vector와 벡터공간모델을 활용한 VA 추천방식으로 평균 만족도는 3.3882이다. 벡터공간모델을 사용한 방식과 클러스터링을 활용한 방식 등 키워드 기반 방식들이 대체로 카테고리 방식이나 무작위 추천 방식보다 좋은 성능을 보였다. 통계적인 기법을 통한 개별 추천 방식의 성과에 대한 비교를 위해서 사용한 가설은

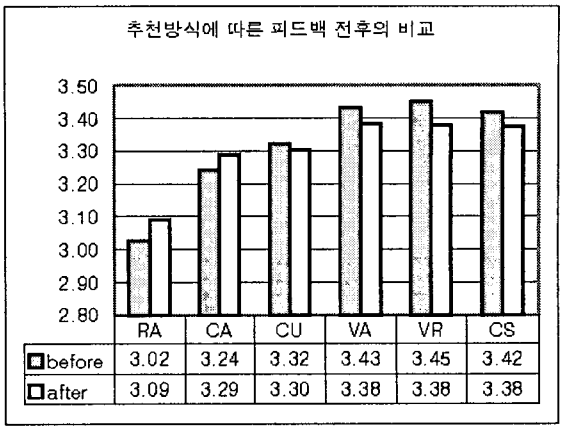
$$H_o : \mu_{CS} = \mu_{CA} = \mu_{CU} = \mu_{VA} = \mu_{VR} = \mu_{RA}$$

이다. 이를 검정하기 위해서 repeated measure analysis 방법을 사용하였다. 그 결과 신뢰수준 95%에서 p-value의 값은 0.001로써 귀무가설을 기각하게 되어 각각의 추천 방식간에는 차이가 있다고 말할 수 있다. 아래의 표에서 알 수 있듯이 사후 검정을 통해 추천방식 개별간의 차이를 검정해본 결과 무작위에 의한 추천방식과 나머지 5가지 방식에 의한 추천방식에 차이가 있고 5가지 방식간에는 만족할 만한 차이는 존재하지 않고 있음을 알 수 있다. 즉 카테고리 방식이나 키워드 방식 간에는 차이가 통계적으로 유의하지 않았으며, 사용자 프로파일을 Average document vector, Rocchio algorithm, Clustering algorithm을 활용하여 생성한 키워드 방식간에도 추천성과의 큰 차이가 없다고 볼 수 있다.

[표 4] 추천 방식의 만족도 비교 검정

귀무가설	p-value	검정결과
$H_o : \mu_{CS} = \mu_{CA} = \mu_{CU} = \mu_{VA} = \mu_{VR}$	0.643	H_o 채택
$H_o : \mu_{CA} = \mu_{CU}$	0.757	H_o 채택
$H_o : \mu_{CS} = \mu_{VA} = \mu_{VR}$	0.951	H_o 채택

3.4.3 각 추천 모델의 피드백 전후 결과 비교



[그림 3] 피드백 전후 비교

[그림 3]은 피드백 전후의 결과를 나타낸 것이다. 피드백의 효과가 나타난 것은 CA방식에 의한 추천방식뿐이고 다른 방식은 피드백 반영 전보다 개선된 결과가 나타나지 않았다. 좀더 통계적으로 분석하기 위해 대응표본 T-test를 실시하였다. 그 결과 5가지 추천방식 모두 피드백 전후의 추천에서 만족도 차이는 95% 신뢰구간에서 존재하지 않음을 알 수 있다.

[표 5] 각 방식의 피드백 검정 결과

추천모델	P-value	결과
$H_0 : \mu_{CU,initial} = \mu_{CU,feedback}$	0.655	H_o 채택
$H_0 : \mu_{CA,initial} = \mu_{CA,feedback}$	0.216	H_o 채택
$H_0 : \mu_{VA,initial} = \mu_{VA,feedback}$	0.219	H_o 채택
$H_0 : \mu_{VR,initial} = \mu_{VR,feedback}$	0.093	H_o 채택
$H_0 : \mu_{CS,initial} = \mu_{CS,feedback}$	0.256	H_o 채택

3.5 결과의 해석

추천 방식 중 키워드 기반의 추천 방식들이 카테고리 방식에 비하여 대체로 좋은 성능을 보이는 것을 알 수 있었다. 하지만 그 차이는 통계적으로 유의하지는 못했다. 피드백 전후의 결과 비교에서는 실험 전에 기대했던 것은 피드백 전의 결과보다는 후의 결과가 보다 향상된 결과가 나올 것이라는 예상이었다. 하지만 분석결과 사용자 프로파일 갱신 전후의 차이 존재하지 않음을 알 수 있다. 이는 피드백이 한번만 이루어졌기에, 사용자 프로파일의 변경에 미치는 정도가 미미하였고 피드백이 거의 같은 시점에서 이루어졌기 때문에 큰 효과가 없었다고 보여진다. 따라서 장시간에 걸쳐 피드백 정보가 누적되면 사용자 프로파일이 정교화 되어 추천성과의 개선이 있으리라고

예상된다.

4. 향후 연구 방향 및 결론

본 연구에서 기업지식포탈의 지식추천을 위한 5가지 지능형 정보제공 모델을 구현하여 비교하였다. 이를 위하여 웹 설문을 통해서 피실험자가 각 추천 방식에 의해 추천되는 기사에 대해서 선호/비선호의 정보를 나타냄으로써 실증적인 조사를 수행하였다. 실험 결과 키워드 기반 방식(벡터 공간 모델 및 클러스터링 모델)이 카테고리 기반 방식에 비해서 나은 성능을 보였다. 피드백 효과에 대한 것도 실험에 포함하였으나, 피드백의 효과는 거의 없는 것으로 나타났다. 본 연구를 통해서 구현된 지식 추천 모델들은 향후 과학기술지식포탈 시스템에 통합될 예정이며, JAVA 컴포넌트 형태로 모듈화하여 다른 지식포털 시스템에도 활용이 가능하게 할 예정이다. 또한 지식 추천 기법을 향상을 위해서 어의적(semantic) 정보를 사용하는 LSI (Latent Semantic Index) 기법의 활용 및 비교 등의 연구를 계속적으로 수행할 예정이다.

Acknowledgments

본 연구는 과학기술부의 지원에 의한 '과학기술 연구를 위한 지식포털 구축'과제의 수행을 통해 얻어진 결과물이다.

References

[1] 김상수, 김용우 "지식관리시스템의 특성과 성공 요인에 관한 탐색적 연구", Hanyang Business Review, Vol. 12, 2000, pp.65~82.

[2] 신동호, "Latent Semantic Analysis를 이용한 내용기반 정보 검색 시스템", 석사학위 논문, 서울대학교, 1999.

[3] Kjersti Aas, " A survey on personalized information filtering systems for the World wide web", Norwegian Computing Center, 1997.

[4] R. Baeza-Yates and B. Ribeiro-Neto, "Modern information retrieval," Addison Wesley, 1999.

[5] Claudia Dias, "Corporate Portals: a Literature Review of a New Concept in Information Management," International Journal of Information Management, Vol. 21, 2001, pp. 269-287.

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

[7] Beerud Dilip Sheth, " A learining approach to personalized information filtering," Master thesis, Department of Computer science and engineering, MIT, 1994.