전공분류표, 사용자 프로파일, LSI를 이용한 검색 모델

우 선 미†

요 약

현재 대부분의 도서관 정보검색 시스템들은 키워드 정합매칭(exacting matching) 방법으로 검색 서비스를 제공하고 있으므로, 검색 결과의 양이 방대하고 비적합한 결과가 많이 포함되어 있다. 따라서 본 논문에서는 키워드기반 검색 엔진의 단점을 보완하고 현재 도서관 검색 환경을 고려하여 보다 적합한 결과를 사용자에게 신속하게 제공하기 위하여 전공분류표와 사용자 프로파일을 이용한 검색 모델 SULRM(Retrieval Model using Subject Classification Table, User Profile & LSI)을 제안한다. SULRM은 키워드 검색 결과로 얻은 자료들을 분류된 자료의 경우와 미분류된 자료의 경우로 나누어, 분류된 자료의 경우에는 전공분류표를 생성하여 자료 필터링을 수행하고, 미분류된 자료의 경우에는 사용자 프로파일과 LSI(Latent Semantic Indexing)을 이용하여 자료의 순위를 결정해서 사용자에게 제시한다. 실험평가는 우리 대학의 디지털 도서관을 실험환경으로 하여 필터링 방법, 사용자 프로파일 갱신 방법, 그리고 문서순위결정 방법의 성능을 측정한다.

키워드 : 전공분류표, 정보 필터링, 사용자 프로파일, SVD, 잠재적 구조 분석, 문서순위결정

Retrieval Model using Subject Classification Table, User Profile, and LSI

Seon-Mi Woo†

ABSTRACT

Because existing information retrieval systems, in particular library retrieval systems, use 'exact keyword matching' with user's query, they present user with massive results including irrelevant information. So, a user spends extra effort and time to get the relevant information from the results. Thus, this paper will propose SULRM, a Retrieval Model using Subject Classification Table, User profile, and LSI(Latent Semantic Indexing), to provide more relevant results. SULRM uses document filtering technique for classified data and document ranking technique for non-classified data in the results of keyword-based retrieval. Filtering technique uses Subject Classification Table, and ranking technique uses user profile and LSI. And, we have performed experiments on the performance of filtering technique, user profile updating method, and document ranking technique using the results of information retrieval system of our university' digital library system. In case that many documents are retrieved, proposed techniques are able to provide user with filtered data and ranked data according to user's subject and preference.

Key Words : Subject Classification Table, Information Filtering, User Profile, SVD, LSA(Latent Semantic Analysis), Document Ranking

1. 서 론

현재는 정보검색 시스템을 이용하여 얻게 되는 검색결과의 양이 매우 많아 그 중에서 적합한 결과를 찾기 위하여 또다시 노력을 기울여야만 하는 실정이다. 그 이유는 대부분의 검색 시스템이 사용자가 입력한 질의와의 정합매칭(Exact Matching)으로 검색을 수행하기 때문에 사용자가 입력한 질의가 비적합 자료의 키워드와 일치하더라도 그 자료까지 결과로 제공하기 때문이다. 전공분야가 점차적으로 세분화되고 검색 결과가 대량화됨에 따라 사용자 중심의 정보검색이 발달하게 되었다. 사용자 중심의 정보검색 방법으로서 사용자에게 검색 편의성을 제공하기 위하여 정보 필터링(Information Filtering)[1, 2, 3, 4] 방법과 문서순위결정(Document Ranking)[5, 6, 7] 방법이 여러 분야에서 활발히 진행되고 있다. 그러나 이런 방법들이 도서관 시스템에는 직접적으로 응용되지 못하고 있는 실정이다. 대부분의 도서관에서 사용하고 있는 정보검색방법은 MARC(MAchine Readable Cataloging) 포맷으로 색인파일을 생성하고, 검색질의와 일치된 결과를 제시하는 이른바 정합매칭 방법을 사용하고 있다. 최근 웹을 통한 이차검색(결과 내 검색)을 지원하고 있기는 하지만, 이 방법은 검색을 다시 해야 하는 번거로움이 있다. 대부분의 도서관에서 단행본은 주제 분류표에 따라 분류되어 있지만, 학위논문이나 기사와 같은 문서들은 분류되어 있지 않다. 또한 일부 도서관에서 자동 색

※ 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2004-003-D00324).
† 정 회 원 : 전북대학교 전북지역전자정보사업단 기금교수
논문접수 : 2005년 3월 24일, 심사완료 : 2005년 8월 11일

인 시스템을 도입하여 사용하고 있으나, 전문을 대상으로 하지 않고 주로 제목을 대상으로 하고 있으며 동의어나 동음이의어 처리가 부족하여 검색을 위한 충분한 색인을 지원하고 있지 못하다. 색인 방법이나 검색 방법의 단점이 해결되더라도 데이터의 양이 방대하므로 검색결과의 양 또한 매우 클 수밖에 없고, 사용자가 모든 검색결과를 참조하기에는 너무 많은 시간과 노력이 소비된다.

본 논문에서는 키워드기반 검색 엔진의 단점을 보완하고 현재 도서관 검색 환경을 고려하여 보다 적합한 결과를 신속하게 제공하기 위하여 전공분류표(Subject Classification Table)와 사용자 프로파일(User Profile), LSI(Latent Semantic Indexing)을 이용한 검색 모델 SULRM(Retrieval Model using Subject Classification Table, User Profile & LSI)을 제안한다. SULRM은 키워드 검색 결과로 얻은 자료 중에서 분류된 자료의 경우에는, 전공분류표를 생성하여 전공과 관련 있는 분류에 속하는 결과를 제공함으로써 필터링 효과를 얻는다. 그리고 미분류된 자료의 경우에는 사용자 프로파일과 LSI를 이용하여 자료의 순위를 결정해줌으로써 사용자가 비적합한 결과까지도 모두 검토해야하는 수고를 덜어준다.

실험 평가는 분류된 자료의 필터링 성능과 미분류된 자료의 순위결정 성능으로 나누어 평가 한다. 실험 데이터는 우리 대학의 디지털 도서관에서 소장하고 있는 색인 데이터베이스와 검색 엔진을 이용한다.

2장에서 정보 필터링, 사용자 모델링, 문서순위결정, LSI에 관하여 간단히 기술한다. 3장에서는 본 논문에서 제안하는 SULRM의 구성과 방법을 예를 들어 설명하고, 4장에서 본 논문에서 제안하는 방법을 실험 평가한다. 5장에서 결론과 향후 연구 방향에 대하여 기술한다.

2. 관련 연구

2.1 정보 필터링과 문서순위결정

정보 필터링 방법은 사용자의 기호를 저장해 놓고 이를 참조하여 필요가 없는 정보를 여과시켜 줌으로써 검색된 결과의 수를 줄여 주는 방법이다[1, 2, 3, 4]. 정보 필터링은 사용자의 요구를 한 번 표현해 놓으면 사용자의 관심이 변하지 않거나 정보자원의 커다란 변동 없이는 계속 사용할 수 있으나, 특정 관심분야로 한정된다.

정보 필터링 시스템에서 정보요구의 표현을 보통 프로파일(profile)이라고 한다. 프로파일이 정보검색과 데이터베이스 시스템에서 사용되는 질의(query)와 같은 역할을 하므로 질의어가 정보 필터링 분야에서 프로파일 대신 사용되기도 한다. 사용자 프로파일은 정보 필터링 분야의 핵심이 되고 있다[8]. 이러한 프로파일 모음을 사용자 모델이라고 한다. 현재 이용되고 있는 필터링 기법을 크게 구분하면 규칙기반 필터링(Rule-based Filtering), 내용기반 필터링(Content-based Filtering), 그리고 협조적 필터링(Collaborative Filtering)으로 구분할 수 있다. 규칙기반 필터링 기법은 미리 정의된 규칙에 맞는 적합한 정보만을 걸러서 간단히 제시하는 시스템에 적용된다. 내용기반 필터링 기법은 사용자 프로파일과 정보들의 내용간의 유사성을 고려하여 필터링하는 방법으로서 가장 많이 사용되는 기법이다. 협조적 필터링은 유사한 관심분야의 다른 사용자의 학습된 선호도를 참고하여 사용자 피드백을 통해 확장 및 축소를 수행하여 개인의 선호도를 구축해 나가는 방식이다. 인터넷상에서 서비스하는 대부분의 정보 필터링 시스템은 사용자의 관심을 표현하기 위해 키워드 벡터를 사용하고 있다[8, 9]. 이러한 시스템들은 보통 정보 필터링을 위한 사용자 프로파일 학습에 점진적 적합성 피드백(incremental relevance feedback) 방법을 사용한다.

문서순위결정(Document Ranking) 방법은 사용자의 질의어와 검색된 문서들이 얼마나 유사한가에 따라 문서의 순위를 결정하고, 이 순서에 따라 사용자가 가장 적합한 문서를 참조하도록 하는 방법이다[5, 6, 7, 10]. 문서순위결정 기법에는 가중치와 유사도를 이용한 방법, 적합성 피드백을 이용한 방법, 데이터 퓨전(Fusion)을 이용한 방법, 불리안 모델(Boolean Model)을 이용한 방법 등이 있다. 가장 기본적인 문서순위결정 기법은 벡터 모델에서 가중치와 유사도를 이용한 방법이다. 이 방법은 전체 문서에서 추출한 용어로 구성된 벡터로 문서와 질의를 표현하고, 두 벡터를 곱하여 각 벡터 요소를 더한 값이 큰 순서대로 문서의 순위를 결정하는 방법이다.

2.2 LSI

키워드 기반 정보검색 시스템은 적합한 문서일지라도 질의어를 포함하지 않는 문서는 검색결과로 얻을 수 없는 문제점이 발생한다. 이런 문제를 해결하기 위해서 문서들 간의 그리고 문서를 구성하고 있는 용어들 간의 의미론적 유사성(Semantic Similarity)을 기본으로 하여 내재되어 있는 잠재적(Latent) 근접 구조를 찾는 방법이 등장하였다. 본 논문에서는 LSI를 이용하여 분류된 문서들의 색인어 가중치를 결정한다. 잠재적 의미의 구조를 해석하는 방법은 먼저, 문서와 문서에 포함된 용어로 행렬을 구성한다. 구성된 행렬을 잠재적 의미 구조 모델(Latent Semantic Structure Model)로 만들기 위해 SVD(Singular Value Decomposition) 기법으로 분석한다[11]. 분석할 문서-용어 행렬을 X라고 했을 때, 3개의 행렬로 분해되는 SVD 방법은 (식 1)과 같고, 축소화된 SVD(Reduced SVD) 방법은 (식 2)와 같다.

$$X = T_0 S_0 D_0' \tag{1}$$

단, T_0 : $T_0' T_0 = I$인 직교행렬(orthogonal),
D_0 : $D_0' D_0 = I$인 직교행렬
S_0 : 대각(diagonal) 정방행렬,
D_0' : D_o의 전치행렬(transpose)

$$\hat{X} = TSD' \tag{2}$$

단, $\hat{X}$: 분석결과 행렬(t×d) ,
T : $T'T=I$인 직교 행렬(t×k)
D : $D'D=I$인 직교 행렬(k×d) ,
S : 대각 정방 행렬(k×k)
k : 행렬의 축소화된 계수 ($k \leq m$)

분석 결과로 얻을 수 있는 용어-문서 행렬에서 값은 색인으로서의 가중치를 나타낸다. 질의와 분석 결과 문서들과 비교하여 유사성을 계산할 때 질의를 비교 가능한 벡터로 변환시키기 위한 공식은 (식 3)과 같다.

$$D_q = X'_q T S^{-1} \quad (3)$$

단, X_q : 질의로 구성된 용어 벡터,
D_q : X_q를 D의 행과 비교 가능한 형태로 유도한 행렬

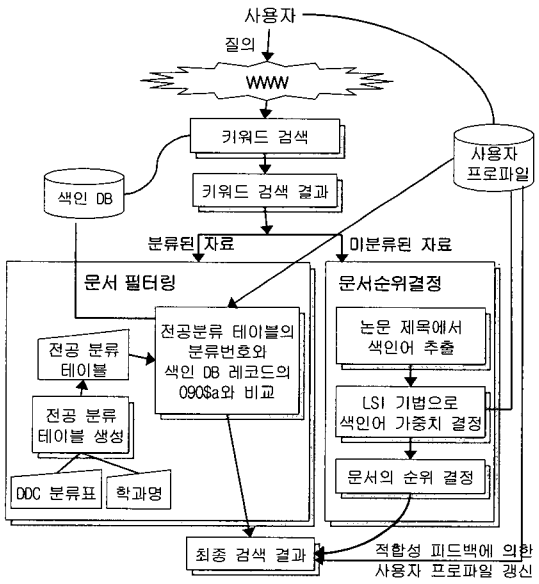
두 문서들 간의 유사정도는 (식 4)와 같이 행렬의 두 열을 연산함으로써 얻을 수 있다.

$$\hat{X}\hat{X}' = DS^2T' \quad (4)$$

3. SULRM 설계

본 논문에서 제안하는 SULRM(Retrieval Model using Subject Classification Table, User Profile & LSI)의 검색 과정을 도식화하면 (그림 1)과 같다.

(그림 1)에서 키워드 검색은 사용자가 입력한 질의와 색인 DB의 키워드와 일치하는 검색결과를 제공하는 것으로서 기존 도서관 검색 모형을 의미한다. 일반적으로 키워드 검색을 실시하면 대량의 검색결과를 얻게 되고, 검색결과 중에는 본인의 관심분야와 다른 결과들도 있을 수 있다.



(그림 1) SULRM의 검색 과정

이런 문제점을 해결하기 위하여 도서관 자료의 특성에 맞게 분류된 자료와 미분류된 자료로 나누어 SULRM을 설명한다.

3.1 분류된 자료의 필터링

단행본과 같이 분류된 자료의 키워드 검색결과 중에서 전공분류표를 이용하여 적합하지 않은 자료들을 제외한 검색결과를 사용자에게 제공한다. 분류된 자료를 필터링하여 사용자에게 제공하는 알고리즘은 다음과 같다.

【알고리즘 1】 분류된 자료의 필터링
(입력) ① 키워드 매칭 방법으로 검색된 분류된 자료
② DDC(Dewey Decimal Classification)
③ 학과에 대한 커리큘럼
(출력) 전공분류에 속하는 분류된 자료의 키워드 검색결과

1. 전공과목에 해당하는 분류번호를 DDC 분류표를 보고 선택하여 분류번호리스트를 작성한다.
2. 작성된 분류번호리스트를 참조하여 사서가 경험지식에 의해 전공분류번호리스트를 생성하여 전공분류표에 추가한다.
3. 키워드 검색결과 중에서 전공분류표의 전공분류번호의 범주에 포함되는 결과를 사용자에게 제공한다.

단계 2에서 전공분류번호리스트를 생성할 때, 각 분류번호는 중복되지 않도록 하며, 하나의 전공이 한 개 이상의 서로 다른 전공분류 번호를 가질 수 있다. (그림 2)는 전공과 관련된 분류번호와 최종 전공분류번호를 생성하는 과정을 나타내고 있다. 분류번호는 커리큘럼에 포함된 과목들을 대상으로 하여 각 대학교의 중앙도서관 분류체계에 따라 과목마다 분류번호를 부여한다.

과목명(전공)	분류 번호	전공분류번호
정보처리 연습	004	004
도서관문화사	020.9	020.9
한지와 인쇄사	655 / 676.2	025.4
컴퓨터 정보처리	004	025.52
정보서비스론	025.52	001.61
정보시스템 분석	001.61	384
지식정보화사회	384	951
한국학정보	951	027.4
공공도서관 경영	027.4	655.6
독서교육론	028	010
서지학 개론	010	020
문헌정보학연구법	020	
:	:	:

(그림 2) 문헌정보학과 전공분류번호 생성 과정 예

(그림 2)에서 과목명 '정보처리 연습'과 '컴퓨터 정보처리'에 대한 분류번호는 004로서 같다. 분류번호가 같을 경우, 전공분류번호에 004는 한 개만 존재하게 한다. '한지와 인쇄사'는 655와 676.2라는 두 개의 분류번호를 갖게 되는데 전공분류번호에는 두 개 모두를 수용한다. 이러한 방법으로 각 학과별 전공분류번호리스트를 생성한다. <표 1>은 학과

〈표 1〉 전공분류표의 일부

학과	간호학과	…	문헌정보학과	…	환경·화학공학부
분류번호 리스트	131.322 610 : :	: :	***001.61*** ***004*** : ***010*** :	: :	301.3 628 : :

별 전공분류번호를 포함하고 있는 전공분류표의 일부를 나타내고 있다.

3.2 미분류된 자료의 순위결정

사용자 프로파일과 LSI 기법을 이용하여 미분류된 자료에 대한 순위를 결정하는 알고리즘은 다음과 같다.

【알고리즘 2】 미분류된 자료의 순위결정
(입력) ① 키워드 매칭 방법으로 검색된 미분류된 자료
② 사용자 프로파일
(출력) 적합성 정도를 기준으로 순위가 결정된 미분류된 자료

1. 사용자 프로파일을 생성한다.
2. 사용자 프로파일을 사용자 위주로 학습시킨다.
3. LSI 기법을 사용하여 미분류된 자료의 잠재적인 구조를 분석한다.
   3.1 키워드와 키워드 출현빈도를 이용하여 용어-자료 행렬 X를 생성한다.
   3.2 세 개의 행렬 T_0, D_0, S_0로 분해한다.
   3.3 문서순위결정을 위하여 가장 큰 영향을 미치는 하나의 차원으로 축소화하여 최종 분석결과 행렬 $\hat{X}$를 구한다.
4. 사용자 프로파일을 이용하여 분석결과 행렬과 비교 가능한 행렬인 슈도문서 DP를 생성한다.
5. 슈도문서와 분석결과 행렬을 비교하여 유사 정도를 계산한다.
6. 5의 유사성 정도를 기준으로 미분류된 자료의 순위를 결정한다.

본 논문에서는 사용자의 관심분야와 선호도를 표현하기 위하여 사용자 식별자, 사용자 암호, 전공(subject), 관심분야, 용어열 T_i(Term Array) 그리고 선호도 벡터 $\vec{P_i}$(Preference Vector)로 사용자 프로파일을 구성한다. 전공은 학과명을 뜻한다. 관심분야 별로 용어열과 선호도 벡터를 구성한다. 해당 관심분야는 사용자가 원하는 용어로 저장가능하고, 용어열 T는 해당 관심분야의 색인어(자료의 제목에서 추출)로 구성한다. 선호도 벡터(Preference Vector) $\vec{P}$는 용어열 T에 대응하는 사용자의 선호도를 나타낸다. 사용자의 선호도는 처음엔 0으로 시작하고, 갱신을 통해 0과 1 사이의 값으로 정규화하여 표현한다.

적절한 갱신 방법을 이용하여 사용자 프로파일의 선호도 벡터를 특정 사용자 위주로 학습시킨다. 본 논문에서는 사용자 프로파일을 갱신하기 위하여 '사용자 입력에 의한 갱신' 방법과 '사용자 적합성 피드백에 의한 갱신' 방법을 이용한다. 사용자 입력에 의한 갱신 방법은 사용자가 특정 용어에 중요성을 부여하고자 할 때, 직접 용어에 해당하는 선호도 벡터의 값을 변경하는 방법이다. 이 갱신 방법은 사용자 프로파일의 초기 학습에 유용하지만, 전문가에게 친숙한 방법으로서 관심분야에 관한 정보를 잘 모르는 사용자에게는 적절한 방법이 아니다. 따라서 이 방법의 적용은 사용자의 선택사항으로 한다. 사용자 적합성 피드백에 의한 갱신 방법은 널리 사용되는 학습 방법이다. 본 논문에서는 순위를 결정하여 제시한 결과 자료들 중에서 상위 5%에 대한 사용자의 평가를 사용자 프로파일의 선호도 갱신에 이용한다. 적절한 사용자 프로파일의 갱신 횟수는 실험을 통하여 결정한다. 사용자는 0~6 사이의 값으로 적합성 정도를 평가하는데, 값 0은 "비적합"을, 값 3은 "보통"을, 값 6은 "적합"을 나타낸다. 사용자가 평가한 자료 중에서 적합성 정도가 5이상인 자료들만을 대상으로 용어를 추출하여 용어열과 가중치 벡터를 갱신한다. 이 갱신 방법은 자료의 제목에서 색인어를 추출하고 추출된 용어의 가중치를 tf*idf 방법으로 구한다. 이렇게 구성된 용어의 가중치 벡터를 참조하여 사용자 프로파일의 선호도 값을 변경하는데, 갱신 공식은 식 5a~식 5c과 같다.

$$0.7 \le w_{rm} : P_i = | \ 2p_{ij} + w_{rm} \ |_{j=1,\ldots,n} \tag{5a}$$

$$0.5 \le w_{rm} < 0.7 : P_i = | \ p_{ij} + w_{rm} \ |_{j=1,\ldots,n} \tag{5b}$$

$$w_{rm} < 0.5 \ : \ P_i = |p_{ij}|_{j=1,\ldots,n} \tag{5c}$$

단, ${}_{ref}D_r$: 사용자가 적합하다고 평가한 r번째 자료
$= (w_{r1}, w_{r2}, \ldots, w_{rm})$
w_{rm} : ${}_{ref}D_r$을 구성하는 m번째 용어의 가중치,
$n = m$

사용자의 선호도와 자료를 비교하여 적합한 순서대로 순위를 결정하기 위해서는 자료들을 대표하는 색인어들의 가중치가 필요하다. 키워드 검색결과로 얻은 자료들 제목에서 색인어를 추출하여 가중치를 계산한다. 분류되지 않은 자료들은 보고서나 논문들처럼 제목이 자료 내용의 주제를 충분히 표현하는 것들이 대부분이므로, 제목에서만 추출한 키워드도 색인으로서의 가치가 충분하다고 볼 수 있다. 키워드 검색 결과로 얻은 자료들은 대부분이 서로 관련이 있는 자료들이다. 그러나 동음이의어 문제로 인하여 질의를 포함하고는 있지만, 관심분야의 자료가 아닐 수도 있다. 따라서 색인어 가중치를 결정할 때, 전체 자료들의 성향을 분석해서 그 성향과 거리가 가까운 자료들일수록 색인어 가중치를 높게 주어야 한다. 이와 같이 자료들의 성향을 분석하고 그 성향에 따라 가중치를 결정하기 위하여 본 논문에서는 LSI (Latent Semantic Indexing)을 이용한다. 분석 과정을 <표 2>와 같은 예로 설명한다. <표 2>는 저자의 대학교 디지털

〈표 2〉 질의 "평가 시스템"에 대한 검색 결과의 일부

자료	제목
D1	웹기반 교육 상호작용 시스템 설계 및 학업성취도 평가
D2	웹기반 학습자 개별적응 평가 시스템의 설계 및 구현
D3	학습을 위한 평가 시스템의 설계 및 구현
D4	제조에서 수 조립공정의 조립효율 평가 시스템
D5	전자계산실무 교과의 수준별 학습을 위한 웹기반 평가시스템

도서관 홈페이지의 학위논문 검색에서 질의 "평가 시스템"에 대한 검색결과 중 일부이다.

LSI 분석 단계를 예를 들어 설명하면 다음과 같다.

(1) <표 2>를 용어-자료 행렬 X로 구성하면 (그림 3)과 같이 8×5 행렬이 된다. 행(row)은 색인어, 열(column)은 자료, 값은 자료 내에 색인어가 출현한 빈도수를 나타낸다.

(2) 행렬 X를 식 1에 의해 분해하면 (그림 4)~(그림 6)과 같은 3개의 행렬 T_0(18×5), D_0(5×5), S_0(5×5)이 된다. 행렬 S_0의 값은 내림차순으로 정렬되는데, 본 논문에서는 검색결과를 대상으로 하므로 하나의 주제로 집중될 수 있다. 따라서 가장 큰 값 하나를 선택하여 분석을 수행한다.

(3) (2)에서 얻은 행렬을 1차원으로 축소화하면 행렬 T(18×1), S(1×1), D'(5×1)(D는 1×5)이 된다. (식 2)에 의해 축소화된 SVD를 수행하면 (그림 7)과 같은 행렬 $\hat{X}$(18×5)가 된다. $\hat{X}$는 색인어가 자료를 대표하는 정도 즉, 공통된 주제를 나타내는 색인어로서의 가중치를 나타낸다.

$$\begin{bmatrix}
0&1&0&0&0\\0&0&0&0&1\\1&0&0&0&0\\0&1&1&0&0\\1&0&0&0&0\\1&1&1&0&0\\0&0&0&1&0\\0&0&0&0&1\\1&1&1&1&1\\1&1&0&0&1\\0&0&0&0&1\\0&0&0&1&0\\0&0&0&1&0\\0&0&0&1&0\\1&1&1&1&1\\0&0&1&0&1\\0&1&0&0&0\\1&0&0&0&0
\end{bmatrix}$$

(그림 3) 색인어-자료 행렬

$$\begin{bmatrix}
0.1226322 & -0.153174 & -0.134719 & -0.263443 & 0.4213705\\
0.1070369 & 0.0705188 & 0.4173682 & 0.0394157 & 0.1210433\\
0.1158445 & -0.132489 & -0.123751 & 0.4262962 & -0.066571\\
0.2239265 & -0.189466 & -0.148419 & -0.48351 & -0.186574\\
0.1158445 & -0.132489 & -0.123751 & 0.4262962 & -0.066571\\
0.339771 & -0.321955 & -0.272169 & -0.057214 & -0.253145\\
0.0749465 & 0.4037579 & -0.165843 & 0.0132777 & 0.0622251\\
0.1070369 & 0.0705188 & 0.4173682 & 0.0394157 & 0.1210433\\
0.5217544 & 0.152322 & -0.020645 & -0.004521 & -0.069877\\
0.3455136 & -0.215144 & 0.1588983 & 0.2022693 & 0.4758424\\
0.1070369 & 0.0705188 & 0.4173682 & 0.0394157 & 0.1210433\\
0.0749465 & 0.4037579 & -0.165843 & 0.0132777 & 0.0622251\\
0.0749465 & 0.4037579 & -0.165843 & 0.0132777 & 0.0622251\\
0.0749465 & 0.4037579 & -0.165843 & 0.0132777 & 0.0622251\\
0.5217544 & 0.152322 & -0.020645 & -0.004521 & -0.069877\\
0.2083312 & 0.0342272 & 0.4036688 & -0.180652 & -0.486901\\
0.1226322 & -0.153174 & -0.134719 & -0.263443 & 0.4213705\\
0.1158445 & -0.132489 & -0.123751 & 0.4262962 & -0.066571
\end{bmatrix}$$

(그림 4) $X = T_0S_0D_0'$에서 T_0

$$\begin{bmatrix}
4.2336026 & 0 & 0 & 0 & 0\\
0 & 2.1804865 & 0 & 0 & 0\\
0 & 0 & 2.0613016 & 0 & 0\\
0 & 0 & 0 & 1.8218286 & 0\\
0 & 0 & 0 & 0 & 1.324411
\end{bmatrix}$$

(그림 5) $X = T_0S_0D_0'$에서 S_0

$$\begin{bmatrix}
0.4904398 & -0.288891 & -0.255087 & 0.7766386 & -0.088168\\
0.5191759 & -0.333994 & -0.277697 & -0.479947 & 0.5580677\\
0.4288398 & -0.079133 & -0.028239 & -0.400926 & -0.805168\\
0.3172935 & 0.8803887 & -0.341853 & 0.0241898 & 0.0824116\\
0.4531517 & 0.1537653 & 0.8603218 & 0.0718087 & 0.160311
\end{bmatrix}$$

(그림 6) $X = T_0S_0D_0'$에서 D_0

$$\begin{bmatrix}
0.2546245 & 0.2695436 & 0.2226433 & 0.1647311 & 0.2352654\\
0.2222436 & 0.2352654 & 0.1943295 & 0.1437821 & 0.2053465\\
0.2405312 & 0.2546245 & 0.2103201 & 0.1556134 & 0.2222436\\
0.4649446 & 0.4921869 & 0.4065469 & 0.3007992 & 0.429595\\
0.2405312 & 0.2546245 & 0.2103201 & 0.1556134 & 0.2222436\\
0.7054758 & 0.7468113 & 0.616867 & 0.4564126 & 0.6518386\\
0.1556134 & 0.1647311 & 0.1360681 & 0.1006752 & 0.1437821\\
0.2222436 & 0.2352654 & 0.1943295 & 0.1437821 & 0.2053465\\
1.0833327 & 1.1468079 & 0.9472646 & 0.7008698 & 1.0009672\\
0.7173993 & 0.7594335 & 0.6272929 & 0.4641266 & 0.6628556\\
0.2222436 & 0.2352654 & 0.1943295 & 0.1437821 & 0.2053465\\
0.1556134 & 0.1647311 & 0.1360681 & 0.1006752 & 0.1437821\\
0.1556134 & 0.1647311 & 0.1360681 & 0.1006752 & 0.1437821\\
0.1556134 & 0.1647311 & 0.1360681 & 0.1006752 & 0.1437821\\
1.0833327 & 1.1468079 & 0.9472646 & 0.7008698 & 1.0009672\\
0.4325637 & 0.4579087 & 0.3782331 & 0.2798502 & 0.399676\\
0.2546245 & 0.2695436 & 0.2226433 & 0.1647311 & 0.2352654\\
0.2405312 & 0.2546245 & 0.2103201 & 0.1556134 & 0.2222436
\end{bmatrix}$$

(그림 7) X의 성향 분석 결과 $\hat{X}$

검색 결과로 얻은 자료들의 잠재적인 성향을 분석하여 색인어 가중치를 결정한 후, 사용자의 선호도를 반영하는 사용자 프로파일과 비교한다. 즉, 행렬 $\hat{X}$와 사용자 프로파일의 유사 정도를 계산하여 유사성 정도에 의해 자료의 순위를 결정한다. 이때 사용자 프로파일을 분석결과 행렬의 열과 같은 형태로 재구성해야 한다. 따라서 자료는 아니지만 분석결과 행렬의 자료 형태로 변환되기 때문에 질의를 슈도 문서(Pseudo Document)라고 부른다. 사용자 프로파일(P_i)의 선호도가 (0.8 0.5 0.9 0.2 0.8 0.1 0 0.6 0.1 1.0 0 0 0 0 0.9 0.9 0.9 0.95)이라고 했을 때, 선호도 벡터로 구성된 행렬은 P(18×1)이 되고, 행렬 P의 전치행렬은 P'(1×18)이 된다. 사용자 선호도를 보면 사용자가 "웹기반 학습"에 관심이 많다는 것을 알 수 있다. (식 3)을 응용한 (식 6)[7]에 의해 슈도 문서 DP를 생성하면 (그림 8)과 같다.

$$DP = P'TS^{-1} \tag{6}$$

단, P : 사용자 프로파일의 i번째 관심분야의 선호도 벡터 P_i의 값으로 구성된 행렬

P' : P의 전치행렬,

T : 행렬 T_0의 축소화된 SVD의 결과 행렬

S^{-1} : 축소화된 SVD 결과로 생성된 행렬 S의 역행렬

$$\begin{bmatrix} 0.4173117 \end{bmatrix}$$

(그림 8) 슈도 문서 $DP(1\times1)$

슈도 문서가 생성되면 자료의 순위를 결정하기 위하여 분석결과 행렬 $\hat{X}$과 슈도 문서 DP의 유사성 정도를 계산한다. (식 4)를 응용한 (식 7)[7]에 의해 슈도 문서 DP와 용어-문서행렬의 유사성 정도를 비교하면 그 결과는 (그림 9)와 같다.

$$DR = ES^2E' \tag{7}$$

단, E : D와 DP가 결합된 확장된 행렬, $E'E=1$

4.3111347	4.5637349	3.7696495	2.7891191	3.9833595
4.5637349	4.8311356	3.9905227	2.9525406	4.2167545
3.7696495	3.9905227	3.2961757	2.4388014	3.4830433
2.7891191	2.9525406	2.4388014	1.8044403	2.5770626
3.9833595	4.2167545	3.4830433	2.5770626	3.6805049

(그림 9) DR : 사용자 선호도와 자료들 간의 유사성을 나타냄

(그림 9)는 대칭행렬이므로 진한 이탤릭체로 표현된 값을 정렬하여 자료의 순위를 결정할 수 있다. 따라서 유사성 정도에 따른 자료의 순위는 D2, D1, D5, D3, D4 순서로 결정된다.

4. 실험 및 평가

본 논문에서 제안하는 방법을 실험 평가하기 위하여 우리 대학교의 디지털 도서관의 검색엔진(Fulcrum 3.x)과 색인 데이터베이스(DBMS : Oracle)를 이용한다. 분류된 자료의 경우, 대학교 학과(전공)별 커리큘럼에 따라 전공분류표를 생성하는데, 컴퓨터학과전공 외 5개 전공을 대상으로 실험한다. 미분류된 자료의 경우, SAS 6.12과 IML(Interactive Matrix Language)을 사용한다.

4.1 분류된 자료의 필터링 성능 평가

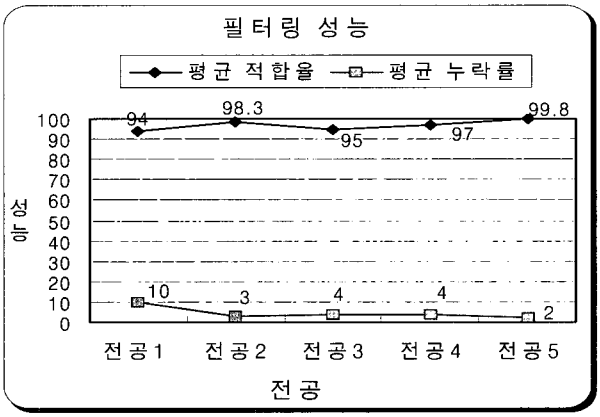
정보검색 분야에서 많이 사용하는 평가 척도에는 정확률(Precision Ratio)과 재현율(Recall Ratio)이 있다. 본 논문에서는 키워드 검색결과를 대상으로 필터링을 실시하므로 재현율 측정은 거의 의미가 없다고 볼 수 있다. 따라서 분류된 자료의 필터링 성능 평가는 정확률 공식을 변형한 (식 8)과 같은 적합률 공식을 제안하여 실시한다.

$$\text{적합률} = \frac{\text{필터링 결과 적합한 자료의 수}}{\text{키워드 검색 결과 적합한 자료의 수}} \tag{8}$$

본 논문에서 제안하는 필터링을 수행한 결과 적합한 결과를 적합하지 않다고 걸러 버리는 경우도 발생할 수 있으므로, 그 비율을 측정하기 위하여 일반적으로 사용하는 누락률 공식을 변형한 (식 9)와 같은 필터누락률를 제안하여 사용한다.

$$\text{필터누락률} = \frac{\text{필터링 결과 제시되지 않은 적합한 자료의 수}}{\text{키워드 검색 결과 중에서 적합한 자료의 수}} \tag{9}$$

정확률과 재현율과의 관계처럼 필터누락률은 적합률과 반비례를 이룬다. 이와 같은 측정 방법을 이용하여 두 가지 측면, 즉 적합률 측정과 필터누락률 측정을 통하여 본 논문에서 제안한 자료의 필터링 방법을 실험 평가한다. 첫 번째 실험은 5개 전공 분야의 10사람이 각각 10번에 걸쳐 질의를 입력하여 검색을 실시하게 하였다. 그 결과 분야별로 적합률의 평균을 구하여 도표로 나타내면 (그림 10)과 같다.



(그림 10) 분류된 자료의 필터링 성능

첫 번째 실험 평가 결과, (그림 10)과 같이 최고 평균 적합률이 99.8, 최하 평균 적합률이 94로서 적합률이 매우 높음을 알 수 있다. 그리고 누락되는 적합한 결과가 다소 발생하지만 사용자가 비적합한 자료들을 검토해야하는 노력에 비하여 미비한 수치이다. 향후 누락률을 최소화할 수 있는 지속적인 연구가 필요하다.

4.2 미분류된 자료의 순위결정 평가

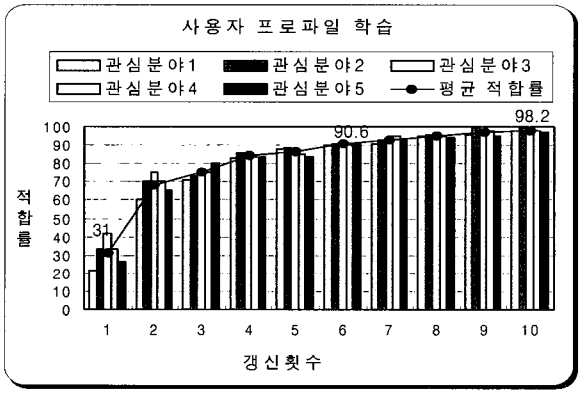
사용자의 선호도가 적절히 표현되기까지의 사용자 프로파일의 갱신횟수 결정과 상위 순위로 제시된 자료의 적합성 정도를 실험으로 알아봄으로써 순위결정의 성능을 평가한다.

평가를 위하여 정확률에 기반을 둔 순위 적합률(Relevance Ratio)을 제안하여 평가에 사용한다. 순위 적합률 공식은 (식 10)과 같다.

$$\text{순위 적합률} = \frac{\sum_{i=1}^{n} R_{score}}{\sum_{i=1}^{n} R_{\max}} \times 100 \tag{10}$$

단, R_{score} : 사용자가 평가한 논문의 적합성 정도로서 표현 범위는 0~6 값이다.
0 : 비적합, 3 : 보통, 6 : 적합
$R_{\max}$: 최고 적합한 정도로서 값은 6
n : 순위가 결정된 논문의 상위 5% 이내 순위를 갖는 자료의 갯수

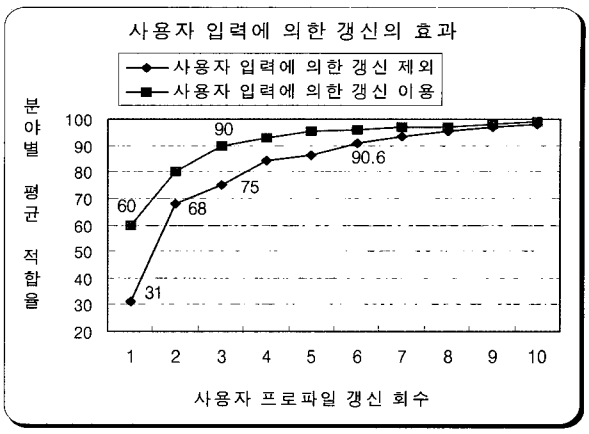
첫 번째 실험 평가는 사용자 프로파일이 사용자의 선호도를 잘 표현하기까지의 갱신 횟수를 알아보기 위해 실시한다. 사용자의 선호도 반영 정도를 알아보기 위하여 "사용자 적합성 피드백에 의한 갱신"방법으로 갱신된 사용자 프로파일을 자료의 순위결정에 사용한다. 사용자 프로파일의 갱신을 거듭하면서 순위가 결정된 결과 중 상위 5%, 20%, 30%내의 자료를 사용자에게 0(비적합)~6(적합) 값으로 평가하게 함으로써, 사용자 프로파일이 몇 번의 갱신 후에 사용자의 선호도를 제대로 반영하는지 알아본다. 5개 분야의 각 10명이 10번 이상의 질의를 입력하여 사용자 프로파일을 학습시킨 후 사용자가 평가한 값을 도식화하면 (그림 11)과 같다.



(그림 11) 사용자 프로파일의 갱신 성능

초기 갱신 기간에는(갱신 횟수 약 4회 정도) 사용자의 선호도가 급속히 증가하고 그 이후에는 완만히 증가함을 알 수 있다. 이와 같이 초기 학습 속도가 빠르고 일정 수준에서는 완만하게 학습되는 사실은 일반적인 "사용자 적합성 피드백"의 특성이다. 이와 같은 실험 결과를 통해 사용자 프로파일의 적절한 갱신횟수를 결정할 수 있다. 즉, 상위 5% 내의 자료들의 적합률이 90% 이상이 되는 시점은 갱신횟수가 6이 되는 시점이다. 그리고 6회의 갱신이 아니더라도 4회의 갱신을 수행하면 80% 이상의 적합률을 얻을 수 있다. 실험 집단이 키워드 검색결과 얻은 자료들이므로 비교적 짧은 갱신 횟수로 사용자 프로파일을 충분히 학습시킬 수 있음을 알 수 있다. 사용자 프로파일 갱신 방법 중에서 "사용자 입력에 의한 갱신"방법을 사용하면 보다 적은 갱신 횟수 동안 사용자의 선호도를 반영할 수 있다. "사용자 입력에 의한 갱신" 방법을 같이 사용한 경우와 그렇지 않은 경우를 비교하면 (그림 12)와 같다.

실험 결과 (그림 12)와 같이 "사용자 입력에 의한 갱신"



(그림 12) 사용자 입력에 의한 사용자 프로파일 갱신 효과

방법을 사용자 프로파일의 학습 초기에 이용하면 단 한 번의 갱신으로 인해 적합률이 61%에 도달하고, 2회의 갱신 수행 후에는 적합률이 80% 이상이 되었다. 그리고 적합률 90% 이상으로 사용자 프로파일이 학습되기 위해서는 약 3회 정도의 갱신이 필요함을 알 수 있다. 이 경우 적절한 사용자 프로파일의 갱신 횟수는 3이 된다.

두 번째 실험평가는 자료의 순위결정의 성능을 측정하기 위하여 실시한다. 5개의 각 분야별로 학습시킨 사용자 프로파일을 이용하여 자료의 순위를 결정한 후, 20명의 사용자들이 10회 이상의 검색을 수행하여 평가한 순위별 평균 적합률은 <표 3>과 같다.

〈표 3〉 자료의 순위별 평균 적합률

자료 순위 / 관심분야	1~5	6~10	11~20	21~30
관심분야 1	100	100	98	100
관심분야 2	99	98	95	95
관심분야 3	100	99	99	99
관심분야 4	100	97	98	99
관심분야 5	98	98	97	95
적합률	99.4	98.4	97.4	97.6

이때 한 명의 사용자는 복수개의 전공을 가질 수 있게 하였다. 평가에 적합률 공식을 사용하였는데, 이 실험의 경우에 식 4에서 n은 순위별 제시되는 자료의 개수가 된다. 예를 들어 자료의 순위 1~5에서 n은 5이고, 순위 11위~20에서의 n은 10이다. 실험 결과, 순위 1~5의 결과에 대해서는 99.4% 이상의 적합률을, 순위 30 미만에서 97.6% 이상의 적합률을 보임으로써 본 논문에서 제안한 순위결정 기법이 사용자의 요구를 만족시키기에 충분히 우수함을 알 수 있다. 그리고 대부분의 기존 도서관에서 적용하고 있는 키워드 정합 매칭 방법에 의한 검색과 본 논문에서 제안하는 SURLM과의 비교 결과는 <표 4>와 같다.

〈표 4〉 키워드 검색과 SULRM과의 비교

평균적합률 / 관심분야	키워드 검색	SULRM
관심분야 1	71	93
관심분야 2	67	96
관심분야 3	55	90
관심분야 4	71	96
관심분야 5	76	97

5. 결 론

본 논문에서는 사용자가 불필요한 검색결과까지 살펴보는 불편을 제거하기 위하여 전공분류표, 사용자 프로파일, 그리고 LSI를 이용한 검색 모델 SULRM(Retrieval Model using Subject Classification Table, User Profile & LSI)을 제안하였다. SULRM은 도서관 시스템의 특성을 고려하여 단행본 자료와 같이 분류된 자료의 경우와 학위논문, 기사 등과 같이 미분류된 자료의 경우로 나누어, 필터링과 자료의 순위결정을 수행하였다. SULRM은 키워드 검색 결과로 얻은 자료들을 대상으로 하여 사용자가 참조할 필요 없는 자료는 필터링을 통해 걸러주고, 순위결정을 통해 비적합 자료의 순위를 낮게 함으로써 사용자에게 편의성을 제공한다.

실험 데이터 집단은 우리 대학교 학과의 커리큘럼과 디지털 도서관에서 사용 중인 검색엔진을 통해 검색된 결과를 이용하였다. 분류된 자료의 필터링 성능평가 결과, 최고 평균 적합률이 99.8%, 최하 평균 적합률이 94%로서 필터링을 수행하여 제공된 자료의 적합률이 매우 높음을 알 수 있다. 그러나 필터링 방법이 분류번호에 기반하고 있고, 학문 분야가 서로 공유하는 부분이 있으므로 적합한 자료의 누락이 발생하였다. 두 번째 성능 평가는 사용자 프로파일이 사용자의 선호도를 충분히 학습하기 위한 갱신 횟수와 학습된 사용자 프로파일을 이용하여 자료의 순위를 결정했을 경우의 적합성 정도를 평가하였다. 실험 결과, 사용자 프로파일은 "사용자 적합성 피드백에 의한 갱신" 방법만으로도 갱신 횟수 6번째부터 적합률 90%이상의 성능을 얻을 수 있었다. 사용자가 초기에 직접 관심분야 용어를 입력하여 사용자 프로파일을 학습시키면 보다 적은 갱신 횟수 3회만으로도 순위 적합률 90%이상을 얻을 수 있다. 그리고 자료의 순위결정 결과, 상위 5%의 자료의 적합성이 평균 99.4%로서 매우 우수함을 알 수 있다.

본 논문에서 제안하는 검색 모델을 단순한 키워드 검색엔진을 사용하는 도서관에 적용한다면 사용자에게 검색 편의성을 제공함으로써 보다 향상된 상질의 서비스를 제공할 수 있을 것이다. 향후 전공분류표의 생성과 수정을 자동화할 수 있는 방법과 필터링 결과 발생하는 누락률을 최소화할 수 있는 방법에 관한 연구를 계속할 것이다.

참 고 문 헌

[1] Bracha Shapira et., "Information Filtering: A New Two-Phase Model Using Stereotypic User Profiling", Journal of Intelligent Information Systems, Vol.8, pp.155-165, 1997.

[2] Douglas W. OARD, "The State of the Art in Text filtering," User modeling and User-adapted Interaction, Vol.7, pp. 141-178, 1997.

[3] Foltz, P. W, "Using Latent Semantic Indexing for Information Filtering", Proceedings of the Conference on Office Information Systems, Cambridge, MA, pp.40-47, 1990.

[4] Sheth B., Maes P, "Evolving Agents for Personalized Information Filtering," In Proceedings of the Ninth IEEE Conference on Artificial Intelligence Applications, 1993.

[5] Czeslaw Danilowicz, Jaroslaw Bali ṅski, "Document Ranking based upon Markov Chains", Information Processing and Management, Vol.37(2001) : pp.623-637.

[6] Michael Persin, "Document Filtering for Fast Ranking," ACM-SIGIR, pp.339-348, 1994.

[7] 우선미, 사용자 프로파일과 잠재적 구조 분석을 이용한 검색된 문서의 순위결정 기법, 박사학위논문, 전북대학교 대학원, 2001.

[8]] Passani, M. and Billsus, D., "Learning and Revising User Profiles: The Identification of Interesting Web Sites", Machine Learning, Vol.27, pp.313-331, 1997.

[9] Dwi H. Widyantoro, Thomas R. Ioerger, John Yen, "An Adaptive Algorithm for Learning Changes in User Interests", 8th International Conference on Information and Knowledge Management(CIKM'99), November 2-6, Kansas city, 1999.

[10] Crestani, F. et., "Is This Document Relevant?...Probably: a Survey of Probabilistic Models in Information Retrieval", ACM Computing Surveys, Vol.30, No.4, pp.528-552, 1998.

[11] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman, "Indexing by Latent Semantic Analysis," Journal of the American Society for Info rmation Science, Vol.41, No.6, pp.391-407, 1990.

[12] Dwi H. Widyantoro, Thomas R. Ioerger, John Yen, "Learning User Interest Dynamics with a Three-Descriptor Representation", Journal of the American Society for Information Science and Technology, Vol.53, No.3, pp.212-225, 2001.

[13] Geoffrey I. Webb, Michael J. Pazzani, Daniel Billsus, "Machine Learning for User Modeling," User Modeling and User-Adapted Interaction, Vol.11, pp.19-29, 2001.

[14] Ingrid Zukerman, David W. Albrecht, "Predictive Statistical Models for User Modeling", User Modeling and User-Adapted Interaction, Vol.11, pp.5-18, 2001.

우 선 미

e-mail : smwoo@chonbuk.ac.kr

1991년 서남대학교 전자계산학과(이학사)

1995년 전북대학교 전산통계학과 (이학석사)

2001년 전북대학교 전산통계학과 (이학박사)

2001년~2003년 전북대학교 중앙도서관 전산실 조교

2004년~현재 전북대학교 전북지역전자정보사업단 기금교수

관심분야 : 사용자 위주의 정보검색, 문서순위결정, 정보 필터링, XML 응용, 디지털 도서관 등