용어 연관성 분석을 이용한 사용자 위주의 문서순위결정 기법

(User-Centered Document Ranking Technique using Term Association Analysis)

우 선 미 † 유 춘 식 † 김 용 성 ††
(Seon-Mi Woo) (Chun-Sik Yoo) (Yong-Sung Kim)

요 약 정보의 가치와 사용자의 정보획득 요구가 증대됨에 따라 특정 개인 위주의 서비스를 제공하는 정보검색 시스템의 필요성이 증대되고 있다. 그러나 현재의 정보검색 시스템들은 사용자의 선호도를 반영하고 편의성을 제공하는 면에서 매우 미흡한 점들이 많다. 따라서 본 논문에서는 적합성 정도에 따라 최적의 문서를 제공하기 위하여 사용자 위주의 문서순위결정 기법을 제안한다. 특정 개인의 선호도(preference)를 반영하기 위하여 사용자 프로파일(User Profile)을 구성 및 갱신하고, LSA(Latent Semantic Analysis)를 적용하여 적합율에 따라 문서의 순위를 결정한다.

Abstract As the value of information and user's desire for information acquisition increases, the need for information retrieval systems that provide adaptive services for specific person is emerging. However, information retrieval systems of today are rather weak in the respect of reflecting the user's preference and offering convenience. Thus this paper proposes a method for document ranking using User Profile which are reflecting preference of the specific user and term association analysis(Latent Semantic Analysis) in order to provide more relevant document in rank.

1. 서 론

정보가 기하급수적으로 증가하고 정보의 가치가 증대됨에 따라 특정 개인의 관심과 선호도(preference)를 파악하여 보다 만족스러운 결과를 제공해주는 사용자 위주의 정보검색 시스템의 필요성이 증대되고 있다. 이러한 사용자의 요구를 만족시키기 위하여 문서순위결정(Document Ranking), 정보 필터링(Information Filtering), 기계학습(Machine Learning) 이론을 이용하여 적응성을 부여하는 방법에 관한 연구가 활발히 진행되고 있다. 그러나 이러한 방법들도 사용자 위주의 요구를 만족시키는 측면에서는 미흡한 점들이 많다.

† 학생회원 : 전북대학교 전산통계학과
smwoo@cs.chonbuk.ac.kr
csyoo@cs.chonbuk.ac kr
†† 종신회원 : 전북대학교 컴퓨터과학과 교수
yskim@moak.chonbuk.ac.kr
논문접수 : 1998년 8월 31일
심사완료 : 2000년 11월 9일

이러한 문제점들을 해결하기 위하여 본 논문에서는 사용자 위주의 문서순위결정 기법을 제안한다. 사용자 프로파일(User Profile)을 구축하여 사용자의 선호도를 반영하고, 통계적 분석 방법인 LSA(Latent Semantic Analysis)를 적용하여 문서순위결정을 수행함으로써 사용자의 요구에 적합한 문서를 적합성의 정도에 따라 제공하도록 한다.

2. 관련 연구

문서순위결정[1,2,3]은 사용자의 질의어와 검색된 문서들이 얼마나 유사한가에 따라 문서의 순위를 결정하고, 이 순서에 따라 사용자가 가장 적합한 문서를 참조하도록 하는 방법이다. 정보 필터링[1,4]은 사용자의 기호를 저장해 놓고, 이를 참조하여 필요 없는 정보를 여과시켜 줌으로써 검색된 결과의 문서의 수를 줄여 주는 방법이다. 기계학습을 활용한 적응형 시스템[5,6]은 사용자 행위 관찰을 통한 학습, 사용자 피드백(feedback)을 통한 학습, 훈련을 통한 학습, 충고(advice)를 통한 학습 등을 이용하여 사용자 위주의 정보검색에 필요한

지식을 습득해 나간다. 이러한 기법들을 제공하는 여러 검색 시스템과 에이전트(agent)들도 개발되고 있다[7,8, 9]. LSA(Latent Semantic Analysis)를 응용한 LSI (Latent Semantic Indexing)를 필터링에 적용한 연구들이 있는데[10,11], 이 방법은 문서들간의 연관성이나 시멘틱을 발견하여 정보검색에 이용하는 기법이다.

[1]은 6가지의 문서순위결정 알고리즘(F4point-5, F4modified, EMIM, Porter's, $w(p-q)$, ZOOM term frequency ranking)을 소개하고 있다. [12]에서는 WWW 환경에서 문서의 순위를 결정하기 위하여 키워드 기반의 문서순위결정 알고리즘(Boolean Spread Activation, Most-cited, vector space model, Vector Spread Activation)을 기술하고 있다. [13]은 LSI의 SVD 분석 기법을 질의확장에 사용하여 사용자가 의도하는 관련된 문서를 더 많이 검색할 수 있게 하고 있다.

3. 사용자 프로파일과 문서순위결정

3.1 사용자 프로파일의 구조

그림 1은 본 논문에서 제안하는 사용자 프로파일의 구조이고, 그 예는 그림 2와 같다.

용어열 T는 해당 관심분야의 색인어(문서의 제목에서 추출)로 구성하고, 선호도 벡터(Preference Vector) $\vec{P}$는 용어열 T에 대응하는 사용자의 선호도를 나타낸다.

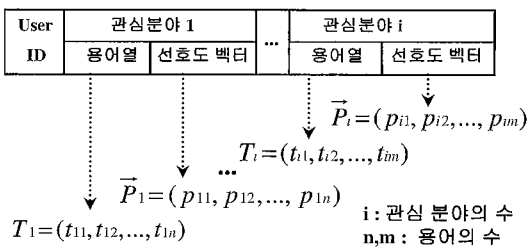


그림 1 사용자 프로파일의 구조

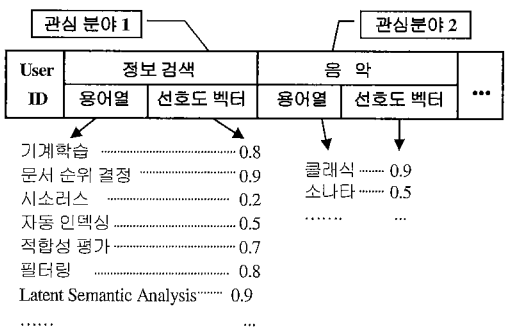


그림 2 사용자 프로파일의 예

그림 2의 선호도는 객관적인 용어 관계를 기술해 놓은 시소러스(thesaurus)에 비해 매우 주관적임을 알 수 있다. '정보검색' 분야의 경우, 'Latent Semantic Analysis'라는 용어가 통계학 분야의 용어임에도 불구하고 사용자가 관심을 갖고 있기 때문에 0.9라는 높은 가중치를 갖는다. 사용자의 선호도는 0으로 시작하고, 갱신을 통해 0과 1 사이의 값으로 표현한다.

3.2 사용자 프로파일의 갱신

(1) 사용자 접근에 의한 갱신

사용자가 특정 용어에 중요성을 부여하고자 할 때, 직접 선호도 벡터의 가중치를 변경하여 선호도를 반영하는 방법이다. 전문가에게 친숙한 방법으로서 사용자 프로파일의 학습에 요구되는 비용을 절약할 수 있다.

(2) 확장된 질의의 확인을 통한 갱신

확장된 질의어에 대한 사용자의 확인을 통해 최종 선택된 질의의 가중치를 사용자 프로파일의 선호도에 반영하는 방법이다. 질의확장은 사용자 프로파일을 사용하는데, 초기 학습(갱신회수 5회 정도)에는 시소러스를 같이 사용한다.

① 사용자 프로파일을 참조한 질의확장

사용자 프로파일의 선호도가 0.9이상인 용어를 추가하여 질의를 확장한다. 확장된 질의의 가중치는 선호도 벡터의 선호도와 같다.

② 시소러스를 참조한 질의확장

확장된 질의가 사용자 프로파일의 용어열에 있는 경우에는 선호도 벡터의 선호도를 질의의 가중치로 한다. 그리고 확장된 질의가 사용자 프로파일의 용어열에 없는 경우에는 가중치를 갖지 않는 시소러스를 사용하는데, 이때 질의의 가중치는 가중치의 최소값과 최대값의 중간값인 0.5로 한다.

본 논문에서 제안하는 확장된 질의의 확인을 통한 갱신 공식은 다음과 같다.

i) q_{ij} 가 최종 질의로 선택된 경우

$$P_i = \left| p_{ij} + \sum_{k=1}^{n} q_{ikl}/n \right|_{j=1,\ldots,n}$$

ii) q_{ij} 가 최종 질의로 선택되지 않은 경우

$$P_i = \left| p_{ij} - \sum_{k=1}^{n} q_{ikl}/n \right|_{j=1,\ldots,n}$$

단, P_i : 사용자 프로파일의 i번째 관심분야의 선호도 벡터 $=(p_{i1}, p_{i2}, \ldots, p_{in})$

n : 용어열에 있는 용어의 수

p_{ij} : 사용자 프로파일의 i번째 관심분야 용어열의 j번째 용어에 대응되는 선호도

Q_i : 사용자 프로파일의 i번째 관심분야에서의 확장된 질의 벡터 $=(q_{i1}, q_{i2}, \ldots, q_{in})$

q_{ij} : Q_i를 구성하는 용어열의 j번째 용어에 대응되는 가중치

위 공식에서 정적인 갱신을 피하기 위하여 질의 가중치들의 평균($\sum_{k=1}^{n} q_{ik}/n$)을 이용한다.

(3) 사용자 적합성 피드백에 의한 갱신

문서순위결정 결과로 제시된 상위 5%의 문서들에 대한 사용자의 적합성 평가에 따라 사용자 프로파일을 갱신하는 방법이다. 사용자는 0~3(0:비적합, 1:보통, 2:적합, 3:매우 적합)의 값으로 적합성 정도를 평가한다. 평가된 문서 중에서 적합성 정도가 2 이상인 문서들만을 대상으로 색인을 추출하여 용어열과 가중치 벡터를 구성한다. 갱신 공식은 다음과 같다.

i) $0.5 \leq w_{ik}$

$$P_i = \left| 2p_{ij} + \sum_{k=1}^{n} w_{ik}/n \right|_{j=1,\ldots,n}$$

ii) $0.2 \leq w_{ik} < 0.5$

$$P_i = \left| p_{ij} + \sum_{k=1}^{n} w_{ik}/n \right|_{j=1,\ldots,n}$$

iii) $w_{ik} < 0.2$

$$P_i = |p_{ij}|_{j=1,\ldots,n}$$

단, D_i : 사용자가 적합하다고 평가한 i번째 문서 $=(w_{i1}, w_{i2}, \ldots, w_{in})$

w_{ij} : D_i을 구성하는 용어 j번째의 가중치

위 공식에서 w_{ij}는 TFxIDF 방법을 이용하여 구한다. 사용자 프로파일의 선호도 값에 대한 계산을 보다 편리하게 하기 위하여 0과 1 사이의 값으로 정규화한다.

3.3 사용자 위주의 문서순위결정

사용자의 요구에 의해 검색된 문서들은 사용자의 관심을 반영하므로 한가지 성향을 가지게 된다. 그러므로 검색된 문서들을 하나의 성향인 사용자의 요구를 기준으로 분석하면 분석 기준과 문서들간의 관계를 파악할 수 있다. 이렇게 분석된 결과를 사용자 프로파일과 비교하여 적합한 순서대로 문서의 순위를 정할 수 있다. 본 논문에서 제안하는 사용자 위주의 문서순위결정 알고리즘은 [알고리즘 1]과 같다.

[알고리즘 1] 사용자 위주의 문서순위결정

[입력] ① 사용자 프로파일 ② 검색결과 문서

[출력] ① 순위가 결정된 검색결과 문서
② 갱신된 사용자 프로파일

user_centered_document_ranking

begin

1. 관심분야의 프로파일 벡터를 선택한다.
2. 질의확장
   IF (갱신 회수 > 5)
   then 사용자 프로파일을 이용하여 질의를 확장한다.
   else 사용자 프로파일과 시소러스를 이용하여 질의를 확장한다.
3. '확장된 질의의 확인을 통한 갱신'을 수행하여 사용자 프로파일을 갱신한다.
4. 검색엔진을 이용하여 문서순위결정을 수행할 데이타(논문)를 얻는다.
5. 검색결과로 얻은 문서들을 분석한다.
6. 분석결과와 사용자 프로파일을 비교하여 문서의 순위를 결정한 후, 사용자가 원하는 범위 내에서 결과를 제시한다.
7. '사용자 적합성 피드백에 의한 갱신'을 수행하여 사용자 프로파일을 갱신한다.

end.

3.3.1 용어 연관성 분석

본 논문에서는 검색결과 문서들을 분석하기 위하여, 다변량 분석의 일종인 LSA(Latent Semantic Analysis)을 적용한다. 표 1은 '필터링 AND 문서순위결정'에 의해 검색된 검색결과의 일부분이다. 본 논문에서 제안하는 방법론에 대한 보다 쉬운 설명을 위하여 표 1를 예를 들어 설명한다.

표 1에 의해 구성된 용어-문서 행렬은 표 2와 같다. 표 2에서 각 행은 컬렉션 내의 색인어(문서의 제목에서 추출)를 나타내며, 각 열은 검색결과 문서를 나타낸다. 그리고 값은 제목 내의 용어 빈도(term frequency)를 나타낸다.

표 1 용어연관성 분석을 수행할 문서들의 예

논문	제목
D1	적합성 피드백을 이용한 필터링 에이전트
D2	사용자 적응형 에이전트를 이용한 정보검색과 필터링
D3	빠른 순위결정을 위한 문서 필터링
D4	정보 필터링
D5	향상된 인터넷 정보 검색을 위한 사용자 모델과 필터링 에이전트
D6	정보 필터링 에이전트
D7	정보 필터링과 검색

표 2 표1에 대한 용어-문서 행렬

용어	문서						
	D1	D2	D3	D4	D5	D6	D7
검색	0	1	0	0	1	0	1
모델	0	0	0	0	1	0	0
문서	0	0	1	0	0	0	0
사용자	0	1	0	0	1	0	0
순위결정	0	0	1	0	0	0	0
에이전트	1	1	0	0	1	1	0
인터넷	0	0	0	0	1	0	0
적응형	0	1	0	0	0	0	0
적합성	1	0	0	0	0	0	0
정보	0	1	0	1	1	1	0
피드백	1	0	0	0	0	0	0
필터링	1	1	1	1	1	1	1

용어-문서 행렬 X가 구성되면, 성분별(본 논문에서는 용어와 문서) 특성을 잘 나타내는 행렬로 분해한다. 분해 방법은 다음과 같은 SVD 방법이다[11].

$$X = T_0 S_0 D_0'$$

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)

표 2를 SVD하면 그림 3과 같이 $T_0(12\times7)$, 대각행렬 $S_0(7\times7)$, $D_0(7\times7)$를 얻을 수 있다.
값은 소수점 넷째 자리에서 반올림한 값이다.

이때 S_0는 정방행렬로서 값이 클수록 문서들이 많이 모이는 대표 인자(factor)가 된다.

$$\begin{bmatrix}
0.335 & -0.322 & 0.165 & 0.389 & -0.458 & -0.378 & -0.188 \\
0.144 & -0.224 & 0.005 & 0.355 & 0.429 & 0 & 0.104 \\
0.041 & 0.292 & 0.430 & 0.164 & 0.056 & 0.378 & -0.030 \\
0.278 & -0.362 & 0.07 & 0.189 & -0.118 & 0.378 & 0.348 \\
0.041 & 0.292 & 0.430 & 0.164 & 0.056 & 0.378 & -0.030 \\
0.439 & 0.058 & -0.464 & -0.014 & 0.062 & 0.378 & -0.664 \\
0.144 & -0.224 & 0.005 & 0.355 & 0.429 & 0 & 0.104 \\
0.134 & -0.138 & 0.002 & -0.166 & -0.546 & 0.378 & 0.243 \\
0.072 & 0.329 & -0.405 & 0.208 & -0.053 & 0 & 0.350 \\
0.430 & -0.196 & 0.056 & -0.625 & 0.306 & 0 & 0.270 \\
0.072 & 0.329 & -0.405 & 0.208 & -0.054 & 0 & 0.350 \\
0.600 & 0.465 & 0239 & -0.052 & -0.032 & -0.378 & 0.053
\end{bmatrix}$$

그림 3-a 직교행렬 T_0

$$\begin{bmatrix}
4.060 & & & & & & \\
& 1.895 & & & & & \\
& & 1.599 & & & & \\
& & & 1.297 & & & \\
& & & & 1.200 & & \\
& & & & & 1.000 & \\
& & & & & & 0.501
\end{bmatrix}$$

그림 3-b 대각 정방 행렬 S_0

$$\begin{bmatrix}
0.291 & 0.623 & -0.647 & 0.270 & -0.064 & 0 & 0.175 \\
0.546 & -0.261 & 0.003 & -0.215 & -0.656 & 0.378 & 0.122 \\
0.168 & 0.554 & 0.554 & 0.213 & 0.068 & 0.378 & -0.015 \\
0.254 & 0.142 & 0.142 & -0.522 & 0.228 & -0.378 & 0.643 \\
0.584 & -0.425 & -0.325 & 0.460 & 0.541 & 0 & 0.052 \\
0.362 & 0.173 & 0.173 & -0.533 & 0.280 & 0 & -0.682 \\
0.230 & 0.075 & 0.075 & 0.260 & -0.408 & -0.756 & -0.268
\end{bmatrix}$$

그림 3-c 직교행렬 D_0

$$\begin{bmatrix}
0.335 \\ 0.144 \\ 0.041 \\ 0.278 \\ 0.041 \\ 0.439 \\ 0.144 \\ 0.134 \\ 0.072 \\ 0.430 \\ 0.072
\end{bmatrix}$$

그림 4-a T_0 를 축소화한 행렬 T

$$\begin{bmatrix} 4.060 \end{bmatrix}$$

그림 4-b S_0 를 축소화한 행렬 S

D′=

$$\begin{bmatrix}
0.291 & 0.546 & 0.168 & 0.254 & 0.584 & 0.362 & 0.230
\end{bmatrix}$$

그림 4-c D_0 를 축소화한 행렬 D

SVD를 수행한 결과 중에서 용어나 문서들의 성향을 가장 잘 반영하는 대표 행렬을 이용하여 최종 분석결과 행렬 $\hat{X}(t\times d)$를 생성한다. 축소화된 SVD 공식은 다음과 같다[11].

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

단, 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 ($k \leq m$)

앞 식에서 생성된 행렬 $\hat{X}$는 인수(계수:rank)가 k로 축소되었을 뿐이지, 용어-문서 행렬 X와 근사적으로 일치하게 된다($X \approx \hat{X} = TSD'$)

그림 4는 주요 인자를 기준으로 행렬 T_0, S_0, D_0를 축소화시킨 결과를 나타내고 있다.

본 논문에서 하나의 인자를 선택하여 축소화된 SVD를 수행하는 이유는 첫째, 분석 데이타가 사용자의 한 가지 관심분야에 대한 검색결과 문서들이고, 둘째, 적합성이라는 한가지 기준으로 문서의 순위를 결정하기 위함이다.

분석결과 행렬 $\hat{X}$는 그림 5와 같다.

그림 5는 각 7개의 문서(열)에 포함된 12개의 용어(행)이 해당 문서를 대표할 수 있는 정도를 나타낸다.

0.396	0.742	0.229	0.345	0.794	0.492	0.313
0.170	0.319	0.098	0.148	0.341	0.211	0.134
0.490	0.092	0.028	0.043	0.098	0.061	0.039
0.329	0.617	0.190	0.186	0.660	0.409	0.260
0.049	0.092	0.028	0.043	0.098	0.061	0.039
0.520	0.973	0.300	0.452	1.040	0.645	0.410
0.170	0.319	0.981	0.148	0.341	0.211	0.134
0.159	0.298	0.092	0.138	0.319	0.198	0.126
0.085	0.159	0.049	0.074	0.170	0.105	0.067
0.508	0.953	0.293	0.443	1.018	0.621	0.402
0.085	0.159	0.049	0.074	0.170	0.105	0.067
0.709	1.329	0.409	0.617	1.421	0.880	0.561

그림 5 용어연관성 분석결과 $\hat{X}$

3.3.2 문서순위결정

사용자 프로파일과 분석 결과 행렬 $\hat{X}$를 비교하여 문서의 순위를 결정한다. 분석결과 문서의 크기와 사용자 프로파일의 선호도 벡터의 크기가 일치하지 않으므로 본 논문에서는 사용자 프로파일을 이용하여 의사문서(pseudo document)를 생성한다. 의사문서(DP)는 다음 식에 의해 생성된다.

$$DP = P_i T_0 S_0^{-1}$$

단, P_i : 사용자 프로파일의 i번째 관심분야의 선호도 벡터

T_0 : SVD 결과로 생성된 직교행렬

S_0^{-1} : SVD 결과로 생성된 대각 정방 행렬 S_0의 역행렬

의사문서 생성에 대한 설명을 위하여 표 2에 나타난 용어와 동일한 용어로 구성된 사용자 프로파일이 그림 6과 같은 상태로 선호도가 학습되어 있다고 가정한다. 의사문서 생성 공식에 의해 생성된 의사문서는 그림 7과 같다. P_i(1×12), T(12×7), S^{-1}(7×7)을 연산하여 DP(1×7)를 구할 수 있다.

의사문서가 구해지면, 문서의 순위를 결정하기 위하여 분석결과($\hat{X} = TSD'$)와 의사문서(DP)와의 유사성 정도를 계산한다.

축소화된 SVD를 수행하기 전 $\hat{X}$내에서 두 문서들을 비교하는 공식은 다음과 같다[11].

$$\hat{X}'\hat{X} = DS^2D'$$

본 논문에서는 문서순위결정을 위하여 차원을 "1"로 축소시켜 분석을 하였으므로, 분석결과 문서들 $\hat{X}$과 의사문서와의 유사성 정도를 계산하기 위하여 앞의 공식을 응용한 다음 식을 제안하여 비교연산을 수행한다.

$$DR = ES^2E'$$

단, E : 축소화된 E_0, E' : E의 전치행렬

E_0 : D와 DP가 결합된 확장된 행렬

P_i	용어	검색	모델	문서	사용자	순위결정	에이전트	인터넷	적응형	적합성	정보	피드백	필터링
	선호도	0.80	0.45	0.75	0.90	0.60	0.65	0.80	0.80	0.90	0.90	0.95	1.00

그림 6 사용자 프로파일의 예

DP	분석값	0.5583	0.1877	-0.0221	0.5975	-0.0944	0.7181	1.9088

그림 7 그림 6을 공식 $DP = P_i T_0 S_0^{-1}$ 에 의해 작성한 의사문서의 예

표 3 의사문서와 검색결과와의 관계(DR)

	DP	D1	D2	D3	D4	D5	D6	D7
DP	5.136	2.679	5.023	1.547	2.333	5.370	3.328	2.118
D1	***2.679***	1.398	2.620	0.807	1.217	2.801	1.736	1.105
D2	***5.023***	2.620	4.912	1.513	2.282	5.251	3.256	2.072
D3	***1.547***	0.807	1.513	0.466	0.703	1.617	1.002	0.640
D4	***2.333***	1.217	2.282	0.703	1.060	2.439	1.512	0.962
D5	***5.370***	2.801	5.251	1.617	2.439	5.614	3.480	2.215
D6	***3.328***	1.736	3.255	1.002	1.512	3.480	2.160	1.373
D7	***2.118***	1.105	2.072	0.638	0.962	2.215	1.373	0.874

앞의 공식에 의해 E(8×1), S^2(1×1), E′(1×8)를 연산하면, 표 3과 같이 의사문서 DP와 검색결과 문서들(D1~D7)의 유사성 정도를 나타내는 대칭행렬 DR(8×8)을 구할 수 있다.

본 논문에서는 의사문서 DP와 검색 결과 문서들(D1~D7)간의 비교(표 3의 진한 이텔릭체로 표시된 값)를 사용하는데, 값은 의사문서와 검색결과 문서간의 유사성 정도의 크기를 나타낸다. 그러므로 사용자의 선호도에 따른 문서 순위는 D5, D2, D6, D1, D4, D7, D3 순으로 결정된다.

순위가 결정된 문서를 사용자가 원하는 범위 내에서 제공하기 위하여 사용자로부터 임계값(threshold)을 입력받는다. 임계값은 결과문서들의 가중치 분포에서 사용자에게 제공될 상위 문서들의 백분율을 나타낸다. 다음과 같은 공식에 의하여 순위가 결정된 문서들 중 W 값보다 큰 유사성 정도를 갖는 문서를 최종 제시한다.

$$W = DR_{\min} + (DR_{\max} - DR_{\min}) * T_{\alpha}$$

단, W: 제시할 문서의 최하 유사성 정도 값

$DR_{\min}$: 의사문서와 결과 문서간의 최하 유사성 정도

$DR_{\max}$: 의사문서와 결과 문서간의 최고 유사성 정도

T_{α} : 사용자가 입력한 임계값

4. 실험 및 평가

4.1 실험 환경

① 실험 영역(domain) : 컴퓨터 과학 분야의 각 10개의 세부 분야의 연구자가 초기에 분야의 대표 키워드로 인터넷에서 얻은 논문
(각 분야별 약 50편씩 총 498편의 논문과 508개의 용어)

② 키워드 추출

- 범위 : 관심분야 논문의 제목
- 방법 : 본 연구팀에서 개발한 자동색인 방법[14]을 사용

③ 사용자 적합성 피드백

- 범위 : 논문의 전문(full text)
- 방법 : TFxIDF[15]

④ 분석 도구 : SAS,
IML(Interactive Matrix Language)

4.2 실험 평가

본 논문에서 제안한 사용자 위주의 문서순위결정 기법의 목적에 따라 두 가지 측면에서 실험 평가를 실시한다. 첫 번째 실험은 본 논문에서 제안한 사용자 프로파일 갱신 방법이 얼마나 효과적인지를 알아보는 실험이다. 두 번째 실험은 제안한 문서순위결정 기법을 이용하여 결정된 문서의 순위별 적합율을 구하여 본 논문의 기법이 사용자의 요구를 얼마나 만족시키는지를 검증한다.

정보검색 분야에서 가장 널리 사용되는 성능 평가의 척도에는 재현율(recall ratio)과 정확율(precision ratio)이 있다. 본 논문에서는 다음과 같이 일반적인 정확율 공식을 변형한 적합율(relevance ratio)공식을 제안하여 사용한다.

$$\text{적합율} = \frac{\sum_{i=1}^{n} R_{score}}{\sum_{i=1}^{n} R_{\max}} \times 100$$

단, R_{score} : 사용자가 평가한 논문의 적합성 정도로서 표현 범위는 0~3 값이고, 값에 따른 의미는 다음과 같다.
3:매우 적합, 2:적합, 1:보통, 0:비적합

$R_{\max}$: 최고 적합 정도로서 값은 3.

n : 순위가 결정된 논문의 상위 10% 내의 순위를 갖는 논문의 개수

4.2.1 사용자 선호도의 반영

사용자 프로파일이 몇 번의 갱신 후에 사용자의 선호도를 제대로 반영하는지, 그리고 사용자의 요구를 어느

표 4 사용자 프로파일의 갱신에 따른 적합율

분야 / 갱신 횟수	1	2	3	4	5	6	7	8	9	10	분야 평균
1	20	30	40	35	25	20	25	25	30	26	27.6
3	65	70	80	70	60	60	65	60	70	70	67.0
5	83	85	89	80	83	80	86	85	85	85	84.1
7	85	90	90	91	90	92	93	92	94	93	91.0
9	90	92	93	93	92	93	95	95	96	95	93.4
11	92	94	94	94	94	95	96	96	96	96	94.7
13	91	95	95	95	95	96	97	97	97	98	95.6
15	93	96	96	96	95	97	98	97	98	98	96.4
17	96	97	96	97	95	98	99	98	99	99	97.4
19	97	98	98	98	97	99	99	99	98	98	98.1

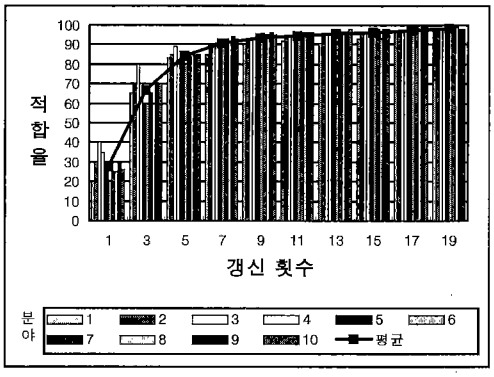


그림 8 사용자 선호도 반영과 검색성능

정도 만족시키는지를 실험을 통해 알아본다.

본 논문에서 제안한 사용자 프로파일 갱신 방법으로 사용자의 선호도를 반영하였을 경우, 사용자 프로파일의 갱신 회수별 적합율을 측정하면 표 4와 같다.

표 4의 값은 10개의 세부 분야별 5명이 평가한 평균 적합율로서, 사용자 프로파일이 갱신되어감에 따라 적합율이 높아짐을 알 수 있다.

그림 8은 표 4를 도식화한 것으로서 꺽은선 그래프는 10개 분야의 평균 적합율을 나타내고 있다. 실험 결과, 적은 갱신 횟수(19회) 동안 사용자의 선호도를 충분히 반영할 수 있었으며(적합율 98% 이상), 이를 통해 본 논문에서 제안한 사용자 프로파일 구성 방법이 사용자의 선호도를 반영하기 위한 방법으로서 상당히 우수함을 알 수 있다.

4.2.2 문서순위의 결정

각 분야별로 5 개씩의 사용자 프로파일을 20회(첫 번째 실험에서 검증된 회수만큼)의 갱신을 거쳐 학습시키면서, 의사문서를 작성하고 이를 이용하여 문서의 순위를 결정하였다.

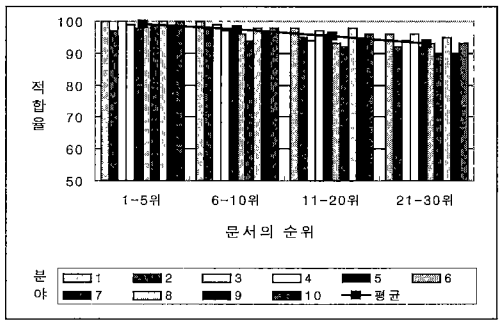


그림 9 문서순위결정과 검색 성능

적합율 공식을 평가에 사용하였는데, 이 실험의 경우의 n은 순위 분류별 제시되는 문서의 개수이다. 즉 예를 들어서 1위~5위에서 n은 5이고, 11위~20위에서의 n은 10이다.

그림 9는 세부분야별 각 5명의 사용자들이 평가한 평균 적합율을 도식화한 것이다.

실험 결과, 1~5순위의 결과에 대해서는 99.1% 이상의 적합율을, 20순위 이내에서 95% 이상의 적합율을 보임으로서, 본 논문에서 제안한 문서순위결정 기법이 사용자의 요구를 만족시키기에 충분히 우수함을 알 수 있다.

한편, SVD 실행 과정에서 발생하는 공간복잡도나 시간복잡도 문제는 iterative algorithm과 sparse matrix를 이용하여 최소화시킬 수 있다.

5. 결론 및 향후 연구 방향

본 논문에서는 사용자 위주의 효율적인 문서순위결정을 위하여 사용자의 선호도를 반영하는 사용자 프로파일을 구성하고, 검색 결과에 대한 용어 연관성을 분석하여 문서순위결정을 수행하는 기법을 제안하였다. 성능 검증을 위하여 사용자 선호도 반영 측면과 문서순위결정 측면에서 실험을 수행하였다. 첫 번째 실험 결과, 사용자 프로파일의 갱신에 따른 적합율이 최고 98.1% 이상을 보임으로써, 본 논문에서 제안한 사용자 프로파일 구성·갱신 방법을 이용하면 사용자의 선호도를 충분히 반영할 수 있음을 알 수 있었다. 또한 두 번째 실험의 결과, 문서의 순위별 적합율이 최고 99.1%(1~5순위)의 결과를 얻게 되어 본 논문에서 제안한 문서순위결정 기법이 사용자에게 적합한 검색결과를 제공할 수 있음을 알 수 있다. 향후에는 동일한 관심분야의 다른 사용자들을 그룹화하고, 그룹 프로파일을 생성하여 상호 참조하는 방법에 대한 연구를 수행할 것이다. 또한 본 논문의 기법을 수행하는데 발생하는 오버헤드를 최소한으로 줄이기 위하여 전처리 과정으로 필터링 기법에 대한 연구를 수행할 것이다.

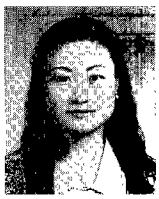
참 고 문 헌

[1] Efthimis N. Efthumiadis, "A User-Centered Evaluation of Ranking Algorithms for Interactive Query Expansion," ACM SIGIR'93, pp. 146-159, 1993.

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

[3] Joon Ho Lee, Yoon Joon Lee, et al., "Ranking

Documents in Thesaurus- Based Boolean Retrieval Systems," Information Processing & Management, Vol. 30, No. 1, pp. 79-91, 1994.
[4] Bracha Shapira, et al., "Information Filtering: A New Two-Phase Model using Stereotypic User Profiling," Journal of Intelligent Information systems, Vol. 8, 1997.
[5] Alistair MacFarlane, Heriot-Watt University, "Information, Knowledge and Learning," Higher Education Quarterly, Vol. 52, No. 1, pp. 77-92, 1998.
[6] Masahiro Morita, Yoichi Shinoda, "Information Filtering Based on User Behavior analysis and Best Match Text Retrieval," SIGIR '94, pp. 272-281, 1994.
[7] Marko Balabanovic et al., "An Adaptive Agent for Automated Web Browsing," http://elib.stanford.edu/Dienst/UI/2.0/Describe/stanford.cs%2fcs-TN-97-52., 1997.
[8] Sima C. Newekk, "User Models and Filtering Agents for Improved Internet Information Retrieval," User Modeling and User-Adapted Interaction, Vol. 7, pp. 223-237, 1997.
[9] 최중민, "인터넷 정보 가공을 위한 에이전트", 정보처리학회지 제4권, 제5호, pp. 101-109, 1997.
[10] Foltz, P. W, "Using Latent Semantic Indexing for Information Filtering," Proceedings of the Conference on Office Information Systems, Cambridge, MA, pp. 40-47, 1990.
[11] Scott Deerwester, Susan T. Dumais, et al., "Indexing by Latent Semantic Analysis", Journal of the American Society for Info rmation Science, 41(6), pp. 391-407, 1990.
[12] Budi Yuwono, Dik L. Lee, "Search and Ranking Algorithms for Locating Resources on the World Wide Web," Proc. of the 12th Int'l Conf. on Data Engineering, New Orleans, Louisiana, Feb, pp. 164-171, 1996.
[13] 임재현, 배희진, 김영찬, "용어 분포에 기반한 지능적 정보검색", 정보과학회 논문지 제25권 제4호, pp. 707-713, 1998.
[14] 유춘식, 우선미 외 3인, "자연어 처리, 통계적 기법, 적합성 검증을 이용한 자동색인 시스템에 관한 연구", 한국정보처리학회 논문지, 제5권, 제6호, pp. 1552-1562, 1998.
[15] 정영미, 정보검색론, 구미무역(주) 출판부, pp. 1-354, 1993.
[16] Czeslaw Danilowicz, "Modelling Of User Preferences and Needs in Boolean Retrieval Systems," Information Processing & Management, Vol. 30, No. 3, pp. 363-378, 1994.



우 선 미
1995년 서남대학교 전자계산학과 이학사. 1997년 전북대학교 대학원 전산통계학과 이학석사. 1997년 ~ 현재 전북대학교 대학원 전산통계학과 박사과정. 관심분야는 인터넷 기반 정보검색, 인공지능, 3D 컴퓨터 애니메이션 등.

유 춘 식
1991년 8월 전북대학교 전산통계학과 이학사. 1994년 전북대학교 대학원 전산통계학과 이학석사. 1994년 ~ 현재 전북대학교 대학원 전산통계학과 박사과정. 관심분야는 구조적 멀티미디어 문서 처리(SGML/XML), 정보검색, 게임공학 등.

김 용 성
정보과학회논문지 : 소프트웨어 및 응용 제 28 권 제 1 호 참조