사용자 위주의 문서순위결정 기법에 관한 연구

우선미[o†], 유춘식†, 이미경††, 김용성†
† 전북대학교 컴퓨터과학과, †† 서울정수기능대학 정보기술과

Study on User-Centered Document Ranking Technique

Sun-Mi Woo[o†], Chun-Sik Yoo†, Mi-Kyung Lee††, Yong-Sung Kim†
† Department of Computer Science, Chonbuk National University
†† Department of Information Technical, Seoul Jungsu Polytechnic College

요약

정보의 가치 증대와 사용자의 정보획득 욕구가 증대됨에 따라 개인 위주의 정보검색 시스템의 필요성이 대두되고 있다. 따라서 본 논문에서는 특정 개인의 관심(interest)과 선호도(preference)를 반영하여 최적의 검색결과를 제공하기 위하여 사용자 프로파일을 구축하고, 통계적 분석 방법 이용하여 문서순위결정을 수행하는 방안을 제안한다.

1. 서론

사회가 복잡해지고 획득할 정보가 많아지며 정보가 특성화됨에 따라 특정 개인의 관심과 선호도 등을 파악하여 보다 만족스러운 결과를 제공하고, 편의성까지 갖춘 사용자 위주의 정보검색 시스템의 필요성이 증대되고 있다. 이러한 사용자 위주의 정보검색을 위한 연구로서는 크게 문서순위결정 방법[16], 정보필터링 방법[3,5,7,9], 기계학습 이론을 활용하여 적응성을 부여하는 방법[7] 등에 관한 연구가 있다 그러나 이러한 방법들도 사용자의 개성이나 선호도 등에 관한 고려가 미흡하기 때문에 사용자 위주의 검색 결과를 효과적으로 제공하지 못하고 있다 그리고 문서순위결정과 필터링을 제공하는 여러 에이전트(agent)들이 개발되고 있으나[1,7,9], 아직까지는 사용자가 검색 도메인에 관한 지식을 반영해야 하는 경우가 많고, 지원하는 기능도 사용자 위주라는 측면에서 미흡한 점이 많다 최근에는 통계적 분석 방법을 이용한 필터링에 관한 연구가 진행되고 있는데[6]. 이 방법은 문서들간의 연관성을 이용하여 완전 정합(exact matching) 기법으로 검색되지 않은 적합한 문서를 찾아줄 수 있으나, 사용자의 선호도를 반영하지는 못한다 이와 같은 문제점들을 해결하기 위하여 본 논문에서는 사용자 프로파일(User Profile)을 구축하여 사용자의 선호도를 반영하고, 통계적 분석 방법을 이용하여 문서의 순위를 결정함으로써 사용자에게 최적의 검색결과를 제공하는 기법을 제안한다.

2. 사용자 위주의 문서순위결정 기법

2.1 사용자 프로파일

본 논문에서는 사용자의 선호도를 반영하기 위하여 그림 1과 같은 구조의 사용자 프로파일을 생성하고 체계적으로 관리한다

사용자 식별자는 서버에 있는 여러 사용자의 프로파일들을 식별하기 위한 것이고, 용어 벡터 $\vec{T}$는 분야별 컬렉션(collection) 내의 색인어(문서의 제목에서 추출함)로 구성되며, 선호도 벡터 $\vec{P}$ 는 $\vec{T}$ 에 대응하는 사용자의 선호도(0과 1사이의 값)를 나타낸다 관심 분야가 변경되었을 경우에는 해당 용어 벡터와 선호도 벡터를 삭제함으로써 사용자 프로파일을 관심 분야에 따라 적절히 관리할 수 있다.

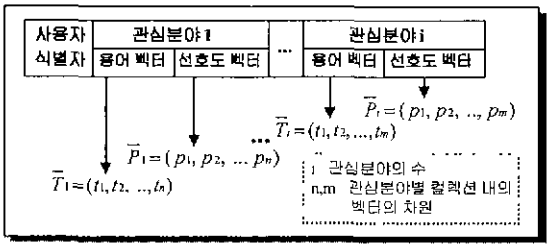

그림 1 사용자 프로파일의 구조

사용자 프로파일을 사용자 중심으로 만들기 위하여 본 논문에서는 다음과 같은 갱신 방법을 제안한다

(1) 사용자 접근에 의한 갱신 방법

사용자가 특정 용어에 중요성을 부여하고자 할 때, 직접 선호도 벡터의 선호도를 변경하는 방법이다

(2) 사용자 적합성 피드백에 의한 갱신 방법

검색결과에 대한 사용자의 적합성 평가에 따라 사용자 프로파일을 갱신하는 방법으로서 갱신 공식은 다음 식과 같다

i) $0.5 \le w_{jk}$ 인 경우, $P_i = \left| 2p_{ij} + \left(\sum_{k=1}^{n} w_{ik}/n \right) \right|_{i=1}^{n}$

ii) $w_{ik} < 0.5$인 경우, $P_i = |p_{ij}|_{i=1}^{n}$

단, D_i 사용자가 적합하다고 평가한 i번째 문서

$= (w_{i1}, w_{i2}, \ldots, w_{in})$

w_{ij} : D_i을 구성하는 용어 j번째의 가중치

사용자 프로파일의 계산을 보다 편리하게 하기 위하여 갱신을 수행할 때마다 0과 1 사이의 값으로 정규화한다

2.2 사용자 위주의 문서순위결정 기법

[알고리즘 1] 사용자 위주의 문서순위결정
[입력] ① 관심분야와 질의어
② 사용자 프로파일, 시소러스 ③ 검색대상
[출력] ① 순위가 결정된 검색결과 ② 갱신된 사용자 프로파일
user_centered_document_ranking( )
begin
1 관심 분야의 프로파일 벡터를 선택한다.
2. 질의를 확장한다.
if (사용자 프로파일의 갱신 전후의 차이가 작다)
then 사용자 프로파일을 참조한다
else 시소러스와 사용자 프로파일 모두를 참조한다.
3. 검색 엔진을 이용하여 문서를 검색한다.
4. 검색결과를 분석하여 적합성 정도를 구한다
4-1 용어-문서 행렬(X)을 구성한다
4-2 행렬 X를 3개의 행렬로 분해한다(SVD)
4-3 성향을 대표하는 차원으로 축소화시켜 분석결과 행렬 $\hat{X}$를 얻는다(reduced SVD).
5. 문서의 순위를 결정한다.
5-1 슈도문서(pseudo document) 생성한다.
5-2. 분석결과 행렬 $\hat{X}$와 슈도문서를 비교한다.
5-3 유사성 정도를 기준으로 하여 내림차순으로 정렬한다.
6. 사용자 적합성 피드백에 의해 선호도를 학습시킨다.
7 사용자에게 최종 검색결과를 제시한다.
end.

2.2.1 검색결과의 적합성 분석

사용자의 질의에 따른 각 문서의 적합성 정도를 알아내기 위하여 잠재적 구조 해석법(LSA:Latent Structure Analysis)을 이용한다. 분석 과정은 먼저 행(raw)은 컬럼션 내의 색인어를, 열(column)은 검색결과 문서를, 값은 문서 내의 용어 빈도를 나타내는 용어-문서 행렬 X를 구성한다.

표 1 검색결과의 일부분

논문	제목
D1	적합성 피드백을 이용한 정보 여과 에이전트
D2	적응형 사용자 인터페이스에 관한 연구
D3	사용자 관심도를 이용한 웹 에이전트
D4	개인 적응형 에이전트를 이용한 정보검색과 필터링
D5	적응형 에이전트
D6	웹 에이전트를 위한 사용자 관심도 학습
D7	학습 기능을 가진 개인 웹 에이전트 구축
:	:

표 2 표1에 대한 용어-문서 행렬

용어	문서						
	1	2	3	4	5	6	7
관심도	0	0	1	0	0	1	0
사용자	0	1	1	1	0	0	0
에이전트	1	0	1	1	1	1	0
웹	0	0	1	0	0	1	0
인터페이스	0	1	0	0	0	1	1
적응형	0	1	0	1	1	0	1
적합성	1	0	0	0	0	0	0
정보검색	0	0	0	1	0	0	1
정보여과	1	0	0	0	1	0	0
피드백	1	0	0	0	0	0	0
학습	0	0	0	0	0	1	1

행렬 X가 구성되면, 행렬의 구조를 이해하기 쉽도록 하기 위하여 다음과 같은 SVD 방법[6]을 이용하여 성분별(본 논문에서는 용어와 문서) 특성을 잘 나타내는 행렬들로 분해한다.

$$X = T_0 S_0 D_0'$$

T_0 : $T_0'T_0 = I$인 직교 행렬, D_0 : $D_0'D_0 = I$인 직교 행렬
S_0 : 대각정방행렬, D_0' : D_0의 전치행렬(transpose)

표 2를 SVD하면, 용어들간의 관계를 나타내는 $T_0(11\times7)$, 문서들의 관계를 나타내는 $D_0(7\times7)$, 값(내림차순으로 정렬됨)이 문서들의 성향을 나타내는 대각행렬 $S_0(7\times7)$를 얻을 수 있다

분석결과 행렬 $\hat{X}$을 구하기 위하여, SVD한 결과 중에서 값들의 성향을 잘 나타내 주는 대표 행렬을 가지고 연산을 수행한다 즉, 표현 공간상의 차원(인자)을 줄이기 위하여 다음 식과 같이 축소화된 SVD를 수행한다[6].

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

$\hat{X}$ 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 $(k \le m)$

위 식에서 생성된 행렬 $\hat{X}$는 계수(rank)가 k로 축소되었을 뿐이지 행렬 X와 근사적으로 일치하게 된다($X \approx \hat{X} = TSD'$) 본 논문에서는 문서들의 성향을 대표할 수 있는 인자를 하나 선택하여 축소화된 SVD를 수행한다. 이유는 분석 데이터가 사용자의 한가지 관심분야에 대한 선호도를 반영한 질의로 검색된 문서들이고, 본 논문의 목적이 적합성을 기준으로 하여 문서의 순위를 결정하는 것이기 때문이다 이때 하나의 인자는 문서들의 공통적인 성향인 사용자 요구에 대한 적합성을 뜻한다. 축소화된 SVD 수행 결과로서 $T(11\times1)$, $S(1\times1)$, $D(7\times1)$(D_0'는 1×7)을 얻을 수 있고, TSD'를 계산하여 최종 결과 행렬 $\hat{X}(11\times7)$를 구할 수 있다. $\hat{X}$은 각 7개의 열(문서)에 포함된 11개의 행(용어)이 해당 문서를 대표할 수 있는 정도를 나타낸다. 본 논문에서는 하나의 인자로 분석하였으므로, 행렬의 값이 문서 7개가 공통으로 갖는 성향인 사용자 요구에 대한 적합성 정도를 나타낸다.

2.2.2 문서의 순위결정

앞 단계에서 수행한 사용자 프로파일과 분석 결과 $\hat{X}$를 이용하여 문서의 순위를 결정한다. 슈도문서는 사용자의 선호도와 검색결과와의 유사성 정도를 구하기 위한 것으로서 다음 식을 이용하여 생성한다[6].

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

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

사용자 프로파일의 용어가 표 2와 같을 때 $P_i(1\times11)$, $T(11\times7)$, $S^{-1}(7\times7)$을 연산하여 슈도문서 $DP(1\times7)$를 구할 수 있다

P_i	용어	관심도	사용자	에이전트	웹	인터페이스	적응성	적합성	정보검색	정보여과	피드백	학습
	선호도	0.90	0.80	0.92	0.50	0.30	0.95	0.50	0.90	0.60	0.50	0.70

↓ $P_i T_0 S_0^{-1}$

DP	분석값						
	0.6561	0.0656	0.0046	0.2328	0.0836	0.4714	0.7193

그림 2 사용자 프로파일과 슈도문서

행렬 $\hat{X}$내에서 두 문서들간의 비교는 다음 식으로 계산할 수 있다[6]

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

본 논문에서는 위의 공식을 응용한 다음 식을 이용하여 분석결과 $\hat{X}$과 슈도문서와의 유사성 정도를 계산한다.

$$DR = ES^2E'$$

단, E : 축소화된 E_0,
E_0 D와 DF가 결합된 확장된 행렬

E_0는 첫 행이 슈도문서이고 나머지가 D_0인 행렬이며 E는 E_0를 1차원으로 축소화시킨 행렬이다. 즉, $E(8\times1)$, $S^2(1\times1)$, $E'(1\times8)$을 연산하면, 표 3과 같이 슈도문서와 검색결과 문서들(D1~D7)의 유사성 정도를 나타내는 대칭행렬 $DR(8\times8)$을 구할 수 있다.

표 3 슈도문서와 검색결과와의 관계(DR)

	DP	D1	D2	D3	D4	D5	D6	D7
DP	4.995	2.258	2.213	3.310	4.111	2.217	3.422	1.868
D1	2.258	1.021	1.000	1.496	1.858	1.002	1.547	0.844
D2	2.213	1.000	0.980	1.466	1.821	0.982	1.516	0.828
D3	3.310	1.496	1.466	2.194	2.725	1.469	2.268	1.238
D4	4.111	1.859	1.821	2.725	3.384	1.825	2.816	1.537
D5	2.217	1.002	0.982	1.469	1.825	0.984	1.519	0.829
D6	3.422	1.547	1.516	2.268	2.816	1.519	2.344	1.280
D7	1.868	0.844	0.828	1.239	1.537	0.829	1.280	0.699

표 3의 진하게 표시된 값이 문서간의 유사성 정도의 크기를 나타내므로, 사용자의 선호도에 따른 문서순위는 D4, D6, D3, D1, D5, D2, D7 순으로 결정된다

4. 실험 및 평가

본 논문에서는 검색결과를 대상으로 문서순위결정을 수행하므로 정확율에 중점을 두고 성능 평가를 한다 실험 영역은 컴퓨터과학 분야의 논문이고, 실험 대상자는 10개의 세부전공 분야별 연구원 5명씩(총 50명)이다 그리고 키워드 추출은 본 연구팀에서 개발한 자동색인 시스템[29]을 사용하여 문서의 제목에서 추출하였으며, 통계적 분석을 위하여 SAS 6.12와 IML (Interactive Matrix Language)을 사용한다 두 가지 측면 즉, 사용자 선호도 반영 측면과 문서순위결정 측면에서 실험 평가를 수행한다 검색은 사용자 프로파일을 참조하여 확장된 질의와 널리 사용되고 있는 Altavista를 이용한다

먼저 본 논문에서 제안한 사용자 프로파일 갱신 방법으로 사용자의 선호도를 반영하면서 정확율을 측정하면 그림 3과 같고, 꺾은선 그래프는 분야 10개의 평균 정확율을 나타내고 있다

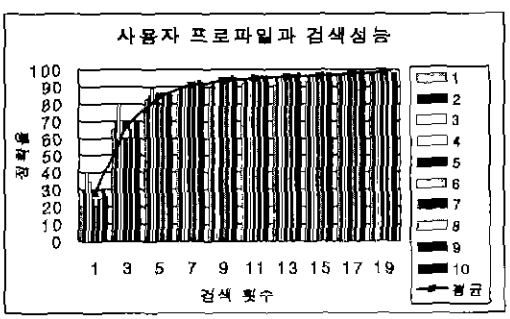

그림 3 사용자 선호도 반영과 검색성능

실험 결과, 적은 검색횟수 동안 사용자의 선호도를 충분히 반영할 수 있고 검색 성능 또한 98% 이상인 실험 결과를 통해, 본 논문에서 제안한 사용자 프로파일 구성 방법이 사용자의 선호도를 반영하기 위한 방법으로서 우수함을 알 수 있다

두 번째 실험은 문서 순위별 세부전공이 다른 사용자들을 대상으로 검색 정확율을 측정하였다 각 분야별 5개씩의 사용자 프로파일을 20회 정도의 검색횟수를 거쳐 각 사용자 위주로 학습시킨 다음, 질의확장과 슈도문서 작성에 이용하여 문서순위를 결정하였다. 그림 4는 세부분야별 각 5명의 사용자들이 평가한 평균 정확율을 나타낸다. 실험 결과 1~5위의 결과에 대해서는 99.1% 이상의 정확율을, 20순위 내에서 95% 이상의 정확율을 보임으로서, 본 논문에서 제안한 문서순위결정 기법이 사용자의 요구를 만족시키기에 충분히 우수함을 알 수 있다

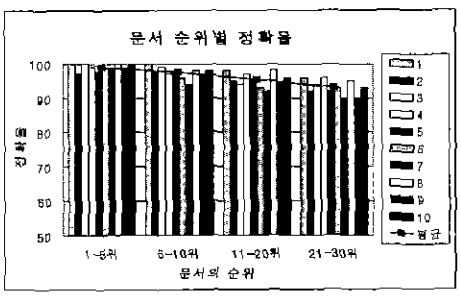

그림 4 문서순위결정과 검색 성능

5. 결론 및 향후 연구 방향

본 논문에서는 사용자 위주의 효율적인 문서순위결정을 위하여 사용자의 선호도를 반영하는 사용자 프로파일을 구성하고, 통계적 분석 방법을 이용하여 문서순위결정을 수행하였다 사용자 프로파일의 성능평가 실험에서 정확율이 98.1% 이상을 보임으로써, 본 논문에서 제안한 사용자 프로파일 구성 방법으로 사용자의 선호도를 충분히 반영할 수 있음을 알 수 있다. 그리고 문서순위결정 방법의 성능평가 실험에서 문서의 순위별 정확율이 최고 99.1%(1~5순위)가 되어, 본 논문에서 제안한 문서순위결정 기법을 이용하면 사용자에게 적합한 검색결과를 제공할 수 있음을 알 수 있다. 향후 연구 과제로는 여러 사용자들의 프로파일을 참고로 하여 유사한 관심분야의 사용자들을 그룹화하고, 각각의 그룹 프로파일을 생성하여 상호 참조하는 방법을 연구하는 것이다 그리고 검색기법에 관한 연구를 계속하여 본 논문에서 제안한 기법과 연계시킨 정보검색 시스템의 개발을 통해 보다 효율적인 검색 시스템을 개발하는 것이다.

참고문헌

[1] Altavista, http://altabista.digital. com/

[2] Eric Bloedorn, Inderject Mani, T Richard MacMillan, "Representational Issues in Machine Learning of User Profiles,"

[3] Masahiro MORITA, Yoichi SHINODA, "Information Filtering Based on User Behavior analysis and Best Match Text Retrieval," SIGIR '94, 1994.

[4] Joon Ho Lee, Yoon Joon Lee, et al., "Ranking Documents in Thesaurus-Based Boolean Retrieval Systems," Information Processing & Management, Vol 30, No 1, 1994

[5] Point Subscription, http://www pointcom.com.

[6] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. "Indexing by Latent Semantic Analysis", JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 41(6), 1990

[7] WebFilter, http://ils.unc.edu/webfilter

[8] 유춘식, 우선미, 유원중, 이종득, 권오봉, 김용성, "자연어 처리, 통계적 기법, 적합성 검증을 이용한 자동색인 시스템에 관한 연구", 한국정보처리학회 논문지, 제5권, 제6호, 1998.

[9] 이정수, 오경환, "직합성 피드백을 이용한 검보여과 에이전트", 검보과학회 발표 논문집, 제25권 1호,1998