사용자 프로파일을 이용한 문서 순위 결정방법

김용호* · 김형균** · 최광미**

조선대학교* · 동강대학**

Ranking Decision Method of Retrived Documents Using User Profile

Yong-ho Kim* · Hyeong-gyun** · Gwang-mi Choi**

Chosun University* · Dongkang College**

E-mail : kimyh@mail.chosun-c.ac.kr

요 약

본 논문에서는 검색된 수많은 결과 중에서 특정 사용자의 선호도를 고려한 최적의 문서만을 제공하기 위하여 사용자 프로파일을 이용한 문서순위 결정기법을 제안한다. 그럼으로써 사용자에게 최적의 문서를 제공하는데 목적이 있다

ABSTRACT

This dissertation proposes a technique of user centered document ranking using User Profile to provide more satisfied results whitch feflect preference of speccific user. User Profile is comstructed to represent his reference of the user.

키워드

User Profile, SVD, Search Engine

Ⅰ. 서 론

인터넷 검색엔진의 성능은 검색엔진 출현 초기에는 검색된 데이터의 양과 검색 속도에 의해 평가되었지만 인터넷상의 정보량이 방대해지고 변화 사이클이 짧아진 최근에는 검색결과의 유효성과 신뢰성으로 평가된다. 일반적으로 검색엔진의 검색결과는 웹 에이전트 또는 검색로봇을 통해 검색엔진이 관리하는 색인 데이터베이스의 크기에 영향을 받는다. 검색엔진들은 방대한 색인데이터베이스를 구축함으로써 많은 검색결과를 제공하지만, 색인 데이터베이스 구축시간의 비현실성으로 인해 검색결과의 신뢰도가 상대적으로 떨어지는 문제점을 갖고 있으며, 이것은 불필요한 인터넷 항해로 인한 시간의 낭비의 원인이 되고 있다.

색인 방법이나 검색방법의 단점이 해결되더라도 데이터의 양이 방대하므로 검색결과의 양 또한 매우 클 수밖에 없고, 사용자가 모든 검색 결과를 참조하기에는 너무 많은 시간과 노력이 소모된다. 따라서 개인의 관심분야나 선호도를 고려하여 일반 정보 검색 시스템이 제공하는 수많은 검색결과 중에서 보다 적합한 결과만을 선택·제공할 수 있는 방법이 요구되고 있다. 이러한 요구에 따라 사용자의 질의와 검색결과 문서와의 유사성 비교를 통해 적합성 정도를 알아내고, 적합성 정도에 따라 순위를 결정하여 사용자에게 문서 순위결정(document ranking decision)에 관한 연구가 진행되고 있다.

문서 순위결정 방법에는 일반적으로 문서를 대표하는 용어 추출과 추출한 용어에 가중치를 부여하는 방법 등 여러 가지가 있다. 용어에 가중치를 부여하기 위하여 가장 많이 사용되는 방법은 용어의 출현 빈도와 같은 문서의 통계적 정보를 이용하는 방법이다. 그러나 이 방법으로 결정된 용어의 가중치는 용어가 대표하는 해당 문서를 표현하기는 적당하지만, 순위를 결정할 전체 문서들 중에서 해당 문서를 표현하기 위한 가중치로

서는 부적절하다. 따라서 문서 전체의 관계성을 고려하여 각 문서의 특성을 표현할 수 있도록 용어의 가중치를 결정할 수 있어야 한다. 현재 문서들 간의 잠재적인 의미(semantic)를 분석하여 이용하는 연구는 간단한 뉴스나 의학 메모를 대상으로 하는 자동색인과 정보 필터링에 적용되고 있다. 그리고 특정 사용자의 관심분야나 선호도를 고려하여 문서의 순위를 결정하기 위해서는 사용자의 요구가 몇 개의 용어 조합이 아닌 상세하면서 선호도를 반영할 수 있는 형태로 표현되어야 한다. 사용자의 요구를 표현하기 위한 연구는 사용자 위주의 정보 검색 기법의 필요성이 증가하면서 정보 필터링 분야에서 사용자 프로파일(user profile)이라는 이름으로 연구되고 있다.

본 논문은 사용자 프로파일을 구축하여 사용자의 선호도를 표현하고 검색결과 문서들을 대상으로 잠재적 구조를 분석한 다음, 사용자 프로파일과 분석결과로 표현된 문서들과의 유사성을 비교한다. 그리고 적합성 정도에 따라 사용자에게 최적의 문서를 제공하는 데에 목적이 있다.

II. 사용자 프로파일구조

사용자 위주의 문서 순위결정을 수행하기 위해서는 먼저 사용자의 관심과 선호도가 표현되어 있어야 한다. 본 논문에서는 사용자의 관심분야와 선호도를 표현하기 위하여 "사용자 식별자(User ID)", "용어열 T_i(Term Array)" 그리고 "선호도 벡터 P_i(Preference Vector)"로 구성된 사용자 프로파일을 제안한다.

사용자 프로파일의 구조를 그림으로 표현하면 그림 1과 같다.

User ID	관심분야 1		....	관심분야 k	
	용어열	선호도백터		용어열	선호도벡터

그림 1. 사용자 프로파일의 구조

기계학습	0.8	클래식	0.9
문서 순위결정	0.9	소나타	0.5
시소러스	0.2		
자동 인덱싱	0.5		
적합성 평가	0.7	...	
필터링	0.8		
유전자	0.9		

III. 유전자 알고리즘을 이용한 문서 순위결정

[알고리즘 1] 사용자 위주의 문서 순위결정
[입력] 사용자 프로파일, 순위를 결정할 문서들
(출력) 순위가 결정된 문서, 갱신된 사용자 프로파일

begin
1. 사용자 프로파일의 관심분야를 입력한다.
/* 단계 2 ~ 단계 5
: 검색결과로 얻은 문서들의 유전자 코드형태 변환 및 관계성 분석함. */
2. [정의 3.2]와 [정의 3.3]을 만족하는 용어-문서 행렬 X를 구성한다.
3. SVD에 의해 행렬 X를 세 개의 행렬 T_0, S_0, D_0로 분해한다
/* 식(3.4)에 의해 행렬 X를 잠재적 구조 모델로 구성함 */
4. 단계 3에서 구한 행렬 S_0의 값들 중에서 가장 큰 값을 가진 차원을 선택한다.
/* 문서들의 성향을 가장 잘 대표하는 하나의 인자를 선택함 */
5. 검색된 문서를 개체와 개체군으로 생성 /* 식(3.6) 이용 */
6. 단계 4에서 구한 S_0의 가장 큰 값에 해당하는 용어를 선택하고 단계 5에서 생성한 개체군에 속하는 문서들에 대해 적합도 검사
/* 식(3.7)~식(3.12) 이용)
/* 단계 7~단계 16 : 사용자 프로파일의 선호도를 갱신. */
7. 갱신 횟수 ← 0
while (갱신 회수 ≤ m) begin
while (갱신 회수 < 1) begin
8. "잠재적 구조를 반영한 갱신" 방법을 수행한다. /* 식(3.1) */
9. "사용자 접근에 의한 갱신" 방법을 수행한다.
end.
if (갱신 회수 ≤ n) then begin
10. "사용자 프로파일을 이용한 갱신" 방법을 수행한다. /* 식(3.2), 식(3.3) 참조 */
end
11. 사용자 프로파일의 선호도 벡터 P_i를 이용하여 단계 5에서 생성한 개체군을 대상으로 단계 6 수행 새로운 개체군 생성
12. 단계 6과 단계 11의 개체군을 Blind Crossover수행 /* 식(3.13) */
13. 단계 6과 단계 11의 개체군을 단어 가중치 기반 교배 수행하여 사용자 선호도에 따라 적합

한 문서와 비적합 문서를 분리. /* 식(3.13) */
14. 단어 적합성 기반 돌연변이 수행 /* 식 (3.17) */ 15. 단계 8, 단계 9 수행
end.
16. 갱신 회수 ← 갱신 회수 + 1
17. 사용자가 입력한 임계값에 해당하는 최존 결과 문서 제시
end.

[알고리즘 1]에서 갱신 횟수 n은 문서 순위결정 결과로 제시된 상위 10% 내의 문서들의 적합률이 60% 이상이 되는 순간을 뜻한다. 그리고 갱신 횟수 m은 사용자 프로파일의 마지막 갱신 횟수로서 문서 순위결정결과로 제시된 상위 10% 내의 문서들의 적합률이 95% 이상이 되는 순간을 의미한다. 갱신 횟수 n과 m은 본 논문의 실험 평가에서 결정한다.

IV. 실험 및 평가

본 논문에서 제안한 사용자 위주의 문서 순위결정 기법은 사용자의 선호도 반영과 문서의 순위결정을 목적으로 하고 있으므로, 성능 평가를 위해 두 가지 측면에서 실험 평가한다. 첫 번째 실험은 본 논문에서 제안한 사용자 프로파일의 갱신 방법이 얼마나 효과적인지를 알아보기 위한 실험이다. 본 논문에서 제안하는 사용자 프로파일 갱신 방법의 성능을 실험하고, 구성과 갱신방법이 다른 사용자 프로파일과의 성능을 비교 분석한다. 두 번째 실험은 제안한 문서 순위결정 기법을 이용하여 결정된 문서의 순위별 적합률을 구하여 본 논문의 기법이 사용자의 요구에 대해 얼마나 적합한 결과를 제공하는지를 검증한다.

A. 실험 환경

본 논문에서 제안한 사용자 프로파일과 잠재적 구조 분석을 이용한 검색된 문서의 순위결정을 위한 실험 환경은 다음과 같다.

1. 실험 데이터

· 실험 데이터 1 : 일반 용어 기반 정보 검색 시스템으로 검색한 컴퓨터
과학 분야의 10개의 세부 분야에 해당하는 논문

· 실험 데이터 2 : 100개의 문서집합

2. 용어 추출

a. 통계적 분석(용어 추출범위)
· 실험 데이터 1 - 논문의 제목
· 실험 데이터 2 - 논문의 제목과 요약
b. 사용자 적합성 피드백
· 용어 추출 범위 : 논문의 제목과 요약
· 용어 가중치 결정 방법 : $tf^* idf$

사용자 위주의 문서 순위결정은 유사한 관심분야라도 사용자의 선호도에 따라 개인별로 문서 순위결정이 다르게 이루어져야 하므로, 세부 전공이 다른 사용자들을 대상으로 순위별 적합률을 측정하였다. 각 분야별로 사용자 프로파일을 13회의 갱신을 거쳐 학습시키면서 의사문서를 작성하고 이를 이용하여 문서의 순위를 결정하였다. 표 1의 값은 세부 분야별 각 3명의 사용자들이 평가한 평균 적합률이다.

표 1. 문서 순위별 평균 적합률

순위 / 관심분야1	1~5	6~10	11~20	21~30
1	100	100	98	96
2	97	98	95	92
3	100	99	94	94
4	99	97	97	96
5	98	98	96	92
6	99	96	93	93
7	99	94	92	90
8	100	98	98	95
9	99	97	95	90
10	100	98	96	93
평균	99.10	97.50	95.40	93.10

평가에 적합률 공식을 사용하였는데, 이 실험의 경우에 n은 순위별 제시되는 문서의 개수이다. 예를 들어 문서의 순위 1~5에서 n은 5이고, 순위 11~20에서의 n은 10이다.

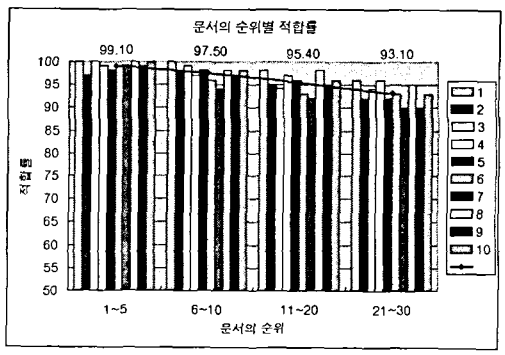

그림 2. 사용자 위주의 문서 순위결정 기법의 성능

V. 결 론

본 논문에서 제안한 사용자 위주의 문서 순위 결정 기법에 대한 성능 검증을 위하여 두 가지 측면, 즉 사용자 선호도 반영 측면과 문서 순위결정의 성능 측면에서 실험을 수행하였다. 사용자 프로파일의 성능평가 실험에서 상위 10% 내의 논문들을 평가했을 때 3회라는 매우 적은 회수의 갱신으로도 평균 적합률 67%를 얻을 수 있었다. 그리고 9회의 갱신으로 적합률 92%를 얻을 수 있었다. 또한 사용자 프로파일의 갱신에 따른 적합률이 최고 98.5% 이상을 보임으로써 본 논문에서 제안한사용자 프로파일 구성·갱신 방법을 이용하면 사용자의 선호도를 충분히 반영 할 수 있음을 알 수 있었다. 사용자 프로파일의 갱신에 음의 사용자 피드백과 양의 사용자 피드백을 모두 사용하는 관련연구와의 비교 평가에서 동일한 분야의 문서들을 대상으로 할 경우에는 음의 피드백 효과가 매우 적음을 알 수 있었다. 본 논문에서 제안한 갱신 방법에 의해 학습된 사용자 프로파일과 본 논문에서 제안한 문서 순위결정 기법을 이용하여 실험을 실시한 결과, 문서의 순위별 적합률이 최고 99.1%의결과를 얻게 되어 본 논문에서 제안한 사용자 위주의 문서 순위결정기법이 사용자에게 적합한 검색결과를 제공할 수 있음을 알 수 있었다.

실험 결과 정보 검색 시스템에서 유전자알고리즘을 사용하면 보다 향상된 정보 검색을 할 수 있다는 것이 증명되었다.

참고문헌

[1] 김창민, 김용기, "퍼지 관계곱을 이용한 내용 기반 정크 메일 분류 모델", 정보과학회 논문지, 제 29권 제 10호, 2002, 10

[2] 신봉기, 김영환, "인터넷 정보 검색 서비스 동향," 정보과학회지, 정보과학회, 16권 8호, pp16~20, 1998.8.

[3] Chen, L., and sycara, K., "WebMate: Personal Agent forBrowsing and Searching," In Proceeding of the 2nd International Conference on Autonomous Agents, pp. 132-139, 1998.

[4] Czeslaw Danilowi,z, la.oslaw Bali tsC···· Document Ranking based upon Markov Chains,"Information Processing and Management, Vol. 37, pp. 623-637, 2001.

[5] Geoff.ey I. Webb, Michael J. Pazzani, Daniel Billsus, "Machine Loaming for User Modeling," User Modeling and User-Adapted Interaction, Vol. 11, pp. 19-29, 2001.

[6] Martin Wechsler, Peter Schauble, "The Probability Ranking Principle Revisited," Information Retrieval, Vol. 3,pp. 217-227, 2000.