사용자 프로파일을 이용한 문서 검색순위 결정에 관한 연구

*김형균, **김용호, **이상범
*조선대학교 컴퓨터공학과, **조선이공대학 컴퓨터정보과

A Study on Ranking Decision of Retrieved Documents Using User Profile

Hyeong gyun Kim*, Yong Ho Kim**, Sang Beom Lee**
*Dept. of Computer , Chosun University, **Chosun College of Science & Technology
E-mail : multikim87@hanmail.net

요 약

본 논문에서는 동일한 분야의 검색된 문서가 갖는 하나의 성향을 중심으로 문서들 자체가 가지고 있는 관계성을 분석하여 용어의 가중치를 결정하였다. 그리고 사용자의 관심분야와 선호도를 적절히 표현하기 위하여 질의가 아닌 사용자 프로파일을 구축하여 이용하였다. 사용자 프로파일은 관심 분야별로 용어열과 선호도 벡터로 구성하고, "사용자접근에 의한 갱신", "사용자 프로파일을 이용한 갱신" 방법을 이용하여 사용자 프로파일을 사용자 위주로 학습시킨다. "사용자 접근에 의한 갱신" 방법은 주제 분야에 대한 지식이 있는 경우에 적용할 수 있는 방법으로서 실험 결과, 사용자 프로파일이 사용자의 선호도를 제대로 표현하기까지의 갱신 회수를 상당히 감소시킬 수 있었다. "사용자 프로파일을 이용한 갱신" 방법은 갱신초기에 수행하는 방법으로서 선호도 값의 차이를 명확히 해주는 결과를 가져온다.

키워드

User Profile, Retrieved Documents

Ⅰ. 서 론

인터넷 검색엔진에서 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 가중치로서는 부적절하다. 따라서 문서 전체의 관계성을 고려하여 각 문서의 특성을 표현할 수 있도록 용어의 가중치를 결정할 수 있어야 한다. 현재 문서들 간의 잠재적인 의미를 분석하여 이용하는 연구는 간단한 뉴스나 의학 메모를 대상으로 하는 자동색인[1]과 정보 필터링[2]에 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 논문에서는 검색된 수많은 결과 중에서 특정 사용자의 선호도를 고려한 최적의 문서만을 제공하기 위하여 사용자 프로파일과 유전자 알고리즘(genetic algorithm)을 이용한 문서 순위결정 기법을 제안한다. 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 프로파일의 구조

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 구조는 I 정의 1]과 같다.

[정의 1]
각 개인별 사용자 프로파일을 UP, UP가 표현할 수 있는 k개의 관심분야를 $I=\{I_1, I_2, ..., I_k\}$, 용어의 개수가 n개인 UP의 i번째 관심분야의 용어열을 $T_i=\{t_{i1}, t_{i2}, ..., t_{in}\}$ 그리고 T_n에 대응하는 UP의 i번째 관심분야의 선호도 벡터를 $P_i=\{p_{i1}, p_{i2}, ...p_{in}\}$라고 할 때, 사용자 프로파일 $UP=\{(I_1, T_1, P_1), (I_2, T_2, P_2), ..., (I_k, T_k, P_k)\}$이다.

사용자 프로파일의 구조를 그림으로 표현하면 [그림 1]과 같다.

User ID	관심분야 1		...	관심분야 k	
	용어열	선호도 벡터		용어열	선호도 벡터

[그림 1] 사용자 프로파일의 구조

[그림 1]에서 관심분야 1의 용어열은 $T_1=(t_{11}, t_{12}, ..., t_{1n})$로 선호도 벡터는 $P_1=(p_{11}, p_{12}, ..., p_{1n})$로 나타낼 수 있으며 관심분야 k의 용어열은 $T_k=(t_{k1}, t_{k2}, ..., t_{km})$로 선호도 벡터는 $P_k=(p_{k1}, p_{k2}, ..., p_{km})$로 나타낼 수 있다. "User ID"는 여러 사용자의 프로파일들을 식별하기위한 것이다. 그리고 관심분야 별로 구성된 용어열 T_i는 관심분야 i에 속한 문서의 제목에서 추출한 용어로 구성한다. 선호도 벡터 P_i는 용어열 T_i에 대응하는 사용자의 선호도를 나타낸다. 선호도 값은 본 논문에서 제안하는 사용자 프로파일 갱신 방법에 의해 사용자 위주로 학습된다.

[정의 2]
검색결과로 얻은 문서가 $d_1, d_2, ..., d_m$이고 $d_1=\{t_{11}, t_{12}, ...t_{1i}\}, d_2=\{t_{21}, t_{22}, ...t_{2i}\}, \cdots,$ $d_m=\{t_{m1}, t_{m2}, ...t_{mi}\}$일 때, 문서의 집합을 $D=\{d_1, d_2, ..., d_m\}$라고 하고 문서에서 추출한 용어의 집합을 $T=\{t_1, t_2, ..., t_n\}$라고 한다.
단, m은 문서의 개수이고 n은 문서집합 D에서 추출한 모든 용어의 개수이다.

i번째 관심분야의 사용자 프로파일의 초기 용어열은 i번째 관심분야의 검색결과 문서들에서 용어-문서 행렬의 용어열과 동일하다. 분석 초기에 구성하는 용어-문서행렬의 구조는 [정의 2]와 [정의 3]을 따른다.

[정의 3]
용어-문서 행렬 X의 행과 열은 [정의 3.2]의 D의 T이고, 행렬 X의 원소 값은 t_n이 d_m에 출현하는 용어의 출현빈도이다.

III. 사용자 위주의 문서 순위결정 알고리즘

본 논문에서 제안하는 문서 순위결정 기법의 기본 개념은 순위를 결정할 문서들은 일반 정보검색 시스템에 의해 얻어진 동일 분야의 문서들이고, 이렇게 얻은 문서들은 사용자의 관심을 반영하므로 한 가지 성향을 갖게 된다는 점이다. 따라서 한 가지 성향을 중심으로 차원을 줄여 분석을 수행할 수 있다. 검색결과 문서들은 사용자가 입력한 질의와 관련된 문서들이므로 분석 과정에서 발견된 하나의 성향은 관심분야라고 할 수 있다. 따라서 동일 분야의 문서들을 잠재적 구조 분석 방법으로 분석하면 문서들 자체가 갖는 관계성을 고려한 용어의 가중치를 얻게 된다. 이와 같이 분석된 결과를 사용자의 선호도를 반영하는 사용자 프로파일과 비교하여 적합한 순서대로 문서의 순위를 결정할 수 있다.

본 논문에서 제안하는 사용자 프로파일은 여러 개의 관심분야를 허용하므로 새로운 관심분야일 경우에는 관심분야를 입력하고, 학습이 끝난 관심분야일 경우에는 원하는 관심분야를 선택한다. 새로운 관심분야가 생겼을 경우에는 해당 관심분야에 대한 사용자의 선호도가 충분히 학습될 수 있도록 사용자 프로파일의 갱신에 사용자가 다소 관여할 필요가 있다. 본 논문에서 제안하는 사용자 위주의 문서 순위결정 알고리즘은 [알고리즘

1]과 같다. 사용자가 원하는 관심분야는 [알고리즘 1]에 의해 사용자의 선호도를 충분히 반영하면 문서 순위결정을 위하여 사용자는 원하는 관심분야만 선택하면 된다.

[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 2]와 [정의 3]을 만족하는 용어-문서 행렬 X를 구성한다.
3. SVD에 의해 행렬 X를 세 개의 행렬 T0, S0, D0로 분해한다
/* 행렬 X를 잠재적 구조 모델로 구성함 */
4. 단계 3에서 구한 행렬 S0의 값들 중에서 가장 큰 값을 가진 차
   원을 선택한다.
/* 문서들의 성향을 가장 잘 대표하는 하나의 인자를 선택함 */
5. 검색된 문서를 개체와 개체군으로 생성
6. 단계 4에서 구한 S0의 가장 큰 값에 해당하는 용어를 선택하
   고, 단계 5에서 생성한 개체군에 속하는 문서들에 대해 적합도
   검사
/* 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.
     9. "사용자 접근에 의한 갱신" 방법을 수행한다.
   end.
   if (갱신 회수 ≤ n) then begin
   10. "사용자 프로파일을 이용한 갱신" 방법을 수행한다.
   end
   11. 사용자 프로파일의 선호도 벡터 P1를 이용하여 단계 5에서
       생성한 개체군을 대상으로 단계 6 수행 새로운 개체군 생
       성
   12. 단계 6과 단계 11의 개체군을 Blind Crossover수행
   13. 단계 6과 단계 11의 개체군을 단어 가중치 기반 교배 수행
       하여 사용자 선호도에 따라 적합한 문서와 비적합 문서를
       분리.
   14. 단어 적합성 기반 돌연변이 수행
   15. 단계 8, 단계 9 수행
end.
16. 갱신 회수 ← 갱신 회수 + 1
17. 사용자가 입력한 임계값에 해당하는 최종 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.

1. 실험 환경

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

- 실험 데이터 1 : 일반 용어 기반 정보 검색 시스템으로 검색한 컴퓨터 과학 분야의 10개의 세부 분야에 해당하는 논문
- 실험 데이터 2 : 100개의 문서집합

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 : $tf^* idf$

2. 문서 순위결정 기법의 성능 평가

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 1]의 값은 세부 분야별 각 3명의 사용자들이 평가

한 평균 적합율이다.

[표 1] 문서의 순위별 평균 적합율

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

본 논문에서 제안안한 방법을 사용하면 사용자 위주의 문서검색에서 상위 1~10번째의 문서는 적합률이 평균 98% 이상이 되므로 질의에 따른 방대한 자료에서 사용자가 관심 있는 정보를 제공할 수 있으며 알맞은 정보검색을 위한 시간과 노력을 줄일 수 있을 것으로 사료된다.

V. 결 론

본 논문에서는 동일한 분야의 검색된 문서가 갖는 하나의 성향을 중심으로 문서들 자체가 가지고 있는 관계성을 분석하여 용어의 가중치를 결정하였다. 그리고 사용자의 관심분야와 선호도를 적절히 표현하기 위하여 질의가 아닌 사용자 프로파일을 구축하여 이용하였다. 사용자 프로파일은 관심 분야별로 용어열과 선호도 벡터로 구성하고, "사용자접근에 의한 갱신", "사용자 프로파일을 이용한 갱신" 방법을 이용하여 사용자 프로파일을 사용자 위주로 학습시킨다. "사용자 접근에 의한 갱신" 방법은 주제 분야에 대한 지식이 있는 경우에 적용할 수 있는 방법으로서 실험결과, 사용자 프로파일이 사용자의 선호도를 제대로 표현하기까지의 갱신 회수를 상당히 감소시킬 수 있었다. "사용자 프로파일을 이용한 갱신" 방법은 갱신초기에 수행하는 방법으로서 선호도 값의 차이를 명확히 해주는 결과를 가져온다. 즉, 선호도가 높은 것은 더욱 높게, 낮은 것은 더욱 낮게 만들어 적합과 비적합의 경계를 넓혀주는 방법이다. 실험 평가 결과, 이 갱신 방법은 순위가 결정된 문서들 중 상위 10%의 적합률이 60% 이상이 되는 시점인 갱신 회수 3회까지 실시하는 것이 적당함을 알 수 있다. 이때 "사용자 접근에 의한 갱신" 방법을 함께 수행하면 1회의 갱신으로 적합률 60% 이상을 얻을 수 있었다. 그리고 동일한 관심분야의 논문들은 한 가지 성향을 나타내므로 검색된 논문들을 분석하는 과정에서 한 개의 인자로 축소화하여 SVD를 수행하였다.

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 제안한 사용자 프로파일 구성·갱신 방법을 이용하면 사용자의 선호도를 충분히 반영 할 수 있음을 알 수 있었다.

참고문헌

[1] Baeza-Yates, R. and Ribeiro-Neto, Berthier. Moderm Information Retrieval, addison-wesley Pub. Co(sd), 1992.

[2] Bracha Shapira, Uri Hanani, Adiraveh, Peretzshoval, "Information Filtering: A New Two-Phase Model Using Stereotypic User Profiling." Journal of Intelligent Information systems, Vol. 8, 1997.

[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] Chris Buckley, Gerard Salton, and James allan. The effect of adding relevance information in a relevance feedback environment. In Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 292-300, Dublin, Ireland, 1994.

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