개인 및 그룹 기반 검색의 혼합을 통한 개인화 검색 효율 향상*

한기준[O]1, 이문용1, 정영임2, 김재훈2, 김환민2
1한국과학기술원 지식서비스공학과, 2한국과학기술정보연구원
Keejun.han@kaist.ac.kr, munyi@kaist.ac.kr, acorn@kisti.re.kr, jay.kim@kisti.re.kr, mrkim@kisti.re.kr

A Hybrid Approach of Personalized and Groupize Search for Improving Effectiveness of Personalized Search

Keejun Han[O]1, Mun Y. Yi1, Youngim Jung2, Jayhoon Kim2, Hwanmin Kim2
1Department of Knowledge Service Engineering, KAIST, 2KISTI

요 약

기존의 개인화 검색 기법은 사용자가 남긴 데이터의 수가 적을 경우 그 정확성이 떨어지고, 또한 모든 사용자에게 각각 다른 프로파일을 구축할 때 발생하는 물리적 시간적 비효율이 발생한다. 따라서 본 논문에서는 서비스를 많이 사용하는 사용자에게는 개인화 서비스를, 그 외의 사용자들에게는 해당 사용자가 속한 그룹의 프로파일을 대신 제공하여 서비스 구현시의 부담을 줄이고 실제 검색 성능을 향상시키는 혼합 방식을 제안하였다. 제안하는 기법의 효용성 검증을 위해, 실제 데이터셋을 사용한 성능 평가 실험을 진행하였고, 그 결과 제안한 혼합 모델이 기존의 개인화 및 그룹 기반 검색 기법보다 더 뛰어난 검색 성능을 보임을 확인하였다.

1. 서론 및 관련 연구

개인화 검색은 동일한 질의어에 대해 사용자의 선호도에 따라 서로 다른 분석 결과를 제공하는 검색 기법으로 다양한 개인의 정보 요구에 맞춤화된 검색 결과를 제공하여 사용자의 검색 만족도를 높이는 기법이다. 개인화 검색은 일반적으로 비개인화 검색 기법보다 사용자의 검색 만족도가 높은 것으로 알려져 있으며, 궁극적으로 사용자는 높은 검색 만족도를 기반으로 더 많은 시간 해당 서비스에 체류하여 서비스 활성화에도 크게 기여하는 것으로 알려져 있다 [1, 2].

그러나, 일반적으로 개인화 검색은 해당 서비스를 많이 방문하지 않는 개별 사용자들의 프로파일까지 모두 분석하여 개인화 서비스를 제공하다 보니, 실제 검색 성능 향상에 비하여 물리적 시간적 자원 소비를 요구한다. 또한 개인화 검색 서비스는 사용자 데이터가 충분하지 않을 경우 구축된 사용자 프로파일의 정확도가 낮아 만족할 만한 개인화 성능이 나오지 않는 콜드 스타트 (Cold Start) 문제를 공통적으로 지니고 있다. 비록 이러한 문제점을 해결하기 위해서 최근 여러 개량된 개인화 알고리즘들이 제안되었지만, 이에 비례하여 점점 더 복잡해지는 알고리즘들은 오히려 해당 알고리즘들의 물리적, 시간적 자원 요구로 실제 서비스에 적용하는 것이 점점 더 어려워지고 있다. 이로 인하여, 최근의 개인화 검색 서비스는 검색 성능 향상과 더불어 실제 서비스에 적용 및 확장가능한지를 검증하는 것의 중요성 또한 점점 더 부각되고 있다 [3, 4].

* 본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신 방송 연구개발사업의 일환으로 수행하였음. [10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발]

본 논문은 기존의 개인 프로파일 구축을 통한 개인화 검색 알고리즘이 해당 서비스를 활발하게 사용하는 사용자에게 효과적인 문제점을 해결하기 위해 개인 프로파일의 정확성이 낮거나 서비스를 처음 사용하는 사용자들에게는 그들이 속한 그룹의 프로파일을 대신 사용하는 방법을 제안한다.

개인 프로파일에 비해 그룹 프로파일은 공통된 흥미를 갖고 있는 사용자들의 관심 주제를 비교적 빠르게 파악할 수 있다는 장점이 있다. 개개인이 관심 있는 주제들 중 상당수는 현재 자신이 속한 그룹의 관심사로부터 비롯되는 것이 많다 [5]. 비록 개개인이 가지고 있는 다양한 주제를 그룹 프로파일만으로 모두 파악하는 것은 어렵지만 적어도, 해당 사용자에게 가장 중요한 몇 가지 주제는 그룹 프로파일을 분석하여 얻어낼 수 있다. 본 논문에서 제안하는 방법을 통해 얻을 수 있는 이득은 다음과 같다.

- 제안하는 방법은 기존의 개인화 검색 기법에 비해 보다 가볍게 개인 프로파일들을 구축하고 운영할 수 있다. 실제 매우 적은 수의 사용자가 절대 다수의 데이터를 생성하는 거듭제곱 법칙 (power-law)이 강하게 나타나는 서비스 특성상 그러한 소수의 사용자들에게는 개인 프로파일을 구축하여 관리하고 다수의 사용자들에게는 그룹 기반의 프로파일을 제공하여 보다 효율적으로 서비스 자원을 구축하고 관리할 수 있다. 또한, 그룹 기반의 검색 기법은 기존의 개인화 검색 기법과 비교하여 통계적으로 큰 차이가 없는 성능을 보였다.
- 제안하는 방법은 상기 제시한 '콜드 스타트' 문제에 보다 가볍게 대응할 수 있다. 서비스에 축적된

사용자 데이터가 적어 정확한 사용자 프로파일을 구축하기 어려운 경우 사용자가 속한 그룹의 프로파일을 대신 사용하여 검색 결과의 정확도를 향상시킬 수 있다.

2. 개인화 검색 및 그룹 기반 검색 알고리즘 구현

개인화 검색의 일반적인 수행 과정은 다음과 같다. (1) 주어진 문서를 설명하는 단어들로 구성된 문서 프로파일을 구축한 뒤, (2) 사용자의 관심사를 표현하는 단어들로 구성된 사용자 프로파일을 구축하고, (3) 이를 바탕으로 사용자 및 문서 프로파일의 유사도를 측정하여 그 유사도가 높은 문서들의 검색 순위를 높여주는 재정렬 순서로 진행된다.

가장 먼저 각 문서를 표현하는 단어 벡터 모델을 만든다. 한 단어 벡터는 문서를 표현하는 단어들과 각 단어의 중요도를 의미하는 가중치들로 구성된다. 본 논문에서는 문서를 표현하는 단어들을 해당 문서의 초록에서 추출하여 사용하였다. 단어들의 가중치로는 tf-idf (Term Frequency - Inverse Document Frequency) 값을 사용하였다. 수식 (1)은 문서 d에 대한 단어 프로파일 P_d를 나타낸다. t_n은 문서 d에 포함한 단어들이며, n은 문서 d가 포함하고 있는 단어들의 총 수를 의미한다. 그리고 $w_{t_n,d}$는 문서 d에서 단어 t_n이 가지고 있는 tf-idf값을 나타낸다.

$$P_d = \{(t_1, w_{t_1,d}), \cdots (t_n, w_{t_n,d})\} \quad (1)$$

문서 프로파일 구축이 완료되면 다음으로 사용자 별로 개인 프로파일을 만든다. 사용자가 클릭한 문서들에 등장한 단어들의 가중치들이 합쳐져 하나의 단어벡터로 개인 프로파일이 표현된다. 결과적으로 사용자가 클릭한 문서들에서 자주 발견되는 단어일수록 더 높은 가중치를 가지게 된다. 수식 (2)는 사용자 u의 프로파일을 나타내며 수식 (3)은 사용자 u의 프로파일에서 단어 t의 가중치 $w_{t,u}$를 계산하는 방법을 보여준다. 이때, D_u는 사용자 u가 클릭한 모든 문서들의 집합을 나타낸다.

$$P_u = \{(t_1, w_{t_1,u}), \cdots (t_m, w_{t_m,u})\} \quad (2)$$

$$w_{t,u} = \sum_{d \in D_u} tf - idf_{t,d} \quad (3)$$

하지만 서론에서 기술한 바와 같이 사용자 u에게 정확한 사용자 프로파일을 구축할 만큼의 데이터가 존재하지 않는다고 판단된다면, 사용자 u가 속한 그룹 g의 그룹 프로파일로 사용자 u의 개인 프로파일을 대체할 수 있다. 그룹 g에 속한 사용자들이 클릭한 문서들에 등장한 단어들의 가중치들이 합쳐져 하나의 단어벡터로 그룹 프로파일이 표현된다. 즉, 동일한 그룹 내에 속한 사용자들이 많이 클릭한 문서일수록 해당 그룹이 공통된 관심을 갖고 있는 주제를 내포하는 문서일 가능성이 크다는 것을 의미한다. 예를 들어, 사용자 u가 속한 그룹 g가 의학전문대학원이라면 해당 그룹에 속한 사용자들이 많이 클릭한 문서는 의학 전공과 관련된 지식을 포함한 문서일 가능성이 크다고 볼 수 있다.

수식 (4)는 그룹 g의 프로파일 P_g를 나타내며, 수식 (5)는 P_g에서 단어 t의 가중치 $w_{t,g}$를 계산하는 방법을 보여준다. 이때, D_g는 그룹 g에 속한 사용자들이 클릭한 모든 문서들의 집합을 나타낸다. 또한, 수식 (6)은 D_u의 개수가 k보다 적을 때, 사용자 프로파일 P_u는 P_g로 대체됨을 의미한다.

$$P_g = \{(t_1, w_{t_1,g}), \cdots (t_l, w_{t_l,g})\} \quad (4)$$

$$w_{t,g} = \sum_{d \in D_g} tf - idf_{t,g} \quad (5)$$

$$P_u = P_g \quad (if\ n(D_u) \le k) \quad (6)$$

마지막으로 사용자와 문서간의 코사인 유사도(Cosine Similarity)를 계산하여 유사도 값이 높은 문서부터 상위 순위에 놓는 재정렬을 실시한다. 수식 (7)은 문서 프로파일 P_d와 사용자 u의 프로파일 P_u간의 유사도를 구하는 식이다.

$$sim(P_d, P_u) = \frac{\sum_{t \in P_d \cap P_u} tfidf_{t,d} * tfidf_{t,u}}{\sqrt{\sum_{t \in P_d} tfidf_{t,d}{}^2} \sqrt{\sum_{t \in P_u} tfidf_{t,u}{}^2}} \quad (7)$$

3. 실 험

3.1 실험 데이터

우리는 공통된 관심사를 갖는 그룹의 프로파일을 구축하기 위해 오직 한 분야에만 전문화된 교육을 제공하는 특성화 대학 [1]중 한곳의 전자 도서관 서비스를 선정하였다. 이는 인문 사회부터 이공계까지 많은 주제를 내포하여 공통된 주제를 파악하기 힘든 종합 대학과는 달리 특성화 대학의 사용자들은 보다 더 세분화되고 구체적인 주제를 바탕으로 서비스를 활용할 것이라는 가정 하에 수행되었다.

해당 대학의 구성원들은 전자 도서관에서 논문, 책, 리포트 등의 다양한 학술자료를 검색하여 열람하였는데 그 중 개인정보 보호를 위해 익명 처리된 한 달간의 도서관 이용 로그 정보를 실험에 활용하였다. 해당 기간 동안 3,478명의 사용자가 8,606개의 문서를 클릭하였으며, 총 클릭수는 10,164번이 발생하였다. 이는 문서당 평균 1.18번의 클릭이 있었음을 의미하지만, 실제로 사용자와 문서 각각의 클릭 횟수를 그림 1과 같이 분석하였을 때 실험 데이터셋은 거듭제곱 법칙의 경향을 만족하여 본 연구에서 가정하는 실제 서비스 환경에 잘 부합함을 의미한다.

아쉽게도, 실험에 사용된 로그 데이터는 사용자의 질의어 정보를 제공하지 않아 질의어와 문서간의 적합성은 해당 실험에서 고려하지 않았다. 하지만, 추후 질의어와 적합한 문서들 중 제안하는 기법에서 상위에 반환된 문서들을 보다 높은 순위로 재정렬하는 과정을 통해 실제 검색 서비스에 쉽게 적용할 수 있을 것이며, 본 연구에서는 개인화 및 그룹 기반 중 보다 더 효율적인 성능을 갖는 검색 기법이 무엇인지 파악하는 것에 초점을 맞춘다. 각 사용자의 클릭 데이터들은 클릭 시간에 따라 처음 90%는 학습 데이터로 마지막 10%는 평가 데이터로 사용되었다.

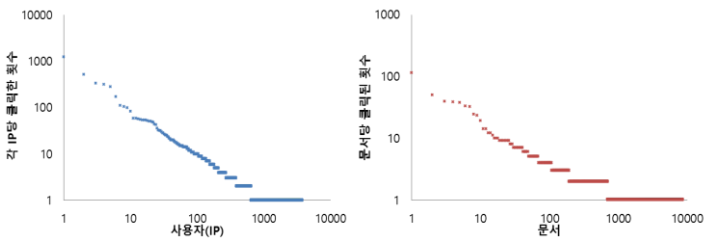

그림 1. 사용자 및 문서 별 클릭 횟수

3.2 평가 절차

평가는 다음의 순서로 진행되었다. 먼저 사용자 별로 학습 데이터를 사용해 프로파일을 구성하였다. 또한 전체 사용자의 학습 데이터를 그룹 데이터로 가정하여 그룹 프로파일을 생성하였다. 그리고 사용자와 문서간의 유사도를 비교하여 유사도가 높은 순부터 반환된 문서들 중 상위 50개 문서들을 추렸다. 마지막으로, 사용자의 평가 데이터안에 포함된 문서가 있는지, 있다면 그 문서가 얼마나 상위 순위에 있는지를 판단하는 척도를 바탕으로 평가를 진행하였다.

평가 데이터에 속한 문서를 정답셋으로 가정하여 평가를 진행하는 방법은 개인화 검색에 만족도를 평가하는 관련 연구에서 널리 사용되는 방식으로 [6], 본 연구에서는 Normalized Discounted Cumulative Gain (NDCG), 그리고 Precision @N이 사용되었다. 두 척도 모두 적합한 문서들이 높은 순위에 있을수록 높은 수치를 가진다. 또한 제안하는 기법을 개인화 검색 방법과 그룹 기반 검색 방법과 비교하였다. 실험을 위하여, 우리는 제안하는 개인화 및 그룹 기반 혼용 검색 기법에서 수식 (6)의 k 값을 변경시켜 얻은 최적 성능을 (그림 2 참조) 상기 두개의 검색 기법과 비교하였다.

3.2 실험 결과

표1의 실험 결과는 제안하는 방법이 두 척도 모두에 대해서 비교하는 기존의 방법들에 비해 더 뛰어난 성능을 가진 것을 보여준다. 우리는 또한 각각의 성능에 대한 통계 분석도 추가적으로 실시하였다. 기호 †는 개인화에 비하여 통계적으로 유의한 성능 향상, ‡은 개인화 및 그룹 기반에 비하여 통계적으로 유의한 성능 향상을 보였음을 의미한다 (Wilcoxon Test, $p<0.05$). 실험에 사용한 데이터의 양이 크지 않았기 때문에 전반적인 평가 데이터의 부재로 전체적인 성능은 낮지만, 그럼에도 불구하고 우리가 제안하는 혼용 기법의 성능이 기존의 방법들에 비해 훨씬 더 뛰어난 성능을 가진 것을 보여준다.

주목할 것은 일반적으로 그룹 기반보다 개인화 검색의 성능이 높음에도 [3] 본 실험에서는 그 반대의 결과가 나타났다는 점이다. 이는 실험에 사용한 데이터가 매우 큰 거듭제곱의 법칙을 따르기 때문에 개인화 프로파일을 사용하는 경우 거의 대부분의 사용자들이 부정확한 프로파일을 갖게 되어 오히려 그룹 기반 검색 기법보다 개인화 검색 기법의 성능이 하락함을 의미한다. 이에 반하여, 제안하는 혼용 기법은 분석할 데이터가 충분한 소수의 사용자들에게는 개인화 검색을, 그 외의 다수의 사용자에게는 그룹 기반의 검색을 제공함으로써, 보다 더 안정적인 검색 성능의 향상을 보임을 알 수 있다.

또한 그림 2는 제안하는 방법에서 적절한 k값을 선정하기 위해 분석한 데이터로 k값이 약 20 이상이 되면 가장 최적의 성능을 보이는 형태로 수렴함을 보여준다. 이는, 사용자가 적어도 20개정도의 문서를 클릭할 때까지는 그룹 기반의 프로파일을 사용하는 것이 좋은 성능을 보이는 것에 좋으며, 20개 이상의 문서를 클릭한 사용자에게는 그룹 프로파일 대신, 개인 프로파일을 구축하여 제공하여, 최종 개인화 검색 성능을 향상시킬 수 있음을 보여준다.

표 1. 각 검색 기법에 따른 성능 변화

	개인화	그룹 기반	개인화 및 그룹 혼용
NDCG	0.089	†0.115	‡0.138
P@1	0.080	†0.110	‡0.140
P@5	0.033	†0.072	‡0.088
P@10	0.002	0.003	†0.006

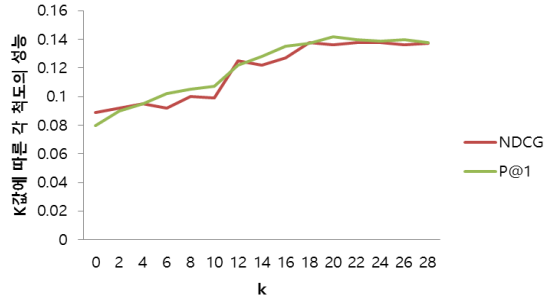

그림 2. k값 변화에 따른 NDCG 및 P@1 값의 변화

4. 결론

본 논문에서는 기존 개인화 검색 기법 및 그룹 기반 검색 기법의 단점을 보완하기 위해 개인화 및 그룹 기반을 혼합한 검색 기법을 적용하는 방법을 제안하였다. 제안하는 그룹 기반 및 개인화 기법은 기존의 기법과 비교하였을 때 더 뛰어난 성능을 보였다. 실험에 사용한 적은 양의 데이터셋 및 제한된 그룹 정보에도 불구하고, 제안하는 알고리즘은 실제 서비스에서 활발한 사용자는 개인화 서비스를, 활발하지 않은 사용자에게는 그룹 기반 서비스를 제공함을 통하여 검색 성능이 향상될 수 있음을 보였다. 향후 연구에서는 보다 더 큰 실제 데이터상에서 제안하는 기법의 성능을 평가하는 것이 필요하다.

참고 문헌

[1] N. J. Belkin, “Some(what) grand challenges for information retrieval,” *SIGIR Forum*, 42(1), 2008.

[2] P. A. Chirita, C. S. Firan, and W. Nejdl, “Personalized query expansion for the Web,” In *Proc. of SIGIR*, pp. 7-14, 2007.

[3] Z. Dou, R. Song, and J. R. Wen, “A Large-scale evaluation and analysis of personalized search strategies,” In *Proc. of CIKM*, pp. 969-978, 2010.

[4] J. Teevan, S. T. Dumais, and E. Horvitz, “Personalizing search via automated analysis of interests and activies, In *Proc. of SIGIR*, pp. 449-456, 2005.

[5] J. Teevan, M. R. Morris, and M. Bush, “Discovering and using groups to improve personalized search, In *Proc. of WSDM*, pp. 15-24, 2009.

[6] S. Xu, S. Bao, B. Fei, Z. Su, and Y, Yu, “Exploring folksonomy for personalized search,“ In *Proc. of SIGIR*, pp. 155-162, 2008.