개인화 정보 검색에 대한 연구

A Study of Personalized Information Retrieval

김태환, Taehwan Kim*, 전호철, Hochul Jeon**, 최중민, Joongmin Choi***

요약 사람들은 월드 와이드 웹(World Wide Web)상에서 사용자가 원하는 정보를 검색하는 여러 알고리즘들을 구현해 왔다. 이렇게 구현된 검색 알고리즘 중 가장 좋은 기술을 가지고 있는 곳은 페이지랭크(PageRank)방식의 구글이다. 하지만 페이지랭크 방식, 즉 외부에서 참조하는 링크가 많은 문서로 검색하여 가장 많은 링크를 가지고 있는 문서를 상위에 보여주는 방식으론 사용자가 원하는 문서를 찾기 힘들다. 개인에게 가치가 있는 문서를 찾기보다 대중에게 가치가 있는 문서를 찾기 때문이다. 이러한 문제를 해결하기 위하여 본 논문에서는 대중적 가치와 개인적 가치를 혼합한 개인화 검색 엔진을 제안한다.

Abstract Many search algorithms have been implemented by many researchers on the world wide web. One of the best algorithms is Google using PageRank technology. PageRank approach, computes the number of inlink of each documents then represents documets in order of many inlink. But It is difficult to find the results that user needs. Because this method finds documents not valueable for a person but valueable for public. this paper propose a personalized search engine mixed public with personal worth to solve this problem

핵심어: Personalized, Information Retrieval, 개인화, 정보검색

*주저자 : 한양대학교 컴퓨터 공학과 박사과정; kimth@cse.hanyang.ac.kr
**공동저자 : 한양대학교 컴퓨터 공학과 박사과정; hcjeon@cse.hanyang.ac.kr
***교신저자 : 한양대학교 컴퓨터 공학과 교수; e-mail: jmchoi@hanyang.ac.kr

1. 서론

월드 와이드 웹(World Wide Web)의 사용자가 폭발적으로 증가함에 따라 대량의 정보가 개인, 기업 홍보 또는 상업적인 목적으로 생성되고 있다. 현재 국내에는 수백 만개 수준의 웹 문서가, 세계적으로는 억 단위 수준의 웹 문서가 산재해 있으며 그 수는 빠르게 증가하고 있다[1]. 이와 같이 방대한 정보를 하이퍼링크(Hyperlink)만을 이용해서 사용자의 정보 욕구를 충족시키기란 불가능하다. 현재 야후, 알타비스타, 구글 등 다양한 정보 검색 엔진이 개발됨으로써 사용자의 정보 욕구를 어느 정도 충족시키고 있으나 아직도 원하는 정보를 찾기가 쉽지 않은 실정이다.

이러한 문제를 해결하기 위하여 추천 방법 즉 개인화에 대한 연구가 진행되었다. 이전 개인화 방법에는 내용기반(content-based) 방법과 협업(collaborative)방법이 있다. 내용기반 방법은 사용자의 관심사를 표현한 프로파일의 내용과 필터링 대상 항목의 내용을 비교하여 사용자에게 흥미로운 또는 유익한 항목들을 선택하는 방법이다. 이 방법은 텍스트 형태의 항목을 다루는 데 아주 유용한 것으로 알려져 있으며 불리안 모델, 벡터공간 모델, 확률 모델, 뉴럴 네트워크 모델, 퍼지집합 모델 등에 기초한 방법들이 있다. 그러나 사용자 특히 초보자는 자신의 원하는 정보를 정확히 표현하기가 어렵다. 즉, 내용기반 필터링의 기본이 되는 사용자 프로파일 구성에 어려움이 있다.

협력 필터링은 타 사용자의 관심사를 예측하는데 동일한 생각을 갖는 사람들의 의견을 이용하는 방법이다. 이방법은 이력(history) 데이터 베이스를 조사하여 대상 사 용자와 유사한 관심사를 갖는 사용자들을 찾고 이들이 대상 항목에 대해 어떻게 평가했는지에 대한 정보를 이용하여 대상 항목을 필터링하는 방법이다. 제록스 팔로 알토 연구소에서 개발된 Tapestry 텍스트 필터링 시스템[3,4]과미네소타 대학의 Group Lens 시스템[5]들이 대표적 협력 시스템이다. 협력 필터링 방법은 다양한 분야에서 활용되고 있는데 Ringo 시스템[6]에서는 음악 앨범을 추천하는데 사용하고 있으며, MovieLens 시스템[7]에서는 영화를, Jeter 시스템[8]에서는 유머를, Flycasting[9]에서는 온라인 라디오를 추천하는데 사용하고 있다.

본 논문에서는 현재 구현된 검색 알고리즘 중 대표 알고리즘인 페이지랭크 즉 대중적 가치를 이용한 검색 엔진과 사용자가 이전에 사용한 검색어 즉 개인적 가치를 혼합한 개인화 검색 엔진을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제시하고 있는 시스템의 구조와 개인화 검색 방법에 대해 설명한다. 3장에서는 실험 결과를 통하여 기존의 방법과 본 논문에서 제시한 방법을 비교 평가하고 4장에서 향후 연구를 기술하고 결론을 내린다.

2. 관련연구

본 장에서는 관련 연구로서 개인화 방법과 링크분석을 이용한 순위 결정 알고리즘인 페이지랭크 알고리즘에 대해 설명한다.

2.1 내용기반 필터링(content-based filtering)

내용 기반 필터링은 콘텐츠의 속성들에 대한 정보를 시스템이 보유하고 있다가 고객이 그 속성에 해당하는 키워드를 입력할 때 그 키워드에 해당하는 속성을 지닌 콘텐츠들을 추천해주는 것이다. 내용 기반 필터링의 단점으로는 첫째가 콘텐츠의 정보가 풍부하다면 추천의 결과가 좋지만 그렇지 않으면 추천의 결과의 신뢰성이 떨어진다는 점이고 둘째는 추천되는 콘텐츠가 사용자 자신이 가진 취향에서 벗어나지 못하는 특수화 경향을 보이기 쉽다는 것이다. 마지막으로 사용자의 의사 표현을 얼마나 많이 하느냐에 따라 추천의 질이 달라진다.

2.2 협력기반 필터링(collaborative filtering)

협력적 필터링은 사용자의 선호도에 대한 데이터를 기반으로 새로운 사용자가 관심을 가질 것으로 생각되는 아이템(상품, 광고, 웹 페이지 등)을 추천해 주는 기법이다. 협력적 필터링은 아이템에 대한 다른 사용자들의 선호를 기반으로 하기 때문에 협력적이라는 용어를 사용하게 된다. 협력적 필터링 시스템은 시스템이 사용자의 암묵적인 데이터를 사용하는지 명시적인 데이터를 사용하는지에 따라 구분을 한다. 명시적인 데이터는 사용자에게 특정 아이템에 대한 선호를 0에서 5까지 정도의 이산형 척도로 입력받는 경우를 말한다. 암묵적 데이터를 사용하는 경우는 사용자의 선호를 대변하는 사용자의 웹 사이트 클릭 패턴이나 구매 패턴 등을 웹 로그나 구매 이력 데이터에서 발견하여 특정 아이템에 대한 선호를 예측하는 것을 말한다. 따라서 협력적 필터링을 사용하면 위의 내용 기반 필터링의 문제점을 보완할 수 있다. 하지만, 연구와 실무분야에서 모두 매우 성공적으로 평가되어 온 협력적 필터링 기법을 활용한 추천 시스템에 있어서는 기본적으로 자신의 취향과 비슷한 사람이 적을 경우 추천의 질이 떨어지고 추천의 대상이 새로 주어지는 경우 이에 대해 누군가가 평가를 하기 전에는 추천이 이루어지지 않는 문제점을 가지고 있다.

2.3 페이지랭크 알고리즘(pagerank algorithm)

1998년 Bring과 Page에 의해 제안된 페이지랭크 알고리즘의 기본 아이디어는 다음과 같다. '문서 u 가 v 를 가리킨다면, u 의 저자는 문서 v 에게 묵시적으로 중요도를 준다'[13]. 즉, 문서 u 의 중요도를 $Rank(u)$, N_u 는 u 의

아웃링크 수(out-link-degree)라 하면, $Link(u,v)$ 는 문서 v 에게 $Rank(u)/N_u$ 만큼의 중요도를 준다. 모든 문서의 수를 n 이라 할 때, 모든 문서에 대한 중요도의 초기값을 $1/n$ 으로 설정하고, 계산을 반복하면 모든 문서의 중요도는 수렴하게 된다. 각 단계에서의 중요도 계산식은 다음과 같다.

$$\forall_V Rank^{(k+1)}(v) = \sum_{u \in B_v} Rank^{(i)}(u)/N_u \quad (1)$$

식 1을 이용하여 모든 문서에 대한 중요도를 반복적으로 계산함으로써 페이지랭크 값을 얻을 수 있다. 또한, 페이지랭크 값을 계산하는 과정은 아이겐 백터(eigen vector)를 계산하는 것으로 표현될 수 있다. M 을 아웃링크 수에 대하여 1로써 평준화(normalization)를 한 인접 행렬(adjacent matrix) 일 때, 식 1은 다음과 같은 행렬식으로 표현 할 수 있다[13].

$$\overrightarrow{Rank} = M^T \times \overrightarrow{Rank} \quad (2)$$

즉, 식 2를 반복적으로 계산하면, 주어진 행렬 M^T 에 대하여 아이겐 값 1을 가지는 아이겐 벡터 $\overrightarrow{Rank}$ 를 계산하는 과정이 된다. 또한, M^T 를 상태변화 확률 행렬이라고 가정하면, 식 2는 웹 그래프에서의 '랜덤 워크(random walk)'에 의한 모델로서도 해석 할 수 있다.

또한, 페이지랭크 값이 수렴하기 위해서는 M이 불규칙 성질을 포함하고 있어야 한다[10]. 불규칙 성질을 만족하기 위하여 식 1과 2를 다음과 같이 수정 할 수 있다.

$$\forall_V Rank^{(k+1)}(v) = d \bullet \sum_{u \in B_v} Rank^{(i)}(u)/N_u + (1-d) \bullet 1/n \quad (3)$$

$$\overrightarrow{Rank} = d \bullet M^T \times \overrightarrow{Rank} + (1-d) \bullet \vec{e}/n \quad (4)$$

3. 시스템 구조와 개인화 검색 방법

3.1 시스템 구조

본 논문에서 사용하고 있는 시스템의 구조는 그림 1과 같은 구조로 이루어져 있다.

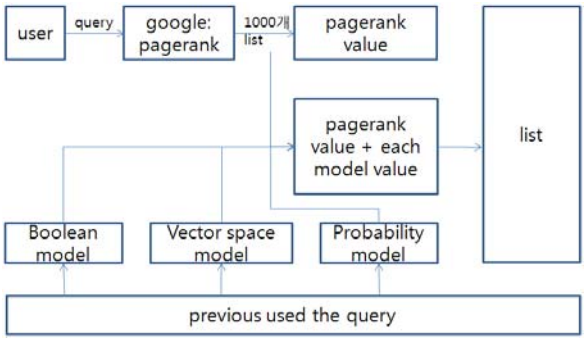

그림 22 시스템 구조

시스템 구조는 크게 3가지로 분류할 수 있다. 첫 번째는 사용자의 질의에 의해 페이지랭크 값을 구하는 부분이며 두 번째는 이전에 사용된 사용자의 질의를 이용하여 불리안 모델, 벡터 공간 모델 또는 확률 모델의 값을 구한다. 세 번째는 이전에 나열되었던 목록을 위의 페이지 랭크 값과 불리안 모델 이나 벡터 공간 모델 또는 확률 모델의 값을 구하여 목록을 다시 나열한다.

3.2 개인화 검색 방법

사용자가 문서 검색할 때 사용자의 질의를 이용해서 검색하게 되는데 , 각각의 질의는 사용자가 원하는 정보를 검색하기 위한 용어이기 때문에 사용자가 관심을 가지는 분야라 할 수 있다. 사용자가 새로운 질의를 할 때 검색된 문서에 사용자가 이전에 검색하기 위해 사용했던 용어들을 문서상에 가중치를 둔다. 또한 각각의 문서에 질의의 용어와 이전에 사용했던 용어들의 가중치를 곱하여 가장 높은 점수를 가진 문서를 검색해 준다.

새로운 질의를 할 때 검색 된 문서를 구하는 페이지랭크 값은 식 1와 같다.

$$PR(A) = 1 - d\left(1 - \frac{RP(t_1)}{C(t_1)} - \frac{RP(t_2)}{C(t_2)} - \ldots - \frac{RP(t_N)}{C(t_N)}\right) \quad (5)$$

이전에 검색하기 위해 사용했던 용어들을 문서상에 가중치를 두는 방법은 식2, 식3과같다.

$$sim(d_j, q) = \frac{\overrightarrow{d_j} \bullet \vec{q}}{\left|\overrightarrow{d_j}\right| \times \left|\vec{q}\right|} \quad (6)$$

$$sim(d_j, q) = \frac{P(\overrightarrow{d_j} \mid \overline{R}) \times P(R)}{P(\overrightarrow{d_j} \mid \overline{R}) \times P(\overline{R})} \quad (7)$$

식 1와 식2 또는 식 1와 식 3을 결합하여 질의와 유사한 문서의 값을 구해 높은 값을 가지는 문서를 상위에 링크시켜준다.

$$sim(d_j, q) = v(PR(A)) + (1-v)\frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \quad (8)$$

$$sim(d_j, q) = v(PR(A)) + (1-v)\frac{P(\vec{d_j} \mid \overline{R}) \times P(R)}{P(\vec{d_j} \mid \overline{R}) \times P(\overline{R})} \quad (9)$$

4. 실험결과

인터넷 정보 검색에서 중요한 문제는 검색 결과의 질과 관련된 것으로 검색 결과에 순위를 정하는 문제이다. 일반적으로 검색을 수행하여 나온 검색 결과로 나오는 문서 조차 도저히 다 읽어 볼 수 없을 정도로 많기 때문이다. 사용자들이 검색 엔진을 사용하는 경향을 살펴보면 모든 검색 결과를 참조하는 것이 아니고 상위에 랭크된 몇 개의 문서만을 보는 경향이 있다. 따라서 검색 엔진의 성능 평가의 주요 척도는 검색 결과의 상위 부분에 사용자의 요구와 일치하는 문서의 수에 초점이 맞춰져야 한다. 즉 재현도(Recal)보다는 정확도(Precision)가 검색 엔진의 성능에 더 중요한 요소이다. 이를 근거로 본 논문에서의 실험은 정확도에 따라 페이지랭크와 제안하는 방법을 비교한다.

본 논문이 제안하여 결정된 순위는 페이지랭크를 중요도로 한 결과와 비교 하였다. 비교 방법은 사용자 질의에 대하여 검색된 문서 50개를 사용자가 의도한 문서가 상위에 나오는지를 비교하였다.

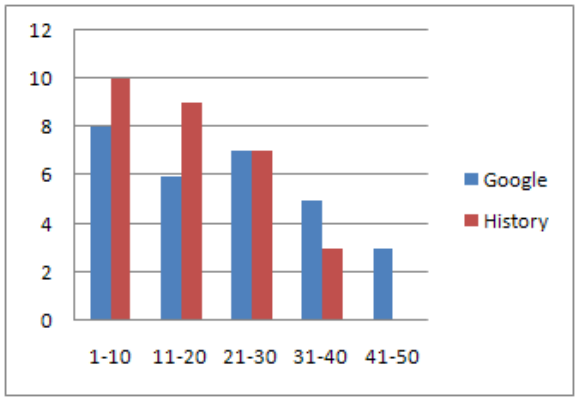

그림 2 First 50 pages (google and history re-rank)

검색된 50개의 페이지 중 29개의 문서가 사용자가 원하는 문서이다. 그림 2는 사용자가 원하는 29개의 문서가 얼마나 상위에 나타내는지 구글(google)과 이전에 사용된 사용자의 질의(history)를 비교한 실험이다.

5. 결론

개인화 정보 검색에 가장 큰 문제점인 다음의 문제점들을 해결하려 했다.

개인화의 문제점은 다음과 같다.

첫 째는 개인의 정보가 풍부하다면 결과가 좋지만 그렇지 않으면 결과의 신뢰성이 떨어진다.

둘 째는 개인화에 의해 검색되는 문서는 사용자가 가진 취향에서 벗어나지 못하는 경향이 나타난다.

셋 째는 개인화의 가장 큰문제점은 자신의 정보를 제공하거나 수집을 허락하는 것에 대한 사용자들의 반감이다.

이러한 문제점들을 본 논문에서는 다음과 같은 방법으로 해결하려 했고 일정부분 해결 하였다.

첫 째는 기본적인 정보 검색을 토대로 개인의 검색 용어를 이용 함으로써 개인의 정보가 없더라도 이전과 다름없는 결과를 도출해 준다.

둘 째는 제안하는 방법이 대중적 가치와 개인적 가치 모두를 고려하기 때문에 개인적 가치가 적고 대중적 가치가 큰 경우 사용자가 가진 취향에 국한 되어서 나타나던 결과를 피할 수 있다.

셋 째는 사용자를 구분할 수 있는 이 메일 주소와 그 사용자가 이전에 사용했던 용어만을 데이터로 수집하기 때문에 개인의 정보를 수집 이용하는데 한계가 있다.

참고문헌

[1] SouMen Charkrabati. "mining the web Discovering Knowledge from Hypertext Data", Morgan Kaufmann Pubolishers. 2003.

[2] 정경용, 김진현, 정헌만, 이정현, "개인화 추천시스템에서 연관 관계 군집에 의한 아이템 기반의 협력적 필터링 기술" , 정보과학회논문지 : 소프트 웨어 및 응용, 제31권, 제4호, pp. 467-477, 2004.

[3] Douglas B. Terry, "A tour through tapestry," Proc. Of the ACM Conference on Organizational Computing Systems(COOCS), pp. 21-30, 1993.

[4] Donna Harman., "Overview of the third Text Retrieval Conference(TREC-3)," D. K. Harman, editor, Overview of the Third Text Retrieval Conference(TREC-3), pp. 1-19, 1994.

[5] Resnick, P., lacovou, N., Suchak, M., Bergstorm, P. and Riedl, J., "GroupLens: An open architecture for collaborative filtering of Netnews," Proc. Of ACM Conf. on Computer-Supported Cooperative Work, pp. 175-186, 1994.

[6] Upendra S. and Patti M.. "Social Information Filtering : Algorithms for Automating "Word of Mouth"," Proc. Of ACM CHI' 95 Conference on Human Factors in Computing Systems, pp. 210-217, 1995.

[7] http://www.cs.umn.edu/research/GroupLens/.

[8] D. Gupta, M. Digivanni, H. Narita, and K. Goldberg, "Jester 2.0: A New Linear-Time Collaborative Filtering Algorithm Applied to Jokes," Proc. Of Workshop on Recommender System : Algorithms and Evaluation, Aug. 1999.

[9] Hauver, D. B. and French,J. C, "Flycasting: Using Collaborative Filtering to Generate a Play list for Online Radio," Proc. Of Int. Conf. on Web Delivery of Music, 2001.

[10] Mooney, Raymond J., Roy, Loriene, "Content-based book recommending using learning for text categorization Proceedings of the ACM International Conference on Digital Libraries,", 2000.

[11] Good, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B., Herlocker, J., Riedl, J., "Combining collaborative filtering with personal agents for better recommendations", Proceedings of the AAAI- '99 Conference, 1999.

[12] Sergey Brin, Lawrence Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proceeding 7th International World Wide Web Conference, Computer Networks and ISDN Systems 30. 1998. pp.1007-117

[13] Taher H. Haveliwala. "Efficient Computation of PageRank", Technical Report, Stanford University. 1999.