블로그 공간에서 링크 기반 유사도를 이용한 게시글 추천

송석순, 윤석호, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mails: {gagler, bogely, wook}@agape.hanyang.ac.kr

Post Recommendation Using Link-based Similarity in Blogosphere

Suk-Soon Song, Seok-Ho Yoon, Sang-Wook Kim
Department of Electronics and Computer Engineering, Hanyang University

요 약

본 논문에서는 링크 기반 유사도 계산을 이용해서 블로그 공간에서 사용자가 관심을 가질만한 게시글들을 사용자에게 추천하는 방안을 제안한다. 제안된 방안은 사용자가 관심을 가졌던 게시글들 중에서 시드 게시글을 선택하고 링크 기반 유사도를 계산하여 시드 게시글과 가장 유사하다고 판단되는 k개의 게시글들을 사용자에게 추천한다. 또한, 시드 게시글들 중에서 추천하고자 하는 주제가 아닌 다른 주제의 게시글들이 잘못 추천되는 문제를 해결하기 위해서 시드 게시글과 동일한 주제라고 확실시되는 게시글들만을 점진적으로 찾아 추천하는 방안을 제안한다. 실제 블로그 데이터를 이용한 실험을 통하여 제안하는 추천 방안의 우수성을 검증한다.

1. 서론

추천 시스템(recommender system)은 자동화된 정보 필터링 기술을 이용하여 사용자의 취향에 맞는 상품이나 정보 등을 찾아 사용자에게 추천하는 시스템이다[1]. 추천 시스템의 핵심 기술 중 하나는 사용자의 선호도를 분석하여 사용자가 원하는 적절한 상품이나 정보를 찾는 정보 필터링 기술이다. 대표적인 정보 필터링 기술로는 협업 필터링(collaborative filtering)이 있으며 아마존닷컴과 같은 유명한 전자상거래 사이트들에서 많이 사용된다[2].

협업 필터링에는 사용자 기반 협업 필터링과 아이템 기반 협업 필터링이 존재한다[3]. 사용자 기반 협업 필터링은 추천 받고자하는 사용자와 유사한 사용자 집단을 찾고 해당 사용자 집단이 선호했던 상품을 찾아 사용자에게 추천하는 방법이다. 아이템 기반 협업 필터링은 사용자가 선호했던 상품과 유사한 상품을 찾아 사용자에게 추천하는 방법이다. 두 방법 모두 사용자가 상품에 매겼던 선호도를 기반으로 유사한 사용자 집단과 유사한 상품 집단을 찾아 추천 시스템에 활용한다. 그림 1은 협업 필터링의 예이다. 오른쪽 A, B, C는 유사 사용자 집단에 속한 사용자들을 의미한다. 사용자 A에게 자신과 유사한 사용자들이 선호했던 4번 상품을 추천할 수 있다. 또한 왼쪽의 1, 2, 3, 4는 유사 상품 집단에 속한 상품들을 나타낸다. 따라서 사용자 B에게 사용자 B가 선호했던 1, 4번 상품과 유사한 2, 3번 상품을 추천할 수 있다.

최근 들어 블로그 공간을 이용하는 사용자들이 증가하면서 블로그 공간에 추천 시스템을 도입하는 것에 대한 필요성이 증가하고 있다. 이는 기존 불특정 다수가 아닌 사용자 개개인에게 특화된 추천 시스템이 블로그 공간의 활성화 및 편의성을 증대시킬 수 있기 때문이다. 따라서 본 논문에서는 블로그 공간에서 특정 주제에 관심을 가진 사용자에게 자신이 관심을 가졌던 해당 주제의 게시글들과 유사한 주제의 좋은 게시글들을 추천하는 방안에 대하여 다루고자 한다.

기존의 협업 필터링 방법을 블로그 공간에 적용시키기 위해서는 먼저, 사용자가 상품들에 대해서 선호도를 매겼던 것처럼 사용자가 게시글에 대해서 선호도를 매겨야 한다. 이러한 선호도는 유사한 사용자 집단 또는 유사한 게시글 집단을 찾는데 사용된다. 그러나 블로그 공간에서는 게시글에 대한 선호도가 명시적으로 존재하지 않는다.

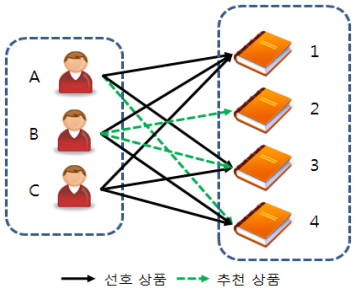


그림 1. 협업 필터링의 예.

반면, 블로그 공간에서는 사용자들이 관심 있는 게시글에 대해 스크랩, 덧글, 엮인글 등의 액션을 취할 수 있다. 이러한 액션은 게시글에 대한 사용자의 관심을 나타낸다. 따라서 사용자가 게시글에 취한 액션은 유사한 사용자 집단과 유사한 게시글 집단을 찾는데 선호도 대신 활용 가능하다. 따라서 본 논문에서는 게시글에 존재하는 액션들을 이용하여 해당 사용자가 주로 관심을 가지는 주제의 게시글과 유사한 주제의 게시글 집단을 찾고 이러한 게시글들을 사용자에게 추천하는 방안을 제안한다.

2. 링크 기반 유사도를 이용한 추천 방안

2.1. 개관

블로그 공간에서 블로그와 게시글을 객체로 보고 블로그가 관심 있는 게시글에 가한 액션을 링크로 정의하면 블로그 공간을 이분 그래프(bipartite graph)로 모델링할 수 있다. 이 결과, 액션을 기반으로 주어진 게시글과 유사한 게시글을 찾는 문제를 링크를 기반으로 주어진 객체와 유사한 객체들을 찾는 문제로 변환시킬 수 있다.

링크를 기반으로 객체들 간의 유사도를 계산하는 것을 링크 기반 유사도 계산이라고 한다. 기존의 링크 기반 유사도 계산 방법에는 Co-Citation[4], Bibliographic Coupling[5], SimRank[6], LinkClus[7] 등이 있다. 본 논문에서는 기존 방법들 중에서 정확도와 성능 면에서 가장 우수하다고 알려진 LinkClus를 이용하여 게시글들 간의 유사도를 계산한다. 또한 계산된 게시글들 간의 유사도 결과를 이용하여 사용자가 관심을 가지는 게시글들과 유사한 게시글들을 찾아 주어진 사용자에게 추천한다.

2.2. 기본 방안

사용자에게 게시글을 추천하기 위해서는 먼저 해당 사용자가 관심을 가져서 액션을 가했던 게시글들 중에서 일부의 게시글들을 시드 게시글로 선택한다. 시드 게시글 선택 기준은 사용자가 액션을 가했던 게시글들 중에서 액션 수가 일정 이상 것을 포함하여 다양한 기준을 사용할 수 있다.

그 다음, 각 게시글에 대하여 시드 게시글들과의 링크 기반 유사도를 계산한다. 이는 해당 게시글과 각 시드 게시글과의 링크 기반 유사도의 평균을 이용한다. 시드 게시글들과의 유사도가 가장 높은 k개의 게시글들을 사용자에게 추천한다.

2.3. 시드 확장 방안

기본 방안은 주어진 시드 게시글들과 모든 게시글들 간의 유사도를 단번에 계산하기 때문에 시드 게시글들을 얼마나 잘 선택하느냐에 큰 영향을 받는다. 만약 시드 게시글들 중에 다른 주제의 게시글이 일부 섞여 있다면 기본 방안은 사용자가 관심이 없는 게시글들을 추천할 가능성이 있다. 따라서 이러한 문제를 해결하기 위해서 본 논문에서는 시드 확장 방안을 제안한다.

일반적으로 시드 게시글들 중에 주제가 다른 게시글이 일부 섞여 있다고 하더라도 게시글들 간의 유사도는 시드 게시글들과의 평균 유사도로 계산하기 때문에 시드 게시글들과의 유사도가 높은 게시글들 일수록 시드 게시글들과 동일한 주제일 확률이 높다. 따라서 시드 확장 방안은 주어진 시드 게시글들과 모든 게시글들 간의 유사도를 계산해서 시드 게시글들과 가장 유사한 n(n<<k)개의 게시글들을 초기 시드 게시글 집합에 추가시킨다. 그리고 다시 확장된 시드 게시글들을 대상으로 모든 게시글들과의 유사도를 계산하고 시드 게시글 집합에 추가하는 절차를 반복한다. 추가된 시드 게시글들의 수가 k개가 되면 해당 절차를 중단하고 k개의 게시글들을 사용자에게 추천한다.

3. 실험

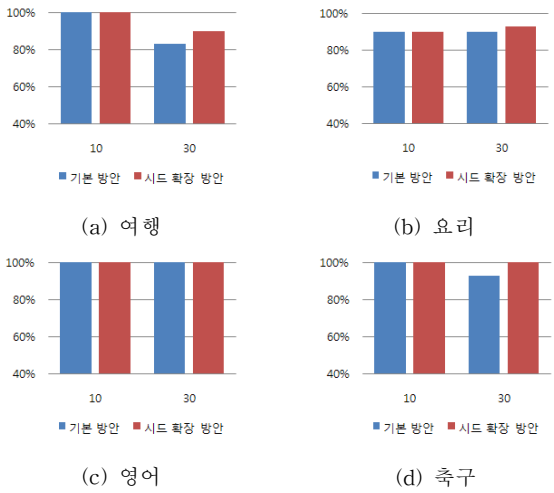
본 논문에서는 실험을 위해 국내 블로그 서비스 중 하나인 네이버 블로그에서 2006년 4월부터 수개월간 수집하여 익명으로 처리한 데이터를 사용하였다.

제안하는 추천 방안을 검증하기 위해서 서로 다른 주제의 시드 게시글들이 주어졌을 때 제안하는 추천 방안을 통해서 추천되는 게시글들이 시드 게시글들과 얼마나 동일한 주제인지를 검사하였다. 시드 게시글들의 주제는 블로그 공간에서 가장 인기가 많은 여행, 요리, 축구, 영어 총 4가지를 사용하였으며, 최종 추천 게시글들의 수는 10, 30으로 설정하였다. 추천되는 게시글들 중 주어진 시드 게시글들과 동일한 주제를 갖는 게시글들의 비율을 정확도로 정의하였다. 동일 주제 여부 판정과 시드 게시글의 선택은 도메인 전문가가 직접 수행하였다.

그림 2는 제안하는 추천 방안의 주제별 정확도를 나타낸다. 모든 주제에 대해 최종 추천 게시글들의 수를 10으로 하였을 때 기본 방안과 시드 확장 방안은 평균 97.5%의 동일한 정확도를 보였다. 그러나 최종 추천 게시글들의 수를 30으로 하였을 때 기본 방안의 정확도가 평균 91.5%인 반면 시드 확장 방안은 기본 방안보다 높은 약 95.75%의 정확도를 보이는 것으로 나타났다.

4. 결론

본 논문에서는 기존 링크 기반 유사도 계산 방안을 이용해서 블로그 공간에서 사용자가 관심을 가질만한 게시글을 추천하는 두 가지 방안을 제안하였다. 먼저, 기본 방안은 사용자가 관심을 가졌던 게시글들 중에서 시드 게시글들을



(a) 여행 (b) 요리

(c) 영어 (d) 축구

그림2. 주제별 추천 게시글들의 정확도.

선택하고 링크 기반 유사도 계산 방안을 통해서 계산된 게시글들 간의 유사도를 이용하여 시드 게시글과 가장 유사한 k개의 게시글들을 사용자에게 추천한다. 또한, 시드 게시글들 중에서 추천하고자 하는 주제가 아닌 다른 주제의 시드 게시글이 섞여 있을 경우, 다른 주제의 게시글들이 추천되는 문제를 해결하기 위해서 시드 게시글과 동일한 주제라고 확신되는 게시글들을 점진적으로 찾아 사용자에게 추천하는 확장 방법을 제안했다. 실험을 통하여 제안하는 추천 방안이 사용자가 관심을 가졌던 게시글들과 동일한 주제의 게시글들을 추천하는가를 검증하였다.

감사의 글

본 연구는 NHN(주)의 지원을 받았습니다. 그러나, 본 논문에서 제시된 의견이나 결론, 또는 권고 등은 온전히 저자(들)의 것이며, 반드시 지원회사의 입장을 대변하는 것은 아닙니다.

참고문헌

[1] P. Resnick and H. Varian, "Recommender Systems," *Journal of the Communication of the ACM* Vol. 40, No. 3, pp. 56-58, 1997.

[2] J. Herlocker et al, "Evaluating Collaborative Filtering Recommender Systems," *Journal of the ACM Transactions on Information Systems*, Vol. 22, No. 1, pp. 5-53, 2004.

[3] B. Sarwar et al, "Item-based Collaborative Filtering Recommendation Algorithms," In *Proc. Int'l. Conf. on on World Wide Web,* pp. 285-295, 2001.

[4] H. Small, "Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents," *Journal of the American Society for Information Science*, Vol. 24, No. 4, pp. 265-269, 1973.

[5] M. Kessler, "Bibliographic Coupling Between Scientific Papers," *Journal of the American Documentation*, Vol. 14, No. 1, pp. 10-25, 1963.

[6] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity," In *Proc. Int'l. Conf. on Special Interest Group on Knowledge Discovery and Data*, pp. 538-543, 2002.

[7] X. Yin, J. Han, and P. Yu, "LinkClus: Efficient Clustering via Heterogeneous Semantic Links," In Proc. *Int'l. Conf. on Very Large Data Bases*, pp. 427-438, 2006.