K-최근접 이웃 추천 엔진에서의 벡터 유사도 사용에 대한 실험적 분석

김혜재*, 손기락

한국외국어대학교 컴퓨터공학과

Empirical Analysis of K-Nearest Neighbor Recommendation Engine using Vector Similarity

Hye-Jae Kim, Kirack Sohn

Dept. of Computer Science & Engineering, Hankuk University of Foreign Studies

요약

인터넷 사용 인구의 폭증으로 인터넷 사이트가 경쟁적으로 유용한 각종 정보를 사용자들에게 제공하여 보다 많은 수의 회원을 확보하기 위해 노력하고 있지만 여러 사이트를 동시에 사용하고 있는 대부분의 인터넷 사용자들에게는 각 사이트에서 날아드는 정보를 매번 일일이 검색해야 하는 일이 여간 번거롭지 않을 뿐만 아니라 이런 무분별하고 획일적인 정보 서비스는 오히려 사용자들의 인터넷 사용을 불편하게 하며 더욱이 그 내용이 관심 밖의 것일 경우 네트워크의 효율적인 사용을 저해하는 정보공해에 지나지 않게 된다. 추천엔진은 기본적으로 끊임없이 유입되는 다량의 정보 중에서 필요한 것을 추천해 주는 것이다. 이에 본 논문에서는 사용자들에게 필요한 정보만을 효율적으로 전달 해주기 위해서 먼저 개인화된 정보의 전달을 위해 사용자의 취향을 파악하여 선택 가능성이 높은 항목을 예측할 수 있어야 한다. 그리고 사용자와 가까운 K 명의 사용자들을 효율적으로 검색하기 위해서 K-최근접 이웃 방식을 사용하고 인덱싱을 사용할 수 있는 세가지 벡터 유사도를 기존의 피어슨 상관계수(Pearson Correlation)와 비교하여 제안한다. 이를 통해 정보의 효율적인 제공방법, 즉 일반적인 검색으로 인한 정보의 제공이 아닌 일반 사용자들의 추천에 의해 정보를 제공하는 K-최근접 이웃 추천 엔진을 세가지 벡터 유사도를 이용해서 분석한다.

1. 서론

최근 인터넷 등의 발달로 온라인 상에서 검색을 통하여 활용할 수 있는 정보의 양이 많아짐에 따라 정보의 홍수 문제가 대두되고 있다. 이러한 문제를 해결하기 위하여 WWW 에서의 정보검색을 도와주는 웹검색엔진들이 개발되고 있다. 사람이 직접 웹사이트를 등록해서 디렉토리서비스를 하는 것도 있으며, 여러 검색엔진들로부터 결과를 받아오는 메타검색엔진 등도 개발되었다.

이러한 검색 엔진들은 자체적으로 웹에 있는 자료들에 대한 인덱스를 구축하고 나중에 사용자의 기호에 가장 가까운 항목들을 이 인덱스를 이용하여 검색하여 제공한다. 하지만 검색의 결과로 제공되는 항목의 수가 너무 많고 또 원하는 정보를 위해서 항목들을 모두 살펴봐야 한다는 점에서 원하는 정보만을 골라보기에는 너무 많은 시간을 요구한다는 단점을 가진다. 특히 사용자가 특정한 항목에 대한 정보만을 원하는 경우 검색을 통해 주어지는 정보의 집합들은 사용자에게 큰 도움을 주기가 어렵다.

이에 따라, 정보를 좀더 압축하고 요약하여 사용자에게 빠르게 읽기 쉬운 형식으로 제공하기 위한 여러 가지 시도들이 있었다. 각 웹사이트의 정보를 수집하여 추천 엔진에서의 개인화된 기호 및 추천 서비스를 제안한다.

본 논문에서는 종래에는 피어슨 상관계수(Pearson Correlation)를 사용했는데 피어슨 상관계수(Pearson Correlation)가 메트릭 거리(metric distance)가 아니여서 색인기법을 사용할 수 없고 따라서 3 가지 벡터 유사도(Vector Similarity)를 제안한다. 사용자 선호를 반영하여 정보를 웹에서 쉽게 찾을 수 있도록 하는 개인화된 추천 엔진을 K-최근접 이웃(nearest Neighbor)방식과 벡터 유사도(Vector Similarity)를 이용해 분석하고자 한다. 이 추천 엔진은 사용자가 빈번하게 검색하는 특정 분야의 정보를 검색하고자 할 때 유용하게 사용될 수 있다.

2. 관련연구

2.1 K-최근접 이웃(nearest neighbor)를 이용한 방법

대용량의 데이터베이스를 다루는데 있어서 주된 문제는 검색의 효율성으로, 보다 효율적인 유사성을 제공하기 위한 색인 기법에 대한 많은 연구들이 이루어지고 있다. Vp-tree 는 메트릭 색인 트리를 이용하는 색인기법이다[1,2].

유사성에 기반한 검색을 효율적으로 지원하는 K-nearest Neighbor 방식은 평가에 참여할 사용자를 선택하고 참여자에 대한 가중치를 부여하기 위해 사용자간의 유사도를 이용한다 [3,4,5].

이웃에 근거한 알고리즘(neighbor-based algorithm)은 예측할려고 하는 사용자의 이웃에 가중치를 부여하고, 그 이웃이 주어진 항목에 대한 등급에 대한 가중치 평균(weighted average)을 예측값으로 사용한다.

Pa,i 는 예측하려는 사용자 a 의 항목 i 에 대한 예측을 나타낸다.

$$P_{a,i} = \bar{r_a} + \frac{\sum_{u=1}^{n} \left(r_{u,i} - \bar{r_u}\right) * W_{a,u}}{\sum_{u=1}^{n} W_{a,u}}$$

Wa,u 는 피어슨 상관계수(Pearson correlation coefficient)에 의해 정의될 때 예측하려는 사용자와 다른 사용자 u 와의 사이에 유사성 가중치를 나타낸다.

$$W_{a,u} = \frac{\sum_{i=1}^{m} (r_{a,i} - \bar{r_a}) * (r_{u,i} - \bar{r_u})}{\sigma_a * \sigma_u}$$

K-최근접 추천 엔진은 데이터베이스에 저장된 대량의 사용자들 중에서 예측하려고 하는 사용자가 원하는 사용자들을 찾는 방법이다. K-최근접 추천 엔진은 데이터베이스의 대량의 사용자들로부터 K 명의 사용자를 추출하여 데이터베이스에 저장한 다음, 이를 이용하여 예측하려고 하는 사용자와 가장 유사한 값을 갖는 사용자를 찾는 방법이다.

2.2 벡터 유사도(Vector Similarity)를 이용한 방법

추천 엔진이 제시하는 정보는 사용자에 의해 평가받게 되는데, 벡터 유사도(Vector Similarity)는 코사인(cosine)값을 측정하고 정보 검색에 효과적으로 이용한다.

기존의 피어슨 상관계수(Pearson Correlation)는 비메트릭 거리(Nonmetric distance)이므로 색인 기법을 사용하지 못한다. 벡터 유사도(Vector Similarity)는 아래의 공식을 이용하여 두 벡터간의 유사도를 결정할 수 있다.

$$W(a,i) = \sum_{j} \frac{r_{a,j}}{\sqrt{\sum_{k \in I.} r^2_{a,k}}} \frac{r_{i,j}}{\sqrt{\sum_{k \in Ii} r^2_{i,k}}}$$

그리고 벡터 유사도(Vector Similarity)는 피어슨 상관계수(Pearson Correlation)의 비메트릭 거리(Nonmetric distance)의 문제점을 해결하기 위해 두 벡터간의 코사인 값을 측정하여 메트릭 거리를 결정하는 것이 가능하다[4].

벡터 유사도(Vector Similarity)는 PVS(Positive Vector Similarity), AVS(Absolute Vector Similarity), RVS(Relative Vector Similarity) 세가지를 제안한다.

PVS(Positive Vector Similarity)는 기존의 사용자 항목표에서 나타나는 등급의 벡터 유사도를 나타낸다.

AVS(Absolute Vector Similarity)는 사용자의 항목등급을 등급의 중간값을 기준으로 해서 선호하는 지의 여부로 나타낸다.

RVS(Relative Vector Similarity)는 사용자 항목등급의 평균을 중간값으로 결정한다.

PVS(Positive Vector Similarity)는 두 사용자의 선호도가 비슷하면 1 로 나타내고, 선호도가 관련이 없다면 0 으로 나타낸다. AVS(Absolute Vector Similarity)와 RVS(Relative Vector Similarity)는 두 사용자의 선호도가 가장 유사성이 있으면 1 로 나타내고, 선호도가 전혀 관련이 없다면 0 으로 나타내며, 선호도가 정반대일 경우에는 -1 로 나타낸다. 이 세가지 벡터 유사도(Vector Similarity)는 메트릭 거리(Metric distance)이므로 데이터베이스의 대량의 사용자 데이터에 색인 기법을 사용할 수 있다.

사용자의 선호도 예측 분야는 유사성이 있는 K 명의 코사인 값을 잘 반영한다. 그러나 기존에 가장 정확하다고 판단되어 일반적으로 사용되고 있는 피어슨 상관계수(Pearson Correlation)는 비메트릭 거리(Nonmetric distance)이므로 색인 기법을 적용할 수 없다. 그러므로 본 논문에서는 색인 기법을 이용할 수 있는 세가지 벡터 유사도(Vector Similarity)를 제안한다.

3. 실험

3.1 실험 방법

추천 엔진의 목적은 웹에서 사용자의 기호에 맞는 정보를 예측해서 찾아 주는 것으로, 유사성을 이용해 사용자의 기호를 추천해 준다. 근접한 사용자의 수가 증가할수록 추천의 정확성을 결정하기 위해 평균 절대 에러(Mean Absolute Error)를 이용한다. 먼저 본 논문에서 제안한 추천 엔진 방법을 이용하는 것이 사용자의 선호항목을 더 효과적으로 예측한 수 있는가를 평가하기 위하여 본 실험은 데이터베이스에 저장된 기존의 매겨진 등급 대신에 다른 데이터를 사용하여 등급을 예측한다. 실험에 사용된 데이터는 1000 개의 테스트 데이터를 이용하여 테스트하였다.

3.2 실험 데이터 집합

실험을 위한 실험 데이터의 집합은 Compaq 에서 수집하여 실험한 EachMovie 데이터 집합을 이용한다. 사용자는 영화항목에 대해 5 등급으로 등급을 매긴다. 가장 좋은 등급은 등급 5 이고, 가장 낮은 등급은 등급 1 이고, 공백은 등급을 매기지 않은 항목을 의미한다.

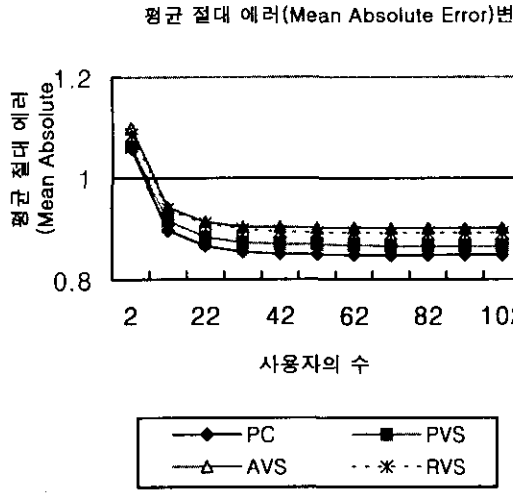
3.3 평가 방법

추천 엔진에서 정확성(accuracy)에 대해 평가하기 위해서 평균 절대 에러(Mean Absolute Error)를 이용한다.

3.4 결과 및 분석

본 추천 엔진의 목적은 웹에서 사용자의 기호에 맞는 정보를 찾아 주는 것으로, 벡터 유사도(Vector Similarity)를 이용해 사용자의 기호를 예측해 나간다. 먼저 본 논문에서 제안한 추천 엔진 방법을 이용하는 것이 사용자의 선호도를 더 효과적으로 예측할 수 있는가를 평가하기 위하여 기존의 피어슨 상관계수(Pearson Correlation)를 사용하는 방법과 벡터 유사도(Vector Similarity)를 사용하는 방법을 비교하였다. 제안한 추천 엔진의 방법이 실제로 사용자의 선호도 정보를 잘 예측하는가를 평가하기 위하여 평균 절대 에러(Mean Absolute Error)를 이용하여 측정하였다. 본 실험에서는 대량의 데이터베이스중에서 일부분의 데이터에 대해 이러한 실험을 실행하여 평균치를 구하여 비교한다.

기존의 피어슨 상관계수(Pearson Correlation)를 사용하는 것이 정확도가 높지만 유사성이 있는 K 명의 사용자를 효과적으로 찾아내기 위해 색인기법을 사용할 수 있는 벡터 유사도(Vector Similarity)를 사용하는것도 정확성(accuracy)이 비슷한 것으로 나타났다.



<그림 1> PC(Pearson Correlation)와 벡터 유사도 (Vector Similarity)를 사용한 평균 절대 에러(Mean Absolute Error)의 변화 비교

<그림 1>을 살펴보면 가까운 사용자의 수가 증가할수록 평균 절대 에러(Mean Absolute Error)값이 감소하므로 정확성(accuracy)이 개선된다는 것을 알 수 있다. 그러나 가장 가까운 사용자의 수가 50 명 이상이면 정확성(accuracy)이 개선되지 않음을 알 수 있다[3]. 피어슨 상관계수(Pearson Correlation)를 이용하는 것이 가장 정확성(accuracy)이 높지만 효과적으로 가까운 K 명의 사용자를 검색하기 위해 색인기법을 사용할 수 있는 벡터 유사도(Vector Similarity)를 이용하는 것도 정확성(accuracy)이 비슷한 것으로 나타났다.

4. 결론

본 논문에서 제안한 추천 엔진에서의 개인화된 정보와 추천 서비스는 개인별 취향과 선호도에 맞는 정보만을 제공해 주는 기술이다. 기존의 피어슨 상관계수(Pearson Correlation)방식의 비메트릭 거리의 문제점을 해결하고 효과적으로 정보를 검색하는 추천 엔진(Recommendation Engine)을 제안하였다. 또한 사용자의 관심 변화에 대해 효과적으로 적응할 수 있는 벡터 유사도(Vector Similarity)를 이용한 방법이 종래의 피어슨 상관계수(Pearson Correlation)방식을 이용한 결과와 거의 비슷하게 정확성을 나타내고 있다.

벡터 유사도(Vector Similarity)를 이용하는 K-최근접 이웃 추천 엔진(Recommendation Engine)은 유사성을 가진 사용자가 종래에 매긴 등급을 이용해서 예측하려고 하는 사용자가 예측하려는 항목의 등급을 예측하는 방법을 이용하였다. 지금까지 일반적으로 사용한 피어슨 상관계수(Pearson Correlation)는 메트릭 거리(Metric distance)의 개념이 아니므로 색인이 불가능하다. 반면 메트릭 거리(Metric distance)를 사용하여 색인이 가능한 벡터 유사도(Vector Similarity)를 이용한 방법이 피어슨 상관계수(Pearson Correlation)를 이용한 방법과 정확성(accuracy)이 유사한 것으로 나타낸다.

. 앞으로 이 방법을 단위 쇼핑몰, 진료 사이트, 사이버 증권거래, 뉴스 서비스 등의 개인화가 필요한 서비스 사이트에 접목시킨다면 개인화된 정보의 제공이 가능해 질 것이다.

5. 참고문헌

[1] Paolo Ciaccia, Marco Patella, Pavel Zezula, "M-tree: An Efficient Access Method for Similarity Search in Metric Spaces", Conference on Very Large Database, 1997.

[2] S.Brin. "Near neighbor search in large metric spaces", Conference on Very Large Database, Zurich, Switzerland, 1995.

[3] J.L. Herlocker, J.A, Konstan, A. Borchers, and J.Riedl "An Algorithmic Framework for Performing Collaborative Filtering", Conference on Research and Development in Information Retrieval, 1999.

[4] P. Resnick, N. Iacovou, M. Suchack, P. Bergstrom, J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", ACM CSCW, 1994.

[5] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering", In Proceedings of the 14th Conference on Uncertainty in Articial Intelligence (UAI-98), pages 43-52, San Francisco, July 24-26 1998