코사인 유사도 기법을 이용한 top-*k* 관련쌍 검색 방법 조사

김성철*, 김정환**, 김나영*, 김태훈*, 유환조*
* 포항공과대학교 컴퓨터공학과
**포항공과대학교 창의 IT 융합공학과

e-mail : sukim@adobe.com, jhkim@dblab.postech.ac.kr, kimnay@postech.ac.kr, ryan@buzzni.com, hwanjoyu@postech.ac.kr

Survey on Top-k Related Pair Search Method Using Cosine Similarity

Sungchul Kim*, Jeong-Hwan Kim**, Na-Yeong Kim*, Taehoon Kim*, Hwanjo Yu*
* Dept. of Computer Science, POSTECH
**Dept. of Creative IT Engineering, POSTECH

요 약

유사도 검색은 전통적으로 데이터베이스 그리고 웹검색 분야의 핵심이었으나, 대용량 데이터의 등장으로 검색의 정확도뿐만이 아니라 효율성 측면에서의 요구가 증가하며 여전히 다양한 분야에서 활발히 연구되고 있다. 아이템간의 유사도를 측정하기 위한 방법론 중 코사인 유사도 방법론은 고차원공간에서의 활용이 유리하다는 이점때문에 가장 널리 활용되고 있는 방법론으로, 정보검색, 장바구니 분석, 생물정보학 등 다양한 분야에서 활용되고 있다. 본 논문에서는 코사인 유사도를 소개하고, 연관성 분석 측면에서 코사인 유사도를 사용한 기존의 연구들을 소개한다.

1. 서론

유사도 검색은 전통적으로 데이터베이스 그리고 웹검색 분야에서 핵심이 되는 요소로, 빅데이터 시대가 도래함에 따라 검색의 정확도뿐만이 아니라 효율성에 대한 요구가 증가하고 있다. 데이터테이블의 아이템셋 간의 유사도 혹은 검색어와 문서 간 유사도를 측정하기 위한 다양한 방법론이 존재하지만, 그 중에서도 코사인 유사도방법은 벡터로 표현이 가능한 데이터를 대상으로 다차원의 양수 공간에서의 유사도 측정이 용이하여 정보검색, 장바구니 분석, 생물정보학 등 다양한 분야에서 널리 활용되고 있다.

본 논문에서는 기존의 연구 중 코사인 유사도를 사용하여 top-*k* 관련쌍을 검색하는 문제를 해결한 기존의 연구 사례를 소개하고자 한다; 1) TOP-DATA[1], 2) TOP-MATA[2]. 먼저 TOP-DATA 는 연관성분석(association mining) 측면에서 의미 있는 관련쌍을 효율적으로 찾기 위한 방법론으로 다양한 실데이터에서 무차별 대입(brute force) 방법론과 비교하여 유의미한 성능향상을 보였으며, 2) TOP-MATA 는 TOP-DATA 방법론의 효율성을 보다 개선하기 위한 방법론이다. 자세한 내용은 다음 장에 소개한다.

이 논문은 2017 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2014R1A2A2A01004454)

2. 방법론

2.1 코사인 유사도

코사인 유사도(cosine similarity) 방법론은 두 벡터간 각도의 코사인값을 이용한 벡터간 유사도를 측정하는 방법론으로 두 벡터 x 와 y 가 주어졌을 때, 다음과 같이 계산한다.

$$\cos(X,Y) = \frac{<X,Y>}{\|X\|\|Y\|} \quad (1)$$

본 식에서 <,>는 두 벡터간 내적연산을 의미하며, ‖ ‖는 L-2 정규화를 의미한다. 본 식에 의하면 두 벡터가 완전히 반대되는 경우 -1 완전히 같은 경우 1 이 되는 -1 부터 1 까지의 값이 허용되나, 아이템셋 간의 유사도를 계산하는 경우 각 차원이 음의 값을 가질 수가 없으므로 결과값은 0 에서 1 까지만 가능하다. 또한 많은 경우에는 각 차원이 바이너리값을 가지는 벡터로만 데이터를 구성하기도 한다. 예를 들어 장바구니 분석의 경우 각 차원은 i 번째 아이템의 구매여부를 나타내어 해당 유저가 아이템을 구매한 경우 1 아닌 경우 0 으로 표현한다.

2.2 문제 정의

Top-k 관련쌍 검색문제는 각 아이템 X_i 가 바이너리

벡터인, 아이템집합 D = {X_1, X_2, …, X_n} 이 주어졌을 때, 코사인 유사도 값이 상위 k 에 해당하는 아이템쌍들을 검색하는 것을 목적으로 한다.
해당 문제를 푸는 가장 기본적인 방법론은 무차별 대입 방법론으로 이는 모든 아이템쌍의 코사인 유사도를 계산하고 비교하여 상위 k 개의 아이템을 검색하는 방식으로 n 개의 아이템이 주어졌을 때 $\frac{n \times (n-1)}{2}$ 쌍의 유사도를 검색하여야 한다. 이는 주어진 아이템집합의 크기가 커질수록 계산하여야 하는 양이 지수승으로 증가하기에 검색코스트가 크다는 단점이 있다.

2.3 TOP-DATA

먼저 대상 벡터가 바이너리 값을 가지는 경우, 위에서 정의한 코사인 유사도는 연관성 분석의 지지도(support) 개념을 이용해 다음과 같이 표현이 가능하다.

$$\cos(X,Y) = \frac{supp(XY)}{\sqrt{supp(X)supp(Y)}} \quad (2)$$

이 때 공기성(co-occurrence)의 정의에 따라 X 와 Y 의 공기성 수는 X 혹은 Y 의 공기성보다 커질 수 없다. 즉, supp(X,Y)는 supp(X) 혹은 supp(Y)보다 항상 작거나 같게 되며 따라서 다음과 같은 상계값(upper bound value)을 계산 할 수 있게 된다.

$$\frac{supp(XY)}{\sqrt{supp(X)supp(Y)}} \leq \frac{supp(Y)}{\sqrt{supp(X)supp(Y)}} = \sqrt{\frac{supp(Y)}{supp(X)}} \quad (3)$$

$$upper(\cos(X,Y)) = \sqrt{\frac{supp(Y)}{supp(X)}} \quad (4)$$

TOP-DATA 는 상계값을 대상으로 하며 기본적인 동작 원리는 다음과 같다. Top-k 리스트가 주어졌을 때, 새로운 관계쌍마다 만약 상계값이 현재 top-k 리스트의 코사인 유사도 최소(min_cos)보다 작다면 해당쌍을 전정(pruning) 리스트에 추가하고, 그렇지 않다면 실제 실제 코사인 유사도를 계산하여 이 값이 최소(min_cos)보다 크다면, 현재 리스트에서 최소값을 가지는 관계쌍와 교체한 후 min_cos 값을 갱신한다. 이는 지지도값을 기준으로 정렬된 아이템 매트릭스(sorted item-matrix)를 대상으로 하며 대각 여행 방법론(Diagonal Traversal Procedure)로 진행하며 자세한 내용은 [1]에서 확인할 수 있다.

2.4 TOP-MATA

TOP-DATA 를 이용하면, 상계값을 이용한 전정효과로 인해 굉장히 많은 수의 관계쌍들의 코사인 유사도를 직접 계산하지 않고도 top-k 관계쌍을 얻을 수 있다. 하지만 이 때 사용되는 값(min_cos)의 최적값은 실제 top-k 리스트를 얻게 되어야 알 수 있으며 그 이전 단계에서는 실제 계산하지 않아도 되는 관계쌍이 많아지며, 이는 데이터가 클수록 효율성을 저하하는 결과를 초래한다. 또한 대각 여행 방법론은 알고리즘상으로는 많은 전정효과를 주지만, 실제 굉장히 높은 인풋-아웃풋(I/O)를 발생하기도 하는 문제가 있다.

위 문제를 방지하기 위해 TOP-MATA 방법론이 제시되었다. 이 방법론은 동작 원리는 TOP-DATA 와 거의 같지만, 대각 여행 방법론 대신에 최대값 우선 여행 방법론을 채택하여 위 문제를 해결하였다.

TOP-MATA 방법론은 정렬된 아이템 매트릭스를 대상으로 한다는 점은 같으나 가장 큰 지지도 상계값을 가지는 행부터 시작하여, 같은 최대 지지도 상계값을 가지는 모든 관계쌍을 먼저 비교한다. 이후 해당 작업은 행단위로 반복하여 최대 상계값이 min_cos 값에 변화를 주지 않을때까지 진행한다. 이는 최대 히프(Max heap) 데이터 구조를 기반으로 진행되며, 자세한 내용은 [2]에서 확인이 가능하다.

3. 실험결과

제안 방법론들은 UCI repository[3]를 비롯하여 실데이터까지 총 7 개의 데이터셋을 대상으로 top-k 관계쌍 검색을 위해 사용되었고, 실제 수행 시간을 측정하여 비교하였다. 실험결과를 요약하면 다음과 같다.

1) TOP-DATA 와 무차별 대입 방법론 비교: 두 방법론의 경우 TOP-DATA 를 사용하는 경우, 상계값의 전정효과로 인하여 검색 속도 측면에서 적게는 약 30% 에서 k 가 작은 경우 데이터셋에 따라서 최대 약 300% 이상의 성능향상을 보였다.

2) TOP-DATA 와 TOP-MATA 비교: 데이터셋에 따라 정도의 차이는 있었지만, 일반적으로 TOP-MATA 가 더 빠르게 동작하였으며 (최대 약 40%), 성능향상은 k 가 커짐에 따라서 더 두드러지게 나타났다. 또한, 성능향상은, 실제 전정효과에 따라 계산하지 않은 관계쌍의 개수나 I/O 코스트 측면(최대 46% 향상)에서도 확인 할 수 있었다.

보다 자세한 결과는 [1, 2]에서 확인할 수 있다.

4. 결론

본 논문에서는 연관성 분석 측면에서 코사인 유사도 기반 top-k 결과를 빠르게 검색하기 위한 TOP-DATA 그리고 TOP-MATA 방법론을 소개 및 비교하였다. 코사인 유사도를 기반으로 한 검색 및 연관성 분석 방법론들은 이미 다양한 분야에서 활발히 활용되고 있지만, 최근에는 많은 사용자들이 쉽게 다양한 디바이스에 접근하는 실정이며 이로부터 생성되는 수많은 컨텐츠 및 유저정보는 단순히 데이터의 양이 아닌 비정형 데이터에 대한 수요를 강요하고 있다. 추후에는 이를 고려한 새로운 방법론이 연구되어야 할 것이다.

참고문헌

[1] Zhu, Shiwei, Junjie Wu, and Guoping Xia. "TOP-K cosine similarity interesting pairs search." Fuzzy Systems and Knowledge discovery (FSKD), 2010 Seventh International Conference on. Vol. 3. IEEE, 2010.

[2] Zhu, Shiwei, et al. "Scaling up top-k cosine similarity search." Data & Knowledge Engineering 70.1 (2011): 60-83.

[3] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science