클러스터 내부 빈발 지지도를 이용한 개선된 사용 프로파일 평가

안계순[0] 이필규
인하대학교 전자계산공학과
kyesun@im.inha.ac.kr pkrhee@inha.ac.kr

Evaluation Of Improved Usage Profiles Using Frequency Support Threshold In Clusters

Kye-Sun Ahn[0] Phill-Kye Rhee
Dept. of Computer Science, Inha University

요 약

웹 로그 기반의 웹 사용 마이닝은 명시적 평가 의존, 확장성 결여, 그리고 다차원 및 희박한 데이터에 성능이 떨어지는 협력적 여과의 문제를 다소 해결할 수 있다. 그러나 k-Means 군집화 방법으로 생성된 군집속 유사 사용자 이동 패턴으로는 클러스터속 사용자 전체의 선호도를 표현할수 없으므로 사용자 이동 패턴인 트랜잭션들로부터 사용 프로파일을 유도해야 한다. 본 논문에서는 유사 군집 사용자들의 관심과 기호를 표현할수 있도록 클러스터 내부 데이타로부터 평균 가중치 및 빈발 지지도 임계값을 사용하여 개선된 사용 프로파일을 생성하고 실험 데이터를 통한 예측력과 추천에 대한 성능을 평가한다.

1. 서 론

월드 와이드 웹(WWW)의 발달과 빠른 대중화로 인하여 방대한 양의 정보들이 생성되면서 사용자들이 원하는 대부분의 정보들을 컴퓨터 앞에 앉아서도 검색 할 수 있게 되었다. 하지만, 동일한 정보들이 폭 넓게 산재하게 되면서, 사용자는 자신에게 꼭 필요한 정보를 찾는데 많은 시간과 노력을 들이게 되었다 [3]. 따라서, 많은 웹 사이트들 내의 정보들을 효과적으로 검색하고자 검색엔진이 개발되고 있다. 기존의 전문화된 웹 사이트들과 통합 서비스를 제공해주는 많은 포털 사이트들은 사용자들의 확보를 위하여 웹 사이트 내에 사용자들이 원하는 정보들을 찾아서 서비스를 해주고 있다. 자동화된 개인화 추천 서비스를 제공하는 성공적인 전자상거래 시스템 대부분이 협력적 여과에 기반을 두고 있다. 협력적 여과는 사용자가 선호할 만한 아이템을 추천 하기 위해서 현 사용자의 아이템에 대한 평가값을 유사 사용자들의 아이템에 대한 평가값과 일치시키는 방법을 사용한다. 그러나 협력적 여과는 아이템의 수가 증가할수록 평가 데이터 속의 희소성(sparsity)이 증가하고 많은 수의 사용자와 아이템상에서 사용자간의 유사도를 계산하는데 오류가 발생하는 문제점을 가지고 있다. 최근에는 이런 전통적인 기술과 연관된 문제점을 해결하기 위한 대안으로서 웹 사용 마이닝에 대해서 연구해 오고 있다. 일반적으로 웹 사용 마이닝은 사용자들의 이동 행위에서 흥미로운 패턴을 발견하기 위해서 웹 사이트로부터 획득된 사용 또는 클릭스트림 데이터에 여러 가지 데이터마이닝 알고리즘을 수행한다. 그러나 데이터마이닝 알고리즘으로부터 발견된 패턴만을 가지고 개인화 추천을 수행하는 것은 충분하지 않다. 개인화 추천을 가능케 하기 위해서는 발견된 패턴으로부터 사용 프로파일을 유도하는 것이 필요하다.

이에 본 논문에서는 유사한 이동 패턴을 가지는 사용자들의 관심과 기호를 표현할수 있도록 클러스터 내부 데이타로부터 평균 가중치 및 빈발 지지도 임계값을 사용하여 개선된 사용 프로파일을 생성하고 실험 데이터를 통한 예측력과 추천에 대한 성능을 평가한다.

2. 연구 배경

웹 사용 마이닝을 위해서는 웹 서버 로그 파일로부터 사용자 세션 파일을 추출해서 해당 사이트에 대한 정보를 구축하는 것이 필요하다. 사용자 세션 파일을 추출하기 위해서는 데이터 전처리가 필요하다. 웹 로그에 대한 전처리 방법은 Cooley가 제안한 방법을 사용하여 세션 파일을 추출한다 [1][4]. 데이터 전처리 과정을 통해 사용자 트랜잭션 파일을 추출한다.

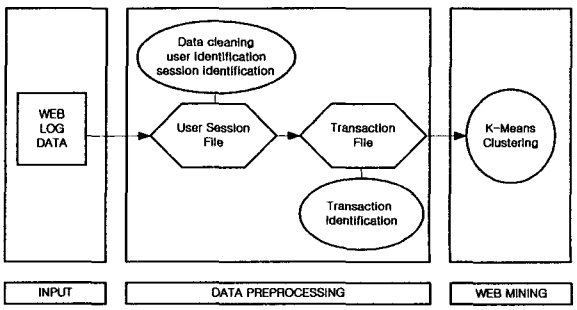

그림 1. 웹 로그 데이터 전처리

k-means 군집화 알고리즘은 클러스트링 기법중 Euclidian Distance(거리)를 이용하여 가깝게 위치한 점들을 찾아 군집으로 묶어 주는 기법으로 차원의 제약이 없는 장점을 가지고 있다. 웹 사용 마이닝의 데이터 전처리로 추출된 m차원의 벡터로 표현된 사용자 트랜잭션들을 입력으로 사용하여 유사한 사용자 이동 패턴을 가지는 사용자들끼리 군집을 형성하는 결과를 가진다[5]. 다음은 간단한 k-Means 클러스터링 알고리즘을 보여준다.

Input : The number of clusters k and a database containing n objects
output : A set of k clusters that minimizes the squared-error criterion

Method
1> choose k objects as the initial cluster centers
2> repeat
3> (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4> update the cluster means, i.e., calculate the mean value of the objects for each cluster;
5> until no change;

K-Means Clustering

3. 사용 프로파일 (Usage Profile)

웹 로그 전처리를 통해 사용자의 이동 패턴을 보여주는 벡터화된 트랜잭션 집합을 생성하고 거리 기반의 유사도 측정에 기반한 k-means 군집화를 통해 서로 유사한 이동 패턴을 가지는 사용자들로 군집을 형성한다. 유사한 이동 패턴을 가지는 사용자들의 군집을 트랜잭션 클러스터라 한다. 식(1)은 클러스터 집합을 나타낸다.

$$TC = \{c_1, c_1, \cdots, c_n\} \quad (1)$$

클러스터화된 사용자들의 페이지에 대한 선호도를 얻기 위해 클러스터속의 트랜잭션을 사용하는 것은 비효과적인 방법이므로 클러스터속 트랜잭션을 바탕으로 공통 관심 정도를 표현하는 사용 프로파일을 유도한다.

식 (1)의 각 클러스터 $c \in TC$ 에서 페이지 선호에 대한 가중치 평균값을 계산한다. 평균값은 클러스터에 속하는 트랜잭션 크기에 대한 클러스터속 트랙잰션들 상에서 나타나는 페이지 가중치들의 합의 비율로 계산한다. 기존의 연구에서는 사용 프로파일을 생성하기 위해서 가중치 평균값을 최대 가중치 1에 맞게 정규화 처리하고 평균가중치 임계값 이하의 페이지는 제거되는 방법을 사용하였다. 평균 가중치만을 사용할 경우 트랜잭션속 페이지의 발생 빈도가 적으면서 머문 시간이 높은 페이지가 사용 프로파일에 영향을 줄수 있는 문제가 발생한다. 그러므로 본 연구에서는 클러스터 내부의 트랜잭션들에서 빈발 지지도를 계산하여 최소 지지도 이상의 평균 가중치를 획득하도록 하였다. 식 (2)에서 트랜잭션 클러스터 $c \in TC$ 로부터 페이지-가중치 쌍의 집합인 사용 프로파일 pf_c 를 생성한다.

$$pf_c = \{< p, w(p, pf_c) > | p \in P, w(p, pf_c) \geq \mu, S(p, pf_c) \geq \beta\} \quad (2)$$

사용 프로파일 pf_c 이 포함하는 페이지 p의 가중치는 식 (3)에 의해서 계산된다.

$$\text{평균 가중치} : w(p, pf_c) = \frac{1}{|c|} \cdot \sum_{t \in c} w(p,t) \quad (3)$$

$$\text{빈발 지지도} : S(p, pf_c) = \frac{\sum_{t \in c} occurrence(p,t)}{|c|} \quad (4)$$

식 (3)의 $w(p,t)$는 트랙잭션 $t \in c$ 에 속하는 페이지 p 의 가중치를 나타내며, 식 (4)의 $occurrence(p,t)$ 는 트랙잭션 $t \in c$ 에서의 페이지 p 의 발생 유무이다.

유사한 이동 패턴을 가지는 사용자 군집으로부터 유도된 사용 프로파일은 정규화된 평균값을 포함하는 n 차원 벡터로 표현되고 동일 사용자 세션상에서 사용자 행위 예측을 위한 추천 알고리즘의 자료로 사용된다.

4. 추천 엔진 (Recommendation Engine)

위와 같은 방법으로 생성된 사용 프로파일은 아래의 추천엔진을 통해 추천 집합을 생성하고 이를 기반으로 추천 성능을 평가할 수 있다[3].

사용 프로파일에 기반한 추천엔진은 현재 활성화된 사용자 세션에 대해서 추천집합을 계산한다. 추천집합은 사용 프로파일들에 대해서 활성화된 사용자의 세션을 일치시켜 얻는다. 활성화된 사용자 세션에서 윈도우 크기가 n인 부분 세션이 추천아이템의 추천 값에 영향을 준다.

주어진 사용 프로파일과 활성화된 사용자 세션의 매칭값 계산은 식 (5)의 두 벡터간 코사인 측정을 사용하여 유사도를 계산한다.

$$match(S,C) = \frac{\sum_k w_k^c \cdot s_k}{\sqrt{\sum_k (s_k)^2 \times \sum_k (w_k^c)^2}} \quad (5)$$

추천 아이템에 대한 추천 값은 프로파일 C 속의 각 페이지 p를 위해서 식 (6)와 같이 계산된다.

$$\text{Rec}(S,p) = \sqrt{w(p,C) \cdot match(S,C)} \quad (6)$$

만약 페이지 p가 현재 활성화된 세션 속에 있다면, 사용 프로파일 속의 페이지 p 추천 값은 0 이 된다. 결론적으로 활성화 세션 S 에 대해서 모든 사용 프로파일로부터 최소 추천 임계값 이상의 모든 추천 아이템 집합의 발견은 아래의 식과 같다.

$$UREC(S) = \{w_i^c \mid C \in UP, \text{Rec}(s, w_i^c) \geq \rho\} \quad (7)$$

식 (7)에서 UP는 모든 사용 프로파일의 집합을 나타낸다. 본 논문에서 추천 아이템은 활성화된 사용자 세션과 상위 일치를 보이는 사용 프로파일들에 의해서 발견되게 된다.

4. 실험

이번 장에서는 개선된 사용 프로파일에 대한 성능 평가를 보인다. 먼저 간단히 실험 환경에 대해서 기술하고 성능 측정 방법을 사용하여 개선된 사용 프로파일의 사용성을 평가한다.

4.1 실험 환경

본 실험은 DePaul CTI Web site 에 2002 년 4 월 2 주간 사이트에 방문한 사용자들의 웹 로그 정보를 통해서 수행했다. 실험 데이터는 5446 명 사용자로부터 683 개의 고유 페이지와 13745 개의 사용자 세션으로 구성되어 있으며 각 사용자가 접속한 페이지에 대해서 머문 시간을 0-999 수치로 표현되어 있다. 실험을 위해서 평균 트랜잭션 크기가 5 이하인 트랜잭션은 제거했으며 전체 트랜잭션 데이타를 학습 데이터와 테스트 데이터의 비율을 8 : 2 로 분할하여 학습데이타는 개선된 사용 프로파일을 생성하는데 사용했고 테스트 데이터는 사용 프로파일의 성능을 평가하는데 사용하였다.

4.2 예측력 평가

클러스터 내부 빈발 지지도 적용후 개선된 사용 프로파일의 예측력을 평가하는 방법으로 가중치를 고려한 WAVP 를 사용하였다[1][2].

그림 2 는 빈발 지지도 임계값을 적용해 개선된 사용 프로파일의 예측력을 비교한 결과를 보인다. W 는 평균 가중치이고 s 는 빈발 지지도를 나타낸다.

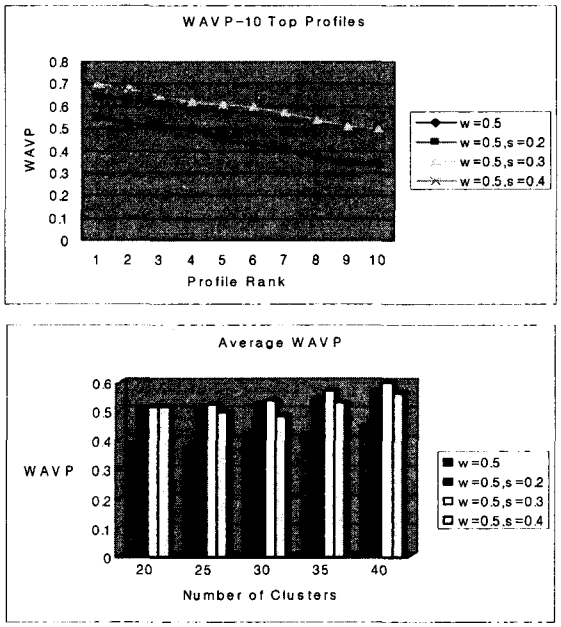

그림 2. Top Ranking 사용 프로파일 비교

군집화가 가장 좋은 40개의 클러스터에서 클러스터 내부의 빈발 지지도 0.3을 적용해 개선시킨 사용 프로파일의 WAVP 값이 최대값을 가지며 평균 가중치만을 적용한 사용 프로파일보다 15% 예측력 상승이 있었다.

4.3 추천 성능 평가

본 실험에서는 개선된 사용 프로파일의 추천에 대한 성능 측정을 위해 F1 방법을 사용하였다[1]. 성능 평가에 앞서 추천에 영향을 미치는 요소인 세션 윈도우 크기는 2 로 설정하고 평균 가중치 0.5 와 빈발 지지도 0.3 의 상위 랭크 사용 프로파일을 사용하여 그림 3 의 결과를 보인다.

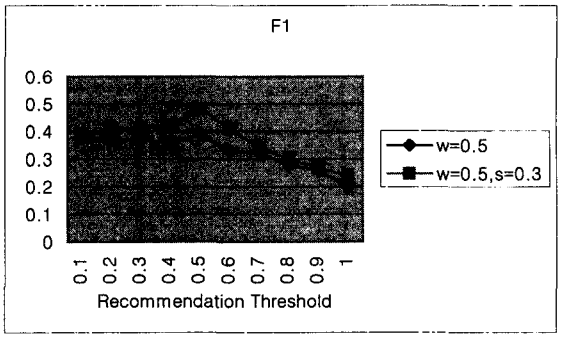

그림 3. 추천 성능 비교

빈발 임계값을 적용한 개선된 사용 프로파일의 추천 정확도가 전체 적으로 다소 높았으며 추천 임계값 0.5 에서 추천 정확도가 가장 높음을 알수 있다.

5. 결론

본 연구에서는 특정 웹 사이트의 웹 로그에 기반한 데이터의 주관성이 배제된 사용 프로파일을 생성하였다.

사용 프로파일 생성시 클러스터속 페이지에 대한 평균 가중치 임계값 및 빈발 지지도 임계값을 통해 일부 트랜잭션에 민감하게 작용한 사용 프로파일의 생성을 피할수 있으며 활성화된 사용자 세션에 대한 예측력 및 추천 성능이 향상 되는 것을 알수 있었다. 하지만 본 연구에서 사용한 k-means 방법은 차원에 제약이 없는 장점이 있음에도 불구하고 수행 속도가 트랜잭션 및 차원의 수에 비례해서 증가하는 비용을 감수해야 하는 문제가 있었다. 향후 연구에서는 차원 감소를 통한 계산 비용의 단축과 다양한 웹 사이트의 로그 데이터 및 기존의 군집화 알고리즘에 의해 생성된 사용 프로파일의 비교 평가를 통해 효과를 검증해 보고자 한다.

참고문헌

[1] Bamshad Mobasher, Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization, In Data Mining and Knowledge Discovery, Kluwer Publishing, Vol. 6, No. 1, pp. 61-82, January 2002

[2] Bettina Berndt, The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis.2002.

[3] R. Cooley, Web Mining: Information and Pattern Discovery on the WWW, ICIA'97 1997.

[4] R.Cooley, Data Preparation for Mining World Wide Web Browsing Patterns, In: Knowledge and Information Systems 1, Springer-Verlag, 00-00, pp. 1-26. 1999

[5] Jiawei Han, Data Mining : Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, p335-354, 550 pages. ISBN 1-55860-489-8,August 2000.