제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

웹 트랜잭션 클러스터링의 정확성을 높이기 위한 흥미도 가중치 적용 유사도 비교방법

강태호*, 유재수**
*충북대학교 정보산업공학과
**충북대학교 정보통신공학과
e-mail:segi21@netdb.chungbuk.ac.kr

Similarity Measurement with Interestingness Weight for Improving the Accuracy of Web Transaction Clustering

Tae-ho Kang*, Jae-Soo Yoo**
*Dept of Information & Industrial Engineering, Chung-Buk University
**Dept of Information & Communication Engineering, Chung-Buk University

요 약

최근 들어 웹사이트 개인화(Web Personalization)에 관한 연구가 활발히 진행되고 있다. 웹 개인화는 클러스터링과 같은 데이터 마이닝 기법을 이용하여 개개의 사용자에게 가장 흥미를 갖을만한 URLs의 집합을 예측하는 것이라 할 수 있다. 기존에는 웹 트랜잭션을 클러스터링 하기 위해서 사용자의 방문여부에 따라 트랜잭션을 비트벡터(bit vector)로 표현하였다. 하지만 이것은 웹 트랜잭션의 클러스터링에 있어서 사용자의 흥미를 배제하고 단순히 방문여부만을 반영하게 된다. 이에 본 논문에서는 사용자의 흥미도(Interestingness)를 반영할 수 있도록 보완된 웹 트랜잭션 모델을 제시하고 제안된 트랜잭션 모델을 적용한 유사도 비교방법을 제안한다. 그리고 성능평가를 통하여 제안한 방법이 기존 방법에 비해 클러스터링의 정확성을 높임을 보인다.

1. 서론

최근 웹사이트의 개인화를 추구하기 위한 노력으로 고객 데이터를 기존의 데이터마이닝 기법을 이용하여 분석하고 활용하는 방법들이 많이 연구되고 있다. 웹서버는 웹사이트에 접속하는 사용자들의 탐색 행위를 그대로 기록하고 있는 웹 로그를 제공하며 이러한 로그는 사용자의 성향을 파악할 수 있는 중요한 정보로 쓰이고 있다. 웹사이트의 개인화는 사용자의 요구를 미리 예측하여 이를 추천하는 것이라 할 수 있다. 즉 사용자의 기대에 부합하는 웹을 제공하는 방법이 일반적으로 사용된다. 이러한 웹사이트의 개인화에는 페이지들 사이의 연관관계를 파악하여 자주 발생하는 빈발항목집단을 알아내는 방법이나[6], 기존사용자들의 웹 탐색 패턴이 유사한 사용자들을 적절히 클러스터링하고 이것들로부터 추천집단을 생성하여 개인화에 이용하는[1]등의 기존의 데이터마이닝 기법들이 활용된다. 이중 유사한 성향의 트랜잭션 집단으로부터 추천을 유도하는 클러스터링을 이용한 개인화방법의 경우 클러스터링의 기초가 되는 유사도 비교는 그 비교의미와 그에 따른 정확성이 매우 중요하다. 기존에는 웹 트랜잭션 데이터를 방문 또는 비방문의 의미로서 이진가중치 벡터로 모델링하여 이들의 비트패턴에 대한 유사도를 비교하는 방식이 일반적으로 사용되었다[1].

이러한 비트벡터의 표현은 방문사용자의 패턴을 알 수 있는 효율적인 방법이긴 하나, 이진가중치를 사용함으로 인해 사용자가 보이는 관심이나 기타 트랜잭션의 특성이 무시되는 경우가 발생하게 되어 정확성이 결여되는 문제점이 있다. 따라서 본 논문에서 이러한 기존방법의 문제점을 지적하고 이를 보완할 수 있도록 웹 트랜잭션에 새로운 가중치를 부여하여 트랜잭션의 특성에 대한 비중을 적용할 수 있는 유사도 비교방법을 제안한다. 이를 통해 보다 의미 있고 정확한 클러스터링을 수행함으로써 활용도 높은 추천 집단 생성의 기반을 마련하고자 하는데 그 목적이 있다.

본 논문의 2장에서 관련연구로 트랜잭션 클러스터링의 기본이 되는 유사도 비교방법에 대한 기존 방법과 그에 대한 문제점을 제시하고, 3장에서 흥미도 가중치 트랜잭션 모델을 제안하여 이를 적용한 유사도 비교방법과 URL의 특성을 고려한 방법을 언급하고 마지막으로 4장에서는 실제 웹 탐색에 대한 성능평가를 통해 정확성의 차이를 보이고 결론을 맺는다.

2. 관련연구

2.1. 웹 개인화(Web Personalization) 과정

웹사이트의 개인화과정은 크게 나누어 데이터 사전처리[1,9]와

데이터들 사이의 관계 및 패턴을 발견하는 오프라인 처리단계와 추천집단을 생성하여 온라인 사용자에게 이를 추천하는 온라인 처리단계의 두 가지 과정으로 나뉘어 수행될 수 있다. 이중 패턴 발견은 가장 중요한 단계라 할 수 있으며, 이는 사용자 데이터들 사이의 특정 패턴을 발견하고 찾아낸 패턴을 이용하여 사용자 성향을 파악하고 행동을 예측하는 과정을 수행한다. 여기에는 웹 추천에 일반적으로 많이 사용되어지는 기법이라 할 수 있는 연관성 규칙을 이용해 발견해낸 빈발항목집단(Frequent Itemset)을 찾아내 이를 추천에 이용하는 방법[6, 7]과 유사한 성향을 보이는 사용자 트랜잭션들을 적절히 클러스터링 하여 교차 추천에 이용하는 방법[1, 7]등이 많이 연구되어지고 있다.

2.2. 사용자세션 구분과 트랜잭션모델

웹 로그파일이 주어지면 분석을 위한 첫 번째 단계로 데이터 정제를 수행한다. 데이터 정제는 페이지뷰당 하나의 성분만 남기고 중복적인 모든 파일을 제거한다는 것을 의미한다. 다음으로 웹 로그로부터 사용자의 세션을 도출해내는 과정이 필요하게 된다. 사용자 세션이란 한 사용자가 웹사이트에 접속하여 웹 탐색을 수행한 후 접속을 종료할 때까지의 일련의 행위라 말할 수 있다. 사용자 세션은 하나의 사용자에 대응하는 순차적인 참조 페이지의 집합으로 나타내질 수 있다.

2.3. 트랜잭션 클러스터링

사전 처리된 로그에 n개의 URL 집합과 m개의 사용자 트랜잭션 집합을 다음과 같이 정의한다.

$$U = \{url_1, url_2, \cdots, url_n\} \quad T = \{t_1, t_2, \cdots, t_m\}$$

여기에서 사용자 트랜잭션들을 클러스터링 할 때 트랜잭션들은 페이지 참조의 벡터들이 다차원 공간에 대응되게 되며 이에 표준적인 클러스터링 알고리즘은 일반적으로 이 공간을 거리라는 척도에 근거하여 상호 가까운 URL 집단들로 분리시킨다. 이때 웹 트랜잭션의 경우 각 클러스터는 URL 참조의 동시 발생 패턴이 유사한 트랜잭션들의 집단을 나타내게 된다.

트랜잭션 $t \in T$ 일 때, 이 트랜잭션에 대한 비트벡터(bit vector) 표현은 다음과 같다.

$$\vec{t} = \langle u_1', u_2', \cdots, u_n' \rangle \qquad u_i' = \begin{cases} 1, & if\ url_i' \in t \\ 0, & otherwise \end{cases}$$

트랜잭션들을 클러스터링하기 위해서 두 트랜잭션들간의 거리 측정이 필요하게 되는데 웹 트랜잭션의 경우 두 트랜잭션 t, s의 유사성 sim(t,s)는 비트벡터의 경우 다음과 같이 트랜잭션에서 일치하는 항목의 크기로 표현된다.

$$sim(t,s) = \frac{|t \cap s|}{\sqrt{|t| \cdot |s|}} \qquad \cdots\cdots \text{(수식 1)}$$

위 정의에 따른 유사도 비교결과는 클러스터링의 기초로 활용되어질 수 있으며 여기에 여러 가지 다양한 클러스터링 알고리즘이 적용될 수 있다.

3. 제안하는 트랜잭션 모델과 유사도 비교방법

웹 트랜잭션에 대하여 방문 패턴에 대한 유사성 비교는 클러스터링에 있어서 중요한 요소이긴 하나 그 의미가 충분하지 못하다. 그 이유는 웹사이트를 방문한 사용자는 각 페이지에 대해 동일한 관심을 가지고 접근하는 것은 아니기 때문이다. 한 예로 웹사이트를 접속한 사용자는 원하는 정보를 얻기 위해 많은 URL을 거쳐 갈 수 있다. 이때 원하는 컨텐츠를 만나게되면 해당 URL에서 머무는 시간은 그냥 거쳐간 URL의 참조시간보다 상대적으로 길 것이다. 이러한 관심의 차이에도 불구하고 비트벡터의 표현은 모든 페이지 접근에 대해 동일한 의미로 취급하고 있다.

웹 트랜잭션의 유사성을 비교하는데 있어서 오히려 흥미로운 컨텐츠에 대한 사용자의 관심이 중요한 경우도 있다. 따라서 방문 패턴뿐만이 아닌 사용자의 관심정도 까지도 적용될 수 있다면 보다 정확한 의미의 비슷한 성향을 가지는 사용자를 클러스터링 할 수 있을 것이다. 이러한 이유로 이번 장에서는 기존의 이진벡터의 문제점을 보완하고자 새로운 가중치를 제시한다.

3.1. 흥미 가중치 적용 트랜잭션 모델

웹 사용자가 참조한 페이지에 대한 관심은 웹 페이지 참조의 중복 횟수 또는 사용자가 참조한 페이지에 머문 시간을 가지고 관심정도를 예측할 수 있다. 이중 해당 URL에서 보낸 시간은 사용자의 관심정도를 추측할 수 있는 좋은 방법이라 할 수 있다 [2]. 본 논문에서는 이러한 특징을 이용하여 사용자의 흥미를 표현하는데 있어서 페이지 참조 시간의 비율을 사용한다.

사용자 흥미도를 적용하여 제안하는 트랜잭션 모델은 다음과 같다. 트랜잭션 $t \in T$ 일 때, 이 트랜잭션은 다음과 같은 흥미가중치(Interestingness Weight)로 표현될 수 있다.

$$\vec{t} = \langle w_1', w_2', \cdots, w_n' \rangle \quad w: \text{흥미 가중치} \quad u_i' = \begin{cases} weight, & if\ url_i' \in t \\ 0, & otherwise \end{cases}$$

기존의 이진 벡터를 변형한 사용자 흥미 관점의 가중치를 부여한 트랜잭션 모델이다. 기존의 이진벡터를 사용했을 때와 각 URL의 매핑은 그대로 유지한다. 다만 방문: 1, 비방문: 0으로 표현하던 방식대신 단지 트랜잭션에서의 각 URL이 가지는 사용자 관심의 비중을 적용한다. 사용자 관심에 대한 가중치는 한 트랜잭션의 전체 탐색 지속 시간에 대한 각 URL에서 소비한 시간의 비율을 사용하며 다음과 같이 정의한다.

$$W[i].Interestingness = W[i].dt / \sum_{j=1}^{n} W[j].dt$$

W[i].dt : 트랜잭션의 i번째 URL에서 머문시간 ..(수식 2)

여기에서 $W[i].dt$는 트랜잭션내의 i번째 URL에서 보낸 시간으로써 이 시간은 트랜잭션 모델로 변환되기 이전의 세션에서

구해질 수 있다. 세션은 동일 호스트에 대한 URL의 순차적인 패턴으로서 각 URL에 대한 참조 지속시간은 다음과 같다.

$$S[k].dt = S[k+1].timestamp - S[k].timestamp$$

s[k].dt : 세션에서의 k번째의 URL에서 보낸시간 ..(수식 3)

여기에서 $S[k].dt$는 세션상태의 k번째에 해당하는 URL에서 소비한 시간으로 이것은 웹서버 로그에 기록되는 URL 요청시간을 이용하여 산출한다. 즉 다음 URL 요청시간으로부터 현재 URL 요청시간의 차이를 계산하면 된다.

3.2. 유사도 비교

기존에 트랜잭션의 이진벡터를 적용하였을 때의 유사도 비교는 트랜잭션이 가지는 전체 URL에 대해 페이지 접속패턴이 서로 일치하는 URL의 비율로 계산되어졌다.

다음의 (수식 4)는 본 논문에서 제안하는 유사도 비교식으로서 트랜잭션 t1, t2 에 대하여 앞에서 제시된 사용자 관점의 흥미도를 어떻게 반영하는지를 보인다.

$$sim(t_1, t_2) = \frac{(|t_1 \cap t_2|) - (|t_1 w - t_2 w| \geq threshold)}{\sqrt{|t_1| \bullet |t_2|}} \quad \text{... (수식 4)}$$

제안하는 비교방법에서도 트랜잭션의 전체 URL에 대해 서로 참조가 일치하는 URL의 비율을 사용한다. 하지만 여기에서는 각 URL에 보인 사용자흥미도의 차이에 대한 제한값(threshold)을 두었다. 즉 앞에서 제시된 흥미도가중치를 이용하여 두 트랜잭션간 각 URL의 흥미도의 차이를 먼저 계산한다. 그리고 계산된 각 URL 흥미도의 두 트랜잭션간 차이가 미리 정해진 제한 값 이하의 경우에 대해서만 유사하다고 인정하는 것이다.

같은 URL접근 패턴을 가지는 경우라 해도 사용자 관심이 전혀 다른 트랜잭션이 존재할 수 있다. 이런 경우 이진가중치 벡터 유사성 비교의 경우는 단지 두 트랜잭션 모두 URL을 참조했다는 사실만으로 트랜잭션의 특징과는 전혀 다르게 유사성을 높게 책정하는 경우가 발생한다. 다음의 경우에서 그 차이점을 보이고있다. 동일한 패턴의 페이지 참조에 대한 관심의 차이가 다음과 같은 경우 이진가중치와 제안하는 흥미도 가중치를 적용하였을 경우 두 트랜잭션간 유사성 비교를 수행하였다.

threshold : 0.5

경우1						경우2					
t1	0.2	0	0.2	0.5	0.1	t1	0.1	0	0.3	0.5	0.1
t2	0.1	0.1	0.1	0.7	0	t1	0.6	0.1	0.1	0.2	0
t1w-t2w	0.1	x	0.1	0.2	x	t1w-t2w	0.5	x	0.2	0.3	x

표 1 사용자 흥미도에 의한 차이

이때 각 페이지의 시간 점유비율이 다음의 표 1과 같다고 하자. 이때 두 트랜잭션을 살펴보면 URL 참조의 비트 패턴은 동일하나 사용자관심측면에서 많은 차이가 있다는 것을 알 수 있다. 즉 두 트랜잭션은 상당히 다른 성향을 가지고 있다. 이럴 때 흥미도의 차이가 현격한 차이를 보이는 URL, 즉 제한 값(threshold) 이상의 URL을 유사성 비교에서 제외시킴으로써 실제 유사한 정도보다 높게 측정될 수 있는 두 트랜잭션사이의 유사성을 조율한다. 앞의 두 가지의 경우에 대하여 다음의 제안된 정의를 적용하여 각 트랜잭션의 유사성을 나타내보면 다음과 같은 차이가 있음을 알 수 있다.

$$sim(t_1, t_2) = \frac{(|t_1 \cap t_2|) - (|t_1 w - t_2 w| \geq threshold)}{\sqrt{|t_1| \bullet |t_2|}}$$

경우1 의 유사도 : 3/5

경우2 의 유사도 : 2/5

이때 제한 값(threshold)의 수치는 웹사이트의 특성에 맞게 조정 되어져야할 필요가 있다.

3.3. 페이지 특성을 고려한 유사도 비교

웹사이트의 각 URL을 그 특성에 따라 분류될 수 있다. 예를 들면 사용자가 원하는 컨텐츠를 가지는 페이지와 그 페이지로 가기까지의 경로를 제공하는 페이지가 있을 수 있다.

이러한 페이지의 특성은 유사성을 비교하는데 중요한 비중을 가질 수 있다. 예를 들어 각 컨텐츠 페이지에 대한 참조시간의 비율은 경로제공 페이지에 비해 상대적으로 크다. 이는 경로제공 페이지에서의 참조 시간의 의미는 비교적 중요하지 않다는 것을 의미하며 다음과 같이 정리할 수 있다.

- 경로제공 페이지
  : 접속패턴의 의미비중 > 사용자의 관심정도
- 컨텐츠 페이지
  : 접속패턴의 의미비중 ≤ 사용자의 관심정도

이러한 특징에 따라 유사성비교시 페이지 특성에 따라 경로페이지의 경우는 이진가중치 유사도 비교로, 컨텐츠 페이지의 경우는 사용자의 관심을 적절히 반영하는 흥미도가중치 유사도 비교로서 보다 정밀한 트랜잭션간 유사성 비교를 가능하게 한다..

4. 성능평가

본 논문에서 제안한 사항을 다음과 같이 성능 평가하였다. 대학의 연구실 사이트를 모델로 한 약 20개의 URL로 구성된 사이트를 만들고 이 사이트에 대하여 웹 탐색을 하도록 수행후 그 결과 얻어진 웹서버로그에 대한 데이터 사전처리와 트랜잭션으로의 변환을 수행하였다. 각 방식의 유사도 비교 결과를 보이기 위하여 10개의 트랜잭션에 대해 이진가중치, 흥미도가중치 그리고 혼합방법의 3가지로 구분하여 유사도 비교 수행하고 그 결과를 표 2와 같이 나타내었다. (예 : 14,12,12) 여기에서는 제한 값 0.3을 실제의 웹 탐색 행위에 있어 명확한 관심의 차이를 나타낼 수 있는 수치라 가정하였고 그 결과 표 2와 같은 각 방식별 유사도의 차이를 보일 수 있었다.

threshold : 0.3

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
t1	20,20 ,20	14,12 ,13	12,12 ,12	16,15 ,15	6, 5, 6	13,13 ,13	7, 6, 7	20,16 ,17	10, 9, 9	8, 8, 8
t2	14,12 ,13	20,20 ,20	18,18 ,18	12,12 ,12	6, 6, 6	5, 5, 5	7, 6, 7	14,12 ,12	12,12 ,12	9, 9, 9
t3	12,12 ,12	18,18 ,18	20,20 ,20	14,14 ,14	8, 8, 8	7, 6, 6	7, 6, 7	12,11 ,11	11,11 ,11	8, 8, 8
t4	16,15 ,15	12,12 ,12	14,14 ,14	20,20 ,20	6, 5, 6	13,11 ,11	9, 8, 9	16,13 ,13	9, 9, 9	9, 8, 8
t5	6, 5, 6	6, 6, 6	8, 8, 8	6, 5, 6	20,20 ,20	11,10 ,11	15,13 ,15	6, 4, 5	14,13 ,14	14,13 ,14
t6	13,13 ,13	5, 5, 5	7, 6, 6	13,11 ,11	11,10 ,11	20,20 ,20	14,13 ,14	13,13 ,13	13,12 ,12	15,14 ,15
t7	7, 6, 7	7, 6, 7	7, 6, 7	9, 8, 9	15,13 ,15	14,13 ,14	20,20 ,20	7, 6, 7	15,14 ,15	17,15 ,16
t8	20,16 ,17	14,12 ,12	12,11 ,11	16,13 ,13	6, 4, 5	13,13 ,13	7, 6, 7	20,20 ,20	10, 9, 9	8, 8, 8
t9	10, 9, 9	12,12 ,12	11,11 ,11	9, 9, 9	14,13 ,14	13,12 ,12	15,14 ,15	10, 9, 9	20,20 ,20	16,16 ,16
t10	8, 8, 8	9, 9, 9	8, 8, 8	9, 8, 8	14,13 ,14	15,14 ,15	17,15 ,16	8, 8, 8	16,16 ,16	20,20 ,20

표 2 유사도 비교표

본 논문에 관한 비교평가의 의미로서 클러스터의 수와 그 중심을 미리 정하고 이를 직접 클러스터링을 수행하였다. 먼저 트랜잭션을 3개의 클러스터로 나누고 이들의 중심을 임의로 t2, t6, t10으로 하고 각 트랜잭션을 가장 가까운 클러스터에 할당하였다.

이진가중치 유사도	
t1	t2
t2	t2
t3	t2
t4	t6
t5	t10
t6	t6
t7	t10
t8	t2
t9	t10
t10	t10

흥미도가중치 유사도	
t1	t6
t2	t2
t3	t2
t4	t2
t5	t10
t6	t6
t7	t10
t8	t6
t9	t10
t10	t10

혼합방식	
t1	t2, t6
t2	t2
t3	t2
t4	t2
t5	t10
t6	t6
t7	t10
t8	t6
t9	t10
t10	t10

표 3 트랜잭션 클러스터링

각 방식에 따라 클러스터링의 결과가 표3과 같이 달라질 수 있음을 알 수 있으며 실제 트랜잭션의 패턴을 비교하여보면 그 정확성을 차이를 확인할 수 있다. 이처럼 제안된 흥미도 가중치 유사도 비교방법은 트랜잭션의 소속 변경을 발생시킬 수 있는데. 이는 명확하게 구분되는 트랜잭션의 클러스터링에 관여하기보다 각 클러스터의 경계부근에 놓여지는 트랜잭션에 대해 정밀하게 소속을 밝혀주는 것에 많은 영향을 미친다 할 수 있다.

5. 결론

본 논문에서는 웹 트랜잭션들에 대한 클러스터링의 정확성을 높이기 위한 흥미 가중치 적용 유사도 비교방법을 제안하였다. 제안한 방법은 이진 가중치 유사도 비교 방법의 정확성 문제를 해결하였고 웹 트랜잭션에 흥미도 가중치를 부여하여 사용자의 관심정도 까지를 고려한 유사도 비교를 수행할 수 있게 하였다. 또한 탐색 패턴과, 사용자 관심의 중요성을 적절히 반영하여 유사도 비교를 수행할 수·있게 하였다. 이것으로 사용자의 유사한 성향을 보다 세밀하게 비교 할 수 있게 하였으며 이것의 정확성을 실제 각 방식에 대한 유사도 비교를 수행하여 결과를 분석해봄으로서 보다 정확한 클러스터링이 유도되는 것을 보였다.

향후 연구로는 여러 가지 클러스터링 알고리즘에 제안한 방식을 실제 적용하여보고 보다 정확한 결과를 얻을 수 있는 방법을 적용하고 이것을 통해 추출된 클러스터로부터 보다 효과적으로 추천집단을 생성하는 알고리즘을 연구하여 웹사이트에 실제 적용하는 것을 목적으로 하고 있다.

참고문헌

[1] R. Cooley and J. Srivastava, "Automatic Personalization Based On Web Usage Mining," Communications of the Association of Computing Machinery (CACM), pp. 142-151, August 2000

[2] Feng Taoand and Murtagh. K, "Towards knowledge discovery from WWW log data," Proceedings of the The International Conference on Information Technology : Coding and Computing (ITCC) , pp. 302 -307, 2000

[3] Robert Cooley, Pang-Ning Tan and Jaideep Srivastava, "Discovery of Interesting Usage Patterns from Web Data," WEBKDD, pp. 163-182, 1999

[4] B. Mobasher, H. Dai and T. Luo, "Discovery of Aggregate Usage Profiles for Web Personalization," Proceedings of the Web Mining for E-Commerce Workshop (WebKDD), August 2000

[5] Alex G. Buchner, Maurice D. Mulvenna, "Discovering internet marketing intelligence through online analytical Web usage mining," ACM SIGMOD Record, Vol. 27, No 4, pp. 54-61, 1998

[6] Rakesh Agrawal, Ramakrishnan Srikant, "Fast Algorithms for Mining Association Rules," Proc. 20th Int. Conf. Very Large Data Bases, VLDB, pp. 487-499, Sept 1994

[7] E-H. Han, G. Karypis, V. Kumar, and B. Mobasher, "Clustering based on association rule hypergraphs", DMKD, 1997

[8] Sanjay Kumar Madria, Sourav S. Bhowmick, Wee Keong Ng and Ee-Peng Lim "Research Issues in Web Data Mining," DaWaK, pp. 303-312, 1999

[9] R. Cooley, B. Mobasher, and J. Srivastava. "Data preparation for mining world wide web browsing pattern," Knowledge and Information Systems, Vol. 1 No. 1, pp. 5-32, 1999

[10] Shahabi,C., A.Zarkesh, and J.Adibi, and V.Shah, "Knowledge Discovery from Users Web-PageNavigation," RIDE, 1997