웹 트랜잭션 클러스터링의 정확성을 높이기 위한 흥미가중치 적용 유사도 비교방법

강 태 호† · 민 영 수† · 유 재 수††

요 약

최근 들어 웹 사이트 개인화(Web Personalization)에 관한 연구가 활발히 진행되고 있다. 웹 개인화는 클러스터링과 같은 데이터 마이닝 기법을 이용하여 각 사용자에게 가장 흥미를 가질만한 URL 집합을 예측하는 것이라 할 수 있다. 기존의 클러스터링을 이용한 방식에서는 웹 트랜잭션들을 웹 사이트의 각 URL들에 방문했는지 안했는지를 나타내는 비트 벡터(bit vector)로 표현하였다. 그리고 이들 비트 벡터의 방문 패턴이 일치하는 정도에 따라 유사성을 결정하였다. 하지만 이것은 유사한 성향을 가지는 웹 트랜잭션을 클러스터링 하는데 있어 사용자의 흥미를 배제하고 단순히 방문 여부만을 반영하게 되는 문제점이 발생하게 된다. 즉 방문 목적 또는 성향이 유사하지 않은 웹 트랜잭션들을 같은 그룹으로 분류할 가능성이 존재하게 된다. 이에 본 논문에서는 기존의 비트 벡터를 이용한 트랜잭션 모델을 사용자의 흥미도(Interestingness)를 반영할 수 있도록 보완하여 새로운 웹 트랜잭션 모델을 제시하고 흥미가중치를 적용한 유사도 비교방법을 제안한다. 그리고 성능평가를 통하여 제안한 방법이 기존 방법에 비해 클러스터링의 정확성을 높임을 보인다.

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

Tae Ho Kang† · Young Soo Min† · Jae Soo Yoo††

ABSTRACT

Recently, many researches on the personalization of a web-site have been actively made. The web personalization predicts the sets of the most interesting URLs for each user through data mining approaches such as clustering techniques. Most existing methods using clustering techniques represented the web transactions as bit vectors that represent whether users visit a certain URL or not to cluster web transactions. The similarity of the web transactions was decided according to the match degree of bit vectors. However, since the existing methods consider only whether users visit a certain URL or not, users' interestingness on the URL is excluded from clustering web transactions. That is, it is possible that the web transactions with different visit purposes or inclinations are classified into the same group. In this paper, we propose an enhanced transaction modeling with interestingness weight to solve such problems and a new similarity measuring method that exploits the proposed transaction modeling. It is shown through performance evaluation that our similarity measuring method improves the accuracy of the web transaction clustering over the existing method.

키워드 : 웹 개인화(Web Personalization), 웹 트랜잭션 클러스터링(Web Transaction Clustering), 유사도 비교(Similarity Measurement)

1. 서 론

인터넷의 발달은 수년 사이 많은 분야에서 엄청난 변화를 가져오게 했으며 이제 웹은 거의 모든 산업의 발달에 있어서 매우 중요한 위치를 차지하게 되었다. 특히 비즈니스 거래가 웹상에서 이루어질 수 있다는 점에서 오는 편리함과 신속성은 전자 상거래의 급속한 성장을 이끄는 핵심 요인이 되어왔다. 이에 따라 웹을 이용한 마케팅에서의 효율증대의 중요성이 부각되고 웹 사이트의 1 : 1 마케팅이라는 목적의 웹 개인화에 대한 연구가 활발히 진행되고 있다[1, 4, 6, 7].

웹 사이트에서 개인화는 사용자와 웹 사이트간의 일련의 가치교환 과정이다. 사용자가 자신의 선호도 또는 관심 등을 등록하면 웹 사이트는 이에 맞게 사용자에게 알맞은 정보를 제공한다[1, 4]. 이렇듯 웹 사이트의 개인화를 위해서는 고객에 관한 정보를 필요로 하고 이러한 고객 데이터를 적절한 방법을 통해 분석해야한다. 이러한 분석을 통해 고객의 성향을 알 수 있으며, 고객들이 바라는 것들에 대해

※ 본 연구는 한국과학기술기획평가원의 생물정보학연구개발 2차 사업의 지원에 의해 수행되었음.
† 준 회 원 : 충북대학교 대학원 정보통신공학과
†† 종신회원 : 충북대학교 전기전자 및 컴퓨터공학부 교수
논문접수 : 2002년 10월 9일, 심사완료 : 2004년 1월 9일

적절히 예측하고 대응할 수 있다. 기존에는 이러한 고객정보를 얻기 위해 사용자가 직접 기입한 프로파일이나, 설문조사를 통한 데이터들로부터 얻어내는 방법이 거의 대부분이었다. 하지만 웹 환경은 이러한 중요한 정보를 개인의 프라이버시 침해의 가능성이나, 별도의 요구에 따른 번거로움을 주지 않고도 자연스럽게 알 수 있는 방법을 제공한다.

사용자가 웹 사이트 방문을 시작하면서부터 끝나기까지의 모든 사용자의 행위를 웹 서버는 그대로 기록하게 되는데 이렇게 기록된 데이터가 바로 웹 서버 로그이다. 이런 웹 서버 로그를 분석함으로서 여러 가지 형태의 중요한 사용자 정보를 유추해 낼 수 있게 되는데, 이들 데이터를 분석하는데 있어 기존의 데이터 마이닝 기법을 활용한 방법들이 현재 활발히 연구되고 있으며 이러한 것들을 웹 마이닝[12]이라 한다. 현재 이러한 웹 마이닝을 통해 웹 사이트의 개인화를 이루려는 노력들이 많이 시도되고 있다.

사용자의 웹 탐색 행동은 웹 사이트 개인화에 이용될 수 있다. 여기에서의 웹 사이트의 개인화 목적은 사용자에게 기대에 부합하는 웹을 제공하는 것이라 할 수 있다. 즉 사용자들에게 맞춤 내용을 미리 예측하여 이를 추천하는 방법이 일반적으로 사용된다. 이러한 추천을 위한 방법으로 첫 번째 페이지별 연관 관계를 파악하여 자주 발생하는 빈발항목 집단을 알아내는 방법이 있다[8]. 이는 웹 페이지들 사이의 연관성을 이용하여 사용자들에게 추천 페이지 집합을 제공해 주는 방식을 사용한다. 하지만 이렇게 제공되는 추천 페이지는 사용자들의 연관성을 고려하지 못한 단지 웹 페이지들간의 연관성만을 이용하여 추천받게 되는 경향이 있다. 이러한 문제를 해결하기 위한 두 번째 방법으로 웹 탐색 패턴이 유사한 사용자들을 적절히 클러스터링 하고 같은 그룹에 포함된 사용자들의 브라우징 패턴을 발견하고 분석하는 것을 통해 추천 집단을 생성하는 방법이 있다[1]. 이중에서 유사한 성향의 트랜잭션 클러스터로부터 추천을 유도하는 클러스터링을 이용한 개인화 방법의 경우 클러스터링의 기초가 되는 유사도 비교는 그 비교의미와 그에 따른 정확성이 매우 중요하다. 기존에는 웹 트랜잭션 데이터를 방문 또는 비방문의 의미로서 이진가중치 벡터로 모델링 하여 이들의 비트 패턴에 대한 유사도를 비교하는 방식이 사용되었다[1]. 이처럼 이진가중치로 표현할 수 있는 방문 패턴은 사용자 성향을 알 수 있는 효율적인 방법이긴 하나, 비트 벡터를 이용함으로 인해 사용자가 보이는 관심이나 기타 트랜잭션의 특성이 무시되는 경우가 발생하게 되어 정확성이 결여되는 문제점이 있다. 따라서 본 논문에서는 이러한 기존방법의 문제점을 지적하고 이를 보완할 수 있도록 웹 트랜잭션에 새로운 가중치를 제안하여 모델링하고 이를 통해 트랜잭션의 특성에 대한 비중을 부여할 수 있는 유사도 비교방법을 제안한다. 이를 통해 보다 의미있고 정확한 클러스터링을 수행함으로써 활용도 높은 추천 집단 생성의 기반을 마련하고자 하는데 그 목적이 있다.

이에 본 논문에서는 먼저 2장에서 관련연구로 웹 로그데이터를 마이닝 기법을 이용해 분석하는 웹 개인화과정에 대한 전반적인 소개를 한다. 또한 본 논문에서 제시하고자 하는 트랜잭션 클러스터링의 기본이 되는 유사도 비교에 대한 기존의 방법과 그에 대한 문제점을 제시하고, 3장에서 이러한 문제점을 보완하고자 제안된 흥미도가중치 트랜잭션 모델을 제안하고 이를 적용한 유사도 비교방법을 제시한다. 그리고 4장에서 더욱 세밀한 유사도 비교를 위해 웹 사이트의 각 페이지를 특성에 따라 분류하고 이들 분류된 페이지에 적합하도록 혼합된 유사도 비교를 제시한다. 마지막으로 5장에서는 비교평가로서 간단한 사이트 및 대규모 웹 로그 데이터를 가지고 각 방식별 트랜잭션 유사도 비교의 차이를 보이고 결론을 맺는다.

2. 관련 연구

이 장에서는 관련연구로서 전반적인 웹 개인화 과정을 웹 서버 로그로부터 여러 가지 정보를 발견하기까지 처리되어야 할 과정들을 소개하며, 본 논문과 관련된 트랜잭션 데이터로의 변환과정과 클러스터링의 기초가 되는 기존의 이진가중치 유사도 비교에 대한 소개를 통해 기존방법에서 개선되어야 할 사항을 언급한다.

2.1 웹 개인화(Web Personalization) 과정

개인화는 크게 나누어 오프라인 처리단계와 온라인 처리단계의 두 가지 과정으로 나뉘어 수행될 수 있다. 먼저 오프라인단계에서 원시 데이터(웹 로그)를 분석이 가능한 형태로의 변환을 수행하는 데이터 사전 준비(Data Preparation)과정[1, 2]와 변환된 데이터로부터 여러 가지 마이닝 기법을 이용하여 특정패턴이나 데이터간의 관계들을 도출해내는 과정으로 이루어진다. 그리고 온라인 단계에서는 웹 서버에 접속한 사용자의 세션을 구분하고 오프라인에서 분석된 내용을 바탕으로 사용자에게 부합하는 추천 집단을 생성하여 이를 추천해주는 과정을 수행함으로써 웹 사이트의 개인화를 수행한다. 이 중 패턴 발견은 웹 마이닝에서 가장 중요한 단계라 할 수 있다. 여러 가지 기법을 이용하여 얻어지는 사용자 데이터 사이에서 패턴을 발견하는 과정으로 찾아낸 패턴을 이용하여 사용자 성향을 파악하고 행동을 예측하는 과정을 수행한다. 이 과정에서는 연관성 규칙(Association rule), 순차적 패턴(Sequential pattern), 군집화(Clustering), 분류(Classification) 등의 기법이 사용될 수 있다. 이중 웹 추천에 일반적으로 많이 사용되어지는 기법이라 할 수 있는 연관성 규칙을 이용해 발견해낸 빈발항목집단(Frequent Itemset)을 찾아내 이를 추천에 이용하는 방법[8]과 유사한 성향을 보이는 사용자 트랜잭션들을

적절히 클러스터링 하여 교차 추천에 이용하는 방법[1, 4] 등이 많이 연구되어지고 있다.

2.2 웹 서버 로그와 데이터 정제

웹 마이닝에서 패턴발견 이전에 수행되어져야 하는 단계는 바로 앞에서 언급했던 데이터 사전준비 단계이다. 이것은 원시 데이터의 정제 및 사용자 세션의 구분과 트랜잭션의 규명을 통해 분석할 수 있는 형태의 의미 있는 트랜잭션 데이터로 변환하는 것을 포함한다. 먼저 원시데이터 즉 웹 서버 로그의 구조를 알아보고 필요한 항목을 추출하는 과정을 살펴본다. 사용자가 웹 서버를 통해 접근을 시도할 경우 사용자의 모든 작업이나 행위들이 기록되는데 이것이 웹 로그이다. 일반적으로 특정 웹 페이지를 보기 위한 사용자의 요구로, 웹 로그는 HTTP 프로토콜의 연결속성 때문에 해당 웹 페이지뿐만 아니라 해당 웹 페이지와 관련된 이미지 파일, 이미지데이터, Include 파일 등이 한번의 사용자 행동으로 요청되어지며 웹 로그에 같이 기록된다. (그림 1)은 거의 모든 웹 로그에서 지원하고 있는 Common Log Format(CLF)형식의 로그이다.

(그림 1)에서와 같이 웹 로그는 사용자의 방문 시마다 원격 사용자의 호스트 이름 또는 IP 주소, 요청한 시간, HTTP 방식(GET, POST, 기타)과 요청된 파일, HTTP 버전, HTTP 응답의 상태코드(예를 들어 200 : 성공적인 웹 페이지 접근, 404 : 없는 페이지 접근에러), 그리고 전송된 파일의 바이트 수의 순서로 로그에 기록한다. 이처럼 로그 파일이 주어지면 첫 번째 단계로 데이터 정제를 해야 한다. 데이터 정제는 페이지 뷰 당 하나의 성분만 남기고 중복적인 모든 파일을 제거한다는 것을 의미하며 이는 사용자의 요청과 직접적인 관련이 없거나 분석에 있어서 불필요한 파일이나 문서를 제거하는 작업으로, 요청된 페이지에 포함된 이미지나 비디오 클립 등이 이에 해당될 수 있다. 이처럼 불필요한 페이지에 대한 필터링은 웹 로그의 요청된 파일에 대한 확장자를 통해 구분되어질 수 있다.

2.3 사용자 세션 구분과 트랜잭션 모델

다음으로 웹 로그로부터 사용자의 세션을 도출해내는 과정이 필요하게 된다. 사용자 세션의 구분은 트랜잭션을 규명하는데 있어서 중요한 요소로 작용하게 된다. 사용자 세션이란 한 사용자가 웹 사이트에 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. 이때 세션에서는 빈도가 아주 낮은 트랜잭션들과 아주 낮은 수준의 페이지 참조들을 제거하기 위해 여과되기도 하는데, 이러한 유형의 필터링은 데이터로부터 노이즈를 제거하는데 매우 중요할 수 있으며, 클러스터링 시 차원을 감소시킬 수 있다.

<표 1>로부터 사용자 세션을 구분한 결과를 <표 2>에서 보이고 있다. 세션을 구분할 때 동일 호스트에 대해 방문 페이지들 사이의 시간적인 공백이 현저하게 많은 경우가 존재할 수 있는데 이때 세션의 연속인지 또는 새로운 방문에 의한 세션의 시작인지를 결정해야 한다.

〈표 1〉 사용자 방문순서

시 간	사용자 호스트 (사용자 ID)	요청페이지
18/Feb/2002 : 00 : 00 : 01	host1	A.html
18/Feb/2002 : 00 : 00 : 04	host2	B.html
18/Feb/2002 : 00 : 00 : 09	host2	C.html
18/Feb/2002 : 00 : 00 : 10	host3	D.html
18/Feb/2002 : 00 : 00 : 12	host1	E.html
....	....	....

〈표 2〉 세션구분

사용자 호스트 (사용자 ID)	세 션
host1	A.html, E.html, ..
host2	B.html, C.html, ..,
host3	D.html, ...
....	....

일반적으로 웹 트랜잭션의 구분에 있어서 세션의 연속으로 인정할 제한시간은 30분으로 정하고 있다. 이 제한시간으로 동일 호스트에 대해 새로운 세션을 구분한다. 즉 한 웹 페이지에서 다른 웹 페이지로 연결과정에서 제한시간이

```
210.115.171.200 … [18/Feb/2002 : 18 : 01 : 48 +0900] "GET / HTTP/1.1" 304 -
210.115.171.200 … [18/Feb/2002 : 18 : 01 : 48 +0900] "GET /home2001/top_logo.shtml HTTP/1.1" 200 8844
210.115.171.200 … [18/Feb/2002 : 18 : 01 : 48 +0900] "GET /home2001/top_menu.html HTTP/1.1" 304 -
210.115.171.200 … [18/Feb/2002 : 18 : 01 : 48 +0900] "GET /home2001/MENU.gif HTTP/1.1" 304 -
210.115.171.200 … [18/Feb/2002 : 18 : 01 : 48 +0900] "GET /home2001/background.gif HTTP/1.1" 304 -
210.115.171.196 … [18/Feb/2002 : 18 : 01 : 48 +0900] "GET /board/main.cgi?board=공지사항 HTTP/1.1" 200 14222
210.115.171.196 … [18/Feb/2002 : 18 : 01 : 49 +0900] "GET /board/image/face/bitimg16.gif HTTP/1.1" 304 -
210.115.171.196 … [18/Feb/2002 : 18 : 01 : 49 +0900] "GET /board/image/button/2/rootkey.gif HTTP/1.1" 304 -
210.115.171.124 … [18/Feb/2002 : 18 : 01 : 50 +0900] "GET /home2001/lecture.html HTTP/1.1" 200 3914
```

(그림 1) 웹 로그

상의 시간 차이를 보인다면 새로운 세션으로 간주하는 것이다. 이러한 세션을 구분하는 인정한도 시간 간격은 각 사이트의 특성에 따라 적절히 정해져야 할 필요가 있다[2]. 이렇게 구분된 세션은 다음과 같이 두 가지의 경우로 변환을 생각할 수 있다. 먼저 여러 페이지의 참조로 이루어지는 하나의 트랜잭션으로 그리고 하나의 페이지 참조로 이루어지는 많은 트랜잭션의 집합으로 변환될 수 있다.

여기에 사용자 웹 탐색 행동의 특징에 근거하여 각 페이지 참조를 컨텐츠 페이지 참조(content page reference), 경로 페이지 참조(navigational page reference) 등으로 범주화하여 적용할 수도 있는데, 이런 식으로 그 응용에 따라 다른 유형의 트랜잭션들이 사용자 세션 파일로부터 얻어질 수 있다.

2.4 트랜잭션 클러스터링

먼저 데이터 사전처리가 이루어졌다고 보고, 사전 처리된 로그에 n개의 URL 집합과 m개의 사용자 트랜잭션 집합을 다음과 같이 정의한다.

$$U=\{url_1, url_2, \cdots, url_n\} \qquad T=\{t_1, t_2, \cdots, t_n\}$$

집합 T의 요소인 t_i은 ($t_i \in T$) 집합 U의 하위집합이다. 여기에서 규명된 사용자 트랜잭션들을 클러스터링을 수행할 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 비트 벡터(bit vector) 표현은 다음과 같다.

$$\vec{t} = \langle u_1^t, u_2^t, \cdots, u_n^t \rangle \quad u_i^t = \begin{cases} 1, & if\ url_i^t \in t \\ 0, & otherwise \end{cases}$$

예를 살펴보면 다음과 같다. 우선 각 세션을 트랜잭션 데이터화 즉 하나의 레코드로 만드는 작업이 필요하다 이를 위하여 각 세션에 방문 페이지를 등록한 후 페이지 방문여부에 따라 방문 : 1 비방문 : 0 이라는 이진가중치(Binary Weight)를 부여하는 작업<표 3>을 하게 된다. 이와 같은 방식에서 URL 참조의 발생여부는 사용자의 흥미를 나타내줄 수 있는 좋은 지표이다.

〈표 3〉 트랜잭션 모델

트랜잭션	페이지A	페이지B	페이지C	페이지D	페이지E
t_1	1	1	0	1	0
t_2	1	0	1	0	1

이를 근거로 웹 사이트 URL의 방문여부에 따라 이진 벡터(binary vector)로 표현함으로서 두 트랜잭션 사이에 동시발생 패턴의 유사성 및 비유사성에 대한 비교가 가능해진다. 이처럼 웹 트랜잭션 데이터의 특성상 이진 벡터 표현은 매우 효과적일 수 있다. 하지만 URL의 방문여부만으로 사용자 트랜잭션의 유사성을 가늠하기엔 부족함이 많다. 여기에서 발생하는 정확성 문제를 3장에서 언급한다.

2.5 트랜잭션 유사성 비교

앞의 과정에서 얻어진 트랜잭션들을 클러스터링하기 위해서 두 트랜잭션들 사이의 거리 측정이 필요하게 되는데 웹 트랜잭션의 경우 두 트랜잭션 t, s가 있을 때 이진 벡터의 경우 유사성 sim(t, s)는 다음과 같이 트랜잭션에서 일치하는 항목의 크기로 표현한다[1].

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

위 수식에 따른 유사도 비교결과는 클러스터링의 기초로 활용되어질 수 있으며 여기에 여러 가지 다양한 클러스터링 알고리즘이 적용된다.

트랜잭션 벡터의 표현과 이들 트랜잭션 사이의 유사성 비교에 있어서 이진가중치 벡터는 사용자의 웹 탐색 패턴을 효과적으로 표현할 수 있는 반면에 트랜잭션의 유사성을 가늠하기엔 그 의미가 충분하지 않으며 이에 따른 정확성에 문제가 발생할 수 있다. 비슷한 사용자 성향을 클러스터링 해서 이를 분석하여 추천집단을 생성하는 시스템에서 중요한 것은 유사한 트랜잭션을 구분할 수 있는 정확성이다. 이를 위해 트랜잭션 사이의 유사성 비교에 있어서 트랜잭션 비교 의미가 충분하여야 한다. 이에 본 논문에서는 이런 문제점을 지적하고 보완하여 비교의 의미를 충분히 부여하고자 사용자의 관심을 포함할 수 있는 가중치를 제안하고, 제안한 가중치를 이용한 유사성 비교 방법을 제시하여 보다 정확한 클러스터링을 유도할 수 있게 한다.

3. 흥미 가중치를 적용한 트랜잭션 유사도 비교 방법

웹 트랜잭션에 대하여 방문 패턴에 대한 유사성 비교는 클러스터링에 있어서 중요한 요소이긴 하나 그 의미가 충분하지 못하다. 그 이유는 웹 사이트를 방문한 사용자는 각 페이지에 대해 동일한 관심을 가지고 접근하는 것은 아니기 때문이다. 한 예로 사용자가 웹 사이트를 접근할 때 원래 해당 사이트에서 얻고자 하는 또는 원하는 컨텐츠에 접근하기 위하여 그 이전에 많은 수의 웹 탐색 경로를 제공하는 페이지를 거치게 되는데 이때 이진 벡터는 모든 페이지 접근에 대해 동일한 의미로 취급하고 있다. 물론 탐색경

로 페이지에 대한 접근이 의미가 없는 것은 아니다. 이러한 탐색패턴 또한 웹 사이트 구조를 최적화 시키고자 하는 같은 응용에 있어서는 매우 중요한 요소가 된다. 하지만 전반적인 웹 트랜잭션의 유사성을 비교하는데 있어서는 그렇게 큰 비중을 차지하지 않으며 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 클러스터링할 수 있을 것이다. 이러한 이유로 이번 장에서는 기존의 이진 벡터의 문제점을 보완하고자 새로운 가중치를 제시한다.

3.1 흥미 가중치(Interestingness Weight)를 적용한 트랜잭션 모델

웹 사용자가 참조한 페이지에 대한 관심을 어떤 방식으로 알아낼 수 있는가 하는 것에 대해 여러 가지의 접근방법이 있을 수 있으나 가장 일반적으로 접근할 수 있는 방법은 다음의 두 가지로 나누어 볼 수 있다. 하나는 웹 페이지 참조의 중복 횟수를 가지고 판단하는 방법이다. 웹 사용자는 한 페이지를 한번 참조하는 경우도 있으나 반면 여러 번 참조할 수도 있다. 이때 여러 번 참조하는 행위는 해당 페이지에 대한 어느 정도의 관심을 표출하는 것이라 할 수 있다. 하지만 이진 벡터에서는 이러한 페이지 중복 참조의 의미는 무시된다. 또 다른 하나의 접근방법은 사용자가 참조한 페이지에 머문 시간을 가지고 관심정도를 예측할 수 있다. 해당 URL에서 보낸 시간은 사용자의 관심을 추측할 수 있는 좋은 방법이라 할 수 있다 [11]. 일반적으로 앞에서 언급한 탐색경로 페이지의 경우는 그 특성에 따라 페이지 참조 시간이 길지 않다. 하지만 컨텐츠를 제공하는 페이지의 경우 그 경과 시간이 상대적으로 길다. 그리고 같은 컨텐츠 페이지의 경우라 하더라도 사용자가 찾고자 했던 페이지와 중간에 경유하는 페이지의 경우는 해당 페이지에서 머문 시간의 차이를 보이게 된다. 이는 성능 평가에서 각 트랜잭션이 가지는 페이지별 시간 소요정도에서도 알 수 있다. 이러한 특징을 이용하여 본 논문에서는 사용자의 흥미를 표현하는데 있어서 페이지 참조 시간의 비율을 사용한다. 물론 이는 대부분의 웹 사용자가 보이는 일반적인 특징을 이용한 것이지만 예외인 경우도 있을 수 있다. 가령 웹 탐색 도중 발생한 다른 일을 처리한다던가, 자리를 잠시 비운다던가 하는 경우이다. 이런 경우 최대 30분의 세션 종료 시간까지 그 페이지에 시간이 할당될 수 있다. 하지만 이러한 경우는 전체 트랜잭션에 대해 빈번하게 발생하는게 아니므로 이를 무시하기로 한다. 반면에 이러한 예외가 경로제공 페이지에서 발생하는 경우 본 논문의 후반부에서 언급될 혼합방식에서 어느 정도 해결할 수 있다.

제안하는 사용자 흥미도를 사용한 트랜잭션 모델은 다음과 같다. 트랜잭션 $t \in T$ 일 때, 이 트랜잭션은 다음과 같은 이진 벡터의 변형으로 표현될 수 있다.

$$\vec{t} = \langle w_1^t,\ w_2^t,\ \cdots,\ w_n^t \rangle \quad w: \text{흥미가중치}$$

$$w_i^t = \begin{cases} weight, & if\ url_i^t \in t \\ 0, & otherwise \end{cases}$$

기존의 이진 벡터를 변형한 사용자 흥미 관점의 가중치를 부여한 트랜잭션 모델이다. 기존의 이진 벡터를 사용했을 때와 각 URL의 매핑은 그대로 유지한다. 다만 방문 : 1, 비방문 : 0으로 표현하던 방식대신 단지 트랜잭션에서의 각 URL이 가지는 사용자 관심의 비중을 적용한다.

여기에서의 사용자 흥미관점의 가중치는 한 트랜잭션의 전체 탐색 지속 시간에 대한 각 URL에서 소비한 시간의 비율을 사용하며 다음과 같이 정의한다.

$$W[i].Interestingness = W[i].dt \Big/ \sum_{j=1}^{n} W[j].dt \qquad (2)$$

W[i].dt : 트랜잭션의 i번째 URL에서 머문시간

여기에서 $W[i].dt$ 는 트랜잭션내의 i번째 URL에서 보낸 시간으로써 이 시간은 트랜잭션 모델로 변환되기 이전의 세션에서 구해질 수 있다. 본 논문에서는 목적상 하나의 세션을 하나의 트랜잭션으로 간주한다.

세션은 동일 호스트에 대한 URL의 순차적인 패턴으로 구성된다. 여기에서 각 URL에 대한 참조 지속시간은 다음과 같이 정의된다.

$$S[k].dt = S[k+1].timestamp - S[k].timestamp \qquad (3)$$

s[k].dt : 세션에서의 k번째 URL에서 보낸시간

여기에서 $S[k].dt$는 세션 상태의 k번째에 해당하는 URL에서 소비한 시간으로 이것은 웹 로그에 기록되는 URL 요청시간을 이용하여 산출한다. 즉 다음 URL 요청시간으로부터 현재 URL 요청시간의 차이를 계산하면 된다. 또한 중복 참조의 경우 트랜잭션의 URL로의 사상하는 과정 중에서 중복된 시간을 더하게 된다. 이런 과정을 통해 계산된 각 URL별 소비시간으로부터 하나의 트랜잭션에서 전체 소비한 시간의 합에 대한 각 URL의 시간 점유비율의 계산이 가능해지게 되고 산출된 트랜잭션내의 각 URL 시간 점유비율을 바로 사용자가 각 URL에 보인 흥미도로 사용한다. 이 과정을 <표 4>, <표 5>, <표 6>에서 나타내었다.

〈표 4〉 세 션

	URL, 머문 시간(단위 : 분)				
세션 1	A, 1	B, 1	E, 2	G, 5	H, 1
세션 2	A, 1	C, 2	D, 2	G, 5	

〈표 5〉 이진가중치 트랜잭션 모델

URL	A	B	C	D	E	F	G	H
$\vec{t_1}$	1	1	0	0	1	0	1	1
$\vec{t_2}$	1	0	1	1	0	0	1	0

〈표 6〉 흥미가중치 트랜잭션 모델

트랜잭션	A	B	C	D	E	F	G	H
$\vec{t_1}$	0.1	0.1	0	0	0.2	0	0.5	0.1
$\vec{t_2}$	0.1	0	0.2	0.2	0	0	0.5	0

지금까지 트랜잭션의 새로운 모델을 제시하고 또한 이를 위해 사용자 관점의 관심정도를 표현할 수 있는 흥미도 라는 가중치를 제안하였다. 이제 트랜잭션 데이터들에 대한 유사성 비교방법 또한 이러한 흥미도를 고려할 수 있도록 새로운 방법이 제시되어야 한다.

3.2 유사도 비교

기존에 트랜잭션의 이진 벡터를 적용하였을 때의 유사도 비교는 트랜잭션이 가지는 전체 URL에 대해 페이지 접속 패턴이 서로 일치하는 URL의 비율로 계산되어졌다.

하지만 본 논문에서 새롭게 제안하는 사용자 관점의 흥미도를 가중치로 가지는 트랜잭션 모델에서는 다음과 같은 유사도 비교식을 가진다. 트랜잭션 t1, t2에 대하여 앞에서 제시된 사용자 관점의 흥미도를 어떻게 반영하는지를 다음 수식에서 보인다.

$$sim(t_1, t_2) = \frac{(|t_1 \cap t_2|) - (|t_1{}^w - t_2{}^w| \geq threshold)}{\sqrt{|t_1| \cdot |t_2|}} \quad (4)$$

제안하는 비교방법에서도 트랜잭션의 전체 URL에 대해 서로 참조가 일치하는 URL의 비율을 사용한다. 하지만 여기에서는 각 URL에 보인 사용자 흥미도 관점의 차이에 대한 threshold를 두었다. 즉 앞에서 제시된 흥미도가중치를 이용하여 두 트랜잭션 사이의 각 URL의 흥미도의 차이를 먼저 계산한다. 그리고 계산된 각 URL 흥미도의 두 트랜잭션 사이의 차이가 미리 정해진 threshold 이하인 경우에 대해서만 유사하다고 인정하는 것이다. 다시 말해 threshold 이상의 차이가 나는 URL에 대해서는 유사하지 않다고 간주하여 제외시키게 된다. 같은 URL접근 패턴을 가지는 경우라 해도 사용자 관심의 방향이 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. 이것은 트랜잭션 사이의 유사성 비교 의미를 떨어뜨리게 하는 것이다. 따라서 제안된 유사도 비교방식에서는 정확한 클러스터링을 유도하기 위하여 트랜잭션 사이의 유사성 비교에 있어서 그 의미부여를 강화함으로서 정확한 유사도 비교를 수행하도록 하고 있다.

다음은 앞에서 언급된 경우에 대한 예를 제시하고 이에 대한 이진가중치 벡터와 제안된 흥미도가중치 벡터의 유사도 비교의 차이를 설명하겠다. 다음과 같은 이진가중치 벡터로 표현된 트랜잭션에 대하여 이진가중치와 제안하는 흥미도가중치를 적용하였을 경우 두 트랜잭션 사이의 유사성 비교를 수행하였다.

$$\vec{t} = \langle u_1^t, u_2^t, \cdots, u_n^t \rangle$$

$$\vec{t1} = \langle 1, 0, 1, 1, 1 \rangle$$

$$\vec{t2} = \langle 1, 1, 1, 1, 0 \rangle$$

$$sim(t_1, t_2) = \frac{(|t_1 \cap t_2|)}{\sqrt{|t_1| \cdot |t_2|}} = \frac{3}{5}$$

이때 각 페이지의 시간 점유비율이 다음의 <표 7>과 같다고 하자. 이때 두 트랜잭션을 살펴보면 URL 참조의 비트 패턴은 동일하나 사용자 관심측면에서 많은 차이가 있다는 것을 알 수 있다. 즉 두 트랜잭션은 상당히 다른 성향을 가지고 있다. 이럴 때 흥미도의 차이가 현격한 차이를 보이는 URL, 즉 threshold 이상의 URL을 유사성 비교에서 제외시킴으로써 실제 유사한 정도보다 높게 측정될 수 있는 두 트랜잭션 사이의 유사성을 조율한다. 이렇게 함으로서 두 트랜잭션에 대해 사용자의 관심까지 반영된 보다 실질적인 유사성 비교가 가능해지고, 정확성을 높일 수가 있게 된다.

〈표 7〉 사용자 흥미도에 의한 차이

경 우 1						경 우 2					
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

threshold : 0.5

앞의 두 가지의 경우에 대하여 다음의 제안된 수식을 적용하여 각 트랜잭션의 유사성을 나타내보면 다음과 같은 차이가 있음을 알 수 있다.

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

경우 1의 유사도 : 3/5

경우 2의 유사도 : 2/5

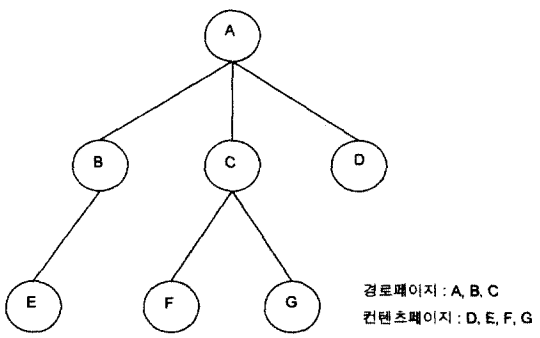
이때 보정의 역할을 수행하는 threshold의 수치는 웹 사이트의 특성에 맞게 평균적인 웹 페이지 접속시간을 가만하여 조정 되어야할 필요가 있다. 이것으로 기존에 웹 트랜잭션 데이터에 대한 이진가중치 모델에서의 문제점을 살펴보고 이를 보완할 수 있도록 본 논문에서 제안된 흥미도가중치 트랜잭션 모델을 가지고 두 트랜잭션 사이의 유사도 비교를 하는 방법에 대하여 설명하였다. 이제 여기에서 한 가지 더 각 URL이 가지는 특성까지를 고려할 수 있는 방법을 다음 장에서 소개한다.

4. 페이지 특성을 고려한 트랜잭션 유사도 비교 방법

웹 사이트의 각 URL은 그 역할에 따르는 특성이 존재한다. 가령 실질적인 사이트의 목적과 결부되는 컨텐츠에 관련된 페이지의 경우나 또는 이러한 컨텐츠 페이지에 도착하기 위한 경로를 제공하는 페이지와 같은 경우를 들 수 있겠다. 이러한 웹 페이지의 특성을 고려하여 웹 트랜잭션의 분석이 이루어진다면 보다 정확한 분석이 될 것이다. 이러한 관점에 따라 웹 사이트의 URL의 특성을 이용한 트랜잭션의 유사성을 비교하는 방법을 두 번째로 제안하고자 한다.

4.1 페이지 특성 분류

웹 사이트의 각 URL을 그 특성에 따라 분류하는 작업은 데이터 사전 처리 단계에서 이루어져야 한다. 이러한 분류 작업은 직접 웹 사이트를 설계한 사항에 의해 수동적으로 분류될 수 있거나 사이트 구조를 이용하여 웹 페이지를 자동으로 분류 할 수도 있다. (그림 2)에서와 같은 간단한 사이트의 구조를 가지고 페이지의 특성을 분류해 보았다.



(그림 2) 웹 페이지 분류

(그림 2)는 사이트의 구조에 따라 페이지의 특성을 예측하여 분류한 것이다. 대부분의 경우 사이트 구조의 가장 하위노드가 웹 사이트의 해당 컨텐츠를 제공하는 컨텐츠 페이지로 구성된다. 이러한 점을 이용하여 경로제공 페이지(Navigation Page)와 컨텐츠 페이지(Content Page)로 구분하였다. 이밖에도 웹 페이지의 특성을 그 특색에 맞게 여러 가지로 구분하여 응용하기도 한다[2]. 하지만 본 논문에서는 경로제공 페이지와 컨텐츠 페이지 두 가지로만 구분하기로 한다.

4.2 특성분류 트랜잭션 모델 및 혼합방식 유사도 비교

웹 페이지의 분류와 함께 중요한 것은 분류된 페이지가 갖는 특성이다. 이에 대해 살펴보면 다음과 같다. 먼저 경로 페이지의 경우 가장 두드러진 특징이라 할 수 있는 것은 거의 모든 트랜잭션에서 볼 때 일반적으로 경로제공 페이지에서 소비하는 시간은 많지 않다, 이것은 소비시간의 차이가 크지 않다는 것을 생각할 수 있다. 이와 같은 사실로 유추해볼 수 있는 것은 경로제공 페이지에서는 사용자의 관심의 차이가 크지 않다는 걸 알 수 있고 따라서 이러한 페이지에서는 사용자의 관심의 정도가 갖는 의미가 크지 않다는 것을 의미한다. 즉 경로제공 페이지에서는 페이지에 대한 방문 여부만으로도 충분한 의미가 부여될 수 있다. 반면에 컨텐츠 페이지의 경우는 해당 컨텐츠에 따라 각 컨텐츠 페이지에서 소비하는 시간의 차이가 큰 경우가 많다. 이런 페이지에서는 사용자의 관심정도가 더욱 큰 비중을 차지할 수 있다. 이것을 다음과 같이 정리할 수 있다.

- 경로제공 페이지 : 접속패턴의 의미 비중 > 사용자의 관심정도
- 컨텐츠 페이지 : 접속패턴의 의미 비중 ≤ 사용자의 관심정도

이러한 특징을 활용하여 트랜잭션 모델을 구성하고 이에 대한 특성별 유사성 비교를 한다면 보다 정확한 클러스터링을 할 수 있을 것이다. 따라서 본 논문에서는 두 번째로 다음과 같은 사항을 제시한다.

경로제공 페이지에 대해서는 기존의 이진가중치 벡터 유사도 비교를 수행하고 컨텐츠 페이지에 대해서는 제안했던 흥미도가중치 벡터 유사도 비교를 한다. 즉 두 가지 방식의 혼합된 모델을 사용하여 유사도 비교를 하는 것이다. 다음은 URL의 집합을 페이지의 특성에 따라 구분한 것이다.

$$U = \{\underbrace{url_A, url_B, url_C}_{\text{경로페이지}}, \underbrace{url_D, url_E, url_F, url_G}_{\text{컨텐츠페이지}}\}$$

이것에 대한 트랜잭션 모델을 다음과 같이 정의한다.

$$\vec{t} = \langle U_N, U_C \rangle \qquad U_N : \text{경로제공페이지의 집합}$$

$$U_C : \text{컨텐츠페이지의 집합}$$

U_N : 이진가중치벡터유사도비교

U_C : 흥미도가중치벡터유사도비교

이처럼 트랜잭션 모델을 구성함으로써 유사성 비교시 페

이지 특성에 따라 경로페이지의 경우는 페이지 접근 패턴을 위주로, 컨텐츠 페이지의 경우는 사용자의 관심을 반영함으로써 조금 더 정밀한 유사도 비교를 할 수 있게 된다.

이제 지금까지의 3가지 웹 트랜잭션 유사도 비교 방식, 즉 이진가중치 벡터 유사도 비교와 제안된 흥미도가중치 벡터 유사도 비교, 그리고 이들을 혼합한 방식의 차이를 다음의 예제를 살펴본다. 먼저 (그림 2)의 사이트 구조에서 두 트랜잭션 t1과 t2에서 방문한 URL들의 집합이 <표 8>과 같다고 가정한다.

〈표 8〉 각방식별 유사도 비교

threshold = 0.5

	A	B	C	D	E	F	G
t_1	0.1	0	0.6	0	0	0.1	0.2
t_2	0.1	0	0.1	0	0.1	0.6	0.1
이진 벡터	○	○	○	○	×	○	○
흥미도 벡터	○	○	×	○	×	×	○
혼합방식	○	○	○	○	×	×	○

<표 8>과 같이 사이트 점유시간의 비율이 주어지고 각 3가지 방식의 정의에 따라 유사하다고 인정될 수 있는 URL을 표기하였다. 유사도 비교의 결과는 다음과 같다.

- 이진가중치 벡터 유사도 비교 : 6/7
- 흥미도가중치 벡터 유사도 비교 : 4/7
- 혼합방식 유사도 비교 : 5/7

먼저 이진가중치 벡터에서의 결과를 살펴보면 두 트랜잭션에 있어서 총 7개의 페이지 중 6개의 페이지가 일치한다. 이것만을 놓고 생각했을 때는 두 개의 트랜잭션 사이의 상당한 유사성을 갖는 것처럼 보인다. 하지만 (그림 2)의 사이트 구조에 따라 이를 비교해보면 이는 상당히 다른 성향을 보이고 있다는 것을 알 수 있다. 이처럼 상당한 차이가 있음에도 불구하고 이진가중치 벡터의 경우 지나치게 높은 유사도로 측정될 수 있다.

다음은 흥미도가중치의 경우이다. 여기에서 threshold를 0.5로 정하였다. 이러한 수치는 예를 쉽게 보이기 위해서 설정한 값이기는 하나 웹 사이트에서 일반적으로 사용자 관심 측면에서 전체 탐색 시간에서 해당 페이지에 대한 점유비율의 차이가 50% 이상이라면 실제 큰 차이라 생각할 수 있다.

<표 8>에서 볼 때 일치되는 항목 중에 이러한 threshold 이상의 항목이 2개가 존재하며 흥미도가중치 유사도 비교에서는 두 항목 모두를 제외시킨다. 마지막으로 혼합방식의 경우 앞에서 페이지 특성에 대해 언급했던 경로 제공 페이지에서의 중요성은 관심사보다는 항목일치에 더 의미가 있다고 정의했었다. <표 8>에서 보면 가중치의 차이가 threshold를 넘는 두 개의 항목 중 하나는 경로제공 페이지이다. 이는 페이지 C의 경우로 경로제공 페이지에 속하는데도 점유시간비율이 다른 페이지들에 비해 상대적으로 크다. 간혹 이런 경우가 존재할 수도 있는데 이는 사용자가 웹 탐색도중 다른 일을 먼저 처리한다던지, 또는 잠시 자리를 비운다든지 하는 경우가 여기에 해당하겠다. 이런 경우에 있어서 흥미도가중치 유사도 비교시에는 이 항목이 threshold에 의해 제외되는데 이는 페이지 특성을 고려해볼 때 항목일치의 의미마저 계산에서 제외시키는 결과를 가져올 수도 있다. 따라서 경로 제공 페이지에서의 이진가중치 유사도 비교와 컨텐츠 페이지에서의 흥미도가중치 유사도 비교를 혼합한 방식을 사용함으로서 페이지 특성에 따른 항목일치 정도의 유사성 비교와 사용자의 관심까지를 적절히 반영한 유사성 비교를 수행할 수 있게 되고, 그로 인해 보다 정확한 유사성을 근거로 한 클러스터링이 가능해진다.

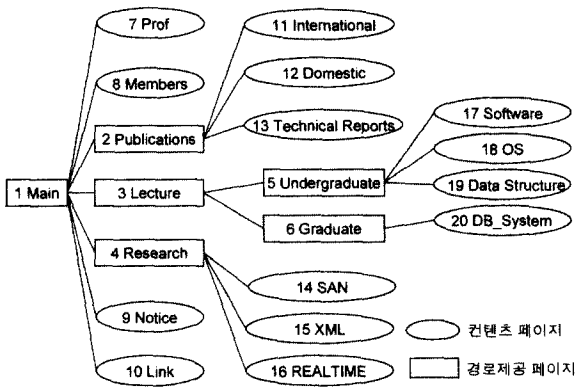
5. 성능 평가

본 논문에서 제안한 사항은 기존의 이진가중치 트랜잭션 모델에 비해 탐색 패턴뿐만 아니라 사용자의 각 URL에 보인 관심도까지를 고려하였다. 또한 혼합방식에서는 이와 더불어 URL의 특성까지 고려하였다. 때문에 기존방식에 비해 더욱 세밀한 정확성을 부여하였다. 하지만 평가에서 정확성에 대한 각 방식의 비교는 객관적인 검증을 보이기에 어려운 부분이 있다. 이는 각 방식에 대한 클러스터링 결과는 상대적으로 분명한 차이점이 발생하나 그 차이는 사용자의 관심이 반영된 것과 그렇지 않은 것의 차이이기 때문이다. 하지만 유사도의 차이에 따라 각 트랜잭션이 소속되는 클러스터가 달라지는 경우가 발생하는데 이는 단순한 방문 여부만을 가지고 분류하던 방식과 달리 사용자의 관심까지 반영한 경우이기 때문에 나타나는 차이점이라 할 수 있다. 따라서 비교 평가로서 각 방식에 따라 트랜잭션이 다른 클러스터에 소속될 수 있다는 것을 보이고 이에 대한 트랜잭션 사이의 실제 유사성을 근거로 정확성을 비교한다.

본 논문에서 제안한 사항을 다음에서 두 가지로 성능 평가하였다. 먼저 제안한 방법의 정확성을 검증하기 위해 대학의 연구실 사이트를 모델로 한 약 20개의 URL로 구성된 사이트를 만들고 이 사이트에 대하여 웹 탐색을 하도록 수행하였다. 그 결과 얻어진 웹 서버 로그에 대한 데이터 사전 준비를 수행하였다. 그리고 각 방식의 유사도 비교 결과를 보이기 위하여 10개의 트랜잭션에 대해 기존의 방법, 제안한 방법, 그리고 혼합방법의 3가지로 구분하여 유사도 비교를 수행하였다. 그리고 위에서 수행한 결과가 대규모의 웹 로그에서 어떻게 적용되는지를 비교하기 위해 실제 웹 사이트 로그를 가지고 평가하였다.

다음은 수행 과정을 보인다. 먼저 성능평가를 위한 사이

트의 구조는 (그림 3)과 같다. 여기에서는 혼합방식의 비교를 위해 URL의 특성에 따라 컨텐츠 페이지와 경로제공 페이지로 구분하였다. (그림 3)의 각 URL옆에 표시된 값은 각 URL들에 대해 유일하게 부여된 번호이다.



(그림 3) 시험 사이트

다음으로 사이트에 대하여 일반적으로 발생할 수 있는 웹 탐색을 수행한 후 결과가 기록된 웹 서버 로그를 이용하여 각 URL에 해당하는 페이지뷰 당 하나의 요소만을 남기도록 데이터 정제를 수행하고 이를 사용자의 호스트(IP)별 각 URL 접근시간을 근거로 순차적인 URL의 집합으로 구성된 세션을 구분하였다. 구분된 세션을 <표 9>에서 보이고 있으며 URL이름과 URL에서 소비한 시간으로 세션을 설명하고 있다. 이때 세션 구분 한도는 30분으로 제한하였고 하나의 세션을 하나의 웹 트랜잭션으로 변환하였다.

세션 단계에서 미리 계산된 각 URL에 해당하는 점유시간의 비율을 흥미도가중치로 부여하여 다음의 <표 10>, <표 11>과 같이 트랜잭션 데이터로 변환을 수행하였다. 이때 혼합방식을 위하여 사이트의 각 URL을 컨텐츠 페이지와 경로제공 페이지로 분리하였다. <표 10>, <표 11>에서 보이는 수치는 각 트랜잭션이 해당 URL에서 머문 시간 점유비율을 가중치로 표시하였다. 이런 트랜잭션에 대해 기존의 이진가중치 유사도 비교와 제안된 흥미도가중치를 적용한 유사도 비교 그리고 페이지 특성에 따른 이들의 혼합방법에 대한 유사도 비교를 수행하여 이들 변환된 트랜잭션들 상호간의 유사도 비교의 측정치를 <표 12>와 같이 나타내었다. <표 12>에서 보이고 있는 수치는 각 트랜잭션 사이의 유사한 정도를 나타내는 유사도로서 앞에서 언급한 각 방식별 유사도 비교 방법을 적용하여 계산하였으며 계산된 결과를 쉽게 비교할 수 있도록 이진가중치, 흥미도가중치, 혼합 방식의 순서로 모두 기재하였다.(예 : 14, 12, 12)

이진가중치 유사도 비교와 흥미도를 적용한 유사도의 차이를 보이기 위해 여기에서는 흥미도 차이의 threshold를 0.3으로 지정하였는데 이 수치는 사이트의 특성에 따라 평균적인 페이지 접속 시간 비율을 고려하여 조정되어져야 할 필요가 있다.

〈표 9〉 추출된 세션

	세션의 URL 순차 {URL, 점유시간(초)}
s1	{main, 3} - {publications, 12} - {domestic, 28} - {technical reports, 9}- {research, 4} - {san, 3} - {real time, 12} - {lecture, 118} - {undergraduate, 3} - {software, 3} - {os, 6} - {db_structure, 112} - {graduate, 3} - {db_system, 9}
s2	{main, 8} - {lecture, 12} - {undergraduate, 8} - {software, 39} - {os, 20} - {db_structure, 146} - {graduate, 8} - {db_system, 194}
s3	{main,3 } - {notice, 27} - {lecture, 5} - {graduate, 2} - {dbsystem, 72} - {undergraduate, 2} - {data structure, 72} - {os, 5}
s4	{main, 3} - {publications, 12} - {domestic, 8} - {research, 4} - {san, 16} - {xml, 4} - {real time, 24} - {lecture, 4} - {undergraduate, 8} - {os, 40} - {db_structure, 124} - {graduate, 20} - {db_system, 132}
s5	{main, 5} - {prof, 10} - {members, 21} - {publications, 193} - {international, 32} - {domestic, 20}-{technical report, 57} - {notice, 84}
s6	{main, 9} - {publications, 5} - {domestic, 65} - {technical reports, 29} - {research, 22} - {san, 9} - {xml, 69} - {real time, 5} - {db_system, 9}
s7	{main, 74} - {porf, 21} - {publications, 12} - {international, 78} - {domestic, 40} - {research, 10} -{xml, 10}
s8	{main, 2} - {publications, 2} - {domestic, 60} - {technical reports, 56} - {research, 3} - {san, 2} - {real time, 2} - {lecture, 3} - {undergraduate, 6} - {software, 4} - {os, 17} - {db_structure, 2} - {graduate, 2} - {db_system, 2}
s9	{main, 8} - {publications, 8} - {international, 8} - {domestic, 41} - {graduate, 11} - {dbsystem, 115}
s10	{main, 8} - {publications, 6} - {international, 30} - {domestic, 9} - {san, 9} - {xml, 143}

〈표 10〉 트랜잭션(경로제공 페이지)

	U1	U2	U3	U4	U5	U6
t1	0.01	0.04	0.38	0.01	0.01	0.01
t2	0.02		0.03		0.02	0.02
t3	0.02		0.03		0.01	0.01
t4	0.01	0.03	0.01	0.01	0.02	0.05
t5	0.01	0.46				
t6	0.04	0.02		0.10		
t7	0.41	0.02		0.06		
t8	0.01	0.01	0.02	0.02	0.05	0.01
t9	0.04	0.04				0.06
t10	0.04	0.03				

〈표 11〉 트랜잭션(컨텐츠 페이지)

	U1	U2	U3	U4	U5	U6
t1	0.01	0.04	0.38	0.01	0.01	0.01
t2	0.02		0.03		0.02	0.02
t3	0.02		0.03		0.01	0.01
t4	0.01	0.03	0.01	0.01	0.02	0.05
t5	0.01	0.46				
t6	0.04	0.02		0.10		
t7	0.41	0.02		0.06		
t8	0.01	0.01	0.02	0.02	0.05	0.01
t9	0.04	0.04				0.06
t10	0.04	0.03				

〈표 12〉 유사도 비교표

	U7	U8	U9	U10	U11	U12	U13	U14	U15	U16	U17	U18	U19	U20
t1						0.09	0.03	0.01		0.04	0.01	0.02	0.36	0.03
t2											0.10	0.05	0.37	0.49
t3			0.18									0.03	0.22	0.48
t4			0.02			0.02		0.04	0.01	0.06		0.10	0.31	0.33
t5	0.02	0.05	0.20		0.07	0.05	0.13							0.01
t6						0.32	0.13	0.04	0.31	0.02				
t7	0.12				0.11	0.23			0.06					
t8						0.40	0.37	0.01		0.01	0.02	0.08	0.01	0.01
t9					0.04	0.21								0.59
t10					0.14	0.05		0.05	0.69					

여기에서는 threshold 0.3을 실제의 웹 탐색 행위에 있어 관심의 차이를 나타낼 수 있는 수치라 가정하였다. threshold를 결정할 때 전체 트랜잭션들에 대한 최대 흥미도 값과 최소 흥미도 값의 차를 평균한 값인 0.43을 근거로 가정 하였으며 이에 대한 내용은 끝부분에서 다시 언급한다. 그리고 실제 실험 사이트의 탐색행위를 가지고 각 방식별 유사도 비교를 수행한 결과 0.3의 차이를 이용함으로서 <표 12>와 같은 각 방식별 유사도의 차이를 보일 수 있었다.

유사하다고 인정된 URL의 개수, 즉 유사도가 10이상의 비교적 높은 유사도를 보이는 트랜잭션 사이에서 기존의 이진가중치 트랜잭션 모델의 유사도와 본 논문에서 제안하는 흥미도가중치 트랜잭션 모델의 유사도 에서 많은 차이가 있는 것을 볼 수 있다. 특히 t1, t8사이와 t4, t8사이를 살펴보면 비트 패턴은 많은 수가 일치하나 사용자의 관심에 있어서 상당부분이 상이하다는 것을 확인할 수 있으며 이것을 유사도에 반영하여 방문여부만으로 유사도 비교를 수행함으로서 실제 트랜잭션의 유사성보다 높게 측정될 수 있는 부분에 대해 적절히 보완하는 역할을 수행하고 있다고 할 수 있다. 이처럼 웹 트랜잭션 유사도 비교에 사용자의 관심도를 부여함으로서 유사성이 지나치게 많이 측정되는 것을 방지하고 결과적으로 각 트랜잭션의 유사성에 근거하여 보다 정확한 클러스터링을 유도할 수 있게 된다.

계산되어진 유사성 매트릭스에 대하여 클러스터링 시 클러스터의 개수를 결정하거나 각 클러스터의 중심을 구하는 데에는 여러 가지 알고리즘이 적용될 수 있겠다. 하지만 여기에서는 본 논문에 관한 비교평가의 의미로서 클러스터의 수와 그 중심을 미리 정하고 이를 직접 클러스터링을 수행하여 유사도의 측정방법에 따라 클러스터가 변경되는 예를 보임으로서 본 논문에서 제안하는 방식의 차이와 그 정확성을 보이도록 한다.

먼저 트랜잭션을 3개의 클러스터로 나누고 이들의 중심을 임의로 t2, t6, t10으로 하고 각 트랜잭션을 가장 가까운 클러스터에 할당하였다. <표 13>은 클러스터링 시 각 방식별 유사도 비교에 근거하여 각 클러스터의 중심에 소속되는 경우를 나타내고 있다. 여기에서 우리는 이진가중치 유사도 비교와 제안된 흥미도가중치 유사도 비교에 따른 클러스터링의 차이점을 t4와 t8을 통해 확인할 수가 있다. 하지만 흥미도가중치 유사도 비교 방식과 혼합방식의 경우 많은 차이가 없음을 알 수 있다. 이는 경로제공 페이지에서 많은 시간을 보내는 경우가 그리 많지 않기 때문이라 할 수 있다.

이제 클러스터의 소속이 변경된 트랜잭션에 대한 실제 비교를 수행하여 본다. 먼저 <표 13>의 t1의 경우 이진가중치 유사도 비교에서는 t2에 흥미도가중치 유사도 비교에서는 t6에 가깝게 표시되었으나 실제 트랜잭션을 비교하여 보면 t2와 t6중 어느 것에 더 유사하다 할 수 없음을 쉽게 확인할 수 있다.

〈표 13〉 트랜잭션 클러스터링

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

흥미도가중치유사도	
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

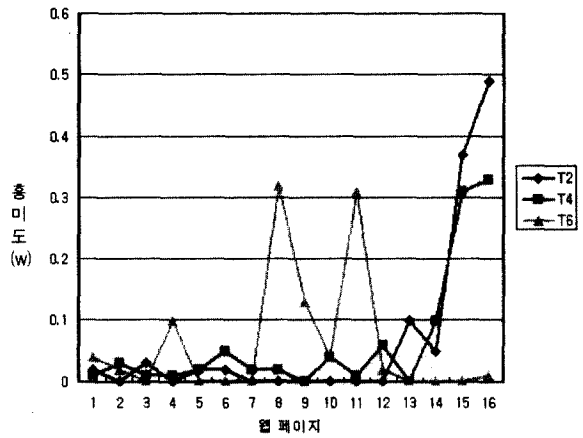
〈표 14〉 트랜잭션 비교

	U1	U2	U3	U4	U5	U6	U9	U12	U13	U14	U15	U16	U17	U18	U19	U20
t2	0.02		0.03		0.02	0.02							0.10	0.05	0.37	0.49
t4	0.01	0.03	0.01	0.01	0.02	0.05	0.02	0.02		0.04	0.01	0.06		0.10	0.31	0.33
t6	0.04	0.02		0.10				0.32	0.13	0.04	0.31	0.02				0.01

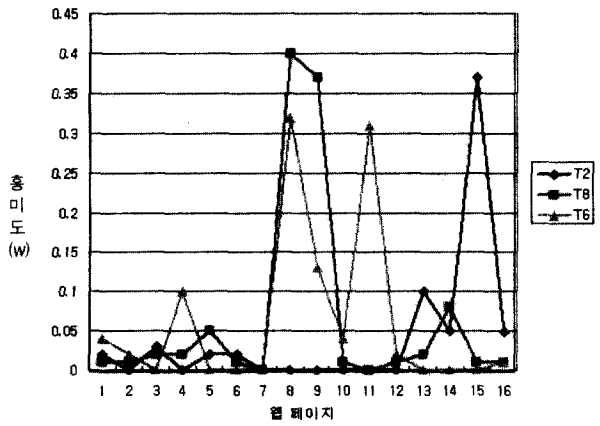
〈표 15〉 트랜잭션 비교 2

	U1	U2	U3	U4	U5	U6	U9	U12	U13	U14	U15	U16	U17	U18	U19	U20
t2	0.02		0.03		0.02	0.02							0.10	0.05	0.37	0.49
t8	0.01	0.01	0.02	0.02	0.05	0.01		0.40	0.37	0.01		0.01	0.02	0.08	0.01	0.01
t6	0.04	0.02		0.10				0.32	0.13	0.04	0.31	0.02				0.01

반면에 확연하게 소속의 차이를 보이는 t4와 t8의 경우를 실제 트랜잭션이 방문한 URL과 각 URL에 보인 관심의 비율을 가지고 비교하여 보면 <표 14> 그리고 <표 15>와 같다. 그리고 내용 비교를 용이하게 하기 위해 (그림 4)와 (그림 5)의 그래프로 나타내었다. 먼저 <표 14>에서 클러스터의 각 중심 t2와 t6에 대하여 트랜잭션 t4의 경우 비트 패턴은 t6에 한 개가 더 일치하는 것을 확인할 수 있다. 이에 반해 사용자의 관심 비율은 전혀 유사하지 않음을 알 수 있다. 하지만 t2에 대하여서는 t6에 비해 비트 패턴의 일치 비율은 약간 낮지만 사용자 관심측면을 바라볼 때 상당히 유사하다는 것을 알 수 있다. 이를 (그림 4)에서 사용자 관심의 정도인 흥미도의 비교를 통해 쉽게 확인할 수 있다.



(그림 4) T2, T4, T6의 흥미도 비교



(그림 5) T2, T8, T6의 흥미도 비교

<표 15>의 경우에서도 t8은 중심 t2, t6과 비교하여 볼 때 이전의 경우와 비슷한 사실을 (그림 5)를 통해 확인할 수 있다. 이렇게 각 방식에 따라 소속의 차이를 보이는 트랜잭션을 실제 비교를 해 봄으로써 제안된 방식이 더 정확한 유사도 비교결과를 나타낸다는 것을 보였다. 결과적으로 제안한 흥미도가중치 유사도 비교가 기존의 비트 패턴의 이진가중치 유사도 비교의 정확성에 대한 문제점을 보완하여 더욱 정확한 클러스터링을 유도하는 것을 확인하였다.

다음으로 앞에서 확인된 결과가 실제 대규모의 웹 로그의 분석에서 어떻게 영향을 미치는지 분석 하였다. 수행에 따른 데이터는 다음과 같다. <표 16>의 원시 로그는 충북대학교 홈페이지의 4일간의 로그 285M을 추출하여 사용하였다. 웹 로그는 연휴기간이면서 충북대학교 수시모집 기간 중에 추출한 로그를 사용하였다. 따라서 웹 로그의 패턴이 어느 정도 구분 되어 있음을 추측할 수 있고 이를 클러스터링 할 때 각 클러스터의 중심으로 활용하였다. 분석에 사용된 데이터는 <표 16>에서 보이고 있다.

〈표 16〉 분석에 사용된 실제 대규모 웹 로그

종 류	수 량
원시 로그	285M
정제 후 로그	165,323 라인
전체 웹 페이지목록	16,868
정제된 웹 페이지목록	440 page
100회 이상 hit된 웹 페이지 목록	50 page
사용자 세션	25,368 세션
사용자 트랜잭션	25,386 트랜잭션

웹 로그는 먼저 분석에 필요하지 않은 그림파일등을 제거하는 등의 사전 준비 작업을 수행하여 정제 후 로그인 165,323개의 웹 페이지 접근을 추출하였다. 추출된 로그에서 발견된 전체 페이지 목록은 16,868개로서 이는 충북대학교 홈페이지가 PHP페이지로 이루어져 있다는 특성에 기인하여 발생된 페이지 목록이다. 따라서 질의에 따라 페이지를 다르게 보여주는 페이지들을 제거하고 해당 메인 페이지만으로 이루어진 440개의 페이지 목록을 추출하였다. 그리고 여기에 웹 페이지 접근 횟수가 매우 미비한 페이지들을 제거함으로서 자주 참조되는 50개의 페이지 목록을 추출하였고 이를 트랜잭션 모델에서 사용하였다.

이렇게 정제된 로그를 통해 발생된 사용자 세션은 총 25,368개이다. 이것을 변환하여 25,368개의 트랜잭션을 생성하였다.

<표 16>의 데이터를 통해 각 트랜잭션들을 4개의 그룹으로 클러스터링 하였다. 이때 각 클러스터의 중심은 로그의 추출 시기를 고려하여 특정 패턴들을 임으로 지정하였다. 각 중심의 방문 페이지의 구성은 <표 17>과 같다.

〈표 17〉 클러스터의 중심 트랜잭션의 방문 페이지 분포

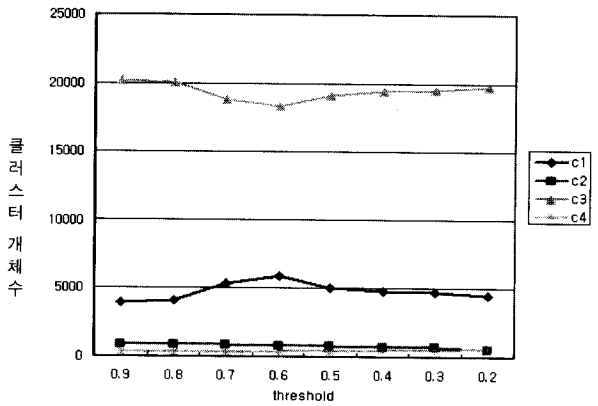
클러스터	중심 트랜잭션	방문 페이지
c1	t1	/index.php, /event/index.php, /notice_main/index.php, /tscenter/index.php, /tscenter/ts_index.php, schedule/index.php, /search/index.php
c2	t2	/index.php, /event/index.php, /entrance/event.window/index.php, /sanchul_2004/index.php, /imagebank/index.php, /univ/index.php, /abroad_info/index.php, /circle/index.php
c3	t3	/index.php, /event/index.php, /gaesin_public/index.php, /house/index_php, /grad/index.php, /onlile_market/index.php
c4	t4	/index.php, /event/index.php, /sanchul_2004/index.php, /week_manu/index.php, /traffic_man/index.php, /grad/index.php, /schedule/index.php

먼저 t1의 경우와 t4의 경우 재학생들이 자주 찾는 페이지들을 구분한 트랜잭션들이다. 그리고 t2의 경우 웹 로그 추출 시기에 관계하여 수시모집에 의해 접근되는 패턴의 트랜잭션이라 할 수 있고, t3의 경우는 검색 패턴이 홈페이지에서 제공하는 주변 자취방 정보, 분실물 센터, 여론 등을 수집하고 제공하는 등의 게시판 서비스에 치중되는 모습을 보이는 트랜잭션이다. 이들 트랜잭션들을 중심으로 기존의 비트 벡터를 사용하여 클러스터링을 수행한 결과와 흥미도가중치에 다양한 threshold를 적용하여 클러스터링을 수행한 결과를 <표 18>에서 나타내었다.

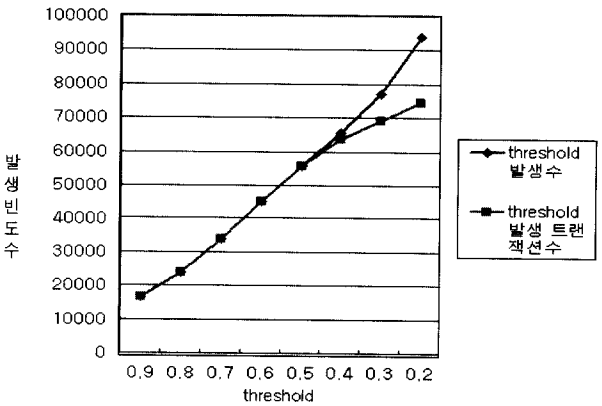
〈표 18〉 가중치 변화에 따른 각 클러스터의 트랜잭션 수 변화

threshold / 클러스터	비트 벡터	threshold 0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
c1	3955	3957	4122	5376	5891	5069	4762	4709	4443
c2	895	894	895	894	821	758	721	683	542
c3	20236	20234	20066	18810	18296	19117	19428	19483	19757
c4	282	283	285	288	360	424	457	493	626
threshold 발생 페이지		16586	23921	33915	45325	55750	65670	76913	93791
threshold 발생 트랜잭션		16586	23921	33913	45287	55675	63579	69088	74566

<표 18>에서는 먼저 비트 벡터 유사도 비교 방식과 흥미 가중치를 적용한 방식에서 threshold를 변화 시켜가며 측정해 보았다. 이때 각 클러스터에 포함되는 트랜잭션의 수가 변화하는 모습을 (그림 6)의 그래프에서 보이고 있다. 그리고 표에서의 'threshold 발생 페이지'는 클러스터링 수행 과정 중 전체 비교 횟수에서 threshold 이상의 차이가 발생하는 수를 나타낸 수치이고, 'threshold 발생 트랜잭션'은 threshold 이상의 차이가 하나라도 발견된 트랜잭션의 수이다. 이들의 변화량을 (그림 7)의 그래프로 나타내었다. (그림 7)에서 threshold가 낮아짐에 따라 발생 횟수가 크게 증가하는 모습을 보이는데 이는 threshold가 낮아질수록 하나의 트랜잭션 안에서 threshold 차이가 많이 발생하는 것을 의미한다.



(그림 6) 각 클러스터의 개체 수 변화



(그림 7) threshold 발생 횟수 및 발생 트랜잭션

<표 18>와 (그림 7)에서 threshold값이 90%와 80% 이상 차이가 나는 경우도 상당수 존재함을 알 수 있다. 이는 관심의 정도가 전혀 다른 웹 페이지 접근들로 구성된 트랜잭션들도 상당수 존재한다는 것을 의미한다. 이러한 변화는 threshold가 낮아질수록 변화의 폭이 넓어지게 되는데 이는 (그림 7)에서 보이고 있으며 threshold가 낮아질수록 threshold 차이 발생 트랜잭션 수가 증가한다는 사실을 통해 알 수 있다. 이때 특정 threshold(0.5)에서 차이 발생 횟수와 발생한 트랜잭션 수가 확연하게 차이를 보이기 시작하는데 이는 해당 threshold 이후부터 차이 발생이 급격하게 증가하는 것을 의미한다. 이처럼 급격하게 차이 발생이 증가함은 해당 시점부터 하나의 트랜잭션에서 많은 페이지들이 threshold 차이를 보이기 시작한다는 것으로서 클러스터링

을 할 때 유사성 오류를 가져올 가능성이 점점 많아진다는 것을 의미한다. 따라서 이러한 시점은 threshold를 결정하는데 있어 중요한 판단 요소가 된다.

threshold의 결정에 고려될 수 있는 또 다른 요소는 전체 트랜잭션들에 대한 최대 흥미도와 최소 흥미도의 차이의 평균값이다. 일반적인 경우 최소 흥미도는 수 초 정도이지만 최대 흥미도는 수십 초에서 수분 또는 수십 분이 되는 경우가 보통으로 상당히 큰 차이를 보이게 된다. 따라서 본 논문에서는 전체 트랜잭션들에서 각 트랜잭션의 최대 흥미도와 최소 흥미도 차이의 평균을 threshold의 최대값으로 한다. 이를 토대로 본 논문에서의 threshold는 앞에서 언급한 threshold차이 발생 횟수의 변화폭을 참고하여 최대 threshold 값보다 작은 값으로 결정한다. 참고로 앞의 실험 사이트의 평균 흥미도 차이는 약 0.43이며, 이를 통해 0.3의 threshold를 설정하여 사용자 관심의 정도가 상이한 트랜잭션들이 높은 유사성을 갖는 것을 보정하였다. 그리하여 관심의 정도까지 유사한 성향의 트랜잭션들을 클러스터링 할 수 있도록 유도하였다. 그리고 충북대학교 홈페이지를 통해 대규모의 웹 사이트 로그를 추출하여 분석한 결과에서의 평균 흥미도의 차이는 약 0.72로서 (그림 7)의 threshold 차이 발생 횟수의 변화폭을 고려하여 threshold를 0.5로 설정하는 것이 적합함을 보였다.

지금까지 간략하게 테스트 사이트의 웹 로그와 실제 대규모의 웹 로그를 가지고 각 방식에 따른 비교 검증을 수행하여 각 유사도 비교 방법에 따른 클러스터링의 차이를 보였다. 이러한 차이는 트랜잭션의 수와 클러스터링 시 클러스터의 수에 따라 본 논문에서 제안한 방법이 미치는 영향이 상당히 다를 수 있다는 것을 보여준다. 또한 클러스터의 수가 많을수록 즉 세부적인 클러스터로 구분할 때 기존의 이진가중치 유사도 비교를 이용한 방법에 비해 흥미도 가중치 유사도 비교방법이 트랜잭션의 소속 변경을 더 많이 발생시킬 수 있다. 또한 이러한 정확성에 따른 소속의 변화는 트랜잭션의 수가 많을수록 비례적으로 많이 발생할 것이다. 이것은 결과적으로 기존의 방법과 본 논문에서 제안한 방법에서 측정된 유사도의 측정치가 많은 차이를 보이지는 않음을 알 수 있는데, 이는 명확하게 구분되는 트랜잭션의 클러스터링에 관여하기보다 각 클러스터의 경계부근에 놓여지는 트랜잭션에 대해 정밀하게 소속을 밝혀줄 수 있는 정밀성을 부여하는데 많은 영향을 미친다 할 수 있다.

결과적으로 방문 패턴의 유사성만을 가지고 클러스터링 할 경우 정확성이 결여될 수 있다는 문제점을 사용자 관심정도를 고려한 흥미도가중치를 적용하여 해결하였다. 또한 패턴 일치의 정도가 어느 정도 유사한 트랜잭션들 사이에서 트랜잭션의 관심성향을 판별하여 보다 유사한 트랜잭션끼리 올바르게 분류될 수 있게 하였다.

6. 결 론

본 논문에서는 웹 트랜잭션들에 대한 클러스터링의 정확성을 높이기 위한 흥미 가중치 적용 유사도 비교방법을 제안하였다. 제안한 방법은 이진가중치 유사도 비교 방법의 정확성 문제를 해결하였고 웹 트랜잭션에 흥미도가중치를 부여하여 사용자의 관심정도 까지를 고려한 유사도 비교를 수행할 수 있게 하였다. 또한 탐색 패턴과, 사용자 관심의 중요성을 적절히 반영하여 유사도 비교를 수행할 수 있도록 웹 페이지의 특성을 분리하여 특성에 알맞은 유사도 비교를 수행할 수 있게 하였다. 이것으로 사용자의 유사한 성향을 보다 세밀하게 비교 할 수 있게 하였으며 이것의 정확성을 실제 웹 데이터를 가지고 각 방식에 대한 유사도 비교를 수행하여 결과를 분석해 봄으로서 보다 정확한 클러스터링이 유도되는 것을 보였다. 그리고 유사도 비교를 수행함에 있어 적용되는 threshold값을 결정할 때 고려되어야 할 사항을 제시하였다. 하지만 유사성 비교의 정확성을 높이는데 중요한 threshold의 최적화를 위해서는 보다 다양한 데이터를 통한 실험 및 분석이 요구된다. 또한 대규모의 데이터를 통한 분석에서 정확성을 입증할 수 있는 방법 또한 다양하게 고려되어야 할 것이다.

이에 따라 향후 연구로는 보다 다양한 웹 로그 데이터에 대한 면밀한 분석을 통해 threshold의 최적화 방법을 구체적으로 제시하고 다양한 방법으로 이들의 정확성을 입증하는 것이다. 더 나아가 여러 가지 클러스터링 알고리즘에 제안한 방식을 실제 적용하여 보고 이것을 통해 추출된 클러스터로부터 보다 효과적으로 추천집단을 생성하는 알고리즘을 연구하여 웹 사이트에 실제 적용하는 것을 목적으로 하고 있다.

참 고 문 헌

[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] 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.

[3] E-H. Han, G. Karypis, V. Kumar and B. Mobasher, "Clustering based on association rule hypergraphs," Data Mining and Knowledge Discovery(DMKD), 1997.

[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, No4, pp.54-61, 1998.

[6] ley, Pang-Ning Tan and Jaideep Srivastava, "Discovery of Interesting Usage Patterns from Web Data," World Wide Web Knowledge and Data mining(WEBKDD), pp.163-182, 1999.

[7] Lin, S. A. Alvarez and C. Ruiz, "Efficient adaptivesupport association rule mining for recommender systems," Data Mining and knowledge Discovery(DMKD), 2002.

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

[9] B. Mobasher, H. dai and T. Luo, "Improving the Effectinveness of Collaborative Filtering on Anonymous Web Usage Data," Proceedings of the IJCAI 2001 Workshop on Intelligent Techniques for Web Personalization(ITWP01), August, 2001.

[10] Shahabi, C., A. Zarkesh and J. Adibi, and V. Shah, "Knowledge Discovery from Users Web-PageNavigation," Research Issues in Data Engineering, 1997.

[11] 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.

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

[13] F. Masseglia, P. Poncelet and M. Teisseire, "Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure," In ACM SigWeb Letters, Vol.8, No.3, pp.13-19, October, 1999.

[14] Mbasher, B., Cooley, R., Srivastaba, J., "web mining : Information and Pattern Discovery on the World Wide Web," In Procedings of the 9th IEEE International Conference on Tools with Artificial Intelligence(ICTAI '97), November, 1997.

[15] B. Mobasher, H. Dai, and T.Luo, "Web Usage and Content Mining for More Effective Personalization," E-Commerce and Web Technologies(ECWeb), September, 2000.

[16] T. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, "From user access patterns to dynamic hypertext linking," WWW5/Computer Networks, Vol.28, No.7-11, 1996.

강 태 호

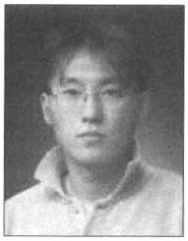
e-mail : segi21@netdb.chungbuk.ac.kr

1999년 호원대학교 정보통신공학과(공학사)

2002년 충북대학교 대학원 정보산업 공학과(공학석사)

2003년~현재 충북대학교 대학원 정보통신 공학과 박사과정

관심분야 : 자료저장 시스템, 데이터베이스 시스템, 웹 컨텐츠 관리 시스템, 웹 마이닝 등



민 영 수

e-mail : minys@netdb.chungbuk.ac.kr

1998년 충북대학교 정보통신공학과(공학사)

2001년 충북대학교 대학원 정보통신 공학과(공학석사)

2001년~현재 충북대학교 대학원 정보 통신공학과 박사과정

관심분야 : XML, 데이터베이스, 데이터마이닝, 무선인터넷 등

유 재 수

e-mail : yjs@cbucc.chungbuk.ac.kr

1989년 전북대학교 컴퓨터공학과(공학사)

1991년 한국과학기술원 전산학과(공학 석사)

1995년 한국과학기술원 전산학과(공학 박사)

1996년~현재 충북대학교 전기전자 및 컴퓨터공학부 부교수

관심분야 : 데이터베이스 시스템, 정보검색, 멀티미디어 데이터베이스, 분산객체 컴퓨팅 등