

웹 마이닝과 협력적 정보 여과를 이용한 개인화 서비스의 성능 개선 방안

°이 치 훈, 고 세 진, 김 용 환, 이 필 규

인하대학교 전자계산공학과

{g1991259, g2001398, g1991260}@inhavision.inha.ac.kr, pkrhee@inha.ac.kr

Improving Web Personalization Service Using Web Mining and Collaborative Filtering

°Chi-Hoon Lee, Se-Jin Ko, Yong-Hwan Kim, Phill-Kyu Rhee
Dept. of Computer Science and Engineering, Inha University

요 약

웹 개인화 기술의 발달은 많은 업체들이 기존 고객의 유지와 신규 고객의 확보를 위한 수단을 제공하였다. 현재의 개인화 기술은 크게 내용 기반 그리고 협력적 정보 여과 방식에 기반한 기술로 나뉘어 질 수 있다. 내용 기반 정보 여과 방식에 기반한 개인화 기술은 멀티미디어 정보로 표현된 대부분의 웹 오브젝트(페이지, 이미지, 동영상, 사운드, 상품 등)에는 적용하기 어렵고, 협력적 정보 여과 방식은 Cold Start Problem과 단일 도메인내에서의 개인화 서비스만이 가능하다는 문제점이 있다. 본 논문에서는 협력적 정보 여과 방식과 데이터 마이닝 기술 중의 연관 규칙 생성 방법을 혼합한 웹 개인화 시스템을 제안한다. 다양한 멀티미디어 형태로 표현되는 웹 오브젝트의 내용 분석이 어려우므로, 각각의 오브젝트를 하나의 아이템으로 인식하고 개인화 서비스를 시도하는 협력적 정보 여과 방식을 채택하였다. 협력적 정보 여과의 결과로 발견된 도메인별 유사 사용자의 웹 오브젝트 사용 정보를 연관 규칙 생성 알고리즘에 적용하여 오브젝트간의 연관성을 발견한다. 발견된 오브젝트간의 연관성은 서로 다른 정보 도메인의 오브젝트가 현재 사용자에게 흥미있는 것인가를 예측할 수 있는 자료로서 사용될 수 있다. 협력적 정보 여과 방식에 의해 생성된 오브젝트의 선호도값과 오브젝트 연관성 정보를 비교하여 사용자에게 개인화된 웹 서비스를 제공한다.

1. 서 론

인터넷의 대중화는 시간과 공간의 제약을 받지 않고 필요한 정보의 획득을 가능하게 하는 편리한 수단을 제공하였다. 이러한 가운데, 인터넷 관련 비즈니스 업체의 기존 고객의 유지와 신규 고객의 확보를 위한 경쟁은 웹의 개인화 기술을 웹 관련 기술로서 중요한 위치를 차지하게 하였다. 개인화 기술은 특정 사용자 또는 사용자 그룹에게 그들이 흥미있어할 만한 웹 오브젝트(컨텐츠, 이미지, 상품 등)를 제공하는 것으로 정의된다. 현재의 개인화 기술은 크게 내용 기반 그리고 협력적 정보 여과 방식에 기반한 기술로 나뉘어 질 수 있다. 내용 기반 정보 여과 방식에 기반한 개인화 기술은 멀티미디어 정보로 표현된 대부분의 웹 오브젝트(페이지, 이미지, 동영상, 사운드, 상품 등)에는 적용하기 어렵고, 협력적 정보 여과 방식은 Cold Start Problem과 단일 도메인내에서의 개인화 서비스만이 가능하다[1].

그러므로, 언급된 문제점을 해결하는 개선된 개인화 시스템이 필요하다. 본 논문에서는 협력적 정보 여과 방식과 데이터 마이닝 기술 중의 연관 규칙 생성 방법을 혼합한 웹 개인화 시스템을 제안한다.

2. 관련 연구

최근의 웹 개인화 기술은 크게 협력적 및 내용 기반 정보 여과에 기반한 시스템으로 나눌 수 있다. 협력적 정보 여과에 기반한 시스템은 일반적으로 웹 서비스에 대한 사용자의 평가값을 명시적으로 받아들이고, 사용자간 상관 엔진(correlation engine)을 통해 구해진 아이템에 대한 사용자의 선호도에 기반하여 개인화를 시도한다. GroupLens, Firefly, NetPerception은 협력적 정보 여과 방식을 채택한 서비스를 제공하고 있다. WebWatcher[2] 와 클라이언트 사이드 에이전트인 Letzia[3]와 같은 내용 기반 정보 여과 시스템은

웹 문서와 사용자 프로파일 간의 유사도에 기반하여 추천을 하고 있다.

3. 웹 오브젝트 개인화 시스템

3.1 개요 및 구조

본 논문에서 제안하는 시스템은 다양한 멀티미디어 형태로 표현되는 웹 오브젝트의 내용 분석의 어렵기 때문에, 각각의 오브젝트를 하나의 아이템으로 인식하고 개인화 서비스를 시도하는 협력적 정보 여과 방식을 채택하였다. 협력적 정보 여과의 결과로 발견된 도메인별 유사 사용자의 웹 사용 정보를 연관 규칙 생성 알고리즘인 Apriori 알고리즘을 사용하여 오브젝트간의 연관성을 발견한다[4]. 발견된 오브젝트간의 연관성은 서로 다른 정보 도메인에서 특정 오브젝트가 현재 사용자에게 흥미있는 것인가를 예측할 수 있는 자료로서 사용될 수 있고 협력적 정보 여과 방식에 의해 생성된 오브젝트의 선호도값과 연관성 정보를 비교하여 사용자에게 개인화된 웹 서비스를 제공한다.

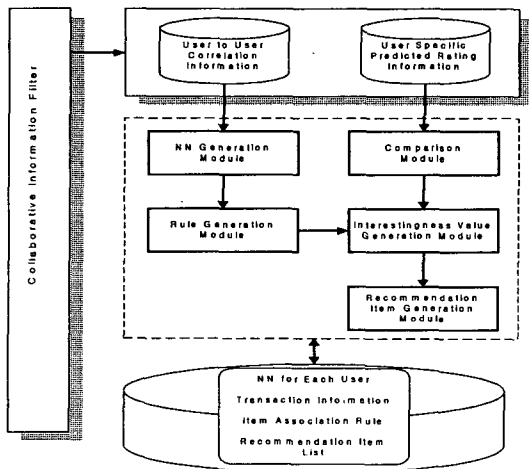


그림 1. 개인화 시스템 구조

3.2 사용자 프로파일 생성

본 논문에서는 협력적 정보 여과 방식을 사용하여 다중 도메인에서 유효한 사용자군을 파악하기 위해서 사용자가 도메인에 따라 서로 다른 프로파일로 표현될 수 있게 하였다. 도메인에 따른 사용자 프로파일을 구성하는 방법은 다음과 같다.

데이터 전처리 과정에서 모든 사용자 세션에서 발생한 트랜잭션을 식별하는 작업 시, 트랜잭션에 포함되어 있는 웹 오브젝트가 속한 도메인을 식별한다. 그러므로, 모든 트랜잭션 내의 오브젝트들은 그들이 속한 도메인에 의해서 분류되어질 수 있다. 도메인에 따라 분류된 오브젝트들은 사용자에게 따라 다시 분류

되어진다. 특정 사용자에게 대해서 수집된 트랜잭션들은 해당 도메인 내에서 사용자를 대표하는 사용자 프로파일을 생성하기 위해 사용된다.

웹 오브젝트가 속하는 도메인은 도메인 전문가(domain expert)에 의해서 정의될 수 있고, 도메인 전문가에 의해 정의된 도메인 d_i 에 속하는 웹 오브젝트는,

$$d_i = \langle obj_{i1}, obj_{i2}, \dots, obj_{in} \rangle$$

같이 정의될 수 있다.

도메인 i 에 속하는 오브젝트에 대한 접근 정보를 사용하여 구성된 사용자 m 의 사용자 프로파일 Up_{im} 는,

$$Up_{im} = \langle (obj_{i1}, l_{i1}), (obj_{i2}, l_{i2}), \dots, (obj_{in}, l_{in}) \rangle$$

같이 표현된다.

여기서 obj_{ij} 은 도메인 i 에 속하는 오브젝트이고,

l_{ij} 는 사용자 m 이 머물렀던 시간(length)으로 표현된다. 이와 같이 표현된 사용자 프로파일은 웹 오브젝트 공간에서 n 차원의 벡터로 표현되므로, 차후 협력적 정보 여과 방식의 유사 사용자 발견을 위하여 사용될 수 있다[5].

3.3 다중 도메인에서의 유사 사용자 발견

기존의 협력적 정보 여과 방식은 단일 도메인내에서 적용가능한 개인화 서비스방법이다. 그러므로, 본 논문에서 제안하려는 여러 도메인에 포함되는 오브젝트에 대한 개인화 서비스를 위해서는 여러 도메인에서 유효한 유사 사용자군의 발견이 필요하다. 본 논문에서는 기존의 협력적 정보 여과 방식에서 유사 사용자군 발견하는 방법을 다중 도메인에서의 유사 사용자군 발견을 위해 확장한다.

도메인 1 (d_1)			도메인 2 (d_2)			도메인 n (d_n)		
사용자	사용자	유사도	사용자	사용자	유사도	사용자	사용자	유사도
a	b	$s_1(ab)$	a	b	$s_2(ab)$	a	b	$s_n(ab)$
a	c	$s_1(ac)$	a	c	$s_2(ac)$	a	c	$s_n(ac)$
...
b	c	$s_1(bc)$	b	c	$s_2(bc)$	b	c	$s_n(bc)$
b	d	$s_1(bd)$	b	d	$s_2(bd)$	b	d	$s_n(bd)$
...
z	x	$s_1(zx)$	z	x	$s_2(zx)$	z	x	$s_n(zx)$
z	y	$s_1(zy)$	z	y	$s_2(zy)$	z	y	$s_n(zy)$

표 1. 도메인별 사용자간 유사도

도메인 1 (d_1)에서의 사용자 A와 B의 유사도값 ($S_1(ab)$)과 도메인 2 (d_2)에서의 유사도값 ($S_2(ab)$), 그리고 도메인 n 까지의 사용자 A와 B의 유사도값 ($S_n(ab)$)이 선형 조합되어 모든 도메인에 걸친 사용자 A와 B의 유사도값으로 결정된다. 같은 방법으로 A와 C, A와 D 등 A를 중심으로한 다른 사용자와의 유사도값을 계산 후, 특정 임계값 이상의 유사도값을 갖는 사용자를 사용자 A와 모든 도메인에서 유효한 유사한 사용자군(nearest neighbor of A over the all domains)으로 결정한다.

3.4 웹 오브젝트의 연관 규칙 생성

모든 사용자는 자신과 유사 사용자군을 갖게 된다. 이러한 특정 사용자 i 의 유사 사용자군의 모든 도메인 오브젝트에 대한 트랜잭션 데이터베이스 정보는 사용자 i 에게 유효한 오브젝트 연관 규칙을 생성하는 데 사용된다. 즉, 모든 사용자들은 자신과 유사 사용자군을 가지고 있고, 그들의 트랜잭션 데이터베이스 정보를 이용하여 생성된 오브젝트 연관 규칙 베이스를 각각 가지고 있다. 그러므로, 사용자가 특정 웹 사이트에 로그인하여 웹을 항해하는 동안 로그인된 사용자의 연관 규칙 베이스가 활성화되어 사용자 세션에 따라, 개인화 서비스가 이루어진다. 개인화 서비스를 위한 오브젝트의 결정에 대한 설명은 다음 절에서 이루어진다.

3.5 개인화 오브젝트 생성

후보 추천 오브젝트 리스트를 생성하는 효과적인 방법은 오프라인상에서 유사 사용자군의 트랜잭션 데이터베이스 정보에 연관 규칙 생성 알고리즘의 적용 결과인 빈발 오브젝트 집합(frequent objectset)을 활용하는 것이다. 현재 활성화 사용자 세션 윈도우(active user session window)에 포함된 오브젝트의 url과 연관 규칙 베이스의 빈발 오브젝트 집합의 항목 리스트를 비교하여 후보 추천 오브젝트 리스트를 결정한다 [6]. 윈도우 크기를 w 라고 한다면, 세션 윈도우에 속하는 오브젝트를 모두 포함하고 명시된 지지도를 만족하는 $w+1$ 개의 항목으로 이루어진 빈발 오브젝트 집합을 발견한다. 발견된 빈발 오브젝트 집합에서 후보 추천 오브젝트의 추천값(recommendation value)은 후보 추천 오브젝트를 결과부에 가지고 있는 빈발 오브젝트 집합의 신뢰도값이 된다. 만약, 명시된 신뢰도값을 만족하는 연관 규칙이 존재하면, 결과부의 후보 추천 오브젝트는 추천 리스트에 등록된다.

<p>[알고리즘 1] 추천 오브젝트 생성 알고리즘의 구조</p> <p>입력 : 고정 길이 활성화 세션 윈도우 w 사이트 그래프 G, 최소 지지도 (S_{min}), 최소 신뢰도 (C_{min})</p> <p>출력 : 추천 오브젝트 리스트 추천 오브젝트 리스트 $\leftarrow \emptyset$</p> <ol style="list-style-type: none"> 1. 최소 지지도 S_{min}를 만족하고 크기가 $w +1$인 모든 빈발 오브젝트 집합(frequent objectset)을 찾는다. 2. $w \Rightarrow \{u\}$를 만족하는 빈발 오브젝트 집합의 신뢰도(c)가 설정된 C_{min}보다 큰 집합을 찾는다. 3. 발견된(1,2를 만족하는) 집합에 대해, 다음 식에 의해 추천값이 계산되고 결과부의 오브젝트는 추천 리스트에 등록된다. <ul style="list-style-type: none"> • $obj.rec_value \leftarrow c * ldf(u,w)$ • 추천 리스트 \leftarrow 추천 리스트 $\cup \{obj\}$
--

위의 방법으로 사용자 세션 정보를 이용하여 개인화 오브젝트를 생성한다. 생성된 오브젝트에 대하여 추천 받을 사용자의 선호도를 참조한다. 즉, 세션정보와 연관 규칙 정보만을 이용하여 얻어진 추천값

(recommendation value)과 협력적 정보 여과 방식에 의해 얻어진 예측 평가값의 합으로 최종적인 추천값(final recommendation value)을 계산한다.

4. 결론

기존의 개인화 방법이 가지고 있는 문제점을 해결하고자, 협력적 정보 여과 방법과 데이터 마이닝의 연관 규칙 생성 알고리즘을 혼합한 개선된 웹 오브젝트 개인화 방안을 제안하였다. 멀티미디어 콘텐츠 분석에 대한 연구가 진행되어 제안된 방법에 내용기반 정보 여과 방식도 적용되는 시스템이 필요할 것이다.

5. 참고 문헌

- [1] Philip S. Yu : Data Mining and Personalization Technologies. In Proc. of the 6th International Conference on Database System for Advanced Application, Taiwan, Taiwan 1999.
- [2] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In the 15th International Conference on Artificial Intelligent, Nagoya, Japan 1997.
- [3] H. Lieberman. Letizia : An agent that assists web browsing. In Proc. of the 1995 International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995.
- [4] Agrawal R., Imielinski T., Swami A. : Mining Association Rules between Sets of Items in Very Large Databases. Proc. of the ACM SIGMOD Conference, 1993.
- [5] R. Cooley, B. Mobasher, and J. Srivastava. : Data preparation for mining World Wide Web browsing pattern. Journal of Knowledge and Information Systems,(1) 1, 1999.
- [6] R. Cooley, B. Mobasher, and J. Srivastava. : Automatic Personalization Through Web Usage Mining. TR99-010 Dept. of Computer Science, Depaul University, 1999.
- [7] B. Mobasher, H. Dai, T. Luo, Y. Sun, J. Zhu : Combining Web Usage and Content Mining for More Effective Personalization. In Proc. of Internaitonal Conference on EC and Web Technology, 2000.