효율적인 카테고리 분류기법에 의한 연관 도메인 추천 서비스

허형욱*, 이은주*, 김응모*
*성균관대학교 컴퓨터공학과
e-mail:yano77@naver.com

Related domain service by effective categorization

Hyung Wook Heo*, Eun Ju Lee*, Ung-Mo Kim*
*School of Information and Communication Engineering, Sungkyunkwan University

요 약

인터넷 사용자 증가에 따라 검색 엔진의 사용 또한 급격히 늘어나고 있는 추세이다. 국내외 다양한 검색 엔진들이 존재하지만 대부분의 자료들이 기본적인 카테고리별로 링크 횟수나 키워드 빈발 횟수에 따라 정렬이 되어 있다. 그러므로 사용자들은 수동적으로 정렬된 도메인들을 따라 가는 실정이다.

본 논문에서는 수동적인 서비스가 아닌 능동적인 서비스에 중점을 둔다. 특정 카테고리 내에서 접속한 사용자에게 최근 시점을 기준으로 가장 빈번하게 접속된 도메인 정보를 제공하여 시간의 단축과 유용한 서비스를 받도록 한다. 본 논문의 서비스 모델은 인터넷 사용자의 로그 데이터베이스와 도메인 데이터베이스를 기반으로 한다. 본 논문에서 제안하는 카테고리 분류 기법으로 두 데이터베이스를 통합하고 정제한다. 정제된 데이터들은 최종적으로 순차 패턴 마이닝 기법에 의해 최종 빈발 패턴을 추출 하게 되고 특정 카테고리에 접속한 사용자에게 도메인 형태로 변환 되어 서비스 하게 된다.

1. 서론

한국 인터넷 진흥원의 조사에 의하면 2008년 4월 까지 초고속 인터넷 통신 가입자의 수는 약 1,503 만 명으로 나타나고 있다[1]. 도메인 서비스를 이용하는 인구도 늘어나고 있다. 이러한 서비스를 받고자 사용자들이 쉽게 사용하고 접속하는 방법으로 검색엔진 (searching engine) 을 이용한다. 검색엔진은 인터넷상에서 정보를 찾기 위한 도구로서 소프트웨어 또는 웹 사이트를 말한다. 국내에서는 네이버(NAVER), 다음(Daum), 야후코리아(YAHOO) 등이 활발히 사용된다. 현재 국내 사용자들 사이에 각광을 받고 있는 네이버 검색 방법으로는 넥서치 시그마 (Nexearch sigma)라는 지능형 검색서비스가 사용되고 있다. 이것은 자연어 검색을 통한 문서의 정확도와 웹 페이지의 링크 인기도를 결합하여 검색된 페이지들의 순위를 매기는 개념을 도입하여 사용자의 의도에 정확하게 부합되는 사이트와 웹문서를 찾아주는 서비스이다[2].

이 밖에도 다양한 검색엔진의 종류가 존재 하지만 공통적인 서비스 제공 방법은 기본적으로 키워드 검색을 통한 서비스를 제공하고 나열된 정보를 따라 사용자가 수동적으로 따라 가는 형식이다. 이러한 방식의 문제점은 단순 링크의 횟수와 키워드 빈발횟수에 따른 정렬이기 때문에 정렬된 순서에 따라 순차적으로 찾아가야 하는 불편함을 초래한다. 또한 특정 카테고리 내에서 어떠한 도메인들이 현 시점에서 사용자들에 의해 빈번하게 접근 되는가에 대한 정보를 알 수 없다.

본 논문에서는 이러한 기존의 수동적인 서비스와 다른 능동적인 서비스 제공에 초점을 둔다. 그 방법으로 각각의 카테고리에 따라 고유한 번호를 적용 한다. 최근시점에서 사용자들의 접속 패턴을 파악하여 유용한 연관된 도메인들을 추천 한다. 이러한 신뢰성 있는 연관도메인들을 추출하기 위해서 본 시스템에서는 카테고리 분류법을 적용한다. 그리고 정제, 통합된 데이터는 순차 패턴 마이닝 (Sequential pattern mining) 기법을 이용하여 선택되는 카테고리에 따라 의미 있는 연관도메인을 추출 한다. 본 논문에서 제안하는 서비스 방식은 이용자의 편의를 도모 할 수 있을 것이며, 사용자의 특정한 검색엔진 사용 횟수의 증가는 서비스 제공자의 이익 실현과도 연결되어 이익을 창출 할 것이라 예상된다.

본 논문의 2 장에서는 연관도메인 서비스에 사용되는 연관규칙과 서비스에 적용되는 관련된 기법을 기술 한다. 3 장에서는 제안된 시스템의 서비스 방식의 모델과 적용된 카테고리에 따른 분류 기법을 기술 한다. 마지막 4장은 본 논문의 전체적인 고찰을 제시 하면서 결론을 맺는다.

2. 관련 연구

본 장에서는 연관 규칙과 순차패턴 마이닝 알고리즘에 대해 살펴본다[4,5].

2.1 연관 규칙(Association Rule)

연관 규칙은 데이터베이스에 존재하는 여러 항목(item)들 간의 연관성을 탐색하여 의미 있는 패턴들을 결과적으로 찾아내는 것을 말한다.

연관규칙은 A→B의 형태로 표현한다. 예를 들어, "Item A가 구매된 경우, Item B도 구매됨" 이라고 해석된다. 연관규칙 탐사과정은 크게 <표 1>과 같이 두 단계로 진행이 된다[3].

1 단계	미리 결정된 최소 지지도를 만족하는 빈발 항목집합(Itemset)을 구성
2 단계	1 단계에서 구성된 빈발항목집합을 이용하여 연관 규칙을 생성. 모든 빈발항목집합에 대해 공집합이 아닌 부분집합들을 찾고, 각각의 부분집합에 대하여 최소 신뢰도를 만족하는 규칙을 생성

<표 1> 연관 규칙 탐색 단계

연관 규칙 탐색 과정에서 중요한 요소인 지지도(Support) 와 신뢰도(Confidence)는 아래와 같이 정의 된다.

정의 1. 지지도(Support)

전체 거래 중 Item A와 Item B 동시에 포함하는 정도를 나타내며 전체적인 구매 도에 대한 정도를 알 수 있다.

$$S = P(A \cap B) \qquad \text{식(1)}$$

정의 2. 신뢰도(Confidence)

Item A를 포함하는 거래 중 Item B가 포함될 확률의 정도를 나타내며 연관성의 정도를 알 수 있다.

$$C = P(A \mid B) = \frac{P(A \cup B)}{P(A)} \qquad \text{식(2)}$$

연관 규칙이 성립되기 위해서는 신뢰도(Confidence)와 지지도(Support)가 최소 경계 값(threshold)을 넘어야 한다.

2.1.2 순차패턴 마이닝(Sequence pattern mining)

순차패턴 마이닝 알고리즘은 각각의 사용자에 대해 일련의 트랜잭션이 발생한 시간별로 정렬된 시퀀스 데이터베이스가 있다면, 각 사용자별로 최소의 지지도를 가진 모든 순차적인 패턴을 찾아내는 것이다[5,6]. 위에서 설명한 연관 규칙이 발생한 횟수에 의해 함께 발생할 가능성이 높은 Itemset을 찾는 것과 다르게 발생순서가 고려되어 시간이라는 요소가 포함된다. 각 Item의 연관성이 측정되고, 이에 따라 유용한 패턴을 찾는 기법이다. 순차패턴 마이닝 알고리즘은 <표 2>와 같은 5 단계로 구성 된다.

1 단계	정렬 단계로 데이터베이스에서 고객 ID와 시간을 각각 주요키와 보조키로 설정하여 정렬하고, 기존의 데이터베이스를 고객 시퀀스 데이터베이스로 변환
2 단계	최소 지지도를 만족하는 빈발 항목을 추출하고, 빈발 1- 시퀀스 집합도 추출
3 단계	빈발 시퀀스를 빠르게 검색하기 위한 형태로 변경
4 단계	빈발 시퀀스를 최소 지지도에 의해 발견
5 단계	발견된 빈발 시퀀스들 중에서 최대 시퀀스를 발견

<표 2> 순차 패턴 마이닝 5 단계

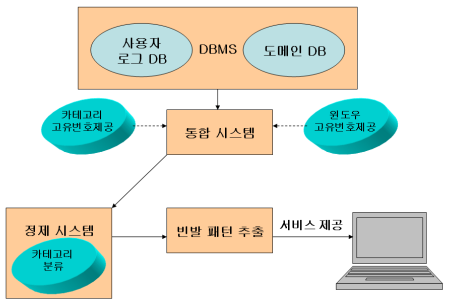
최대 시퀀스(Maximal Sequence)가 발생되어진 고객의 모든 시퀀스를 고객 시퀀스(Customer Sequence)라고 정의 한다. 그리고 각각의 최대 시퀀스들을 순차 패턴이라 하고 빈발 시퀀스란 최소지지도를 만족하는 시퀀스를 의미 한다.

3. 제안된 서비스 모델

본 장에서는 제안하고자 하는 서비스의 전체적인 모델과 각 단계의 모듈에 적용된 기법들에 대해 기술 한다.

3.1 제안된 서비스 모델

본 논문에서 제안하는 서비스 모델 크게 4 단계로 분류된다.

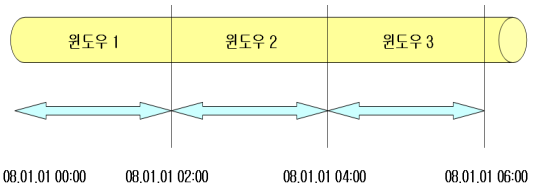


(그림 1) 제안된 서비스 모델

3.1.1 1-단계

1 단계는 사용자 로그 데이터베이스와 도메인 데이터베이스를 통합하는 단계이다. 이 단계에서는 후에 3 단계에서 수행되는 데이터 마이닝 과정의 속도와 처리 되어야

할 사용자 로그 데이터의 용량을 줄이기 위한 작업이 이루어진다. 도메인 서버에 있는 데이터들은 카테고리에 따라 고유한 번호가 주어지며, 로그 데이터베이스에 있는 데이터들은 분류가 된 도메인 서버와 연동을 하여 각각의 로그와 부합하는 번호로 변환 된다. 발생된 시간에 따라 윈도우에 의해 나누어진 항목들은 고유의 윈도우 번호를 가진다.



(그림 2) 시간 간격에 따른 윈도우

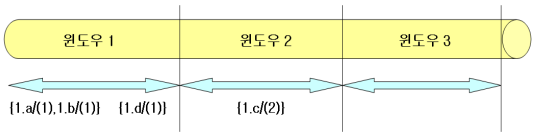
(그림 2)에서 윈도우는 정해진 시간간격에 따라 구별 된다. 제안하는 시스템에서는 2시간이라는 기본적인 설정을 하고, 윈도우 내의 데이터양이 기존에 비해 1.5배에서 2배로 증가 하거나 감소할 경우 30분 간격으로 확장 시키거나 줄일 수 있도록 한다.

예를 들어 음악, 스포츠, 동영상 이라는 큰 카테고리에는 각각 1, 2, 3 의 고유번호가 할당 되며, 카테고리에 속한 도메인은 알파벳순으로 저장된다. 사용자 로그 데이터에 한국 야구 홈페이지가 남아있다면 스포츠 카테고리에 속하게 되며 1.a 와 같은 형태로 변환이 된다. 마지막으로 윈도우 고유 번호를 할당 한다. 데이터는 1.a/(윈도우 고유번호) 와 같은 형식으로 변환 된다.

3.1.2 2-단계

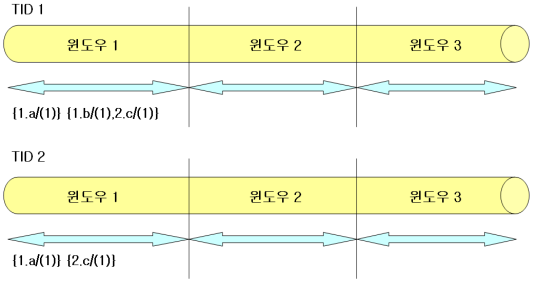
2 단계는 정제 단계로 의미 있는 데이터를 추출하기 위한 사전 단계이다. 시간에 따라 접근된 도메인의 연관성을 판단하여 연관성이 유효하지 않다고 판단되는 항목들이 삭제된다. 기본적으로 카테고리에 따라 분류되므로 같은 고유 카테고리 번호를 가진 항목들은 연관성이 있는 것으로 간주한다. 하지만 같은 카테고리 내의 도메인이라도 시간에 따라 부여된 윈도우의 고유 번호에 따라 연관 유무를 판단하게 된다. 반대로 다른 카테고리에 속하는 항목이라도 윈도우 내에서 최소 지지도를 만족한다면 연관성이 있는 것으로 간주 하여 삭제되지 않는다.

(그림 3)은 정의된 윈도우에 따라 시간이 구분되어 각각의 윈도우 번호를 가지고 있다.



(그림 3) 카테고리 분류법 예1

(그림 3)에서 (1.c)는 같은 카테고리 번호를 가지고 있더라도 {1.a,1.b},{1.d}와 다른 윈도우에 위치하므로 항목에서 제외된다.



(그림 4) 카테고리 분류법 예 2

(그림 4) 에서는 (2.c) 항목은 다른 카테고리 번호를 가지고 있더라도 같은 윈도우 상에 존재 하며 최소 지지도가 2 라고 할 경우 이를 만족함으로 연관성이 있는 항목이라 인정하고 제외되지 않는다.

3.1.3 3-단계

3 단계는 1, 2 단계를 거쳐 정제된 데이터를 대상으로 최소지지도에 입각 하여 빈발 항목을 찾는다. 그리고 빈발 항목을 기준으로 길이 1, 2, 3 ... n 의 빈발 시퀀스를 추출하고 최종적으로 최대 빈발 시퀀스를 추출하게 된다.

3.1.4 4-단계

4 단계는 3단계에서 추출된 최대 빈발 시퀀스를 특정 카테고리에 접속한 사용자에게 서비스를 제공하는 단계이다.

3.2 제안된 서비스 과정

<표 3>의 테이블은 카테고리 분류법으로 나눠지기 전 통합 단계를 나타낸다.

TID	Category Number	Domain List
1	1	{1.a/(1)} {1.b/(1)} {1.f/(1)}
2	1	{1.c/(1),1.j/(2)} {1.a/(1)} {1.k/(1),1.d/(1),2.g/(1)}
3	1	{1.a/(1),1.c/(1)}
4	1	{1.a/(1)} {1.k/(1),1.g/(1)} {1.f/(1)}

<표 3> 1-단계 결과 데이터

다음 <표 4>는 2 단계인 정제 과정을 거친 데이터들로 카테고리 1 에 대한 최대 빈발 시퀀스를 추출 이전의 데이터이다. 최소 지지도를 2로 하였을 경우 <표 4>에서 {1.j}는 같은 카테고리 내에 있지만 윈도우 고유번호가 일치하지 않아 제외된다. {2.g}의 경우 카테고리 1의 항목들과 같은 윈도우에 위치하고 최소지지도가 2를 만족함으로 삭제되지 않는다.

TID	Category Number	Domain List
1	1	{1.a} {1.b} {1.f}
2	1	{1.c} {1.a} {1.k,1.d,2.g}
3	1	{1.a,1.c}
4	1	{1.a} {1.k,1.g} {1.f}

<표 4> 2-단계 결과 데이터

<표 5>에서는 카테고리 1에 대한 최대 빈발 시퀀스 추출 결과를 보여주고 있다. 여기서 {1.a}{1.f}는 TID 1,4 에 의해 지지되고 {1.a}{1.k,2.g}는 TID 2, 4에 의해 지지된다. 다른 {1.a}, {1.c}, {1.k}, {1.g}, {1.f}, {1.a,1.k}, {1.a,2.g} 시퀀스들도 주어진 지지도를 만족하지만 최대 시퀀스가 아니므로 제외된다.

최소지지도가 2일 때 최대 빈발 시퀀스	{1.a} {1.f}
	{1.a} {1.k,2.g}

<표 5> 정제된 데이터로 최대 빈발 시퀀스 추출

고유 번호로 표현된 List를 최종적으로 도메인 DB를 통해 사용자가 알 수 있는 주소로 변환되고, 특정 카테고리에 접근한 사용자에게 서비스된다. 최종적으로 서비스를 받은 사용자는 빈발 시퀀스로 제공되는 도메인을 추천 받아 접근을 하게 된다. 즉, 현 시점에서 접속률이 높은 도메인에서 일정 시간 뒤에 연관성이 높은 도메인으로 이동하면서 단순 링크 횟수에 의한 도메인 정렬보다 의미 있는 정보를 제공 받을 수 있게 된다. 예를 들어, 사용자는 한국 야구로 스포츠 카테고리에 접속하면 한국 야구 홈페이지, 특정 구단 홈페이지, 특정 구단의 경기 일정 등 연관된 도메인 서비스를 제공 받게 된다.

4. 결론 및 향후 연구과제

본 연구는 특정 카테고리에 접속하는 사용자에게 최근 시점을 기준으로 가장 빈발 하게 접속되는 연관 도메인을 카테고리별로 나누어 제공한다. 이런 점에서 각 카테고리에 따라 사용자들이 선호하는 도메인들을 추천 받을 수 있게 되며, 유용한 정보를 보다 신속하고 카테고리 별로 정확한 서비스를 제공 받을 수 있는 장점이 있다. 또한 특정한 카테고리에 국한되지 않고, 다른 카테고리의 정보라도 특정 시점에서 유용한 정보일 경우 사용자에게 접근의 문을 열어 놓음으로써 폭넓은 선택의 서비스를 제공 할 수가 있다.

감사의 말

본 연구는 지식경제부 및 정보통신 연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었으며 (IITA-2008-C1090-0801-0028), 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 지식경제부의 유비쿼터스 컴퓨팅 및 네트워크 원천 기반기술 개발사업의 08B3-B1-10M 과제로 지원된 것임.

참고문헌

[1] http://isis.nida.or.kr/? 한국인터넷 진흥원

[2] http://www.meganews.co.kr/news/internet/0,3903121110037272,00.htm

[3] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," In Proc. of the 20th International Conference on Very Large DataBases (VLDB94), pp.487-499, Santiage, Chile, September 1994.

[4] Jiawei Han, "Miceline Kamber, Data Mining Concepts and Techniques," Simon Fraser University, pp.227-236, 2001.

[5] Mannila, H, Toivonen, H, and Verkamo, L, "Discovery of Freqeunt Episodes in Event Sequences", Department of computer Science Series of Publications Report C-1997-15, 1997

[6] B. Mobasher, et al., "Automatic Personalization on Web Usage Mining,"Technical Report TR99-010,Department of Computer Science, Depaul University,1999.

[7] 이용준, 서성보, 류근호, 김혜규,"시간간격을 고려한 시간 관계 규칙 탐사 기법", 정보과학회논문지, 제28권3호,p301-314,2001.