TV-Anytime 메타데이터 연속 데이터 마이닝을 이용한 시청 선호도 프로파일 생성 기법

신세정, 이원석
연세대학교 컴퓨터과학과
e-mail : starofu@database.yonsei.ac.kr
leewo@database.yonsei.ac.kr

User Behavior Profiling based on Continuous Data Mining

Se-Jung Shin, Won-Suk Lee
Dept. of Computer Science, Yonsei University

요 약

최근 시작된 국내 디지털 지상파 방송으로 이제 본격적인 디지털 방송 시대가 열리게 되었다. 디지털 방송 서비스는 다매체, 다채널을 통한 방송 프로그램의 증가와 양방향 TV 방송 서비스로 인해 사용자에게 다양한 방송 프로그램의 선택과 개인별 맞춤형 시청 기회를 제공함으로써 새로운 방송 서비스 환경을 필요로 하게 되었다. 이에 본 논문에서는 맞춤형 DTV(Digital TV) 방송 서비스를 제공하기 위하여 TV-Anytime 영상 메타데이터에 대한 연속 데이터 마이닝 기법을 이용하여 시청 선호도 프로파일을 생성하는 효율적인 기법을 제안한다. 또한, 내장형 운영체제 기반의 사용자 디스플레이 모듈을 제공하며, 실험을 통하여 본 논문에서 제안하는 방법의 효율성을 고증한다.

1. 서론

최근 시작된 국내 디지털 지상파 방송으로 이제 본격적인 디지털 방송 시대가 열리게 되었다. 기존의 TV 방송 서비스에서는 방송프로그램을 단순히 수동적으로 시청하는 형태가 주를 이루지만, 디지털 방송 서비스는 양방향 TV 방송 서비스를 가능하게 함으로써 정보 공급자와 사용자 간의 수평적 커뮤니케이션이 가능하게 되었다. 다매체, 다채널을 통한 방송 프로그램의 증가와 양방향 TV 방송 서비스는 사용자에게 다양한 방송 프로그램의 선택과 개인별 맞춤형 시청 기회를 제공함으로써 새로운 방송 서비스 환경을 필요로 하게 되었다.

본 논문에서는 맞춤형 DTV(Digital TV) 방송 서비스를 제공하기 위하여 각 사용자의 시청 패턴을 분석함으로써 개별 사용자의 취향에 맞는 방송 정보를 선별적으로 제공할 수 있도록 하는 시청 선호도 프로파일 생성 기법을 제안한다. 이러한 맞춤형 DTV 방송 서비스의 구현을 위해서는 DTV 방송 스트림과 다채널 방송에 대한 EPG(Electronic Program Guide) 데이터 등 매시간 끊임없이 생성되는 실시간 데이터들의 분석을 필요로 한다. 기존의 대용량 데이터 베이스에 대한 데이터 분석기법들로 이러한 정보들을 저장/분석하는 경우 과도한 컴퓨팅 자원이 요구될 뿐만 아니라 분석 과정 동안 새로 생성된 정보들이 축적되어 실시간으로 그 결과를 활용할 수 없다. 따라서 이러한 실시간 정보들을 한정된 컴퓨팅 자원 내에서 실시간 분석을 할 수 있는 연속 데이터 마이닝 기법을 적용한다. 또한, 다채널 서비스를 보다 효율적으로 이용할 수 있도록 시청 선호도 프로파일 디스플레이 모듈을 구현하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구에 대해 알아보고 3 장에서는 제안한 시스템의 블록 다이어그램을 설명한다. 4 장에서는 TV-Anytime 영상 메타데이터 스트림을 실시간으로 분석하여 시청 선호도 프로파일을 생성하는 알고리즘에 대하여 기술하고 5 장에서는 사용자 인터페이스에 대하여 설명한다. 6 장에서는 실험결과 및 성능에 대해서 논의 하고, 마지막으로 7 장에서 결론을 기술한다.

2. 관련연구

사용자 맞춤형 방송을 위해서는 방송 컨텐츠의 내용과 사용 환경을 기술하기 위한 메타데이터가 제공되어야 한다. 멀티미디어의 컨텐츠 서술을 위한 표준으로는 MPEG-7[2]이 있으며 멀티미디어의 전송, 소비 등의 사용 환경 기술에 중점을 둔 표준으로는 MPEG-21[3]이 있다. 메타데이터 기반의 맞춤형 방송을 위한 표준은 현재 TV-Anytime Forum[1]을 중심으로 활발하게 연구되고 있다. TV-Anytime Forum 은 개인 저장장치(PDR: Personal Digital Recorder)를 이용하여 사용자가 원하는 컨텐츠를 원하는 시간에 효율적으로 선택 시청할 수 있는 맞춤형 디지털 방송 제공을 목적으로 설립되었다. TV-Anytime 의 메타데이터는 MPEG-7 에서 정의된 내용 기반 정보와 EPG 정보를 포함하며 크게 프로그램 기술 메타데이터와 사용자 기술 메타데이터로 구별된다. 프로그램 기술 메타데이터는 방송 프로그램에 대한 고유번호 부여 방식과 컨텐츠 내용 표현을 위한 메타데이터들로 이루어져 있다. 사용자 기술 메타데이터는 다양한 값들로 사용자 선호도를 기술할 수 있도록 정의하고 있다.

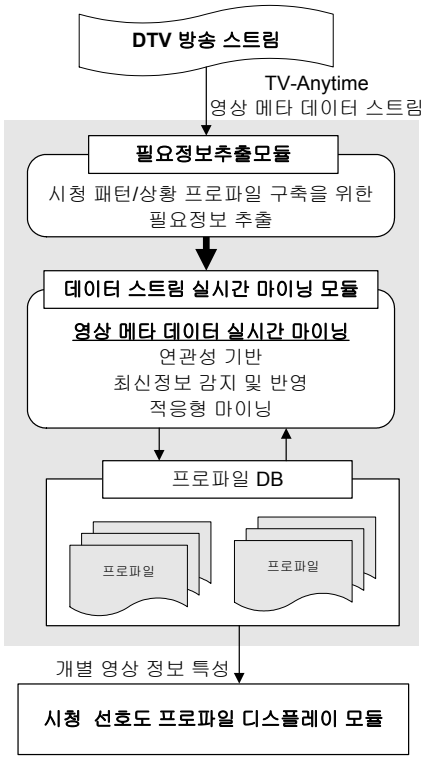
디지털 영상 저장장치를 통한 방송 서비스를 제공하는 *TiVo*[4] 는 협동 필터링(Collaborative Filtering)방법을 사용하여 프로그램의 선택에 유사한 성향을 보였던 사용자들은 새로운 프로그램에 대해서도 유사한 성향을 보인다는 가정을 바탕으로 하는 추천 시스템을 이용한다. 이 방법은 내용분석의 용이성에 관계없이 어떠한 프로그램에 대해서나 분석이 가능하다는 장점이 있지만, 새로운 사용자들이 등장하면, 기존 사용자 관련 데이터를 저장하는 데 드는 비용과 계산을 수행하는 데 걸리는 시간이 급격히 증가하게 되는 단점이 있다. 또한, 사용자의 피드백을 필요로 한기 때문에 데이터에 많은 오차를 포함할 수 있다.

[8]에서는 사용자들의 시청 패턴을 분석하여 표적 광고를 제공하는 *Advertising Delivery System*(*ADS*)을 제안한다. *ADS* 는 PVR(Personal Video Recorder)로부터 사용자의 시청 데이터를 받아서 분류한다. 이러한 정보는 사용자의 인구통계학적/심리적 정보들과 함께 프로파일로 생성되며, 다중 데이터 마이닝을 통해서 정제된 후, 사용자의 투표 시스템에 의해 결과를 조정한다. 이 시스템은 개별적인 사용자보다는 가족단위의 분류에 중점을 두고 있으며 프로파일이 끊임없이 생성되는 프로그램 시청 정보에 즉각적으로 적응하지 못하는 단점이 있다.

Conceptual Matching[9]은 사용자가 이전에 시청한 프로그램들로부터 학습 데이터를 구성하여 추천 프로그램을 제공하는 방법을 제안하였으며 그 밖에도 사용자 맞춤형 서비스를 제공하기 위한 다양한 방법들이 제안되고 있다.

3. 시스템의 블록다이어그램

TV-Anytime 영상 메타데이터 스트림을 마이닝하여 시청 선호도 프로파일을 생성하기 위한 전체 시스템 구성도는 그림 1 과 같다. 미리 정의된 내용 기반 주요 키워드의 종류와 구조에 따라 필요 정보 추출 모듈에서 TV-Anytime 영상 메타데이터로부터 키워드를 추출한다. 빈발 영상 키워드 해쉬 테이블을 이용하여 영상 메타 데이터로부터 추출된 키워드 중 출현 빈도가 높은 빈발 영상 키워드를 마이닝 모듈에 넘긴다.



(그림 1) 전체 시스템 구성도

데이터 스트림 실시간 마이닝 모듈은 지속적으로 발생되는 사용자별 영상 메타데이터 인스턴스의 키워드를 실시간으로 마이닝하기 위한 사용자별 영상 메타 데이터 스트림 마이닝 구조와 빈발 영상 키워드 집합을 관리하는 모듈로 구성된다. 생성된 사용자별 시청 선호도 프로파일은 선호도 프로파일 디스플레이 모듈을 통해 EPG 의 프로그램 키워드를 이용하여 시청 선호도 프로파일 정보에 부합하는 추천 프로그램 리스트와 함께 제공된다.

4. 시청 선호도 프로파일 생성 기법

본 논문에서 제안하는 영상 메타데이터의 실시간 마이닝을 위해서는 TV-Anytime 의 프로그램 기술 메타데이터에 대한 전처리 과정을 거쳐 주요 키워드들이 추출된다. 즉, 사용자의 시청 선호도 프로파일 구성에 이용되는 프로그램의 제목, 방송 시간, 장르, 출연진, 제작진 등과 같은 데이터들이 추출되며, 영상 키워드 해쉬 테이블을 이용해 추출된 영상 키워드와 해당 키워드의 출현빈도수가 함께 관리된다.

4.1 영상 메타데이터 스트림 마이닝

데이터 스트림 형태로 제공되는 영상 메타데이터는 사용자가 시청한 한 프로그램의 기술 메타 데이터가

하나의 트랜잭션을 이루며 각 트랜잭션에 출현하는 키워드들을 각각 개별적인 항목으로 간주한다.

영상 메타데이터 마이닝을 위해 이전 연구에서 제안된 온라인 데이터 스트림 환경에서 빈발항목집합을 탐색하는 *estDec* 방법[5]을 사용하였다. *estDec* 방법은 지연추가와 전지 작업을 통하여 빈발항목집합이 될 가능성이 있는 항목집합들만을 관리한다. 데이터 스트림에서 출현한 모든 항목들은 해당 항목이 가까운 미래에 빈발항목이 될 수 있을 때까지 전위트리에 대한 추가작업이 지연된다. 즉, 잠재적으로 중요한 항목이 충분히 큰 지지도를 갖는 적절한 시점에 전위트리 구조에 추가된다. 또한 효율적인 항목 전지 방법을 통해 이전 데이터 집합에서 중요한 항목이었더라도 해당 항목의 현재 지지도가 사전에 정의된 최소 지지도 보다 매우 작다면 해당항목은 전위트리 구조로부터 전지된다. 이러한 방법을 통해 메모리에 관리되는 전위트리의 크기를 감소시킬 수 있다.

4.2 빈발 영상 키워드 집합 탐색 모듈

영상 메타데이터 스트림이 영상 메타데이터 실시간 마이닝 모듈을 통해 전위트리 구조에 관리되면, 빈발 영상 키워드로부터 생성되는 빈발 영상 키워드 집합을 탐색하여 관리한다. 이는 전위트리를 깊이 우선 탐색(DFS: Depth First Search) 방법으로 순회하며 사전에 정의된 사용자 임계값인 최소 지지도 이상의 값을 갖는 빈발 영상 키워드를 추출하고 이로부터 생성되는 빈발 영상 키워드 집합을 생성한다.

4.3 단일/그룹 사용자 시청 선호도 프로파일

단일 사용자 시청 선호도 프로파일은 사용자별 시청 선호도 특성 분석을 위해 영상 메타 데이터 스트림 마이닝 구조에서 관리되는 빈발 영상 키워드 집합으로부터 시청 패턴을 분석하여 생성된다. 또한, 그룹 사용자 그룹을 구성하는 단일 시청 선호도 프로파일로부터 공통된 빈발 영상 키워드 집합을 추출하여 그룹 사용자의 시청 선호도 특성을 반영하는 그룹 시청 프로파일을 구축하는데 사용된다. 사용자간 선호도 거리란 단일 사용자 시청 선호 키워드 집합들간의 거리 계산 지표로서 식 1 과 같이 계산된다.

$$Distance(A\ to\ B) = \frac{\sum_{e_i \in (A-B)}^{i} S_A(e_i) + \sum_{e_i \in (A \cap B)}^{i} |S_A(e_i) - S_B(e_i)|}{\sum_{e_i \in A}^{i} S_A(e_i)}$$

$$(0 \le \text{Distance} \le 1)$$

A : 사용자 A 의 선호 키워드 항목집합
B: 사용자 B 의 선호 키워드 항목집합
$S(e_i)$: 선호 키워드 항목 e_i의 지지도

(식 1) 사용자간 선호도 거리 계산

5. 구현

본 논문에서는 Set Top Box 와 같은 다채널 방송 수신 장치에서 TV-Anytime 영상 메타데이터를 파싱(parsing)하여 시청 선호도 프로파일을 생성하는 것으로 가정하고 내장형 운영체제인 Pocket PC 2003 기반의 시스템을 구축하였다.

(그림 2) 단일/그룹 시청 선호도 프로파일 생성

그림 2 는 각각 단일 사용자의 시청 선호도 프로파일과 그룹 시청 선호도 프로파일을 나타낸다. 사용자의 시청 로그를 분석하여 프로파일이 생성되며, 이 프로파일에 따라 EPG 와 비교하여 지정한 시간에 방송되는 추천 프로그램을 매치포인트와 함께 제공한다.

(그림 3) 시청 선호도 프로파일의 적응성

그림 3 은 사용자의 시청 패턴의 변화에 따른 시청 선호도 프로파일과 추천 프로그램의 적응성을 보여준다. 사용자의 시청 패턴이 주로 부동산, 바둑 프로그램 위주에서 스포츠 위주로 변해갈 경우, 적응적인 실시간 마이닝을 통해 프로파일과 추천 프로그램이 적응적으로 변화함을 보여주고 있다.

6. 실험 결과 및 성능평가

디지털 방송은 TV-Anytime 영상 메타데이터가 수신되어 EPG 데이터와 함께 제공되어야 하나, 현재 이러한 데이터를 얻기 어려우므로 인터넷에서 서비스되는 EPG 에서 프로그램 정보를 추출하여 실험에 사용하였다. 총 105 개의 채널로부터 표 1 에 가정한 사

용자들의 2 개월간의 시청 로그를 생성하고, 1 주일간의 EPG 데이터를 바탕으로 추천 프로그램 리스트를 생성하도록 하였다.

<표 1> 사용자별 시청 패턴 가정

사용자	시청 방송의 주요 키워드
사용자 1	부동산, 경제, 증권, 주식 재테크, 뉴스, 국제정세, 경제동향, 정치, 생활, 날씨, 바둑, 스포츠, 골프, 축구, 프로야구, 해외스포츠
사용자 2	드라마, 연예/오락, 홈쇼핑, 패션, 뷰티, 부동산, 경제, 증권, 주식, 재테크, 뉴스, 국제정세, 경제동향, 정치, 생활, 날씨
사용자 3	드라마, 연예/오락, 교육, 수능, 영화, 뉴스, 국제정세, 경제동향, 정치, 생활, 날씨
사용자 4	영화, 음악, 연예/오락, 스포츠, 축구, 프로야구, 해외스포츠, 컴퓨터게임, 교육
사용자 5	관광/여행, 바둑, 뉴스, 국제정세, 경제동향, 정치, 생활, 날씨, 시사/다큐, 의료/건강

시스템의 성능평가를 위해 대용량 데이터 베이스 환경에서 빈발항목 탐색을 위한 마이닝 방법인 *Apriori*[6]와 데이터 스트림 환경에서 빈발항목 탐색을 위한 마이닝 방법인 *Lossy Counting* 방법[7]을 본 논문에서 사용한 *estDec* 방법과 비교하였다.

시청프로그램이 끊임없이 생성되는 실시간 데이터임을 고려할 때, 그림 4 에서 시청 프로그램이 늘어날수록 실행시간이 급속하게 증가하는 *Apriori* 방법은 적절하지 못하다는 것을 알 수 있다.

그림 5 은 최소 지지도의 변화에 따른 각 알고리즘의 메모리 사용량의 변화를 보인다. *Lossy Counting* 알고리즘은 버퍼의 크기와 일괄처리하는 트랜잭션의 수가 비례하므로, 버퍼의 크기가 증가함에 따라 높은 효율을 보이지만 결과적으로 빈발항목집합 탐색을 위해 필요한 메모리 사용공간이 증가하는 단점이 있다. 이에 반해 *estDec* 방법은 지연추가와 전지작업을 통해 효율적으로 메모리를 관리하므로 상대적으로 적은 메모리를 사용한다.

그림 6 는 시청프로그램수의 증가에 따른 각 알고리즘들의 오차를 비교한 그래프이다. 제안된 방법들의 상대적인 정확도를 표현하기 위해, [5]와 같은 방법으로 두 빈발항목 결과 집합에 대해 average support error *ASE* 를 사용하였다. *ASE* 값이 작을수록 비교하는 마이닝 결과 집합이 유사함을 나타낸다. 데이터의 양이 증가할수록 *estDec* 방법의 마이닝 오차가 안정화되어감을 알 수 있다.

7. 결론

메타데이터 기반의 맞춤형 디지털 방송 서비스를 위해서는 영상 메타데이터로부터 사용자들의 선호도를 반영하는 정보를 효율적으로 추출하여 관리해야 한다. 본 논문에서는 *estDec* 방법을 이용하여 끊임없이 제공되는 방송 프로그램에 대한 정보를 실시간으로 분석하고 사용자별 시청 선호도 프로파일을 생성하는 방법을 제안하였다. 또한, 내장형 운영체제 기반의 사용자 디스플레이 모듈을 구현하였으며 제안한 방법의 효율성과 정확성을 비교 실험을 통해 입증하였다.

참고문헌

[1] TV-Anytime Forum, http://www.tv-anytime.org/

[2] MPEG-7 Document, ISO/IEC 15938-5, "Information Technology – Multimedia content description interface – Part 5: Multimedia Description Schemes," ISO/IEC JTC1/SC29/WG11, 2004.

[3] MPEG-21 Document, ISO/IEC 21000, "MPEG-21 Overview v.5," ISO/IEC JTC1/SC29/WG11/N5231, 2002.

[4] Ali K., Stam WV. "TiVo: making show recommendations using a distributed collaborative filtering architecture," In Proc. of the 10th ACM SIGKDD Intl. conf. on Knowledge Discovery and Data Mining, pp.394-401, 2004.

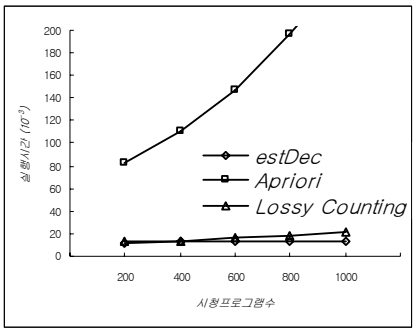
[5] J.H. Chang and W.S. Lee. "Finding recent frequent itemsets adaptively over online data streams," In Proc. of the 9th ACM SIGKDD Intl. conf. on Knowledge Discovery and Data Mining, pp. 487-492, 2003.

[6] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules," In Proc. of the 20th Intl. Conf. on Very Large Data Bases, pp. 487-499, 1994.

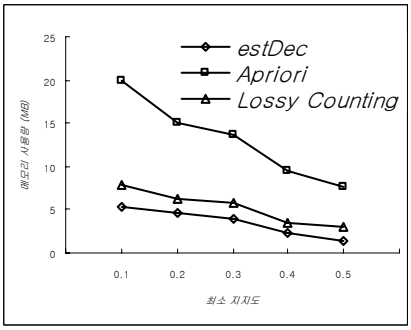
[7] G.S. Manku and R. Motwani. "Approximate Frequency Counts over Data Streams," In Proc. of the 28th Intl. Conf. on Very Large Data Bases, pp. 346-357, 2002.

[8] Spangler W., Gal-Or M., Masy J., Using data mining to profile TV viewers-ACM CACM, December 2003.

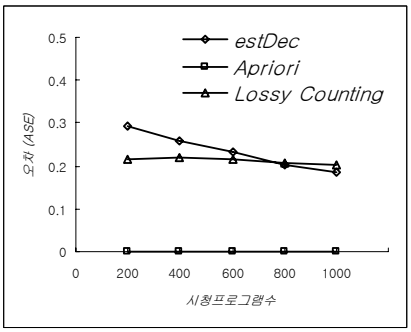
[9] Takagi. T, Kasuya. S, Mukaidono. M and Yanaguchi. T, "Conceptual matching and its applications to selection of TV Programs and BGMs," Systems, Man. and Cybernetic, IEEE SMC. 99. Vol. 3, pp. 269-273, 1999.



(그림 4)실행시간 비교



(그림 5) 메모리 사용량 비교



(그림 6) 마이닝 결과의 오차 비교