한국컴퓨터정보학회 동계학술대회 논문집 제21권 제1호 (2013. 1)

Socially aware computing을 위한 대규모 데이터베이스의 연관 규칙 감축 기법

정휘운*, 박건용*, 박종창°, 윤희용*
*성균관대학교 정보통신대학
°성결대학교 컴퓨터공학부
e-mail:{jeonghw, geonyong}@skku.edu*, gadimen@hanmail.net°, youn@ece.skku.ac.kr*

Association Rule Mining Scheme of Large-Scale Database for Socially Aware Computing

Hwi-Woon Jeong*, Geon-Yong Park*, Jong-Chang Park°, Hee-Yong Youn*
*College of Information and Communication Engineering, Sungkyunkwan University
°Dept. of Computer Engineering, Sungkyul University

● 요 약 ●

연관 규칙 감축 기법은 대규모 데이터를 사용하는 Socially aware computing분야에서 매우 중요한 이슈이다. 본 논문에서는 수집된 각종 데이터들을 각 속성 기준에 따라 이진 변환한 후 가중치를 부여하고 논리식 감축 방법을 이용하여 신뢰성을 보장하는 규칙을 도출하는 새로운 데이터 감축 기법을 제안한다. 이는 컴퓨터 시뮬레이션 결과 기존의 방식들에 비해 지지도, 신뢰도, 규칙 감소율, 연관 규칙 추출 시간에 좋은 성능을 보였으며 이는 빠른 시간 내에 신뢰성 높은 대규모 데이터 처리가 필요한 Socially aware computing분야에 적합하다고 판단한다.

키워드: Association rule reduction, Socially aware computing, Large-scale database.

I. 서 론

오늘날 다수의 사용자의 각종 센서 데이터의 집합을 통해 가치 있는 데이터 규칙을 찾고 그 규칙을 이용하여 사용자의 상황을 판단하여 가장 적합한 서비스를 제공하는 헬스케어 시스템, 지능형 차량 운행 시스템 등의 Socially aware computing 분야가 주목을 받고 있다.[2] Socially aware computing의 실현을 위해 가장 중요한 기술 중 하나는 다수의 사용자로부터 제공된 대규모의 데이터에서 가치 있는 데이터 규칙을 정확하고 빠르게 찾아 서비스를 제공 받는 사용자의 데이터와 비교하여 사용자의 상황을 판단하는 것이다. Socially aware computing 시스템은 마치 실시간 처리 시스템처럼 운용되며 이를 위해서는 대규모 데이터에서의 연관 규칙 감축 기술을 통해 사용자 데이터에 대한 비교군이 되는 가치 있는 데이터의 규모를 최소화하여 데이터 간의 비교 시간을 줄이는 것 또한 매우 중요하다. 따라서 지지도와 신뢰도를 보장하고 데이터 규칙 감축의 효율이 높으며 데이터 처리의 시간적 효율이 좋은 연관 규칙 감소 기법을 연구하는 것은 대규모 데이터를 이용하는 Socially aware computing 분야의 발전에 매우 중요한 일이다.

본 논문에서는 다수의 센서를 통해 수집된 데이터 집합을 각각의 데이터 속성에 따른 기준을 가지고 이진 변환 후, 각 데이터에 대한 사전 확률을 추출, 그에 따른 가중치를 부여하고 논리식 간소화 기법인 Quine Mccluskey method를 사용한 새로운 데이터 감축 기법을 제안한다. 이 방법은 신뢰성이 떨어지는 규칙을 감축시켜 연관 규칙 추출 효율을 향상시킬 수 있다. 또한 컴퓨터 시뮬레이션 결과, 관련 분야에 많이 적용되고 있는 Apriori-algorithm과 FP-growth algorithm에 비해 지지도, 신뢰도, 규칙 감소율, 연관 규칙 추출 시간에 좋은 성능을 보였으며 이는 빠른 시간 내에 신뢰성 높은 대규모 데이터 처리가 필요한 Socially aware computing 분야에 적합하다고 생각한다.

본 논문의 다음 구성은 다음과 같다. 섹션2에서 연관 규칙 감축과 관련 주요 Algorithm에 대해 설명한다. 섹션3에서는 제안하는 연관 규칙 감축 기법에 대한 간단한 설명 및 컴퓨터 시뮬레이션을 이용하여 기존의 방법들과 성능을 비교, 평가한다. 마지막으로 섹션4에서 결론과 향후 연구에 대해서 서술한다.

II. 관련 연구

연관 규칙은 대용량의 트랜잭션들이 누적된 데이터베이스에서 각 트랜잭션간의 상호 관계를 통계적 방법에 의해 연관성이 있는 항목들 사이의 규칙성을 추출하는 과정이다. 즉, "어떤 사건이 일어나면 다른 사건이 일어난다."와 같은 연관성을 말하는 것이

다.[9]

연관 규칙 추출에 있어 널리 사용되는 Apriori algorithm은 각 pass에서 빈발 항목집합들의 후보 항목 집합을 구성하고 난 후에 각 후보 항목 집합의 발생 빈도를 계산하고, 사용자가 정의한 minimum support를 기초로 하여 빈발 항목 집합들을 결정한다.

Apriori algorithm은 최소지지도 이상 항목만으로 이루어진 후보 집합을 생성하여 탐색공간을 줄이고 있지만, 최소 지지도를 만족하는 항목수가 늘어날수록 많은 후보 집합을 생성해야 하고, 후보 집합 생성을 위해 반복적인 데이터베이스 스캔을 필요로 하고 있다. 후보 집합 생성 시 모든 항목집합의 조합을 후보 집합으로 생성하고 데이터베이스 스캔을 통해 후보 집합의 빈발도를 측정하기 때문에 트랜잭션 데이터 내에 존재하지 않는 항목을 후보 집합으로 생성하여 불필요한 공간을 낭비할 수 있다.

또한 본 논문에서 제안한 방법의 성능평가를 위한 비교군으로 사용될 FP-growth algorithm은 빈발 항목을 가지는 데이터베이스를 FP-tree로 압축하고 divide and conquer기법을 사용하여 각 항목에 대하여 연관된 트리를 추출한 조건부 conditional pattern tree를 생성하고 연관규칙을 추출하는 방법이다.

FP-growth algorithm은 후보 집합을 생성하지 않고 빈발 항목을 공유하기 때문에 공간 낭비가 적고, 데이터베이스를 총 두 번만 스캔하므로 속도 면에서 Apriori-algorithm보다 우수하다. 하지만 FP-growth 알고리즘은 빠른 연관규칙추출을 위해 트리 구조를 최대한 간결하게 압축하는 것이 목적이기 때문에 새로운 데이터의 추가가 어려운 단점이 있다.[3]

III. 본 론

이 섹션에서 우리는 대규모 데이터베이스내의 다양한 연관 규칙을 효율적으로 감축하며 신뢰성과 지지도가 보장되는 연관 규칙을 도출하는 방법을 제안한다. 대용량 데이터베이스에서 사용자가 원하는 특별한 데이터들의 신뢰성 높은 규칙을 찾는 것은 대규모 데이터를 다루는 Socially aware computing 분야를 비롯한 많은 IT 분야에 유용하게 쓰일 수 있다.

1. 기본개념

가장 먼저 수집된 각각의 데이터를 각 데이터 속성별 기준을 통해 이진수로 표현한다. 그 후 데이터 집합을 전건부의 규칙의 이상치(1로 표현) 개수에 따라 오름차순으로 나열하며 이 중 이벤트 발생(후건부를 1로 표현)의 규칙을 추출하여 정렬한다. 정렬된 데이터 집합에 사전 확률을 계산하여 그에 따른 가중치(사후 확률)를 도출한다. 이때 최소 가중치에 부합 하지 못하는 규칙은 데이터 집합에서 제거한다. 다음 이렇게 정리된 데이터 집합을 이상치의 개수가 n개인 규칙과 n+1인 규칙을 비교하고 don't care bit가 일치하는 규칙들을 결합한다. 이때 결합된 값은 don't care bit(-)와 care bit(이진)로 표시하며 결합을 통한 규칙 감축을 진행한다. 이는 규칙 감축이 더 이상 진행되지 않을 때 까지 진행되며 감축이 끝난 규칙들의 집합에서 중복 규칙을 제외한다.

예를 들어, 제안하는 방법을 사용하여 표1과 같은 임의의 데이터 집합에 대한 규칙 감소를 실행하면 표2와 같은 최종 규칙을 도출할 수 있다.

표1. 임의의 데이터 집합

Table 1. Random dataset

0	0	0	0	0
0	0	0	1	1
0	0	1	0	0
0	1	0	0	0
0	1	0	1	1
1	1	0	0	1
0	0	1	1	1
0	1	1	0	1
1	0	1	0	0
1	1	1	0	0
1	0	1	1	1
1	1	0	1	0
0	1	1	1	1

표2. 최종 연관 규칙

Table 2. The final association rule

S_1	S_2	S_3	S_4	f
0	-	-	1	1

이렇게 생성된 규칙은 신뢰도와 지지도를 각각 식 (1)과 (2)를 통하여 계산하여 규칙의 연관성을 측정할 수 있다.

$$S(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \quad (1)$$

$$C(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (2)$$

위에서 정의한 식에 따르면 표2에서 규칙 $\{S_4\} \rightarrow f$의 경우 지지도 100% 신뢰도 100%로 나타낼 수 있으며 규칙 감소율은 92%로 계산할 수 있다.

2. 성능평가

이 섹션에서는 본 논문에서 제안하는 방법의 성능을 컴퓨터 시뮬레이션을 이용하여 평가하고자 한다.

연관 규칙 추출기법으로 널리 사용되고 있는Apriori-algorithm 그리고 FP-growth algorithm과 본 논문에서 제안하는 New-Scheme을 비교하여 데이터의 규모에 따른 신뢰도, 지지도, 규칙감소율, 데이터 처리 속도에 대해서 평가를 진행하였다.

평가를 위한 데이터베이스는 임의로 구축한 전건부 7개와 후건부 2개의 속성을 가진 바디센서데이터를 사용하였다. 각 데이터들은 일정 수치를 넘었을 경우 1 그렇지 않은 경우 0인(예를 들어, 몸무게가 정상치면 0 비 정상치면 1) 이진수 형태로 표현되게 하였으며 데이터의 셋의 규모는 1000개부터 10000개까지 점차적으로 늘려가면서 평가를 실행하였다.

그림 1, 2는 각 Algorithm을 이용하여 감축한 결과 중 $\{S_5\} \rightarrow f_1$의 규칙을 선정하여 데이터 규모에 따른 지지도와 신뢰도의 차이를 비교하여 나타낸 것이다.

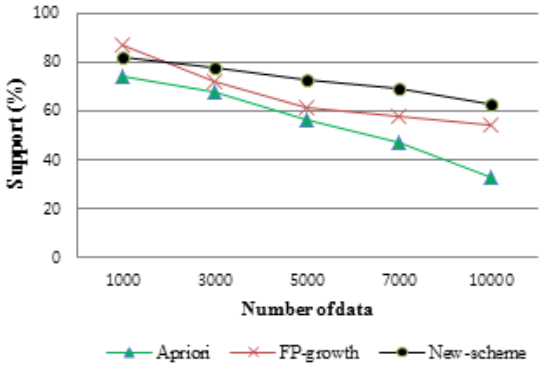

그림 1. 연관 규칙의 지지도에 대한 평가
Fig 1. Evaluation on the support of association rule

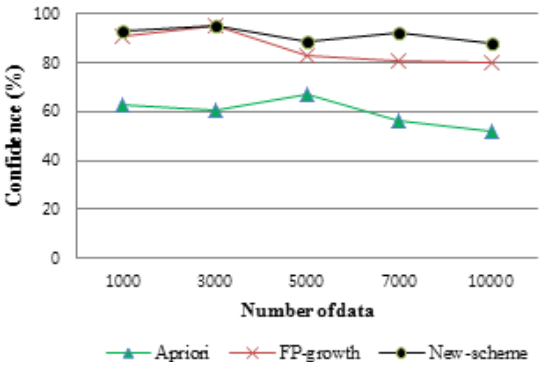

그림 2. 연관 규칙의 신뢰도에 대한 평가
Fig 2. Evaluation on the credibility of association rule

본 논문에서 제안하는 scheme은 비교군인 다른 두 개의 알고리즘에 비해 높은 지지도와 신뢰도를 보장한다는 결과를 알 수 있다.

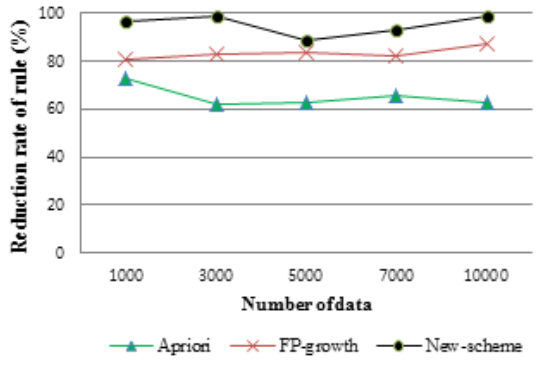

그림 3. 각 Algorithm별 규칙 감소율
Fig 3. Rule reduction rate for each algorithm

또한 그림 3에서 볼 수 있듯이 본 논문에서 제안하는 scheme은 가장 높은 규칙 감축율을 보였고 이는 신뢰성 높은 최소한의 데이터들과 빠른 비교를 통해 정확한 서비스를 도출하는데 효율적임을 보인다.

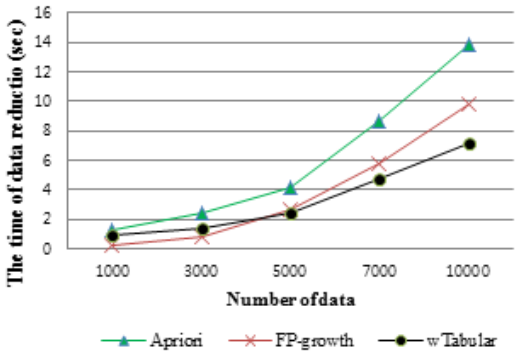

그림 4. 각 Algorithm별 데이터 처리 시간
Figure 4. Data processing time for each algorithm

그림 4는 데이터 규모에 따른 규칙 추출 시간을 비교하여 나타낸 것으로 데이터의 규모가 비교적 작을 때 본 논문에서 제안하는 scheme이 FP-growth algorithm보다 규칙 추출 시간이 오래 걸리지만 데이터 규모가 커질수록 우리의 scheme이 더 효율이 높아진다는 것을 확인하였고 이는 대규모 데이터 처리가 반드시 필요한 Socially aware computing분야에 적절하게 적용될 수 있으리라 생각한다.

IV. 결 론

연관 규칙 감축 기법은 대규모 데이터를 사용하는 Socially aware computing분야에서 매우 중요한 이슈이다. 그동안 대규모의 데이터를 효율적으로 감축하고 연관 규칙을 찾아내기 위해 속성의 계층 그리고 부정 연관 규칙 등을 기반으로한 다양한 Algorithm에 대한 많은 연구가 진행되었다.

본 논문에서는 데이터를 이진 변환하여 가중치를 부여하고 최소 가중치를 만족하지 못하는 데이터를 사전에 제거한 후 논리식 감축 기법을 사용하여 연관 규칙의 감축을 진행하는 새로운 연관 규칙 감축 기법을 제안하였다. 컴퓨터 시뮬레이션을 통해 제안하는 방법의 성능을 평가한 결과 기존의 방식들에 비해 지지도와 신뢰도, 규칙 감소율이 증가하였고 데이터 처리 시간이 감소하는 것으로 나타났다.

향후 연구로는 데이터의 이진 변화 없이 다진 속성 값을 바로 처리 할 수 있는 연관 규칙 감축 방법론에 대한 연구가 요구된다.

ACKNOWLEDGEMENT

본 연구는 방위사업청과 국방과학연구소(UD10070MD), 한국산학연협회(C0017380), BK21 사업, 한국연구재단 기초연구사업(2012R1A1A2040257)의 지원을 받아 수행되었습니다.

참고문헌

[1] wei Zhang,Hongzhi Liao,Na Zhao, “Research on the FP Growth Algorithm about Association Rule Mining,” Business and Information Management ISBIM 2008, pp. 315-318.

[2] Esko Kurvinen, Antti Oulasvirta, “Towards Socially Aware Pervasive Computing: A Turntaking Approach,” Pervasive Computing and Communications, 2004, pp. 346-350.

[3] Sung-Yeol Song, “Incremental association rule mining using FP-tree,” Soongsil Univ., 2010.

[4] Dongme Sun, Shaohua Teng, Wei Zhang, Haibin Zhu, “An Algorithm to Improve the Effectiveness of Apriori,” Cognitive Informatics 2007, pp. 385-390.

[5] Chen Hong-ye, Jin Guo-ying, “Incremental FP_Growth Mining Algorithm Based on Web Information Extraction,” Information and Computing Science, 2009, pp. 91-93.

[6] Paranjape, P.,Deshpande, U, “An Optimistic Messaging Distributed Algorithm for Association Rule Mining,” India Conference (INDICON), 2009, pp. 1-5.

[7] Agrawal,R, R.Srikant, “Fast algorithms for mining association rules,” ACM SIGMOD.

[8] W.G.Teng, M.S.Chen, “Incremental Mining on Association Rules,” in Foundations and Advances in Data Mining, pp. 125-162.

[9] Mi-Yeun Kim, “Study on development of improved association rule algorithm.” Yensei Univ. 2001.

[10] Pang-ning Tan, M.Steinbach, Vipin Kumer, “Introduction to Data mining,” Publisher Addison-Wesley 2006.

[11] John D. Holt and Soon M. Chung, “Mining association rules in text databases using multipass with inverted hashing and pruning,” ICTAI 2002, pp. 46-56.

[12] Bo. Wan, Shiwu.Xu, Lin.Yang, “Combination of Partition Table and Grid Index in Large-Scale Spatial Database Query,” Information Science and Engineering (ICISE), 2009, pp. 2007-2011.

[13] C.Leung, “Can Tree : a canonical-order tree for incremental frequent-pattern mining,” Knowledge and Information Systems, pp. 287-311.

[14] Han J., Kamber M., “Data Mining : Concepts and Techniques,” Morgan Kaufmann, 2006.