대량의 연관규칙에서 의미있는 패턴 추출 기법

이진용*, 문현정, 우용태
창원대학교 컴퓨터공학과
jinmir@orgio.net, {mun, ytwoo}@sarim.changwon.ac.kr

Jin-Yong Lee*, Hyeon-Jeong Mun, Yong-Tae Woo
Dept. of Computer Engineering, Changwon National University

요 약

본 논문에서는 연관규칙 탐사에서 발견된 대량의 패턴 중에서 의미있는 패턴을 효과적으로 추출하기 위한 텍스트마이닝 기법을 제시하였다. Agrawal 등이 제안한 R-interesting 값을 수용하여 의미있는 패턴을 추출하기 위한 방법이다. 대량의 연관규칙중에서 특정 분야에서 추출된 패턴의 빈도수와 다른 분야의 빈도수의 비율에 따른 χ^2값의 A셀에 대한 기여도와 R 값을 비교한 결과 빈도수가 같더라도 다른 분야에 나타나는 비율이 높을수록 기여도와 R 값은 낮게 나타났다. 또한 특정 분야에만 나타나는 패턴에 대해서 빈도수에 따른 기여도와 R 값은 빈도수가 높을수록 기여도는 높아지고 R 값은 변화가 없었다. 이 결과를 이용하여 R 값이 같은 경우 빈도수가 높은 순으로 의미있는 패턴을 추출할 수 있었다.

1. 서 론

인터넷의 확산과 네트워크 기술의 발달로 디지탈 문서의 양이 급격히 증가하고 있다. 이에 따라 대량의 디지탈 문서에서 유용한 지식 정보를 추출하기 위한 지식탐사시스템(KDD, Knowledge Discovery in Database)에 대한 연구가 활발하게 진행되고 있다. 이러한 KDD 시스템을 구성하는 핵심적인 기술은 대량의 문서로부터 연관된 정보를 효과적으로 추출하기 위한 텍스트마이닝 기법이다.

텍스트마이닝을 위한 기존 연구는 단어들간의 의미있는 패턴을 발견하기 위한 연관규칙, 단어간의 순서 관계를 고려한 에피소드 규칙, 관련된 문서끼리 클러스터링하기 위한 개념적 클러스터링, 신경망 기법 등에 관한 연구가 진행되었다[1]. 이 중에서 연관규칙을 이용한 텍스트마이닝 기법은 문서를 자동적으로 분류하기 위한 대표 색인어 추출이나 관련된 문서끼리 클러스터링하기 위한 분야에서 많이 이용되고 있다.

일반적으로 연관규칙 탐사기법에 의해 발견된 연관규칙중에서 실질적으로 의미있는 규칙은 극소수이다. 따라서 대량의 연관규칙 중에서 의미있는 패턴을 효과적으로 탐색하는 문제는 연관규칙에 관한 중요한 연구 방향의 하나로 다양한 형태로 연구가 진행되고 있다. Agrawal 등은 계층 구조에서 상위 계층 규칙에 대한 관측 값과 상위 계층의 규칙에 의해 추정된 하위 계층의 규칙에 대한 기대값의 비율로 정의되는 R-interesting 값을 정의하여 장바구니 분석에서 의미있는 규칙을 찾는 방법을 제안하였다[2].

Cai와 Fu 등은 중요한 아이템에 가중치를 부여하여 의미있는 규칙을 효과적으로 선별하는 방법을 연구하였다[3]. Tan과 Kumar 등은 상관계수를 이용하여 상관관계가 낮거나 음의 상관관계 패턴을 갖는 규칙들을 제거하는 방법을 제안하였고, χ^2 가설 검증방법을 이용하여 패턴을 추출하였다[4]. Klemettinen 등은 템플리트를 이용하여 발견된 연관규칙중에서 무의미한 연관규칙을 제거하거나 관심있는 규칙만 필터링하기 위한 방법을 제안하였다[5]. 또한 계층구조를 이용하여 규칙을 일반화하는 방법과 클러스터링하는 방법도 연구되었다[2, 6].

본 논문은 한국과학재단의 2001년 목적기초연구 (2001-1-30300-015-1) 지원으로 수행되었음.

문서를 대상으로 하는 텍스트마이닝 분야에서도 연관규칙 탐사과정에서 발견된 패턴 중에서 의미있는 패턴을 효과적으로 찾기 위한 문제는 중요한 연구 주제의 하나이다. 특히 텍스트마이닝에서는 아이템이 비정형적이고 불규칙적이어서 정형화된 아이템을 대상으로한 장바구니 분석 기법보다 패턴의 수가 더 많아지는 경향이 있다. 그러나 텍스트마이닝 분야에서 발견된 대량의 패턴으로부터 의미있는 패턴을 찾기 위한 연구는 별로 진행되지 않았다. Ahonen 등은 연관규칙을 수정한 에피소드 규칙을 이용하여 인접해서 출현한 단어 집합을 발견하기 위한 연구를 진행하였다[8].

본 논문에서는 문서를 대상으로 연관규칙 탐사기법을 적용하여 발견된 대량의 패턴중에서 의미있는 패턴을 효과적으로 추출하기 위한 새로운 기법을 제시하였다. 제안한 방법은 Agrawal 등이 제안한 R-interesting 값을 이용하여 의미있는 패턴을 추출하기 위한 텍스트마이닝 기법이다. 컴퓨터 분야의 논문을 대상으로한 실험을 통하여 대량으로 발견된 연관규칙중에서 R 값의 변화에 따라 의미있는 패턴을 효과적으로 추출하기 위한 방법을 제시하였다. 이렇게 구성한 의미있는 패턴은 KDD 시스템에서 문서 분류를 위해 대표 색인어를 효과적으로 구성하기 위한 방법에 적용할 수 있다.

제안한 기법의 효율성을 검증하기 위해, 대표적인 통계적 기법의 하나인 χ^2 가설 검정 기법과 비교 실험을 통하여 제안된 방법의 효율성을 보였다.

2. 기존 연구의 문제점

그 동안 데이터마이닝 분야에서는 발견된 대량의 연관규칙중에서 의미있는 규칙을 효과적으로 탐색하기 위한 연구가 다양하게 진행되어 왔지만 텍스트마이닝 분야에서는 활발하게 진행되지 않았다. 텍스트마이닝 분야에서는 의미있는 패턴을 추출하기 위하여 χ^2 가설 검정기법이나 상관계수와 같은 통계적인 기법을 주로 사용하였다. χ^2 기법은 비교적 좋은 성능을 보이지만 우연성 테이블에서 모든 셀에 대한 정보를 이용하는 관계로 비용이 증가한다. 또한 단순히 χ^2 값이 높다고 해서 단어간 또는 카테고리와 단어간의 상관성이 높다고 판별할 수 없으므로 χ^2 값에 대한 셀의 기여도를 고려하여야 한다[7]. 그리고 지지도가 아주 낮은 패턴도 의미있는 규칙으로 발견될 수 있다.

3. 텍스트마이닝에서 의미있는 패턴 추출 기법

3.1 텍스트마이닝에서 의미있는 패턴 추출 기법

본 논문에서는 관련된 문서끼리 자동적으로 분류하거나 대량의 문서에서 유용한 지식 정보를 정확하게 검색하기 위하여 텍스트마이닝에서 의미있는 패턴을 추출하기 위한 방법을 제안하였다. 제안한 방법은 Agrawal 등이 제안한 R-interesting 값을 수정하여 대량의 연관규칙에서 의미있는 패턴을 추출하기 위한 방법이다. 문서를 대상으로한 다양한 실험을 통하여 카테고리별로 최적의 R 값을 구성하여 의미있는 패턴을 추출하였다.

제안한 방법은 문서로부터 전문 용어를 추출하기 위한 전처리 과정, 전체 문서 집합과 카테고리별로 연관규칙 탐사 알고리즘을 적용하여 패턴을 발견하기 위한 과정 그리고 발견된 패턴 중에서 의미있는 패턴을 추출하기 위한 후처리 과정으로 이루어진다. 다음 그림 1은 본 논문에서 제안한 기법에 대한 전체적인 개념도이다.

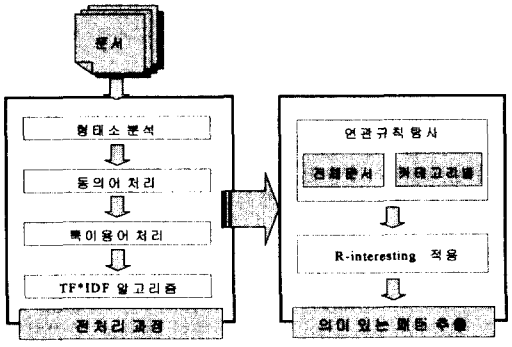

그림 1. 의미있는 패턴 추출 기법에 대한 개념도

3.2 전처리 과정

먼저, 전체 문서 집합에 대한 형태소 분석을 통하여 문서에서 출현하는 모든 용어를 추출하였다. 형태소 분석기는 공개용 형태소 분석기인 HAM4.0a를 사용하였다[9]. 추출된 단어에서 컴퓨터 용어 사전에 수록된 컴퓨터 용어만을 별도로 추출하였다.

전문 용어 중에서 같은 의미를 가진 용어이지만 저자에 따라 영어나 한국어 용어를 혼용하고 있다. 특히 영어로 된 전문 용어를 한글로 표기하는 경우에서 자주 발생한다. 이러한 동의어는 별도의 동의어 사전을 구성하여 용어를 표준화하였다. 예를 들어 '데이터베이스', '데이타베이스', 'database', 'databases', 'db' 등과 같은 용어는 하나의 용어로 통일하였다.

전체 문서에서 출현하는 절대 빈도수가 매우 적은 용어는 연산 시간만 낭비하고 최소 지지도를 만족하지 못하기 때문에 연관규칙으로 발견되지 않는다. 그리고 전문 용어이지만 모든 분야에서 공통적으로 사용되는 전문 용어는 특정 분야를 대표하는 용어로 보기 어렵다. 본 논문에서는 이러한 용어를 특이 용어로 처리하여 연관규칙 탐사 과정에서 제외시켜 무의미한 연관규칙의 양산을 방지하여 대표 색인어를 효율적으로 추출할 수 있도록 하였다.

3.3 R-interesting 기법을 이용한 의미있는 패턴 추출

본 논문에서는 텍스트마이닝에서 카테고리별로 의미있는 패턴을 효과적으로 추출하기 위한 방법을 제안하였다. 먼저, 카테고리별 패턴을 분석하기 위하여 각 카테고리별로 연관규칙을 탐사하였다. 최적의 R-interesting 값을 구성하기 위하여 전체 문서 집합에 대해서도 연관규칙을 탐사하였다.

그리고 전체 문서 집합의 규칙에 대한 지지도 관측 값과 전체 문서 집합에 대한 규칙의 지지도에 의해 추정된 세부 카테고리별 규칙에 대한 지지도 기대값의 비율을 계산하였다. R-interesting 값을 구하기 위해 수정된 식은 다음 식(1)과 같다.

$$E[Freq(C_i \mid Z)] = \frac{1}{n} \times Freq(U \mid Z) \quad (1 \le i \le n,\ n\text{은카테고리수}) \tag{1}$$

전체 문서 집합을 U, 세부 카테고리를 C_i, $Z = X \rightarrow Y$ 라하고, U 에 대한 $X \rightarrow Y$의 지지도가 주어지면, C_i에 대한 $X \rightarrow Y$의 지지도에 대한 기대값을 구한다.

3.4 R-interesting과 χ^2 의 비교

3.4.1 χ^2 테스트

χ^2 테스트는 우연성 테이블을 이용하여 단어들간 또는 카테고리와 단어간의 독립성 여부를 판별하기 위한 대표적인 통계적 기법이다. Yang 등은 DF 기법, IG 기법, MI 기법, TS 기법 등과 같은 다른 대표 색인어 추출 기법에 비해 χ^2 기법의 분류 성능이 우수함을 보였다[10].

Brin 등은 χ^2 테스트를 이용하여 연관규칙의 독립성을 검정하기 위한 상관성 규칙(Correlation Rule)을 정의하였고, χ^2 값에 대한 A의 기여도(Contribution)가 클수록 카테고리와 단어간의 관련성이 높음을 보였다[8]. 여기서 A는 패턴 t 가 출현하면서 카테고리 C_i인 문서의 수를 말한다. A의 기여도에 대한 식은 다음 식 (2), (3)과 같다.

$$Contri_A = \frac{(A - E_A)^2}{E_A} \tag{2}$$

$$E_A = \frac{n \times (A+B) \times (A+C)}{N} \tag{3}$$

여기서 E_A는 기대값, A는 패턴 t가 출현하면서 카테고리 C_i인 문서의 수, N은 전체 문서의 수를 나타낸다.

3.4.2 χ^2 값과 R 값의 비교 분석

일반적으로 텍스트마이닝에서 R 값은 χ^2 값에 대한 A의 기여도와 유사한 경향을 가진다. 여기서 A는 패턴 t가 출현하면서 카테고리 C_i인 문서의 수를 말한다. 첫째, 카테고리 C_i에 나타나는 패턴의 빈도수가 같을 경우 카테고리 C_i 에 많이 나타날수록 기여도는 높아지고, R 값도 높아진다. 또한 빈도수가 같을지라도 다른 분야에 나타나는 비율이 높을수록 기여도와 R 값은 낮아진다. 둘째, 패턴이 카테고리에 C_i에만 나타났을 경우 빈도수가 높을수록 기여도는 높아진다. 즉, R 값이 같을 때는 빈도수가 높을수록 의미있는 패턴이 된다. 셋째, 빈도수도 높고 카테고리 C_i에 많이 나타날수록 일반적으로 기여도가 높은 값을 가진다. 즉, R 값과 빈도수가 동시에 높을수록 의미있는 패턴이라 볼 수 있다.

4. 실험 결과 및 고찰

본 논문에서 제안한 텍스트마이닝에서 의미있는 패턴 추출 기법의 효율성을 검증하기 위하여 컴퓨터 관련 학회에서 발표된 논문을 대상으로 실험하였다. 학회에서 분류한 8개의 세부 분야별로 30편씩 선정하여 분야별로 연관규칙을 적용하여 패턴을 발

견하였다. 세부 분야별로 발견된 패턴의 수는 최소 88,276개, 최대 172,076개가 발견되었다.

먼저, 각 세부 분야별로 연관규칙을 적용하여 발견된 패턴 중에서 임의의 R 값 이상인 패턴을 추출하였다. R 값이 높을수록 특정 분야에서 출현하는 패턴의 비율이 높으므로 의미있는 패턴으로 추출되었다. 예를들면, 데이터베이스 분야의 경우, R 값이 3이상인 경우에는 다른 분야에 나타나는 비율이 높더라도 빈도수가 높은 경우에 의미있는 패턴으로 추출된다. R 값이 6 이상인 경우에는 다른 분야에 나타나는 빈도수가 낮으면서 높은 빈도수를 가지는 경우에 의미있는 패턴으로 추출된다.

제안한 방법의 효율성을 검증하기 위하여 대표적인 통계 기법인 χ^2 가설검정 기법과 비교 분석하였다. 먼저, 각 세부 분야별로 χ^2 에 대한 A 셀의 기여도 순으로 상위 500개의 패턴을 추출하고, 일정 R 값 이상일 때, 빈도수와 R 값순으로 상위 500개의 패턴을 추출하였다. 이 때, R 값의 변화에 따라 χ^2 와 비교 실험을 한 결과, 대부분의 분야에서 R 값이 5.8에서 6.3사이에서 χ^2 와 유사한 결과를 보였다.

데이타베이스 분야의 경우 일정 R 값 이상일 때, 빈도수와 R 값 순으로 상위 500개의 패턴과 χ^2 에서 A셀의 기여도 순으로 상위 500개의 패턴 중에 공통되는 패턴의 수를 비교한 결과이다. R 값이 6.1일 때, χ^2 와 비교에서 공통되는 패턴의 수가 500개 중 388개(78%)로 가장 높았다.

다음 표 2는 데이타베이스 분야에서 R 값이 6.1보다 클 때, 빈도수와 R 값 순으로 상위 25개의 패턴, χ^2 에서 A셀의 기여도 순으로 상위 15개의 패턴, 연관규칙 탐사 알고리즘으로 발견된 상위 15개의 패턴을 나타낸 것이다. 일반적인 연관규칙에서 우선 순위가 높은 패턴들은 χ^2 나 R-interesting 값을 이용한 방법에서는 높은 우선 순위를 가지지 않았다.

표 2. 제안한 방법과 χ^2 에서 패턴 비교

연관규칙 기법	제안한 방법	χ^2
데이타베이스 → 객체	스키마 → 데이타베이스	스키마 → 데이타베이스
객체 → 데이타베이스	데이타베이스 → 스키마	데이타베이스 → 스키마
질의 → 데이타베이스	질의 → 스키마	질의 → 스키마
데이타베이스 → 질의	스키마 → 질의	스키마 → 질의
테이블 → 데이타베이스	관계 → sql	테이블 → 스키마
데이타베이스 → 관계	sql → 관계	스키마 → 관계
데이타베이스 → 검색	스키마 → 검색	데이타베이스 → select
데이타베이스 → 테이블	스키마 → 객체	스키마 → 테이블
관계 → 데이타베이스	검색 → 스키마	관계 → 스키마
검색 → 데이타베이스	객체 → 스키마	select → 데이타베이스
접근 → 데이타베이스	객체 → sql	관계 → sql
데이타베이스 → 접근	sql → 객체	sql → 관계
질의 → 객체	테이블 → 스키마	데이타베이스 → sql
객체 → 질의	스키마 → 관계	sql → 데이타베이스
디자인 → 데이타베이스	데이타베이스 → select	타입 → 스키마

다음 그림 2는 8개의 세부 분야에서 공통되는 패턴의 수가 가장 많을 때의 R 값을 나타낸 것이다. 6개의 분야에서는 R 값이 5.7에서 6.3사이일 때 공통되는 패턴의 수가 가장 많았고, 인공지능 4.5, 전산 수학 및 교육은 5.4일 때 공통되는 패턴의 수가 가장 많았다. 인공지능 분야에서 나타나는 패턴은 전산 수학 및 교육 또는 정보보호 분야에서도 많이 발견되었다. 전산수학 및 교육 분야의 경우에도 멀티미디어 분야와 정보보호 분야와 동일한 패턴이 많이 발견되었다.

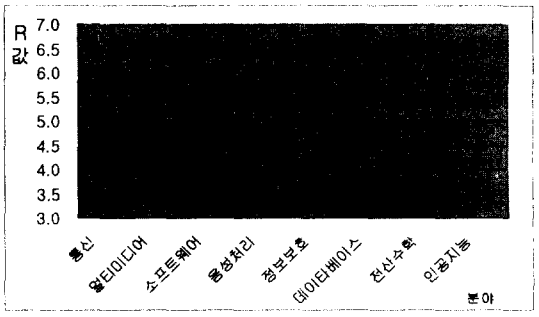

그림 2. 공통되는 패턴의 수가 가장 많을 때의 R 값

5. 결 론

본 논문에서는 문서에서 발견된 대량의 패턴 중에서 의미있는 패턴을 효과적으로 발견하기 위한 텍스트마이닝 기법을 제시하였다. 본 기법은 Agrawal 등이 제안한 R-interesting 값을 수정하여 의미있는 패턴을 추출하기 위한 방법이다. 특정 분야에서 나타나는 패턴의 빈도수와 다른 분야에서 나타나는 패턴 빈도수의 비율에 따른 χ^2 값의 A 기여도와 R 값을 비교한 결과 빈도수가 같을지라도 다른 분야에 나타나는 비율이 높을수록 기여도와 R 값은 낮아졌다. 또한 패턴이 특정 분야에만 나타날 경우, 빈도수에 따른 기여도와 R 값의 비교결과 빈도수가 높을수록 기여도는 높아지고 R 값은 변화가 없었다. 제안한 기법은 KDD 시스템 개발을 위한 문서 분류 과정에서 대표 색인어를 효과적으로 구성하기 위한 기법으로 사용할 수 있다.

참 고 문 헌

[1] Mark Dixon, "An overview of document mining technology," 1997.

[2] R. Srikant and R. Agrawal, "Mining generalized association rules," Proc. of the VLDB, pp.407-419, 1995.

[3] C. H. Cai, A..W. Fu, C. H. Cheng, and W. W. Kwong, "Mining association rules with weighted items," Proc. of the IDEAS, pp.68-77, 1998.

[4] B. Liu, W. Hsu, Y. Ma, "Integrating classification and association rule mining," KDD, pp.27-31, 1998.

[5] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo, "Finding interesting rules from large sets of discovered association rules," Proc. of the CIKM, pp.401-407, 1994.

[6] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," KDD, pp.27-31, 1998.

[7] S. Brin, R. Motwani, and C. Silverstein, "Beyond market basket: Generalizing association rules to correlations," Proc. of the ACM SIGMOD, pp.265-276, 1997.

[8] H. Ahonen, O. Heinonen, M. Klemettinen, and A. I. Verkamo. "Applying data mining techniques in text analysis," Technical Report C-1997-23, Department of Computer Science, University of Helsinki, 1997.

[9] 강승식, "HAM: 한국어 형태소 분석 라이브러리," http://ham.hansung.ac.kr/ham/ham-intr.html

[10] Y. Yang, and J. O. Pdedersen, "A comparative study on feature selection in text categorization," Proc. of the ICML, pp.412-420, 1997.