NFP-Algorithm 알고리즘을 기반한 텍스트 연관 패턴 추출

유수경*, 김교정*
*숙명여자대학교 멀티미디어학과

e-mail:skyou76@sookmyung.ac.kr
kiochkim@sookmyung.ac.kr

Text Assocation Pattern Extraction using NFP-tree Algorithm

Soo-Kung Yu*, Kio-chung Kim*
*Dept of Multimedia, Sookmyung Women's University

요 약

인터넷상에서 존재하는 많은 데이터베이스들 중 현실적으로 유용한 정보를 가지고 있는 것은 텍스트 데이타베이스이다. 텍스트 마이닝 기법에서 비구조적인 특징을 가진 텍스트 데이타로부터 유용한 정보를 분석하고 추출하여 연관된 패턴을 탐색하는 과정은 중요한 연구과제이다. 이에 본 논문은 인터넷에서 저장된 텍스트 데이터를 가지고 기존 텍스트 마이닝 기법 중 연관탐색 기법을 적용하여 사용자 중심의 연관된 패턴을 찾아서 의미있는 정보를 얻고자 한다. 탐색하기 위해 먼저 전처리 작업으로 용어의 객체를 추출하고. 추출된 각 객체들은 대용량 데이터에서 시간적, 공간적면에서 효율적인 연관탐색 기법인 NFP-Algorithm(N-most interesting k-itemsets Using FP-tree and FP-Growth)을 적용시켜서 의미있는 정보를 추출했다. 또한 Apriori계 Algorithm, FP-Algorithm, NFP-Algorithm을 비교하여 NFP-Algorithm이 시간적면에서 효율적임을 보여주었다.

1. 서론

대용량의 정보로부터 유용한 지식 정보를 추출하는 작업인 지식 검색기법에 관한 활발한 연구가 진행되고 있다. 그러나 지식 검색 과정은 산재되어 있는 정보의 방대함과 현재 사용하는 기법으로는 정확한 정보를 발견하기는 어렵다. 특히 유용한 정보들이 주로 저장된 텍스트 데이터베이스에서 의미있는 패턴을 탐색하기가 더욱 어렵다. 왜냐하면 텍스트 데이터는 비정형적이고 불규칙적인 구조이기에 실질적으로 의미있는 패턴을 찾는데는 많은 어려움을 주기 때문이다. 그래서 텍스트 데이타 대상으로한 텍스트 마이닝 분야에서는 의미있는 패턴을 효과적으로 탐색하는 과정은 중요한 연구 문제로 남고 있다.

정보 검색 기술에서 데이타가 대량으로 증가할 경우, 검색 오버헤드 또한 증가하여 정보 검색 시에 수행 시간 증가의 문제점을 가진다. 이에 텍스트 마이닝(Text Mining)기법에서는 대표 키워드를 사용하는 문서색인(Document Indexing)방법을 기반으로 검색 범위를 축소시키고 이에 따른 각 색인들을 데이터 마이닝 기법을 적용시켰다. 특히 검색의 효율성을 높이고 각 색인마다 연관성을 찾기 위한 데이터 마이닝 기법 중 연관탐색 기법을 적용한 사례들이 많다. 그러나 의미있는 패턴을 찾기 위해 마이닝 과정에서도 여전히 많은 연산 시간을 소비하므로 시간적으로 효율적인 알고리즘이 필요하다.

본 논문에서는 데이터베이스에안에 저장된 텍스트 데이타를 텍스트 마이닝 기법인 연관탐색을 적용하여 유용한 지식 정보를 검색하는 방법을 제시한다. 그리고 기존 수치 데이타에서만 적용한 NFP-Tree 알고리즘을 텍스트 데이터에 적용하여 최대 만족할 만한 K개의 연관된 패턴을 추출하고, 시간적, 공간적으로도 효율적임을 제시한다. 이에 인터넷에서 저장된 텍스트 데이타를 대상으로 연관 탐색 기법인 Aprioir, FP-Algorithm, NFP-Algorithm을 가각 적용한 결과, 다른 알고리즘에 비히 NFP-Algorithm이 확장된 연관된 정보를 얻을 수 있었고, 대용량 데이터에서도 시간적인 면에 효율적임을 보여준다.

2. 관련 연구

2.1 텍스트 마이닝

텍스트마이닝 분야에서는 의미있는 패턴을 추출하기 위해여 통계적인 기법과 인공 지능에 의한 기법

들을 주로 사용하였다. 통계적 방법으로는 베이지안 확률에 의한 기법, K 최근 인접(k-Nearest Neighbor)알고리즘 등이 있으며 인공지능 기법으로는 신경망, 유전자 알고리즘, 퍼지 이론 등이 있다.

기존의 연구들은 연관규칙을 수정한 에피소드 규칙을 이용하여 인접해서 출현한 단어 집합을 발견하기 위한 Ahonen 연구, 중요한 아이템에 가중치를 부여하여 의미있는 규칙을 효과적으로 선별하는 방법을 제시한 Cai와 Fu의 연구, 상관계수를 이용하여 상관 관계가 낮거나 음의 상관관계 패턴을 갖는 규칙들을 제거하는 방법을 제안하고, X^2 가설 검증방법을 이용하여 패턴을 추출, 계층주조를 이용하여 규칙을 일반화하는 방법과 클러스터링하는 방법 등을 제안한 Tan과 Kumar 등의 연구가 있다.[2]

그러나 이러한 기법들은 수치 데이터가 아닌 비 정형화된 데이타에서는 적용하기 어렵다. 왜냐하면 통계적 수치가 높다고 해서 단어간 상관성이 높다고 판별할 수 없으며, 다차원의 의미를 추출했을때 많은 시간적 소요와 논리적인 관계에서의 다차원 표현이 어렵기 때문이다. 또한 전체 정보에서 출현하는 절대 빈도수가 매우 적은 객체는 연산 시간만 낭비하고 최소 지지도를 만족하지 못하기 때문에 효율적으로 의미있는 연관 패턴들을 발견하지 못한다.

2.2 비 Apriori계 알고리즘

연관 규칙 탐색 기법에는 Aprior계 알고리즘과 비 Aprior 알고리즘으로 나누어진다. Aprior계 알고리즘은 데이터베이스를 스캔하는 과정마다 후보 항목 집합 생성을 하면서 상호 관련된 객체를 찾는다. 이에 반해, 비 Apriori계 알고리즘은 후보 항목집합 발생 과정 없이 빈발 항목집합을 찾는 것으로 최대 두 번의 트랜잭션 데이터 베이스 스캔 과정을 가진다. 첫 번째 과정은 압축된 빈발 패턴 정보를 저장하기 위해 FP-트리 만드는 과정, 두 번째는 조건부 패턴 베이스를 생성하는 FP-트리 마이닝 하기 위한 트리 탐색 과정이다. 이러한 과정은 빈발한 다차원의 추출이 많을 수록 비 Apriori계는 Apriori계 보다 공간적, 시간적 오버헤드를 제거할 수 있는 장점이 있다. 이에 반해 비 Apriori계의 대표적인 알고리즘으로 [3]에서 소개되어진 FP-tree, FP-growth 알고리즘이 있다. 그러나 비 Apriori계의 마이닝도 특별한 지식을 갖고 있지 않는 한 사용자의 요구되어진 결과들에 대해 적정한 수를 얻을 지지도 값을 결정하기는 어렵다. 다시 말해서, 어떤 값이 너무 크면, 어떠한 것도 발견되기 어렵고 어떤 값이 작으면 너무 많은 아이템 집합들이 생성될 것이다.

3. 텍스트 데이터베이스로부터 의미있는 패턴 추출 기법

그림 1은 본 논문에 제안한 연관된 패턴을 추출한 방법으로 두 단계 과정으로 이루어진다. 첫 번째 는 인터넷의 데이터베이스에서 저장된 텍스트 데이터 속성 객체를 추출하는 전처리 과정이고, 두 번째 는 추출된 속성 객체의 의미있는 패턴을 추출하기위해 연관 탐색 기법을 적용함으로 확장된 연관된 패턴을 추출하는 과정이다.

그림 1 텍스트 데이터안에 의미있는 패턴 추출

3.1 객체 속성 추출하기 위한 전처리 작업

용어 객체 추출은 저장하는 공간의 크기를 줄일 수 있고 키워드에 대해서 정확하게 계산하기 위해 그림 1과 같이 먼저 데이타 안에 있는 텍스트를 형태소 분석으로 통한 색인 추출, 특수용어 제거, 불용어 제거 등의 순서로 전처리를 통한 특징 객체의 속성을 추출한다.

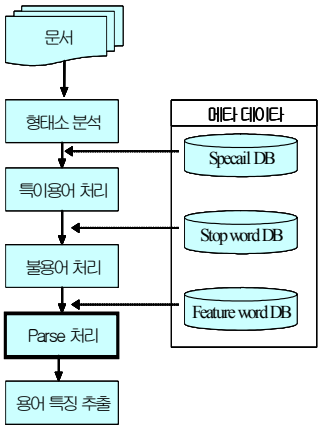

그림 2 객체속성 추출

데이타는 기본적으로 한국어로 텍스트 안에 있는 형태소 분석을 통한 용어의 특징 추출 과정인 전처리 작업이 필요한다. 전처리 단계로 어절분리하고 준말을 본디말로 변경하다. 다음은 특수 문자들 제거, 접속사, 조사와 한글 끝말 등 불용어를 제거 시킨다. 중요 용어는 (식 1)을 적용하여 용어의 빈도수가 높은 용어에 대해 메타 데이터에 저장한다. 마지막으로 중요한 용어를 추출하기 위해 미리 저장된 메타 데이터와 비교하여 형태소가 복잡한 부분을 파싱(Parsing)시키다

T*IDF 가중치 $w_i = \log_2\left(\frac{N_i}{n_i}\right)$ (식 1)

w_i : i번째 용어의 가중치
N_i : i번째 용어의 빈도수(term frequency)
n_i : i번째 용어의 문서 빈도수(document frequency)

3.2 NFP-Tree 기법을 이용한 의미 있는 패턴 추출

본 논문에서는 그림 3의 개념하에 Apriori계의 대표적인 알고리즘인 FP-Tree 와 FP-Growth 메소드를 기반으로 최대 N개와 관련된 항목 집합을 추출하는 NFP-Tree 와 NFP-Growth 의 메소드를 통해 전처리 작업으로 추출된 용어속의 연관된 패턴을 추출하고자 한다.

D : 트랜잭션 데이터베이스
K_{max} : 관련된 항목집합들의 크기
$result_k$: 현재 최대 N개 관련된 k-항목집합들의 결과 집합
ξ : 모든 항목집합들에 대한 초기 지지도
ξ_k : k개 항목집합들에 대한 초기 지지도

그림 3 개념 정의

(1) FP-tree and FP-Growth

FP-Algorithm은 다음 2가지 단계의 과정을 가진다. 첫 번째 단계는 압축된 빈발 패턴 정보를 저장하기 위해 FP-Tree를 만들기 전 과정으로 그림 4과 같이 데이터 베이스를 스캔하여 크기가 1인 빈발항목을 선출하여 지지도순으로 정렬한 헤더 테이블을 만든다.

TID	Items	Sorted Frequent Items
001	a, b, c, d	c, d, a, b
002	b, c, d, e	c, d, b, e
003	a, c, d	c, d, a
004	e, f	e

그림 4 데이타베이스

다음은 그림 4와 같이 헤더 테이블의 기반으로 FP-Tree의 압축구조를 만들기 위해 루트를 null로 하고, 데이터 베이스의 트랜잭션을 읽어들여, 지지도순으로 정렬하고, 이를 루트에서 하나의 트리 구조로 만든다. 트랜잭션을 읽어들일 때 공통된 빈발항목에 대해서는 지지도를 1씩 증가하고, 새로운 빈발항목에 대해서는 노드를 추가한다. 그리고 생성한 헤더 테이블 기초로 해당한 노드들을 차례로 링크시킨다.

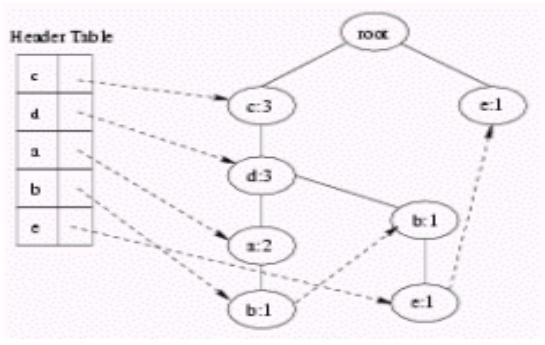

그림 5 헤더테이블 및 트리만들기

두 번째 단계는 빈발한 항목에 대한 패턴 베이스를 생성하는 과정으로 FP-Tree 구조 기초로 FP-Growth 알고리즘을 적용하여 연관된 용어의 패턴을 탐색한다.

(2) N-most Interesting 항목을 위한 FP-tree

본 논문은 위에서 제시한 FP-Algorithm의 과정을 기초로 한 NFP-Algorithm 기반으로 연관된 패턴을 찾고자 한다. NFP-Algorithm은 최대 N-most Interesting 항목을 추출하기 위한 알고리즘으로, [4]에서 제시한 Itemset-Loop 알고리즘 과 FP-tree의 알고리즘의 장점을 보완한 알고리즘이다. Itemset-Loop 알고리즘은 Aprior계 연관탐색 과정에서 후보자를 생성 할 때 마다 다른 지지도를 적용하는 방법이다. 좋은 성능을 가짐을 입증했지만, 부분집합들의 속성을 멈추는 데에만 의존되어서 후보생성이 되는 k-항목집합이 크면 모든 항목 집합들 또한 큰데 각 속성은 최대 N개 관련된 항목집합의 초기값을 가지지 못하는 단점이 있다. 이에 후보생성이 없는 FP-Tree의 알고리즘을 적용 시켰다.

NFP-알고리즘의 전체 알고리즘 과정은 그림 6과 같다. ξ_k의 값은 항목집합이 현재 $result_k$에 포함할 때 비로소 결정된다. 초기화 값 ξ는 0으로 놓고, 마이닝 하는동안 $1 \leq K \leq K_{\max}$ 범위 안에 N-th 최대 빈번한 k-항목집합의 지지도 사이 최소값을 주어짐에 따라 ξ는 1씩 증가시킨다. 수식으로 표현하면 (식 2)와 같다.

$$\xi = \min(result_1, result_1, ..., result_{k_{\max}}) \quad \text{(식 2)}$$

Input: D
Input : k_{max} , N
Output: N-most interesting k-itemsets for $1<k<k_{max}$
(1) $result_k$ 는 interesting k-itemsets의 결과이다 $result_k$ =0 으로 초기화
(2) 데이터베이스 D 스캔하고 각 항목의 지지도를 찾는다
(3) 내림차순 으로 정렬된 항목들을 sorted-list를 만든다.
ξ_1을 결정한다.
(4) $\xi_2 = \xi_3 = \cdots = \xi_{\max} = \xi = 0$
(5) FP-tree를 만든다. "null" 의 이름으로 시작해서 가지를 만든다.
(6) NFP-build(T, D, sorted-lit, ξ) 메소드 호출
(7) $result_1, result_1, ..., result_{k_{\max}} \leftarrow$ NFP-min($T, \varnothing, \xi, \xi_1, \xi_2, ...\xi_{k_{\max}}$)

그림 6 Main Algorithm

그림 8은 트리 만들기 위한 알고리즘이고, 그림 9은 연관 패턴을 탐색하기 위한 알고리즘이다. 그림 4의 데이터 베이스를 기반으로 N은 5, $K_{\max}$는 30이라 가정하고, 1-항목집합인 {a:8, b:8, c:6, d:6, e:4, f:4}을 찾는다면 ξ_1는 4가 된다. 다음 빈발한 2-항목집합들은 {cd:5, ab:8, bd:4, bc:6, ad:6}이 되고, ξ_2는 5가 된다. 그러므로 5보다 더 큰 지지도에 고려하지 않기 때문에 여기서 멈춘다.

NFP-build(T, D, sorted-list, ξ)
(1) 지지도 ≥ ξ 의 정렬된 1개 항목집단들의 리스트 선택
(2) FP-Tree를 업데이트
For each transaction, Trans in D
if selected-list안에 있는 항목으로 구성된 Trans 이면
{
(a) 각각의 지지도에따라 항목들을 정렬하고 선택한다.;
(b) Let[i|I]은 Trans안에 빈번한 항목 리스트를 정렬한다.;
I는 첫 번째 엘리먼트이고 I는 리스트에 남은 엘리먼트이다.;
(c) Inset_NFPtree([i|I], T) 호출;
}

Inset_NFPtree([i|I], T)
(1) if T has a child C such that C.item-name - i.item-name then{
C의 수를 1씩 증가시키다.;
}else{
(a) 새로운 노드 C를 만든다, ;
(b) 항목들에 1씩 계산하고 항목들의 현재 링크 T로한다.;
(c) 항목들의 노드 링크는 node-link structure.로부터 항목 이름 이과 함께 링크한다.;
}
(2) if I is non-empty. invoke Insert_NFPtree(I.C).

그림 7 NFP-tree building

NFP-mine(Tree, a, ξ, ξ_1, ξ_2,..... ξ_{kmax})
{
if Tree contains a single path P
(1) then for each combination , β, of the nodes in the path P do
(a) generate itemset $\beta \sqcup \alpha$ width support = minimum support of nodes in β
(b) if $\xi_{|\beta \sqcup \alpha|} \leqq support$
then insert $\beta \sqcup \alpha$ $result_{|\beta \sqcup \alpha|}$; update $\xi_{|\beta \sqcup \alpha|}$; update ξ if necessarya
(2) elae of each
(c) generage itemset
(b) construct β's conditional pattern base using ξ and
then β's conditional pattern base using ξ and
(c) if $Tree_\beta \neq \varnothing$ then NFP-mine($Tree_\beta$, β, $\xi, \xi_1, \xi_2, \cdots, \xi_{k_{max}}$)
}

그림 8 NFP-tree Mining

4. 실험 결과 및 고찰

본 논문은 윈도우 환경에서 Java4 언어로 구현하였으며, JDBC/ODBC 를 이용한 Access를 이용하여 데이터베이스 작업을 하였다.

암 시민 연단의 싸이트에서 제공된 91,126개의 Q&A 게시판 안에 있는 텍스트 데이터를 가지고 실험하였다. 그림 9은 K_{max} 가 4일때 Apriori와 FP-Algorithm과 NFP-Algorithm 등 연관된 패턴을 추출하는데 소요하는 시간을 보여주고 있다. 즉, NFP-Algorithm이 Apriori와 FP-Algorithm 보다 소요시간이 적음을 보여주고 있다.

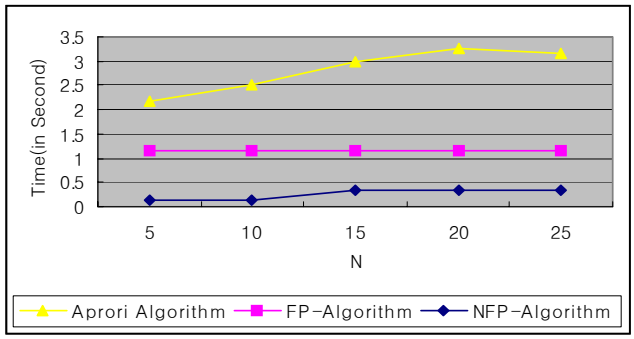

그림 9 Kmax =4

Rank	item of NFP-Alg	Rank	item of NFP-Alg
1	수술	8	건강식품
2	말기	9	병원
3	항암	10	**간암**
4	**통증**	11	환자
5	**소화기계통**	12	효과
6	**아버지**	13	요법
7	치료	14	**어머니**

표 1 추출된 객체 순위 결과

표 1에서는 Kmax =1 일때 "위암" 에 대한 연관된 상위 14개만 보여는 결과이다. NFP-트리 기법으로 적용하여 위암과 연관된 의미 있는 단어로 "항암", "통증", "소화기계통", "간암" 등이 추출됨을 볼 수 있다. 또한 "어머니"보다 "아버지"의 용어가 상위에 있는 결과로 남성이 위암 확률이 높음을 추측할 수 있다.

본 논문의 향후 과제로는 객체들 간의 보다 세밀하고 의미있는 연관성 파악하기 위하여 데이터 내의 객체간 거리 및 공간 관계에 대한 표현과 측정을 위한 기법을 활용해 보겠다.

참고문헌

[1] 문현정, 개인화된 지능적 정보 에이전트 시스템의 사용자 중심 지식 프로파일에 대한 연구, 숙명여자대학교 박사논문, 2001
[2] J. Han, M. Kamber , Data Mining Concepts and Techniques, Morgan Kaufmann,2001
[3] J. Han, J. Pei, and Y. Yin, Mining Frequent Patterns without Candidate Generation, SIGMOD Conference 2000: 1-12
[4] A.W.C. Fu,R, W. W.Kwong, and J.Tang, "Mining N-most Interesting Itemsets" in Proc. of th Intl. Sym. on Methodologies for Intelligent Systems (ISMIS), 2000.
[5] Y.L. Cheung, A. Fu, "Mining Association Rules without Support Threshold: with and without Item Constraints", IEEE Transactions on Knowledge and Data Engineering, 2000
[6] C.C. Aggarwal and P.S.Yu "Mining large itemsets for association rules", in Bulletion of the IEEE Computer Society Technical Committe on Data Engineering, pp.23-31, March 1998.