개인화된 사용자 학습을 위한 연관 객체 추출 설계 및 구현

유수경*, 김교정*
*숙명여자대학교 멀티미디어학과

Assocate Object Extraction Using personalized user Learning

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

요 약

본 논문은 웹 도큐먼트를 기반으로 사용자에게 의미 있는 정보를 찾아주기 위한 연관 객체 추출 기법인 PMPL(Personalized Multi-Strategey Pattern Learning) 시스템을 제안하고자 한다. PMPL 모듈은 인터넷의 정보를 여과하여 필터링하고, 사용자 개인화의 키워드를 중심으로 연관된 객체를 추출한다. 이때 연관된 객체 추출 시 대용량 데이터에서 시간적, 공간적면에서 효율적인 연관 탐색 기법인 Fp-Tree와 Fp-Growth 알고리즘을 적용시켰으며, 연관규칙 탐색을 보완하기 위해 가중치 기법인 만유인력 기법을 적용시켰다. PMPL 시스템을 실행한 결과 개인화된 사용자 중심어 기초로 기존의 단일 학습 기법에 비해 더 많은 의미 있는 연관 지식을 추출한 결과가 보였다.

1. 서론

인터넷과 정보통신기술의 발전은 이용 가능한 전자정보량을 폭발적으로 증가시키고 있으며, 점점 더 많은 양의 정보가 새롭게 생성되고, 정리되고 있으며 또한 디지털 형식으로 전환되고 있다. 데이터와 도큐먼트에 대한 방대한 보고로 폭 넓게 인식되고 있는 인터넷은 사이버 스페이스상에 텍스트, 이미지, 오디오 그리고 비디오등의 다양한 정보를 제공한다[Pea and Gomez, 1992].

웹 도큐먼트를 통해 정보를 추출기 위한 기존의 학습 방법으로는 통계적 기법, 인공지능 기법, 마이닝 기법 등이 있다. 통계적 기법은 통계적 수치가 높다고 해서 단어간 상관성이 높다고 판별할 수 없으며, 다차원의 의미를 추출했을때 많은 시간적 소요와 논리적인 관계에서의 표현이 어렵다. 인공지능 기법이나 마이닝 기법으로는 다양한 표현이 가능하나 복잡한 계산이 필요하며, 전체 정보에서 출현하는 절대 빈도수가 매우 적은 객체는 연산 시간만 낭비하고 그에 따른 효율적인 연관규칙들을 발견되지 못한다.

다중전략학습은 단일 전략 학습 기법들을 적절히 통합하여 서로 다른 전략들이 서로를 상호 보완하고 약점을 상쇄하여 상승효과를 얻을 수 있다[Tecuci, 1995].

따라서 본 논문은 개인화된 다중 학습을 통해 웹 도큐먼트안에서 사용자의 질의에 요구에 따라 학습하는 PMPL(Personalized Multi-Strage Pattern Learning) 모듈을 제안하고자 한다. 또한 네트워크상 환경에서 제공되어지는 PMPL 마이너를 구현하고자한다. PMPL 기법에 대한 보다 효율적인 알고리즘을 적용하기 위해 연관 규칙 탐색 기법은 FP-Growth 알고리즘과, 객체간의 공간적 관계성을 적용한 만유인력의 기법 등을 상호 보완해서 적용한다. 또한 PMPL의 실험을 통해 더 확장된 패턴을 추출할 수 있었다.

2. 관련 연구

2.1 개인화된 다중 전략학습

다중전략 학습 기법은 서로 다른 학습 전략들을 통합하여 학습 하도록 하는 기법으로서 단일 학습 기법들

간의 상호 보완을 통하여 장점을 살리고 단일전략 학습자 능력이상의 임무를 수행하도록 하는 것을 목적으로 한다[Michalski, 2001]. 이러한 다중 전략 학습 기법에 쓰일 단일 기법들은 학습이 적용되는 응용분야에 대한 분석을 기반으로 선택 되어야 한다. 학습을 위한 입력(input) 형식, 학습의 목적 등이 학습 전략 선택을 위한 중요한 요인이다[Michalski, 2001].

2.2 연관규칙 탐색 및 알고리즘

연관규칙 탐색은 2가지 연산자 지지도(Support)와 신뢰도(Confidence)를 사용한다. 데이터 베이스 안에서 빈발항목이 A, B라 할 때, 지지도와 신뢰도는 다음 (식 1), (식 2)와 같이 표현할 수 있다.

$$\text{Support(A,B)} = P(A \sqcap B) \qquad \text{(식 1)}$$

$$\text{Confidence(A,B)=P(B|A)} = \frac{P(A \sqcap B)}{P(A)} \qquad \text{(식 2)}$$

연관 규칙을 탐색하는 과정은 크게 두 부분으로 나눌 수 있다. 빈발 항목 집합(Frequent itemset)들을 찾는 과정과 이를 기반으로 실제 연관된 규칙을 생성해 내는 과정이다.

연관 규칙 탐색 방법은 크게 Aprior계 알고리즘과 비 Aprior계 알고리즘으로 구분한다. 전통적으로 Aprior계 알고리즘을 사용했지만, 비 Aprior계 알고리즘이 시간적, 공간적으로 효율적임을 증명 하면서 오늘날 많이 사용한다. 비 Aprior계의 대표적인 알고리즘으로 [3]에서 제시한 FP-Tree와 FP-Growth 알고리즘이다.

2.3 만유인력 모델에 기반한 가중치

뉴턴에 의해서 형상화된 만유 인력의 법칙으로 공간상에 위치한 두 물체 사이에 작용하는 힘을 기본으로 한다. 만유 인력 법칙의 기본 식은 다음과 같다.

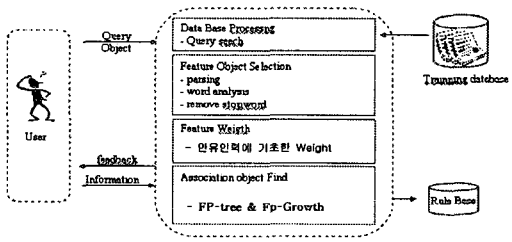
$$F = G\frac{Mm}{R^2} \qquad \text{(식 3)}$$

F : 두 물체 사이에 존재하는 힘
M, m : 두 물체의 질량
G: 만유인력 상수
R: 두 물체 사이의 거리

이와 같은 만유인력 모델은 객체 상호간의 연관성을 고려한 자연스러운 군집화 기법에 적용되고 있다. [Kim,E.,2000]. 또한 웹 도큐먼트의 HTML 태그(tag) 처리에도 적용되고 있다[S대, 2000].

3. 웹 도큐먼트를 위한 PMPL

본 장에서는 개인화된 다중전략학습(PMAL : Personalized Multi-Strage Assocation Learning)의 사용자 중심의 의미있는 연관 규칙을 추출하기 위한 웹 도큐먼트 마이닝 학습 패러다임을 제시한다. 그리고 사용자 중심 연관 객체 추출 기법과 만유인력 모델 기반 연관 객체 가중치 기법을 적용한 개인화된 다중전략학습 기법을 제시한다. 즉, 서로의 약점을 보완할 두 가지 기법을 동시에 적용하여 보다 향상된 결과를 얻도록 한다.



[그림 1] 개인화된 다중전략학습
(PMAL : Personalized Multi-Strage Assocation Learning)

3.1 웹 도큐머트 기반 사용자 중심의 특징 단어 추출 전처리 작업

본 논문은 사용자 중심의 특징 단어를 추출하기위해 [그림 2]의 과정을 적용한다.

```
1. Initialization
   Set F ← 'initial set of n feature'
   Set S ← 'empty set'
   Set min.tf initalize
2. Database Preprocessing
   oD ← Table(wtD)
   oD : object_Recordset_TextData = a set of Tb
   Tb : Tag_Black = {Tags, Tagterm}
3. Load Meta Data
   vSpec ← meta Data of special character
   vWord ← meta Data of stop ward
   vFword ← meta Data of feature word
4. Language word anyalsis
   vW ← ObjectParser(Tb)
   vW ← split(vW)
   vW ← remove_stop_word(vW,vSpec,vWord,vFword)
5. Index Selection
    if tf < min.tf
      remove (vWi)
   else
      Output the set S contating the selected features
```

[그림 2] 연관 객체를 추출하기 위한 전처리 작업 알고리즘

웹 도큐먼트는 웹 구조의 HTML의 파싱을 통해 태그 블럭 단위로 분해하고, 실험하고자 하는 데이터가 주로 한국언어 사용하기 때문에 형태소 분석, 불용어 제거, 특수문자 제거 등의 전처리 과정을 통해 보다 정확한 의미 있는 객체들을 추출하고자 한다. 또한 전체 문서에서 출현하는 절대 빈도수가 매우 적은 객체들은 용어들의 관계가 무의미한 연관 규칙을 발생시키기에 미리 클리닝을 한다.

3.2 사용자 중심의 연관 객체 추출

사용자의 정보요구와 관심 그리고 선호도를 반영하는 매우 중요한 의미를 지니고 있는 사용자 학습을 위한 중심어로 삼아 그와 연관된 객체를 추출한다. 이를 위하여 연관 규칙 탐사 데이터 마이닝 기법을 적용하여 전처리 작업을 통해 추출된 객체로 부터 연관된 객체를 추출한다.

사용자 중심의 연관된 객체를 추출하기 위해 다음 단계의 과정을 가진다.

Step 1. 앞 3.1에서 추출된 S의 집합에서 (식 4), (식 5)와 같이 사용자가 질의한 키워드와 관련된 집합만 S_{term}에 저장한다.

$$S \ni S_{term} \quad \text{(식 4)}$$

$$S_{term} = \{ O_{term} : O_{term1}, \ldots, O_{termi} \} \quad \text{(식 5)}$$

Step 2. min.Support값과 min.Confidence를 초기화한다.

Step 3. FP-tree를 생성하기 위해 내림차순 으로 정렬된 항목들 형태로 가진 Header Table를 만든다.

Step 4. FP-tree를 만든다. "null" 의 이름으로 시작해서 가지를 만든다.

Step 5. FP-Growth (*Tree, a*)메소드 호출를 통해 연관된 객체를 탐색한다. FP-Growth (*Tree, a*)의 알고리즘은 [그림 3]과 같다.

Step 6. 사용자 중심의 객체와 연관된 객체간의 Support와 Confidence 대한 (식 1), (식 2)를 계산하다.

Step 7. Confidence 에 따른 상위 n개의 객체를 추출한다.

본 논문은 사용자 질의에 따른 학습이기 때문에 지지도에 대해 배제한다. 다만 신뢰도가 높을 수록 의미있는 객체임을 가정한다.

```
Procedure FP-growth (Tree, a)
{
  (1) if Tree contains a single path P then {
  (2)   for each combination (denoted as B) of the nodes in the path P
  (3)   generate pattern β⊔α with support = minimum support
          of nodes in B }
  (4) else for each a_i in the header of Tree {
  (5)   generate pattern β = a_i ⊔ a with support = a_i.support;
  (6)   construct β's conditional pattern base and then β's conditional
          FP_tree Tree_β;
  (7)   if Tree_β ≠ 0 then
  (8)     Call FP-growth(Tree_β, β); }
}
```

[그림 3] FP-Growth Algorithm

3.3 연관 객체에 공간상에대한 객체의 가중치

중요한 속성들의 연관된 객체를 추출하는 과정에서 신뢰성을 향상시키기 위해서는 해당 문서의 공통적인 특징을 가려내어 이를 기준으로 각 속성마다 가중치를 차별적으로 두어 더욱 정확한 중요 속성을 추출하는 방법이 이용되고 있다.

[그림 4]는 연관 객체 간의 공간상에 대한 객체 가중치를 계산하기 위해 만유 인력의 모델의 기초로한 가중치 알고리즘이다.[1]

1. Association_object_Tuple $aT^j = \{O^Q, O^{C_1}, \ldots, O^{C_n}\}$
2. Document_Content_Object $O^{C_i} : < O^{C_i}_{term}, O^{C_i}_{m} >$
   $O^{C_i}_{m}$: mass of , tf
   tf : 현재문서의 키워드 빈도수
3. 만유인력(UG) 가중치 생성
   (1) $ugWB \leftarrow GetUGWeight(O^Q, O^{C_i})$
   (2) $ugWB$: a set of ugO^C
   (3) ugO^C :UG_Weighted_Content_Object = $\{ugO^C_{term}, ugO^C_{cw_{ug}}\}$
   (4)
   3.1 UG 힘(Force)를 얻는다.

$$F_j(O^Q, O^{C_i}) = G\frac{O^Q_m \cdot O^{C_i}_m}{R(aT^j)^2} \quad \text{(식 6)}$$

   $R(aT^j)$: unit distance in aT^J

   3.2 만유인력(UG) 가중치 계산

$$ugO^{C_i}_{cw\_UG} = \sum_k D_K(O^Q, O^{C_i})$$

[그림 4] 만유인력 모델의 기초한 연관 규칙 가중치 알고리즘

[그림 3] 알고리즘은 다음과 같은 가정을 가진다.

Lef 1. 객체 연관성 공간 상에서 객체들은 중심어의 질량에 의한 힘 (F)의 작용으로 사용자 질의에 합성되기 위한 순위의 기준인 부 합치(correspondence weight)가 높아지게 된다.

Lef 2. 수식 (식6)에 의하면 인력은 물체의 질량이 증가함에 따라 커지며 문서안의 거리는 단일하다.

4. PMPL기반한 Miner 구현

본 논문은 윈도우 환경에서 Java 4 언어로 Client/Server Version으로 구현되었으며, JDBC/ODBC 를 이용한 MSSQL2000을 이용하여 데이터베이스 작업을 하였다.

[그림 5]의 화면과 같이 Server가 시작하면 소켓이 시작되고 클라이언트가 접속과 로그인이 인증됨과 동시에 클라이언트가 요구하는 메시지에 따라 실행하고 결과 값을 전달한다.

웹 도큐먼트의 연관된 패턴을 추출하기 위해 색인 추출, 마이닝 탐색, 그래프 보기 등을 구현했다. 또한 보안을 위해 본 시스템은 시스템 Admin과 그룹내의 User들 마다 권한을 달리하기위해 로그린 기능, Group 관리, User 관리 등의 기능을 구현했다. 또한 데이터베이스 추가, 삭제, 수정할 수 있는 데이터베이스 관리가 있다. [그림 6]은 객체 추출 리스트와 PMPL 모듈을 통한 나온 연관 객체의 리스트이다.

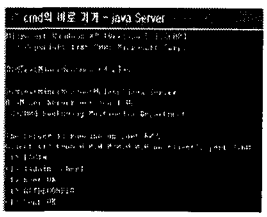


그림 5 Server 화면

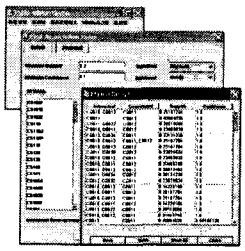


그림 6 Client 화면

5. 실험 및 결과

본 논문은 암 시민 연단의 싸이트에서 제공된 91,126개의 웹 데이터를 가지고 실험하였다. 표 1은 "위암"의 키워드 중심으로 연관규칙, 만유인력의 가중치 , PMPL 학습 모듈을 통해 발견된 상위 15개의 패턴을 추출한 결과이다.

표 1에서 보는 바와 같이 사용자 중심어 "위암"이 대부분의 상위 레벨에 나타나고 있는 것으로 보아 "위암" 에 대한 학습 예제 집합이라는 것을 알 수 있다. "위암"과 연관된 의미 있는 단어로 "항암", "통증", "소화기계통" 등이 연관규칙, 만유인력의 가중치로 사용한 단일 학습보다 PMPL 기법으로 학습한 방법이 더 상위 위치에 있거나 추가 되었음을 볼 수 있다. 따라서 PMPL의 모듈을 통해 연관된 지식들을 더 많이 표현해 주는 것을 볼 수 있다.

Rank	FP-Alg	Universal Gravity Model	PMPL
1	위암	주사약	위암
2	병원	수술	수술
3	수술	효과	알기
4	환자	북한	항암
5	전이	메일	통증
6	지금	항암	소화기계통
7	치료	대체의학	아버지
8	아버지	유황	치료
9	상태	위암	건강식품
10	방법	소화계통	병원
11	부탁	복어알	간암
12	알기	가족	환자
13	현재	건강식품	효과
14	요법	지금	오리
15	복용	대체	요법

표 1 상위 15개의 패턴을 추출한 결과

본 논문의 향후 과제로는 객체들간의 보다 세밀한 연관성 파악을 위하여 데이타 내의 객체간 거리의 관계에 대한 표현과 측정을 위한 기법이 필요하다. 또한 다양한 도메인과 데이터의 실험이 필요하다.

[참고 문헌]

[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 Methodogies 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.