의미있는 정보 검색을 위한 개인화된 다중 전략 학습 모듈의 설계 및 구현

유수경^{O} , 김교정
숙명여자대학교 멀티미디어과학과
{sookyou76^{O}, kiochkim}@sookmyung.ac.kr

Design and Implementation of PMSL for Information Retrieval

Sookoung Yu^{0} Kiochung Kim
Dept. of Multimedia, Sookmyung Women's University

요 약

오늘날 인터넷상에서 존재하는 많은 정보들은 다양한 사용자의 개인 특성에 맞게 새로운 정보의 지식으로 제공되어지기를 원한다. 기존의 연구는 단일 학습 기법을 통해 정보를 추출했으나 사용자에게 보다 의미 있는 정보를 제공하기 위해 다중 전략 학습 기법인 PMSL(Personalized Multi-Strategy Learning) 모듈 시스템을 제안하고자 한다. PMSL 모듈은 인터넷의 정보를 여과하여 필터링하고, 사용자 개인화의 키워드를 중심으로 연관된 객체를 추출한다. 이때 연관된 객체 추출시 대용량 데이터에서 시간적, 공간적면에서 효율적인 연관 탐색 기법인 Fp-Tree와 Fp-Growth 알고리즘을 적용시킴으로 결과의 효율성을 높이고자 하였으며, 연관규칙의 문제점을 보완하기 위해 가중치 기법인 TF*IDF 학습 기법을 적용시켰다. PMSL 모듈을 실행한 결과 기존 학습 기법에 비해 보다 더 의미 있는 연관 지식을 추출하게 되었다.

1. 서 론

인터넷과 정보통신기술의 발전으로 인한 정보량의 증가는 사용자의 정보획득 욕구와 비례하고 있다. 또한 다양한 사용자는 단순히 정보를 얻는 것만이 아니라 개인의 특성에 맞게 양질의 정보를 효과적으로 획득하길 바란다.

사용자가 정보를 제공 받기 위해 전통적으로 불리언 기법을 많이 사용한다. 불리언 기법은 데이터베이스 안에 있는 정보를 검색 하기 위해 불리언 연산자(AND, OR, NOT)를 사용하며 일련의 키워드로 구성된 질의 입력을 요구한다. 그래서 근접한 모든 정보를 찾기 위해서는 사용자에게 원하는 것을 정확히 모두 입력하도록 요구 되어지는 문제점이 있다. 따라서 사용자는 여전히 관련정보를 찾기 위하여 대량의 정보를 살펴보아야 하는 불편함을 가지고 있다. 또한 검색 결과의 양이 많을수록 의도하지 않는 검색 결과가 나올 수 있다. 에 보완한 단일 학습 기법으로 확률 검색, 가중치 검색, 퍼지 집합 검색, 추론 검색 등이 존재하지만 검색 속도면에서 불리언 검색보다 떨어지며 구현상의 난이도 측면에서 논리적 관계의 표현이 어렵다는 문제 때문에 결국 기존의 검색 엔진들은 대부분 불리언 연산만을 지원하고 있다.

다중전략 학습 기법은 서로 다른 학습 전략들을 통합하여 학습 하도록 하는 기법으로서 단일 학습기법들간의 상호 보완을 통하여 장점을 살리고 단일전략 학습자 능력이상의 임무를 수행하도록 하는 것을 목적으로 한다.

따라서 본 논문은 연관탐색 기법과 TF-IDF의 기법의 장단점을 상호 보완한 다중 전략 학습 기법을 통해 개인화에 맞는 효율적인 정보 검색을 위한 PMSL (Personalized Multi Strategey Learning) 모듈을 설계 구현하고자 한다. 그리고 PMSL을 실행한 결과 TF-IDF 방식의 학습 혹은 연관 탐색 방식의 학습보다 PMSL 기법을 적용한 학습 방식이 보다 더 의미 있는 연관 객체로 추출했음을 볼 수 있다.

2. 관련 연구

2.1 정보검색

(1) 데이터 마이닝 기법

의미 있는 정보를 획득하기 위한 데이터 마이닝 기법으로 단어들간의 의미 있는 패턴을 발견하기 위한 연관규칙, 관련된 문서끼리 클러스터링하기 위한 개념적 클러스터링, 단어간의 순서 관계를 고려한 에피소드 규칙, 신경망 기법 등이 있다. 그 중에서 연관규칙을 이용한 마이닝 기법은 문서를 자동적으로 분류하기 위한 대표 색인어 추출이나 관련된 문서끼리 클러스터링 하기 위한 분야 등에 많이 이용 되고 있다[1]. 그러나 전체 정보에서 출현하는 절대 빈도수가 매우 적은 객체는 연산 시간만 낭비하고 최소 지지도를 만족하지 못하기 때문에 의미 있는 연관규칙들을 발견한지 못한다.

(2) 가중치 기법

Cia와 Fu 등은 중요한 아이템에 가중치를 부여하여 의미 있는 규칙을 효과적으로 선별하는 방법을 연구하였다[1]. 개인화를 위한 단일 전략 학습 모델에서는 사용자 중심 가중치(weight)를 적용한 키워드-벡터 모델을 많이 적용한다. 이 기법은 TF-IDF 방식의 기초로 데이터에 나타나는 객체 각각의 가중치만을 고려할 뿐 객체들간의 연관성(relation)을 파악 할 수 없는 문제점을 가지고 있다.

3. PMSL 모듈 과정

본 논문에서 제시된 PMSL의 설계구조는 그림 1과 같이 표현한다. 개인화의 다중전략학습 기법을 기반으로 의미 있는 정보를 추출하기 위한 PMSL기법은 다음 두 단계의 과정을 수행한다. 첫 번째 단계로는 사용자로부터 제시된 정보 데이터와 사용자의 학습의도를 내포하는 중심어를 입력으로 취하여 사용자 중심어와 연관된 데이터의 속성 객체(feature object)를 추출한다. 두 번째 단계는 각 속성들에 가중치를 부여하여 높은 연관성을 갖는 데이터 속성 객체를 이용하여 확장된 의미 있는 정보를 추출한다.

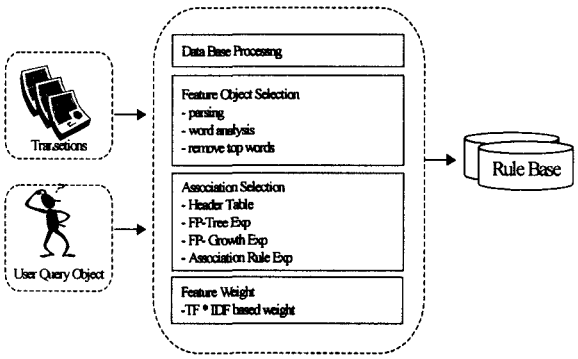

그림 1 PMSL(Personalized Multi-Strategey Learning) 시스템 구조

3.1 사용자 중심의 연관 객체 추출

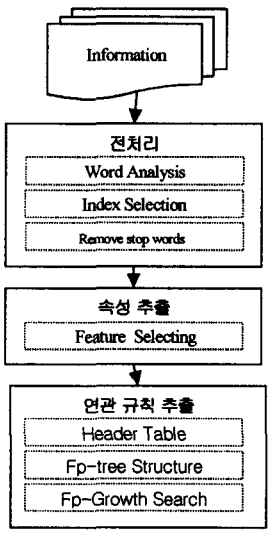

그림 2 PMSL 학습 순서도

연관 객체 추출은 먼저 데이터 안에 있는 텍스트를 형태소 분석으로 통한 색인 추출, 불용어 제거 등의 순서로 전처리 작업을 통해 객체의 속성을 추출하고 마이닝 기법인 연관 탐색 기법을 가동한다.

(1) 전처리 과정

다음 데이터는 기본적으로 한국어로 등록되어 있기 때문에 단어 객체들을 추출하기위한 형태소 분석을 통해 데이터에서 출현하는 모든 용어를 추출하였다. 추출된 단어들은 한국어의 모호성을 배제시키기 위해 불용어 제거 작업으로 저장하는 공간의 크기를 줄일 수 있고 키워드에 대해서 정확하게 계산할 수 있었다.

(2) 연관 규칙 탐색

본 논문은 사용자의 일련의 키워드 입력에 따라 적합한 의미 있는 지식을 추출하기 위해 사용자 질의를 중심으로 연관된 객체 속성을 추출하고자한다.

연관 규칙 탐색은 그림 3의 각각의 트랜잭션 집합 안에 있는 용어들의 경향을 파악해서 상호 연관성을 찾아준다.

연관 규칙 탐색 알고리즘은 빈발항목 집합 발견의 과정에서 크게 Apriori계 알고리즘과 비 Apriori계 알고리즘으로 나눈다. 최근 연구에 의하면 공간적, 시간적인 면에서 후보 항목 집단을 생성하지 않는 비 Apriori계 알고리즘이 더 효율적으로 오버헤드를 제거할 수 있음을 제안했다[3].

비 Apriori계 알고리즘은 최대 두 번의 트랜잭션 데이터베이스 스캔 과정이 필요하다.

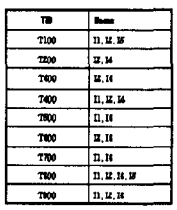

그림 3 트랜잭션 데이터

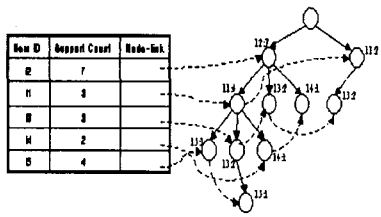

그림 4 압축된 빈발 패턴 정보를 저장하는 FP-tree

첫 번째 단계는 압축된 빈발 패턴 정보를 저장하기 위해 Fp-트리 만드는 과정으로 데이터 베이스를 스캔하여 크기가 1인 빈발항목을 선출하여 지지도순으로 정렬한 헤더 테이블을 만든다. 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.

두 번째 단계는 빈발한 항목에 대한 패턴 베이스를 생성하는 마이닝 과정으로 Fp-Tree 구조 기초로 그림 5의 Fp-Growth 알고리즘을 적용하여 연관된 용어의 객체를 탐색한다.

```
Input: 데이터 베이스를 사용한 FP-tree 구조, 최소 지지도 초기값
Output: The complete set of frequent patterns.
Method: Call FP-growth (FP-tree, null)
Procedure FP-growth (Tree, α)
{
    (1) If Tree contains a single path P then
    (2) for each combination β of the nodes in the path P do
    (3)  generate pattern β∪α with support = minimum support of nodes
    in β
    (4) else generate the frequent set A of α,
    (5) then for each α, according to the order from high frequency to low
    frequency in A
        do {
    (6)     generate pattern β = α_i ∪α with support = α_i.support;
    (7)     call Candidate_generation( );
    (8) construct the projection of 's conditional pattern base on 's
    candidate frequent set and then 's conditional FP-tree Treeβ based on
    the projected conditional pattern base;
    (9)  if Tree β ≠ ∅
    (10)     then call FP-growth (Tree β, β)     }
}
```

그림 5 FP-Growth 알고리즘

3.2 사용자 중심의 TF*IDF 가중치 기법

TF-IDF 방식이란 각 문서에 나타나는 키워드의 빈도수(TF:Term Frequency)와 역문서 빈도수(IDF:Inverse Document Frequency)에 의한 키워드 가중치 계산 방식이다.

본 논문은 텍스트 처리 과정을 거쳐 추출된 모든 단어 속성-불용어 제외-의 객체로 사용 단어 속성들은 그 빈도수에 기반 하여 식(1)에 제시된 TF-IDF 가중치 기법에 의하여 속성의 중요도 가중치를 적용한다.

TF*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)

TF-IDF 가중치 식(1)에 의하면 데이터 집합에 전체에 걸쳐 나타나는 용어들은 중요도가 낮고, 반대로 특정 데이터에 나타나는 용어들은 상대적으로 중요도가 높다는 전제를 갖고 있다. 따라서 임의의 데이터 집합에서 데이터 클러스터의 특성을 추출하기 위하여 일반적인 용어들을 제외시키는데 매우 유용하다. TF-IDF 가중치 값은 연관규칙 탐색의 신뢰도(confidence) 값에 가중치를 부여함으로 의미 없는 용어들을 제외시킨다.

4. 실험 및 결론

본 논문은 윈도우 환경에서 Java4 언어로 구현하였으며, JDBC/ODBC 를 이용한 MSSQL를 이용하여 데이터베이스 작업을 하였다. 그림 6은 PMSL을 구현한 어플리케이션이다. 구현한 어플리케이션은 일련의 키워드 중심으로 "Fp-Algorithm", "TF-IDF", "PMSL-Algorithm" 방식으로 학습한 결과를 보여준다.

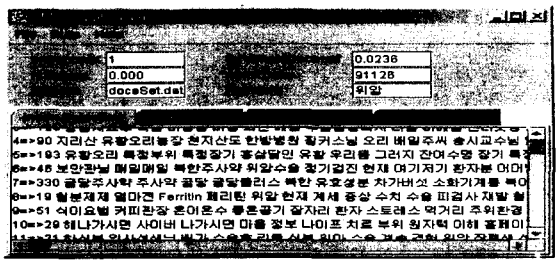

그림 6 PMSL의 Application

본 논문은 '암 시민 연단'의 싸이트에서 제공된 91,126개의 Q&A 게시판 데이터를 가지고 실험하였다. 표 1은 "위암"의 키워드 중심으로 연관규칙, TF-IDF, PMSL 학습 모듈을 통해 발견된 상위 15개의 패턴을 추출한 결과이다.

1	위암	주사약	위암
2	병원	수술	수술
3	수술	효과	말기
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	말기	가족	환자
13	현재	건강식품	효과
14	요법	지금	오리
15	복용	대체	요법

표 1 추출된 객체 순위 결과

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, TF-IDF로 사용한 단일 학습보다 PMSL 기법으로 학습한 방법이 더 상위 위치에 있거나 추가되었음을 볼 수 있다. 따라서 PMSL의 모듈을 통해 연관된 지식들을 더 많이 표현해 주는 것을 볼 수 있다.

본 논문의 향후 과제로는 객체들간의 보다 세밀한 연관성 파악을 위하여 데이터 내의 객체간 거리 및 공간 관계에 대한 표현과 측정을 위한 기법을 활용해 보겠다.

5. 참고 문헌

[1] 문현정, 2001, 개인화된 지능적 정보 에이전트 시스템의 사용자 중심 지식 프로파일에 대한 연구, 숙명여자대학교 박사논문

[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