전위 트리를 이용한 사용자 프로파일 기반의 문서 패턴 검색 기법

우호진*, 이원석
연세대학교 컴퓨터과학과
{judas*, leewo}@database.yonsei.ac.kr

Text Pattern Search Based on User Profile using Prefix Tree

Ho Jin Woo, Won Suk Lee
Department of Computer Science, Yonsei University

요 약

기하급수적으로 증가하는 데이터 중에서 개개인 사용자에게 적합한 정보를 추출하여 제공해야 할 필요성이 증대되고 있다. 본 논문에서는 대용량의 문서 집합으로부터 사용자가 원하는 특정 주제의 정보를 정확하게 추출해 낼 수 있는 문서 패턴 검색 방법을 제시한다. 사용자 선호도를 정확하게 반영할 수 있도록 전위 트리를 기반으로 사용자의 키워드 마이닝 프로파일을 생성하고, 이를 이용하여 문서 집합에서 매치된 패턴을 찾아내는 방법을 제안하였다. 생성된 프로파일을 이용한 검색 기법의 효용성을 실험을 통해 검증하였다.

1. 서론

데이터 검색을 통해 필요한 정보를 추출해 내는 것은 다양하게 연구되어 온 분야이다. 정보 추출이 가지는 의미는 단순히 사용자의 질의에 적합한 문서를 찾아내는 것만이 아니라, 문서들로부터 발견할 수 있는 관계 또는 관심 있는 패턴, 숨겨진 새로운 사실들을 추출하는 행위를 포괄한다[1]. 범용 검색 엔진의 키워드 검색을 이용하거나 SQL 을 작성하여 데이터를 검색하는 것은 사용자에게 질의 입력 및 학습에 대한 부담을 준다. 또한 개개인 사용자의 특성을 파악하지 않기 때문에 각 사용자가 원하는 구체적인 정보를 정확하게 제공하기 어렵다. 따라서 대용량의 데이터베이스로부터 개별적 사용자가 선호하는 정보를 정확히 검색하는 효과적인 기법이 필요하다.

질의에 적합한 문서 또는 구조화된 정보를 찾는 것은 문서 검색의 기본적인 목적이다. DISCOVER[5], BANKS[6]는 텍스트 데이터베이스에서 키워드 입력을 통해 정보를 검색하는 대표적인 시스템이다. 문서 검색의 또 다른 목적은 문서로부터 숨겨진 패턴이나 구조적인 특징을 발견하는 것이다. 문서에서 패턴을 추출하는 기존의 방법[8, 9]은 대용량의 데이터베이스에 대해서는 좋은 성능을 보이지 못한다. QXtract[7]는 대량의 문서에서도 효과적이지만 단순히 문서 구조만을 추출한다는 한계가 있다.

본 논문에서는 문서 내에서의 정확한 패턴 추출을 위해서 사용자가 원하는 특정 주제 또는 특정 분야에 대해 예제 기반의 프로파일(profile)을 작성하고 이를 이용하여 보다 정확하게 문서를 검색하는 기법을 제안하고자 한다.

2. 전위 트리 기반의 사용자 프로파일 생성 기법

프로파일은 사용자가 선호하는 중요 키워드로 구성되며 불필요한 정보들을 여과하여 사용자에게 적합한 정보들만 제공할 수 있도록 한다. 따라서 정보의 질을 위해서는 프로파일을 어떤 특성들을 이용하여 어떤 방식으로 생성하는가 하는 점이 매우 중요하다. 본 논문에서 제안하는 프로파일의 생성은 키워드 추출과 키워드 마이닝의 두 단계로 구분된다. 2.1 절에서는 사용자가 예제로부터 키워드를 추출하는 방법에 대해 설명한다. 이어서 2.2 절에서는 추출된 키워드를 마이닝하여 프로파일을 생성하는 기법에 대해 기술한다.

2.1 프로파일을 위한 키워드 추출

기존의 연구에서는 일반적으로 단어 빈도수(word frequency) 기반으로 프로파일을 생성한다. 문서에 나타난 빈도수가 높은 단어에 가중치를 부여하여 키워드로 추출하는 것이다[4]. 그러나 단어의 중요도가 단지 출현 빈도에 의해 결정된다고 보기는 어렵다. 오히려 출현 횟수는 적어도 소제목이나 강조 효과가 적용된 단어는 문서 내에서 중요한 단어가 될 수 있다[3]. 따라서 [3]에서 제안한 스타일 기반의 변형된 가중치 공식에 따라 단어의 중요도를 파악하고 키워드를 추출한다. 변형된 가중치 공식은 식(1)과 같다.

단어 w_i의 가중치 $f_i = P(Z < a),\ a = \dfrac{fv_i - Avg(fv)}{Std(fv)}$ 식(1)

SW_k = 스타일 k 에 따른 적용 가중치 값,
$Avg(fv) = fv_i$ 의 평균,
$Std(fv) = fv_i$ 의 표준편차,

$$fv_i = \sum_{\text{출현한 } wi} \left(1 + \sum_{k=\text{적용스타일}} SW_k\right)$$

키워드 추출 대상은 사용자에게 예제 문서로 주어진다. 사용자가 문서에서 원하는 내용이 포함된 부분을 선택하면 식(1)에 의해 단어 w_i 와 가중치 f_i 가 나오며, 가중치가 사용자가 지정한 키워드 추출값 K 이상인 단어들로 트랜잭션(transaction) T_k 를 생성한다. 각 트랜잭션은 다음과 같이 출현 순서를 나타내는 TID 와 유한한 단어 집합으로 구성된다.

$$T_k = \{ \langle \text{TID} = k \rangle, \{ w_{k1}, \ldots, w_{kn} \} \}, w_{k1} < \ldots < w_{kn}$$

2.2 프로파일을 위한 키워드 마이닝

키워드를 추출한 뒤에는 그 키워드들이 잠재적으로 가지고 있는 의미를 최대한 유지해야 한다. 키워드를 단순히 축적하기만 하면 문맥 정보를 잃어버리거나 연관이 있는 단어 사이의 의존 관계 정보를 반영하지 못하는 단점이 있다[4]. 단어들 간의 의존도는 결국 그 단어들로 이루어진 연관 규칙(association rule)이라고 할 수 있으며, 연관 규칙은 빈발 항목 집합(frequent itemset)으로부터 찾아낼 수 있다. 여기에서는 전위 트리(prefix tree) 구조[2]를 이용한 키워드 마이닝을 통해 단어 간의 연관 규칙을 찾아내고 프로파일을 생성한다.

전위 트리는 다음과 같은 구조를 가지고 있다.

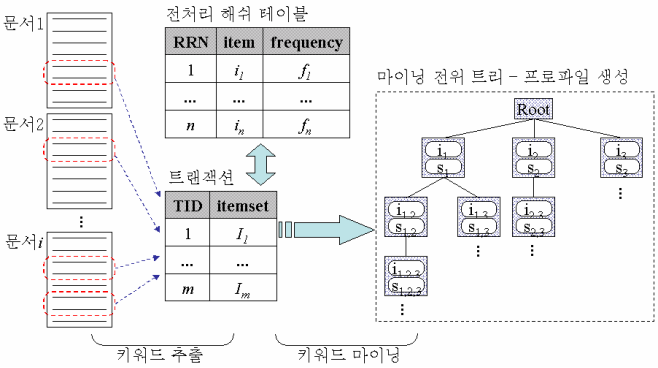
1. 'null' 값을 가지는 하나의 루트 노드를 가진다.
2. 각 노드는 단위 항목, 단위 항목의 출현 빈도수, 자식 노드를 연결하는 링크 필드를 가진다.
3. 항목 집합(itemset) $I = i_1 \cdots i_k$ 에 대해서 각 항목 $i_1, \cdots, i_k$ 는 사전적으로 정렬되어 있으며 I 에 대한 전위 트리의 경로는 root→i_1→⋯→i_k ($i_k \in I$, $k \geq 1$) 로 표현한다. I 의 출현 빈도수는 경로의 마지막 노드 i_k의 출현 빈도수로 나타낸다.

정확한 마이닝을 위해서 전처리를 위한 해쉬 테이블(hash table)을 관리한다. 해쉬 테이블의 각 엔트리는 엔트리의 주소 값, 출현 단어와 단어의 빈도수 필드로 구성되며, 트랜잭션의 단어들을 저장하고 빈도수를 증가시킨다. 단어의 출현 비율이 사용자가 지정한 최소 지지도 S 이하인 동안 해당 단어는 전위 트리의 입력 트랜잭션의 항목에서 제외된다. 이는 마이닝에 대한 사전 전지(pre-pruning) 효과를 가져오며 전위 트리를 최소 지지도 이상인 빈발 항목 집합들만으로 구성한다.

전위 트리를 이용한 마이닝 단계는 다음과 같다.

1. T_k 의 항목들을 해쉬 테이블에 적재하고 S 이상인 항목들로 트랜잭션 T_k' 을 만든다.
2. T_k' 의 모든 부분 집합에 속하는 항목 집합 $\{ I_j \mid I_j \in \{2^{Tk'} - \phi\},\ 1 \leq j \leq 2^{|Tk'|}-1 \}$ 에 대해 동일 항목에 해당하는 전위 트리의 노드 경로를 검색한다.
   ① 전위 트리에 I_j 에 대한 경로가 존재하면 I_j 의 각 항목에 대응하는 노드의 출현 빈도수를 해쉬 테이블의 단어 빈도수로 갱신한다.
   ② I_j 가 전위 트리에 없는 항목 집합이면 해당 항목들을 전위 트리에 삽입하고 출현 빈도수를 해쉬 테이블의 단어 빈도수로 갱신한다.
3. 모든 트랜잭션에 대해 단계 1, 2를 수행한다.

전위 트리에는 사용자가 원하는 특정 주제 또는 특정 분야에 대해 높은 빈도로 출현한 키워드 및 키워드들의 연관 관계가 저장되며, 이 전위 트리를 그 사용자를 위한 프로파일로 사용한다. 프로파일 생성의 전체적인 구성은 [그림 1]과 같다.



[그림 1] 프로파일 생성의 전체 구성

3. 사용자 프로파일을 이용한 문서 패턴 검색

생성된 프로파일에는 중요한 단어들과 그 단어들 간의 연관 관계가 함께 유지되므로, 프로파일을 이용하면 보다 정밀한 문서의 패턴 정보를 찾을 수 있다.

효율적인 검색을 위해 문서들을 문단(paragraph) 단위로 나눈다. 문단은 텍스트 내의 개행 문자(carriage return) 및 HTML 문서의
, <P> 태그를 기준으로 구분한다. 분리된 문단에는 번호를 부여하고 단어들을 사전적으로 정렬한다. 다음과 같이 각 문서 D_i 는 문단의 집합으로, 문단 P_k 는 출현 순서를 나타내는 PID 와 유한한 단어 집합으로 구성된다.

$$D_i = \{ P_k \},\ 1 \leq k \leq m$$
$$P_k = \{ \langle \text{PID} = k \rangle, \{ w_{k1}, \ldots, w_{kn} \} \}, w_{k1} < \ldots < w_{kn}$$

문단 단위의 검색마다 프로파일의 패턴들과 얼마나 적합한지 판단할 수 있는 점수를 얻는다[3]. 높은 점수를 획득한 문단은 그만큼 적합한 문서 부분이라는 뜻이다. 사용자가 지정한 점수 임계값(threshold) θ 이하인 문단들은 추출 과정에서 제외하여 요구 점수를 넘는 부분들만 결과로 제시한다. θ 를 연속으로 넘는 두 개 이상의 문단에 대해서는 재검색을 수행하여 문단들 사이에 걸쳐서 나타나는 패턴들을 간과하지 않도록 한다. 점수 계산식은 다음 식(2)와 같다.

$$Score = \frac{\sum_{i=\text{매치된패턴}} m_i \times s_i}{\sum_{j=\text{전체패턴}} l_i \times s_i} \qquad \text{식(2)}$$

m_i = 패턴 i 에서 일치하는 단어의 수,
l_i = 패턴 i 의 길이,
s_i = 패턴 i 의 지지도

프로파일을 이용한 문서 검색 단계는 다음과 같다.

1. 문서 D_i 를 문단 단위 P_k 로 분리하고 연속된 문단들의 단어를 저장할 확장소 B 를 할당한다.
2. P_k 의 모든 부분 집합에 속하는 단어 집합 $\{ I_j \mid I_j \in \{2^{Pk}\text{-}\phi\},\ 1 \le j \le 2^{|Pk|}\text{-}1 \}$ 에 대해 동일 항목에 해당하는 전위 트리의 노드 경로를 탐색한다.
   ① $Score(P_k) \ge \theta$ 이면 B 와 P_k 를 병합하고 B 의 원소를 재정렬한다.
   ② $Score(P_k) < \theta$ 이면 B 를 비운다.
3. B 가 2개 이상의 문단을 가지고 있으면 모든 단어 집합 $\{ I_j \mid I_j \in \{2^B\text{-}\phi\},\ 1 \le j \le 2^{|B|}\text{-}1 \}$ 에 대해 전위 트리의 노드 경로를 탐색한다.
4. D_i 의 모든 문단에 대해 단계 2, 3을 수행한다.
5. 모든 대상 문서에 대해 단계 1부터 수행한다.

단어 집합 $I = i_1 \ldots i_k$ 에 대응하는 전위 트리의 경로 root→i_1→…→i_k 가 존재할 때 그 경로를 매치된 패턴(matched pattern)이라고 정의하며, 이 매치된 패턴 집합이 사용자가 원하는 검색 결과가 된다. 높은 점수를 획득한 패턴들은 간접적인 피드백(feedback)으로 사용하여 프로파일을 학습시키는 역할을 한다.

4. 실험 및 평가

실험은 크게 두 가지로 나뉘어진다. 첫 번째는 생성된 프로파일이 선호도를 정확히 반영하는가에 대한 실험이며, 두 번째는 문서 검색 결과가 정확하게 제공되는가에 대한 실험이다.

실험 데이터 집합은 <표 1>의 세 가지 종류를 선택하였다. KIET 데이터는 산업연구원에서 제공하는 연구 간행물 문서로, 절(section) 구분이 정형화된 보고서 형식이다. NEWS 데이터는 KBS 웹 페이지에서 구할 수 있는 9 시 뉴스 대본이며 뉴스의 특성상 분량과 문단의 수가 거의 동일한 특징이 있다. COURT 데이터는 법원 도서관에서 제공하는 판례 및 문헌 데이터 집합으로, 문서 크기의 편차가 심하고 동일한 절이라도 문서마다 내용에 많은 차이가 있다.

<표 1> Data sets

데이터 집합	전체 문서수	문서 형식	문서 당 평균 단어수	문서 당 평균 문단수	사용자 선호도
KIET	129	hwp	970.50	7.86	금융 산업 동향
NEWS	730	HTML	2593.33	38.05	북한 핵
COURT	84,531	dbf	1129.71	13.71	저작권 침해 기각

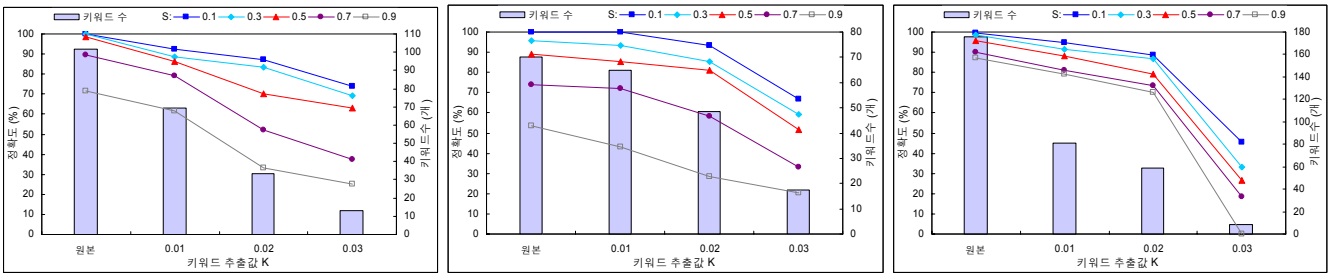
4.1 전위 트리 기반의 사용자 프로파일 검증

프로파일이 사용자의 선호 분야를 정확히 반영하는지 판단하기 위해 프로파일 생성에서 사용한 예제 문서 집합에 대해 판정을 수행하였다. 정확도를 판단하는 기준으로는 F_1 measure[10]를 사용한다. 따라서 정확도는 다음 식(3)과 같다.

$$Accuracy = \frac{2PR}{P+R} \qquad \text{식(3)}$$

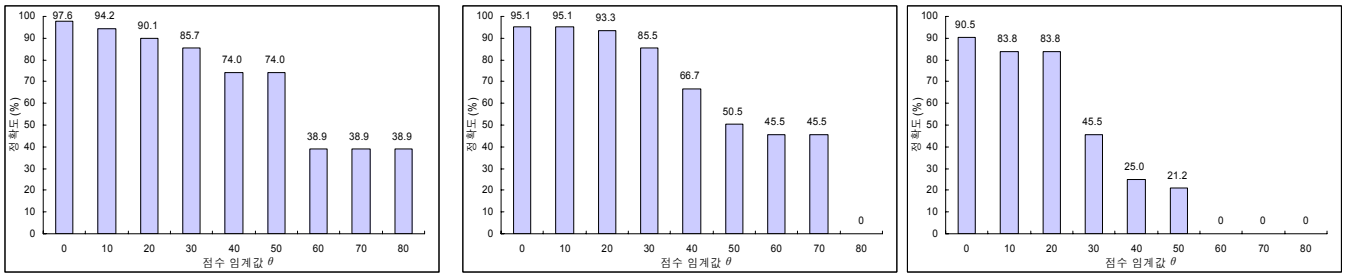
P : 정확율(precision),
R : 재현율(recall)

[그림 2]는 키워드 추출값 K 와 마이닝 최소 지지도 S 를 변화시켰을 때 각각 KIET, NEWS, COURT 데이터의 프로파일 검증 평균 정확도와 추출된 키워드 수의 변화를 나타낸 것이다. (a)에서 K 가 증가하면 상대적인 정확도가 감소하며 동일한 키워드 추출에서도 항목 집합의 최소 지지도 S 가 높아질수록 정확도가 감소하는 것을 알 수 있다. 그러나 S 를 지나치게 높게 잡는 것은 프로파일이 대부분의 연관 관계를 반영하지 못하므로 적합하지 않다. NEWS 데이터도 마찬가지로 키워드가 적어지거나 최소 지지도가 높아질수록 선호도를 잘 반영하지 못하였다. COURT 데이터의 경우 0.03 이상의 가중치를 가질 만큼 중요한 단어가 적게 출현한 것으로 나타났다. K 0.02 이내에서는 키워드들이 높은 연관 관계를 가지고 있으며 프로파일이 선호도를 제대로 반영하였다.



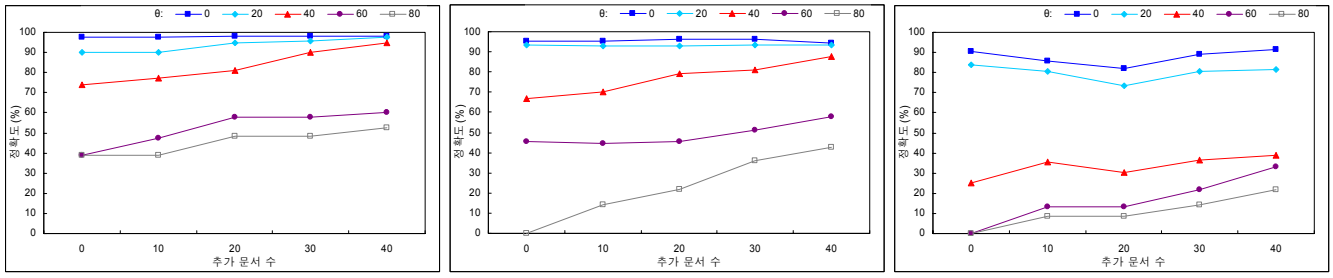
(a) KIET (b) NEWS (c) COURT

[그림 2] K 및 S 의 변화에 따른 각 프로파일의 상대적인 평균 정확도



(a) KIET (b) NEWS (c) COURT

[그림 3] θ 의 변화에 따른 프로파일 검색의 평균 정확도



(a) KIET (b) NEWS (c) COURT

[그림 4] 추가 문서 수의 변화에 따른 프로파일 검색의 평균 정확도

4.2 사용자 프로파일을 이용한 문서 패턴 검색 검증

프로파일을 이용한 검색 성능에 대한 실험을 수행하였다. [그림 3]은 S 를 0.1, K 를 0.01 로 고정하고 점수 임계값 θ 를 증가시켰을 때의 평균 정확도 비교 결과이다. 임계값이 크면 그만큼 패턴이 많이 걸러지므로 정확도가 감소한다. KIET 와 NEWS 의 경우 약 50 점 이내에서 많은 패턴이 추출되었다. COURT 데이터에서는 60 점 이상의 θ 에 대해 패턴을 전혀 찾지 못하는 것으로 나타났다. 이것은 단어들 간의 연관 관계가 짧고 단어 출현 빈도가 적기 때문이다.

문서 수에 따른 정확도를 평가하기 위해 검색 결과 매치된 부분들을 피드백으로 적용한 경우에 대해 추가로 실험하였다. [그림 4]는 S 를 0.1, K 를 0.01 로 고정하고 검색을 수행한 뒤 매치된 부분들을 프로파일에 추가하였을 때 각 θ 마다의 평균 정확도 그래프이다. 문서의 수를 0 부터 40 까지 증가시켰을 때 KIET 와 NEWS 데이터는 정확도가 증가하였다. COURT 의 경우 정확도가 떨어지기도 하는데, 이는 기존의 연관 관계 지지도를 높이지 못하고 새로운 연관 관계들이 많이 생성된 것이라고 볼 수 있다.

5. 결론 및 향후 과제

본 논문에서는 기존 프로파일 기법의 한계를 보완하여 보다 정확하게 사용자의 선호도를 반영하고 정밀한 검색을 수행하는 방법을 제안하였다. 스타일 적용 가중치를 이용하여 키워드를 추출하고 전위 트리를 기반으로 효과적인 사용자 프로파일을 구성하였다. 또한 다양한 도메인에서 프로파일을 이용하여 더욱 세부적인 검색 결과를 제공할 수 있음을 보였다. 향후에는 사용자의 선호도가 다양해질 때 프로파일의 효과적인 구성 방법에 대한 연구가 진행되어야 할 것이다. 정확도나 실행 속도의 측면에서 검색 성능을 향상시키는 방안에 대한 연구도 필요하다.

참고문헌

[1] R. Grishman, “Information extraction: Techniques and challenges,” *In a Multidisciplinary Approach to an Emerging Information Technology*, 1997.

[2] R.C. Agarwal, C.C. Aggarwal and V.V.V. Prasad, “Depth First Generation of Long Patterns,” *In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.

[3] 이준휘, 이원석, “스타일 기반 키워드 추출”, 한국정보처리학회 춘계학술발표회, 2002.

[4] D. Mladenic and M. Globelnik, “Word sequences as features in text learning,” *In Proceedings of the 17th Electrotechnical and Computer Science Conference*, 1998.

[5] V. Hristidis and Y. Papakonstantinou, “DISCOVER: Keyword Search in Relational Databases,” *In Proceedings of the 28th VLDB Conference*, 2002.

[6] A. Hulgeri and C. Nakhe, “Keyword Searching and Browsing in Databases using BANKS,” *In Proceedings of the 18th International Conference on Data Engineering*, 2002.

[7] E. Agichtein and L. Gravano, “Querying Text Databases for Efficient Information Extraction,” *In Proceedings of the 19th International Conference on Data Engineering*, 2003.

[8] E. Agichtein and L. Gravano, “Snowball: Extracting Relations from Large Plain-Text Collections,” *In Proceedings of the 5th ACM International Conference on Digital Libraries*, 2000.

[9] R. Yangarber, R. Grishman, P. Tapanainen and S. Huttunen, “Unsupervised Discovery of Scenario-Level Patterns for Information Extraction,” *In Proceedings of the 6th Conference on Applied Natural Language Processing*, 2000.

[10] Y. Yang, “An Evaluation of Statistical Approaches to Text Categorization,” *Information Retrieval Journal*, 1999.