주요 항목 집합을 이용한 문서 클러스터링 및 연관 탐사 기법†

서 성보○ 김 선철 이 준욱 류 근호
충 북 대 학 교 데 이 터 베 이 스 연 구 실
{sbseo, sckim, junux, khryu}@dblab.chungbuk.ac.kr

Clustering and Association Rule Mining of Transactions using Large Items

Sung Bo Seo○ Jun Wook Lee Sun Cheul Kim Keun Ho Ryu
Dept. of Computer Science, Chungbuk National University

요 약

현재 광범위한 웹 문서를 검색하기 위해서 많은 사용자들이 여러 종류의 검색엔진을 사용하고 있다. 하지만 대부분의 사용자는 검색엔진에 의해 검색된 문서의 순서화가 된 긴 리스트의 검색 문서들과 이들이 갖는 낮은 신뢰도에 대해 검색된 문서 중에 자신이 원하는 타당한 문서를 검색하는 불편함이 있어 왔다. 정보 검색에서 문서의 클러스터링은 검색된 결과를 재구성하는 효율적이고 선택적인 방법이다. 이 연구에서는 문서를 트랜잭션 관점에서 해석하여 하나의 클러스터에 대해 유사성을 측정하기 위해 주요항목과 비 주요항목으로 구분하여 각 트랜잭션의 최소 비용 계산을 통해 자동화된 문서 클러스터링 기법을 제안한다. 또한 클러스터링 단계에서 주요 항목간의 연관 규칙을 생성하기 위하여 문서 클러스터링을 위한 디스크 엑세스 동안 키워드간의 연관성을 찾을 수 있는 효율적인 검색 기법을 제시한다.

1. 서론

90년대가 정보검색의 시대였다면 2000년은 지식검색의 시대가 될 것이다. 좀더 상세히 말하면 3세대 검색엔진의 등장이며 이는 크게 2가지 기술로 나뉘는데 하나는 자연어 검색이며 다른 하나는 자동분류 시스템이다.

실제로 현재 사용하는 검색엔진들의 대부분은 AND, OR 등과 같은 불리언 연산에 의존하고 있다. 이러한 검색 모델은 불리언 연산 이외에 확률검색, 가중치 검색, 퍼지 집합 검색, 추론 검색 등의 형태로 존재한다. 검색 속도면에서 불리언 검색이 좋은 결과를 보이지만 정확도 면에서 가장 떨어지며 또한 질의어에 대한 적합한 순서, 단순한 키워드에 단순 일치하는 문서만을 검색하기 때문에 만족스러운 결과를 얻어내지 못한다. 하지만 대부분의 시스템에서는 구현상의 어려움으로 인해 대부분 이 방식을 사용하고 있다.

이 논문에서는 자동분류와 검색을 위해 다음의 두 가지 방안을 제시하고, 이 방안에 적용할 수 있는 클러스터링 기법과 연관 탐사 기법을 제안한다. 첫째, 광범위한 웹 문서의 자동 분류를 위해 하나의 문서를 트랜잭션 관점에서 해석하여 문서의 최소 발생 빈도를 가지는 단어를 주요 항목이라 하였으며 이 주요 항목을 이용 트랜잭션 집합이 최소 비용 계산을 통해 유사한 클러스터와 비 유사한 클러스터로 동적 자동 분류 하는 기법을 소개한다. 둘째, 주요 항목의 집합을 생성하는 과정에서 최소 신뢰도를 만족하는 발생 항목간의 연관성을 고려하여 키워드 검색시 이들 연관 단어를 이용하는 기법으로 신뢰성있는 재 검색 입력이 가능하다.

† 이 논문은 한국 과학 재단의 99년 특정 기초 연구 사업의 연구비 지원에 의해 수행되었음.

위 두 가지 기법을 이용 효율적으로 문서 클러스터링을 수행하고 분류된 클러스터를 대표하는 주요 항목들간의 연관성을 고려하여 검색시 키워드 연관성을 고려해 검색이 가능하며 검색은 자동 분류되어 있는 클러스터에서 검색되어 진다. 또한 이 두 가지 기법에 사용되는 알고리즘은 한번의 디스크 검색과정에서 두 가지 규칙을 생성할 수 있으므로 대용량의 문서를 엑세스하는 비용을 절감할 수 있다.

2. 관련 연구

데이터 클러스터링 방법에는 두 객체의 쌍 사이의 유사성을 측정하는 관점(행렬, 코사인, Dice and Jaccard Coefficient)에서 논의되어져 왔으며 문서 발생 빈도(용어를 포함한 문서의 부분)관점으로는 문서내 주요 항목 집합을 이용한 텍스트 문서 분류와 클러스터링 방법이 널리 사용되어져 왔다. 그럼에도 불구하고 중요한 차이점은 발생 빈도 측면에서 트랜잭션이 하나의 클러스터 관점에서 지정된 것과는 달리 전체 클러스터는 아니며 주요 항목은 클러스터가 진행되면서 계속 유지되어져야만 한다. 또한 이 방법은 트랜잭션을 읽고 다음 트랜잭션에서 최상의 클러스터에 할당하는 *K*-means 알고리즘과 유사하지만 차이점은 첫째, 초기 클러스터의 개수인 K 값을 요구하지 않으며, 둘째, 클러스터의 거리에 도대를 두지 않으며 클러스터의 전체적인 최상의 해결책을 제시한다 [Wang99].

연관 규칙을 찾는 문제는 [Agra94]에 소개되어 실제 거래 데이터베이스에 기록된 판매 데이터의 분석에 적용되고 있다. 항목들의 집합으로 구성된 트랜잭션이 주어졌을 때 연관 규칙은 X->Y의 형태로 표현되며, 이는 X와 Y가 항목들의 집합이고 하나의 트랜잭션에 X와 Y가 동시에 존재한다는 것이다. 마이닝 연관 규칙 탐사의 문제는 사용자가 지정한 최소 지지도와 신뢰도를 만족하는 모든 연관 규칙을 찾기 위한

시도로 정의한다[Agra94]. 여기서 최소 지지도를 만족하는 항목을 주요 항목이라고 보며 이때 항목간의 연관성을 이용해 키워드 연관성을 찾아낼 수 있다.

3. 주요 항목을 이용한 클러스터링 기법

이 장에서는 연관 규칙 탐사에서의 주요 항목 개념을 토대로 클러스터링의 개념을 정의하며 자동화된 분류를 위한 비용계산 기법과 알고리즘을 제시한다. 트랜잭션의 집합{t_1,t_2, ... ,t_n}을 고려할 때 각 트랜잭션 t_i는 항목{i_1,i_2,... i_p}들의 집합이며 하나의 클러스터 C는 {t_1,t_2, ... ,t_n}를 분할한 {C_1,C_2, ... ,C_n}이다. 이때 각 C_i를 클러스터라 한다.

3.1 주요항목을 이용한 접근 방법

트랜잭션에 대한 클러스터의 유사성을 측정하기 위해 주요 항목 집합을 사용하며 이때 클러스터 C_i의 지지도는 항목을 포함하는 C_i의 트랜잭션 개수이다. 사용자 지정 최소 지지도인 θ ($0<\theta<1$)는 하나의 항목에 대해 C_i가 적어도 $\theta*|C_i|$를 만족할 때 클러스터 C_i에서 주요항목이라고 하며 그렇지 않으면 비 주요항목이라 한다. $Large_i$는 C_i의 주요항목들의 집합이라 하며 $Small_i$는 비 주요항목의 집합이라 표시한다. 클러스터 C에 대해서 C의 비용은 Inter-Cluster와 Intra-Cluster 비용의 두 개의 요소를 고려하여 최소 비용을 고려하게 된다.

Intra-Cluster Cost는 비 유사성 측면에서 비 주요 항목의 총 개수에 의해 측정되어 진다. $Intra(C)=|U_{i=1}^{k} Small_i|$ -- -- -- (식 1)

이 컴퍼넌트는 너무 많은 비 주요항목들의 약 결합 클러스터링을 생성하는 것을 억제하며, 두 클러스터가 결합되어 질 때 비 주요항목이 두 번 반복해서 세지 않기 위해 $\sum_{i=0}^{k}|Small_i|$를 사용하지 않고 위 수식(1)를 사용한다.

Inter-Cluster Cost는 내부 클러스터의 유사성을 측정하기 위한 것이며 주요 항목의 집합이 클러스터의 유사성에 기여하므로 각 클러스터는 가능한 주요항목의 최소 부분이 겹쳐야 한다.

$Inter(C) = \sum_{i=1}^{k}|Large_i| - |U_{i=1}^{k} Large_i|$ -- -- -- (식 2)

두 개의 비용을 함께 계산하기 위한 클러스터 C의 비용 함수는 사용자의 지정 값 w에 따라 유사성과 비유사성의 가중치를 설정 할 수 있으며 이 논문에서는 w=1로 지정하여 실험하였다.

$Cost(C) = w * Inter(C) + Intra(C)$ -- -- -- (식 3)

위의 비용계산에 따라 트랜잭션의 집합과 최소 지지도가 주어졌을 때 $Cost(C)$가 최소인 클러스터 C를 찾는 방법은 아래와 같다.

주어진 최소 지지도는 자동 분류를 위해 초기 클러스터의 수를 지정하지 않고 비용이 최소가 되는 것을 찾는 것이며 또한 이 알고리즘은 명확한 임계값(Threshold)을 사용하지 않지만 대신 좋은 클러스터링의 검색시 부적절한 분류가 되지 않도록 한다. 이러한 특징은 동적 클러스터링에서 클러스터의 유사성이 새로운 트랜잭션이 더해질 때 유용하다. 사실 *K*-means와 같은 근접성을 사용하지는 않지만 전체 클러스터의 질을 평가하는데 주요항목의 집합을 사용한다.

예 3.1은 6개의 트랜잭션을 고려하여 Inter-cost 와 Intra-cost의 전체 비용의 계산을 통해 클러스터 되는 예를 보여 준다.

예 3.1) 6개의 트랜잭션 고려

t_1={a,b,c}, t_2{a,b,c,d}, t_3={a,b,c,e}, t_4={a,b,f}, t_5={d,g,h}, t_6={d,g,i}

사용자 정의 최소 지지도를 60%라 할 때 주요항목의 집합은 적어도 4개의 트랜잭션을 포함되어 있어야 한다 (i.e., 6*60).

1) C_2= {C_1={t_1,t_2,t_3,t_4}, C_2={t_5,t_6}} C_1에 대해 주요항목의 집합은 적어도 C_1에 3개의 트랜잭션에 포함되어 있어야 하며 $Large_1$={a,b,c} $Small_1$={d,e,f} 유사하게 $Large_2$={d,g}, $Small_2$={h,i}이다 그러므로 $Intra(C_2)$=5, $Inter(C_2)$ =0이며 $Cost(C_2)$는 5이다

2) C_3= {C_1={t1,t2}, C_2={t3,t4},C_3={t5,t6}} $Large_1$={a,b,c} $Small_1$={d}

$Large_2$ ={a,b}, $Small_2$={c,e,f}, $Large_3$={d,g} $Small_3$ ={h,i}, $Intra(C_3)$=6, $Inter(C_3)$ =3이며 $Cost(C_3)$는 9이다. 예 1)의 클러스터의 비용이 2)의 클러스터의 비용보다 적으므로 1)과 같은 방법으로 클러스터링 되어진다.

3.2 항목 연관 규칙 탐사

대량의 상품 판매 데이터베이스에서 항목간의 연관 규칙을 찾는 문제에서 최소 지지도와 신뢰도를 만족하는 항목간의 관계분석을 마케팅 전략 수립, 병의 징후 관계 분석과 웹 사용자 패턴분석 등에 유용하게 사용되어진다. 이 논문에서는 클러스터링 되어진 문서의 주요 항목 집합간의 키워드 연관 규칙을 찾아 검색시 신뢰도가 높은 키워드 추천 및 재 검색에 적용한다. 문서 항목의 집합 Υ ={i_1,i_2, ... i_m}이라 가정하면 트랜잭션의 집합 D는 각 트랜잭션 T에 대해 $T \subseteq \Upsilon$인 항목들의 집합을 말한다. 각 트랜잭션과 관련된 유일한 구별자는 TID이며 만약 $X \subseteq T$면 하나의 트랜잭션이 Υ에 몇 가지 항목의 집합 X를 포함한다라고 말할 수 있다. 연관 규칙은 X->Y의 형태로 표현되며($X \subset \Upsilon$, $Y \subset \Upsilon$ and $X \cap Y = \emptyset$) 이는 X와 Y가 항목들의 집합이고 하나의 트랜잭션에 X와 Y가 동시에 존재한다는 것이다. 규칙 X->Y는 트랜잭션 D의 C%가 X와 Y를 동시에 포함할 때 신뢰도 C%를 가진 트랜잭션 D를 가진다. 연관 규칙을 찾는 문제는 두 단계로 분리되어 진다. 첫 단계는, 최소 지지도 이상의 트랜잭션 지지도를 가지는 모든 항목들의 집합을 찾는다. 항목집합에 대한 지지도는 항목들을 포함한 트랜잭션의 개수를 말하며 최소 지지도 이상의 항목들을 주요 항목의 집합이라 하며 다른 것을 비 주요항목이라 한다. 둘째 단계에서는, 규칙을 생성하기 위하여 주요항목의 집합을 사용하며 일반적인 아이디어는 ABCD와 AB가 주요 항목의 집합이라 하면 규칙 AB->CD는 신뢰도(conf) = 지지도(ABCD)/지지도(AB)의 비율을 계산하여 결정하며 만약 신뢰도가 최소 신뢰도보다 크다면 규칙을 포함한다.(이 규칙은 ABCD가 주요항목의 집합이므로 반드시 최소 지지도를 만족해야 한다)

3.3 클러스터링 및 연관 규칙 탐사 알고리즘

```
/* 클러스터 할당 단계 */
1) while not end of file do
2)     read the next transaction <t>;
3)     allocate t to an existing or new cluster Ci to min Cost(C);
4)     write<t,Ci>;

/* 정제 과정 및 주요항목 Count */
5) repeat
6)   not-moved =true;
7)   while not end of file do
8)       read the next transaction <t,Ci>;
9)       move t to an existing non-singleton cluster Cj to min Cost(C);
10)      count Largei to Ci or Cj;
11)      if Ci ≠ Cj then
12)        write  <t,Cj>;
13)        not-moved=false;
```

```
14)        eliminate any empty cluster;
15)     call association rule with Largei parameter;
16) until not_moved;

/* 하나의 클러스터에 대한 주요 항목 연관 탐사 */
17)  L1 ={Large 1-itemsets};
18)  for (k=2; Lk-1 ≠ ∅; k++) do
19)     Ck= apriori-gen(Lk-1);
20)     forall transactions t ∈ D do
21)        Ct = Subset(Ck,t);
22)        forall candidates c ∈ Ct do
23)           c.count ++;
24)     Lk = {c ∈ Ck | c.count ≥ minsup}
25) write UkLk;
```

그림 1. 클러스터링 및 연관규칙 탐사 알고리즘

그림 1은 클러스터링 및 연관 규칙 탐사 알고리즘을 보여주며 각 트랜잭션은 연속적으로 읽고 존재하는 클러스터에 분류되거나 새로운 클러스터를 생성 이 과정에서 현재 클러스터 C에 대해 최소 Cost(C)를 고려한다. 이 과정이 할당 단계이며 정제 과정에서는 할당 단계와 같은 방법으로 각 트랜잭션을 읽으며 존재하는 비 단일 클러스터 중 최소 비용 클러스터에 할당하게 되며 트랜잭션이 존재하지 않는 클러스터는 제거하며 반복 과정동안 모든 트랜잭션에 대해 한번의 이동이 이루어지지 않으면 마치게 된다.

이 정제 과정동안 각 클러스터의 주요항목 집합의 개수를 세어 해당 클러스터의 주요 항목 집합간의 연관 규칙을 생성하게 된다. 이 과정에서 Apriori[Agra94] 알고리즘을 이용한다. 단계 1에서는 클러스터 정제 과정에서 세어진 주요 항목의 집합을 이용하여 주요 1항목을 찾는 것으로 후보 1항목에 대한 발생 빈도수를 계산한 후 주요 1항목으로서 최소 지지도와 신뢰도를 만족하는 항목들을 선택한다. 단계 2에서 주요 k항목 집합은 Apriori-gen[Agra94]알고리즘을 이용하여 $k-1$의 주요 항목으로부터 최소 지지도와 신뢰도를 만족하는 주요 k항목의 집합을 생성해 낸다. 이 단계는 새로운 주요항목 집합이 발견되지 않을 때까지 반복한다. 이 과정을 거치면 각 클러스터를 대표하는 주요항목 집합을 이용하여 항목 연관 규칙이 생성되어 진다.

4. 실험

4.1 실험 환경 및 내용

실험을 위해 웹 로봇인 BDDBot를 이용하여 검색을 수행하였으며 주요 항목을 이용한 문서 클러스터링 및 연관규칙 생성을 위해 데이터는 JDK1.2 문서와 연구실 홈페이지 데이터로 구성하였다. 웹 서버와 자바 환경은 Window NT이고 데이터 저장을 위해 오라클 7.3.4 DBMS를 이용하였다.

4.2 실험 결과 분석 및 결과

그림 2는 실험에 대한 질의 검색 결과를 보여준다. 먼저 BDDDot를 이용하여 JDK1.2 문서와 연구실 홈페이지에 대한 각 웹 문서에 대해 항목에 관한 인덱스 키를 위해 Hash 테이블을 이용하여 주어진 항목의 지지도에 대한 삽입, 삭제, 변경이 이루어진다. 인덱스 키로서 항목에 대한 지지도에 관해 B-트리를 이용하여 주어진 항목의 지지도를 가지는 모든 항목을 찾기 위한 제어 경로를 제공한다 이 논문에서 제시된

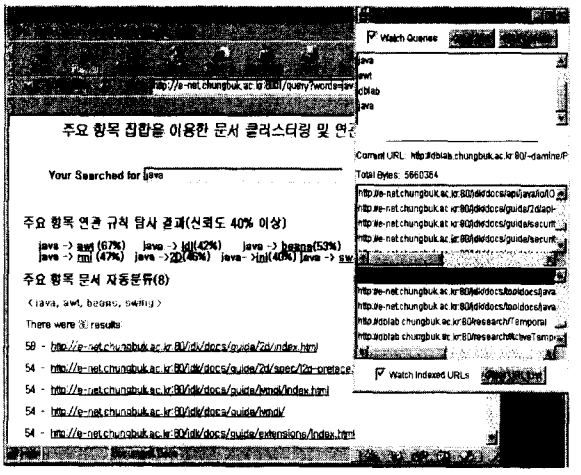

그림 2. 주요항목 집합을 이용한 검색 결과

알고리즘은 검색시 인덱싱 되어진 키를 이용하여 수행되어 진다.

웹 브라우저에서 질의어가 입력되면 질의 결과는 주요 항목 집합에 대한 키워드 연관 규칙을 생성하여 질의 단어와 동시 검색 또는 신뢰도 높은 연관 단어 재 검색이 이루어 질 수 있다. 주요 항목을 이용한 문서 자동 분류는 전체 트랜잭션에 대해 각 클러스터의 주요 항목 집합을 보여 주며 이 주요 항목에 대한 분류 문서의 결과가 출력된다.

5 결론 및 향후 연구

이 논문에서는 웹 문서의 자동 분류 기법을 위해 하나의 문서를 트랜잭션 관점에서 해석하였으며 문서의 최소 발생 빈도를 가지는 단어를 주요 항목이라는 관점에서 트랜잭션의 집합을 유사한 클러스터와 비 유사한 클러스터로 분류한다. 또한 최소 비용 계산을 통해 동적 자동 분류 기법과 최소 신뢰도를 만족하는 발생 항목간의 연관성을 고려하여 키워드 검색시 연관 단어를 이용하여 신뢰성있는 재 검색 입력이 가능한 방안을 제시하였다. 적용분야에 있어 문서 검색 분류와 전자상거래, 의학, 비즈니스 등 산업 전반에 응용이 가능하며 동일한 방식의 분류 및 연관 규칙 생성으로 I/O 효율성이 증대되었다. 향후 자동 분류를 위한 비용 계산의 정확성과 효율적인 저장 구조의 개선이 필요하다.

6. 참고 문헌

[Agra94] R.Agrawal and R.Srikant. "Fast Algorithms for Mining Association Rules," In Proc. of VLDB 1994.

[Wang99] Ke Wang and Chu Xu. "Clustering Transactions Using Large Items," in ACM CIKM-99, 1999.

[Oren98] O.Zamir and O.Etzioni, "Web document clustering : a feasibility Demonstration," In Proc. of ACM SIGIR, 1998.

[Dani99] Daniel Boley and Maria Gini. "Document Categorization and Query Generation on the World Wide Web Using WebACE," In Journal of Artificial Intelligence Review, 1999.