DRFP-Tree를 이용한 e-Book분류방법 제안

김종엽*, 조경수, 김응모
*성균관대학교 정보통신공학부
e-mail : womkjy@gmail.com

Proposal of e-Book Classification Method using DRFP-Tree

Kim Jong Yeup,* Kyung Soo Cho, Ung-mo Kim
*School of Information and Communication Engineering,
Sungkyunkwan University

요 약

2007년 Amazon.com이 미국에서 e-Book 전용 단말기 'Kindle'을 출시한 이래, Sony와 대형 서점 Barnes&Noble등 메이저 업체는 물론 다수의 중소업체들이 e-Book 시장에 진출하고 있다. 최근에는 Apple이 iPad를 출시하고 e-Book 시장에 진출한 가운데, Google 역시 6월 이후 e-Book 시장에 진출할 것을 발표함으로써 e-Book 시장의 경쟁이 더욱 치열해지고 있다. e-Book의 급속한 보급증가와 함께 이런 방대한 도서를 관리하는 곳에서 자동 도서 분류의 필요성도 증가하고 있다. 기존의 문서분류 방법들은 대게 수작업, 텍스트(단어)의 집합으로 간주하여 기계 학습방법을 그대로 적용하거나 약간의 변형을 가한 방법들이 대부분 이었다. 본 제안서에서는 데이터 마이닝 분야에서 사용되는 DRFP-Tree 구조를 이용하여 e-Book 내의 문장들의 패턴을 저장하고 이를 사용하여 e-Book을 분류하는 방법을 제안한다.

1. 서 론

전자책(e-Book) 러시(Rush)다. Amazon의 'Kindle'이 불씨를 당긴 후, iPhone으로 대표되는 스마트폰 열풍이 이를 부채질하고 있으며, iPad의 등장으로 불타오르는 형국이다. 지난 1월 라스베이거스에서 개최된 미국 소비자 가전쇼(CES)에서는 e-Book 단독 부스가 설치되어 그 위상을 널리 알렸다. 닷컴의 붕괴와 함께 침체에 빠졌던 전자책 시장을 2007년 Amazon 'Kindle'이 부활시키면서 전 세계 e-Book 시장 성장을 견인하고 있다. Sony, Google 등 글로벌미디어기업들이 속속 진입하고 있으며, 국내에서도 삼성, LG 등 대기업 중심의 단말기 제조사와 교보, 인터파크 등 대형 유통사 등이 결합된 단말기 출시 등으로 과열 조짐까지 보이고 있다.

종이책 대비 30%, 많게는 70%가량 저렴하며, 유통과 제고관리문제를 해결 할 수 있는 e-Book은 아무리 많은 책도 간단하게 휴대 가능하며, 훼손 및 분실의 위험이 없고, 시간과 장소에 상관없이 언제든 구매하여 바로 읽을 수 있다. 이러한 장점들 때문에 전 세계 e-Book 시장이 연평균 37.2%로 급성장 중이며 2013년에는 09년 대비 4배인 13.4조원의 성장이 기대된다. '전 세계의 모든 책을 인터넷으로 제공 하겠다'는 Google의 슬로건에서 느낄 수 있듯이 바야흐로 '아날로그 미디어의 최후의 보루'로 일컬어지던 '서적'의 디지털화가 본격화 되고 있는 것 이다.

이렇게 다양한 플랫폼을 기반으로 많은 도서들이 e-Book으로 출판되고 있고 기존의 도서들마저 e-Book으로 변환 되고 있다. Google은 저작권이 만료된 e-Book과 출판사들의 보유하고 있는 200만권 이상의 e-Book등 약 700만권 이상을 확보하고 있어 e-Book시장을 독식할 것으로 전망되고 있으며, 현재 Barnes&Noble 100만권, Amazon 40만 권에 앞으로 출판될 e-Book을 고려한다면 그 분량은 실로 어마어마 한 것이다.[1]

본 논문에서는 이렇게 방대한 분량의 e-Book을 분류하기 위하여 기존의 Bayes, k-NN, SVM(support vector maghine)등과 같은 기계학습 관점에서의 문서분류 방법과 이러한 분류 방법을 묶어 성능을 향상시킨 Committee Machine, Boosting, Bagging등 과 같은 방법을 벗어나 좀 더 빠르고 효율적인 '자동화된 도서 분류방법'을 제시하고자한다.

논문의 구성은 다음과 같다. 2장에서 관련연구에 대해서 설명하고, 3장에서는 DRFP-Tree를 이용한 e-Book분류 방법에 대해 제안하고. 4장에서 향후 연구 방향에 대해 제시하며 결론을 맺는다.

2. 관련연구

2.1 연관규칙 마이닝

연관규칙 마이닝은 주어진 데이터 집합에서 흥미로운 연관성이 있는 항목을 찾아내는 방법 중 하나이다. 마케팅에서 고객이 동시에 구매한 장바구니를 살펴봄으로써 거래되는 상품들의 관련성을 발견하고 분석하는 방법으로 장바구니 분석(Market basket analysis)이라고도 알려져 있다. 이렇듯 연관성 분석은 연관성 규칙을 통해서 하나의 거래나 사건에 포함되어 있는 둘 이상의 품목들의 상호관

련성을 발견하는 것이다. 일반적으로 연관성 분석은 수학과 통계학의 확률과 기대치에 대한 개념을 기반으로 하고 있는데, 이러한 연관성 규칙을 해석하는데 있어 원인과 결과의 직접적인 인과관계로 생각해서는 곤란하고 두 개 또는 그 이상 품목들 사이의 '상호의 관련성'으로 해석해야 한다. 본 논문에서는 문장을 구성하는 단어의 상호 관련성을 고려한다.[2]

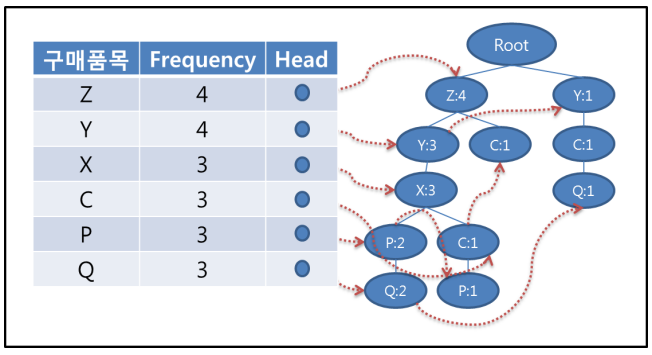
2.2 빈발 패턴 증가 기법(Frequent Pattern Growth)

FP-Growth기법은 연관 규칙 마이닝의 알고리즘 중 하나로 후보생성 없이 분할-정복 기법을 사용하여 빈발항목을 찾아낸다. FP-Growth는 깊이 우선 알고리즘으로 FP-Tree라는 자료구조를 사용한다. 이는 빈번히 발생하는 패턴의 중요한 정보를 효율적으로 압축, 저장하는 자료구조이며, Prefix-tree를 확장한 형태이다. 예를 들어 마트에서 고객들의 구매내역을 분석한 결과 '30대 초반의 남성 고객이 귀저기를 사면 맥주도 같이 사는 경향이 많다'와 같은 구매 패턴을 분석할 수 있다. 이와 같이 빈번히 발생하는 패턴을 쉽게 찾아내는 방법이 FP-Tree이다. [3]

(표 1) FP-Tree를 구현하기 위한 Database예제

구매자	구매품목	자주구매하는 품목
1	Z, X, Y, P, N, U, P, Q	Z, Y, X, P, Q
2	X, C, Y, Z, I, P, P	Z, Y, X, C, P
3	C, Z, H, J, P	Z, C
4	C, Y, U, S, Q	Y, C, Q
5	X, Z, Y, E, U, Q, P, B	Z, Y, X, P, Q

(표 1)은 앞서 예를 든 마트의 구매 분석을 FP-Tree로 설명하기위한 Database의 예이다. 주의 깊게 봐야할 '자주 구매하는 품목'은 모든 구매내역들을 많이 나온 순으로 정렬한 다음 모든 구매고객에게서 50%이상 나타나는 품목들만 나타낸 것이다. 이렇게 만든 '자주 구매하는 품목은' 다음 (그림 1)과 같이 FP-Tree로 구성하게 된다.



(그림1) 표1의 데이터를 FP-Tree로 구현한 결과

자주 구매한 품목을 많이 나타난 순으로 정렬하여 'Header Table'을 만들고, 구매자별 '자주 구매 하는 품목'을 그 순서를 유지한 채 트리구조로 구성하면 된다. 완성된 FP-Tree로부터 Frequent Pattern을 찾는 방법은 다음과 같다. 예를 들어 C품목이 들어간 Frequent Pattern을 찾는다면 그림 1의 Header Table의 품목'C'의 link를 따라가면서 각각의 부모 노드를 살펴보면 된다. 그 결과는 {(C, X, Y, Z : 1), (C, Z : 1), (C, Y : 1) }가 나오게 되고, 이 결과를 종합하면 C는 품목 'Z'와 2번, 품목 'Y'와 2번, 품목 'X'와 1번 함께 나타나는 것을 알 수 있다.

2.3 DRFP-Tree(Disk-resident FP-tree)

FP-Growth기법의 FP-Tree는 기존의 Apriori, CLOSET+, MAFIA와 같은 알고리즘들이 가진 여러번의 DB스캔을 단 2회로 단축하면서 소요시간을 획기적으로 줄일 수 있었다. 하지만 이러한 장점에도 불구하고 DataBase의 크기가 커지게 되면 많은 공간의 메모리를 필요로 하게 되는 단점을 갖고 있다. 이 경우 패턴 마이닝을 수행하지 못하는 경우가 발생하며 이런 문제를 해결하기 위한 방법으로 Muhaimenul Adnan과 Reda Alhajj는 DRFP(Disk-resident FP-tree) 알고리즘을 제안했다.[4]

이는 FP-Tree의 단점을 보완한 것으로 메모리 병목현상이 발생하면 tree 구조를 유지와 동시에 부분화 시켜 2차 저장장치에 나누어 저장장치에 나누어 저장한다. 그래서 낮은 Support threshold나 큰 Database 환경에서도 Frequent pattern mining을 가능하게 하며 데이터 전처리 과정에 트랜잭션들의 순서를 정렬하는 것도 추가되었다.

하지만 FP-Tree에서는 존재하지 않던 Tree의 이원화하는 알고리즘 구조로 인해 계산에 걸리는 시간이 FP-Tree에 비해 크게 증가하는 단점을 갖고 있다. (그림 2)에서 FP-Tree와 DRFP-Tree를 쉽게 비교하였다.

(그림 2) FP-Tree와 DRFP-Tree 비교

3. 제안내용

3.1 e-Book포멧 선택

본 논문에서는 e-Book의 라이센스 문제를 고려하지 않아도 되고, 문서 분류에서 가장 널리 쓰이는 문서 데이타

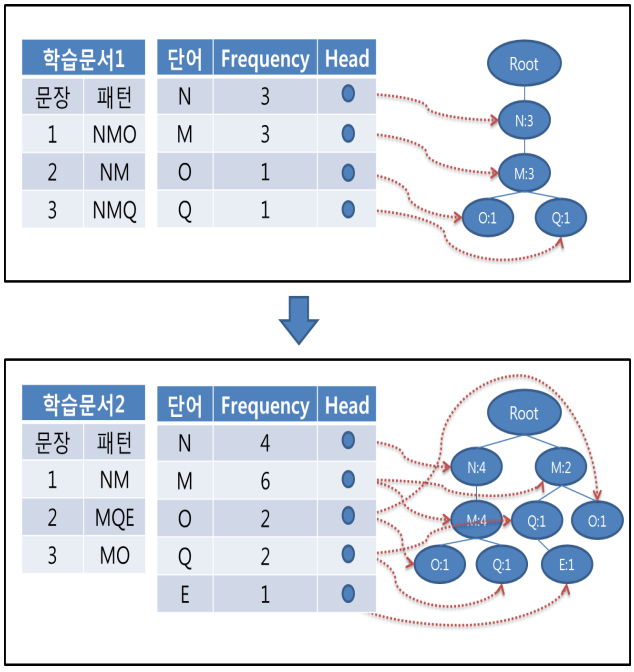
인 reuters-21578를 사용한다. 이는 로이터 통신의 기사 21578개를 모은 것으로 David Lewis 등이 총 135개의 중복이 허용되는 Topic으로 분류한 문서 집합이다.[5]

3.2 DRFP-Tree의 선택

앞서 살펴본 것과 같이 DRFP-Tree 장점은 제한된 시스템자원에 구애받지 않는다. 이는 방대한 분량의 e-Book을 고려할 때 가장 적합한 데이터 마이닝 기법으로 판단 할 수 있다. FT-Tree를 이용 할 경우 최초 속도에 대한 이점은 있을 수 있지만 방대한 분량의 e-Book을 분류하는 과정에서 어떠한 시스템이던지 메모리의 병목현상은 피할 수 없으며 이는 곧 '더 이상 분류가 불가능'하다는 의미로 해석할 수 있다. DRFP-Tree구조에서 기존의 FP-Tree구조를 부분화하여 2차 저장장치에 나누어 저장하는 과정의 계산 소비시간을 고려하더라도 '이용 불가한 FP-Tree'와 기존의 기계 학습방법을 바탕으로 한 문서분류 방법과 비교하여 보다 더 효율적이고 빠른 분류가 가능할 것이란 판단이다.

3.3 전처리와 e-Book의 분류

전처리란 분류할 e-Book을 비교하기위한 비교대상 DRFP-Tree의 생성을 의미한다. 즉, 전처리 과정을 거치지 않으면 문서는 문자들의 의미 없는 나열에 불과하게 된다. 우선 하나의 reuters-21578의 학습문서를 문장들로 나눈다. 그 다음 문장내의 단어들을 추출하고, 추출된 단어에서 의미는 같지만 형태가 다른 Stop-word[6]는 제외한다. 이런 단어는 다른 단어로 구분하기보다 단어의 원형으로 통일하여 하나의 단어로 구분하는 것이 더욱 효율적이며 정확할 것이다.



(그림 3) 개선된 DRFP-Tree 생성법

이후 문서들을 각 카테고리 DRFP-Tree에 할당한다. 이때 방대한 e-Book의 량과 각각의 문장들까지 고려한다면 처음부터 'Header Table'을 만드는 것은 상당한 비용이 발생한다. 따라서 첫 학습 문서에서 나타난 단어들을 빈도순으로 정렬하여 이를 'Header Table'로 만들고 그 순서에 맞춰 각 문장의 단어들을 정렬하여 DRFP-Tree를 생성한다. 이후 각 학습 문서에서 'Header Table'과 비교하여 이미 있는 단어일 경우 그 빈도수를 더하여 주고, 새로운 단어이면 새로운 단어들의 빈도순으로 기존 'Header Table'의 제일 밑에 덧 붙여 준다. 이 후 이 순서로 각 문장의 단어들을 정렬하여 기존의 DRFP-Tree에 덧붙여 주고 이때, 학습문서 내에서 5%이상 나타난 단어만을 사용하여 정확도를 높인다. 이와 같은 방법으로 첫 학습문서에서의 문장이나, 나중에 쓰인 학습문서의 문장이나 단어들의 순서는 일관성을 유지하게 되고 결과적으로 문장의 패턴을 유지한 체 모든 학습문서들을 검색하여 'Header Table'을 생성하는 과정이 생략되어 속도 향상이 가능하다.

전처리 과정에서 얻어진 DRFP-Tree와 분류할 e-Book의 유사도를 구하려면 문서의 각 문장의 패턴이 DRFP-Tree에 얼마나 자주 나타나는지와 얼마나 유사한 패턴인지를 비교하면 된다. 이렇게 얻어진 유사도를 이용하여 해당 e-Book을 DDC(듀이 십진 분류법), LCC(미의회도서관 분류법), KDC(한국십진분류법)와 같이 도서 분류기준에 의거하여 분류한다.

4. 향후 연구 방향

Google북, Amazon의 Kindle, 여기에 도전장을 내민 애플의 iBook과 같이 다양한 플랫폼을 바탕으로 방대한 분량의 e-Book 시장이 형성되어 가고 있다. 연구과정에서 txt와 같이 범용적인 포맷의 e-Book을 지원하는 경우가 있는 반면 각 플랫폼에 적용되는 특정 포맷(e-pub, pdf, yin, gbs, ebk등)만을 지원하는 경우가 대부분이었다. 이는 e-Book의 라이센스 문제와 표준화 되어 있지 않은 포맷을 각 제조업의 편의에 따라 개발한다는 점을 생각해보면 당연한 결과로 받아들일 수 있다. 이러한 상황에서 제안된 e-Book분류 방법을 적용시키기 위해서 포맷들의 표준안 마련에 대한 연구가 계속되어야 할 것이다.[7] 또한 본 논문에서 사용한 reuters-21578 학습문서가 영어로 작성된 점을 고려 할 때 다양한 언어로 출판되는 e-Book의 언어 문제는 많은 연구가 필요 할 것이다.

마지막으로 전처리 후 얻어진 DRFP-Tree와 분류 해야 할 e-Book의 유사도를 비교할 수 있는 객관적 지표에 대한 연구에 대한 부분은 다음 과제로 남겨둔다.

감사의 글

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0075771).

참고문헌

[1] 유선실, 구글의 e-Book 시장 진출과 세계 e-Book 시장 경쟁 현황, 정보통신정책연구원 제22권 9호 통권 485호

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rule," Proc. Int'l Conf. Very Large Data Bases, pp. 487-499, Sept. 1994

[3] Jiawei Han, Jian pet, Yiwen Yin Runying Mao, Mining Frequent Patterns without Candidate Generation, Data Mining and Knowledge Discovery 2004.

[4] Muhaimenul Adnan and reda Alhajj, "DRFP-tree:disk-resident frequent pattern tree", Applied intelligence, 2007

[5] D.D.Lewis, An evaluation of phrasal and clustered representations on a text categorization task, In Proceedings of SIGIR-92, pages 37-50, 1992

[6] Gerard Salton, Chris Buckley, 571 stopword list for the experimental SMART information retrieval system at cornell University http://www.lextek.com/manuals/onix/stopwords2.html

[7] Strabase, e-Book 시장 활성화의 관건 '표준 포맷', 그 현황과 단일 표준 제정의 전망, 디지털 미래와 전략 vol. 35 (2008년 12월) pp.53-58 1976-2607 2008