예측 FP-tree를 이용한 어종별 어장 기법

정희연*, 조경수*, 김응모*
*성균관대학교 정보통신공학부
e-mail : loveforu@skku.ac.kr

Usage of FP-tree for forecasting technique of the fishery

Hui-Yen Jeong*, Kyungsoo Cho*, Ung-Mo Kim*
*School of Information and Communication Engineering,
Sungkyunkwan University

요 약

정보화 사회로의 진입이 본격화 되면서 사회의 전반적인 분야에 걸쳐 다양한 용도로 컴퓨터 시스템이 사용되고 있다. 그에 따라 데이터의 방대한 양적 팽창이 이루어졌고, 이러한 데이터를 유용한 정보와 지식으로 바꿔야 하는 필요성들이 생겨났다. 이에 데이터 마이닝이라는 개념이 등장했고 현재 점점 더 많은 분야에서 사용되고 있고 다양한 각도에서 활발한 연구가 진행되고 있다. 현재 어장의 예측 방법은 주관적인 경험에 대부분 의존하고 객관적인 신뢰성이 떨어진다. 이에 본 논문은 데이터 마이닝 기법을 적용하여 데이터베이스의 정보를 이용해 어종별로 가장 빈번하게 이용되어지는 어장을 선별해 주는 기법을 제안한다.

1. 서론

정보화 사회로의 진입이 본격화 되면서 사회의 전반적인 분야에 걸쳐 다양한 용도로 컴퓨터 시스템이 사용되고 있다. 이런 정보화 사회로의 진입의 정도가 발전됨에 따라 그에 따르는 데이터의 방대한 양적 팽창이 이루어졌고, 이러한 데이터를 유용한 정보와 지식으로 바꿔야 하는 필요성들이 생겼다.

기업경영, 생산관리, 시장 분석에서부터 공학설계와 과학탐구에 이르기까지 광범위한 응용분야에 이러한 정보와 지식들이 사용되기 때문에 데이터 마이닝이라는 개념이 등장하게 되었다.[1] 데이터 마이닝 기법은 현재 다양한 각도에서 활발히 연구가 계속 되고 있으며, 사회의 여러 분야에 적용되고 있다. 일례로, 스마트폰의 맛집 검색에서는 스마트폰 사용자들의 맛집 관련 리뷰나 평가 등을 이용해 정보를 축적, 분류, 분석하여 사용자가 얻고자 하는 정보를 효율적으로 분석해준다.

본 논문은 어업 종사자들에게 어종별로 어장을 예측할 수 있는 시스템에 마이닝 기법을 적용함으로써 실제 어업 종사자들에게 어획량 증가에 도움이 되고자 한다. 현재 어장 예측은 그날 일기예보 등을 보고 어업 종사자들의 경험을 기반으로 어종별 어장을 예측하는 것이 대부분이다. 이러한 정보는 개인적인 경험에 의거된 자료이기 때문에, 그 신빙성이 높지 않고, 환경적인 변수에 의해 변화가 생길 경우 다른 어장을 예측하기 힘들다는 단점을 가지고 있다. 하지만, 마이닝 기법을 이용해서 얻어진 정보는 특정 어종에 따라 어장 분포 패턴을 제공해 주기 때문에 객관적으로 신뢰도가 높은 예측을 할 수 있다. 이에 제안하고자 하는 아이디어는 다음과 같다. 우선 각 어종별로 구별된 데이터베이스를 구축한 후 마이닝 기법을 이용하여, 좌표를 이용한 어종별 어장 분포 정보를 축적한다. 그리고 이처럼 축적된 정보들을 이용하여 패턴 별로 추출해 낸다. 이후 사용자가 어종별 어장 예측에 관한 정보를 얻고자 할 경우 이전에 자주 이용되었던 패턴들을 제공해 줌으로써 효율적으로 어장을 찾아 준다.

본 논문은 다음과 같이 구성된다. 2장에서는 기본적인 마이닝 기법에 대해서 설명한다. 3장에서는 본 논문에서 제안하는 마이닝 기법을 이용한 어종별 어장 예측 기법에 대해 설명한다. 4장에서는 전체적인 고찰 및 발전된 연구를 위한 향후 연구 과제를 제시함으로써 결론을 맺는다.

2. 관련 연구

이 절에서는 어종별 어장 예측 기법에 사용될 수

있는 마이닝 기법 중 가장 잘 알려져 있는 연관 규칙 마이닝 기법과 빈발 패턴 증가 기법에 관하여 설명한다.

2.1 연관 규칙 마이닝 기법

연관규칙 마이닝 기법[2,3]은 주어진 데이터 집합에서 연관성이 있는 항목을 찾아내는 기법이다.

I={$I_1, I_2, ..., I_m$}을 항목들의 집합이라고 하자. 각각의 트랜잭션 T는 T⊆I인 항목들의 집합이고, D는 작업 관련 데이터로서 데이터베이스 트랜잭션의 집합이라고 하자. 각 트랜잭션은 고유한 트랜잭션 번호(TID)를 갖는다. A를 항목들의 집합이라고 하면, 트랜잭션 T가 A⊆T를 만족하는 경우에만 항목 A가 트랜잭션 T에 포함된다고 한다. 이 때, A⊂I, A⊂I, A∩B=∅을 만족하는 경우 연관규칙은 A⇒B의 형식으로 표현된다.

규칙 A⇒B는 트랜잭션 집합 D에서 A, B 두 집합을 동시에 포함하고 있는 트랜잭션의 백분율이 s일 경우 지지도 s를 갖는다고 한다. 지지도는 확률 P(A∪B)를 계산해서 얻을 수 있다. 예를 들어서, 연관규칙에서 지지도 3%가 의미하는 바는 백화점의 전체 구매 고객(D)의 3%가 청바지(A)와 티셔츠(B)를 동시에 구입한다는 것을 의미한다. 백화점 측에서는 이를 이용해 층별 매장 관리, 연관된 상품 배치, 패키지 상품 등의 소비자의 구매를 부추길 수 있는 다양한 전략을 세울 수 있다.

다음으로, 집합 A를 포함하는 트랜잭션 중에서 집합 B도 포함하고 있는 트랜잭션의 백분율이 c일 경우, 규칙 A⇒B가 신뢰도 c를 갖는다고 한다. 신뢰도는 조건부확률 P(B|A)를 계산해서 얻을 수 있다. 예를 들면, 70%의 신뢰도를 가지고 있다는 것이 의미하는 바는 축구공을 구입한 고객의 70%가 축구화도 구입한다는 것을 의미한다.

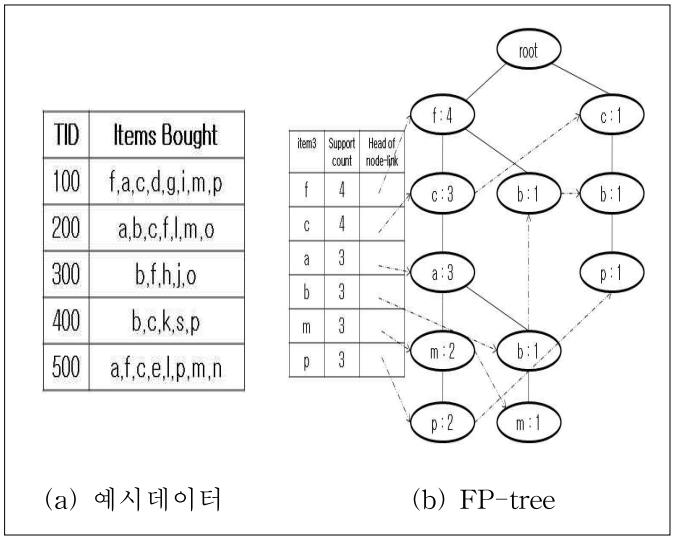
일반적으로 데이터베이스에서 연관 규칙 마이닝은 사용자나 전문가가 정해놓은 최소 지지도 임계값과 최소 신뢰도 임계값을 만족하는 모든 규칙을 찾는 것이다.

2.2 FP-Growth(Frequent Pattern Growth : 빈발 패턴 성장)

FP-growth 기법[4,5]은 FP-tree라 불리는 밀집 자료 구조를 사용하여 빈발 항목을 가지는 데이터베이스를 FP-tree로 압축한다. 이 때 항목들 간 연관 정보는 전혀 손실이 없으며, 그런 후에 압축된 데이터베이스를 하나의 빈발 항목에 대하여 연관된 조건 데이터베이스의 집합으로 분할하고 이와 같이 분할된 각각의 데이터베이스에 대해서 개별적으로 마이닝해서 결과를 얻어내게 된다. 이 기법의 구체적인 절차는 다음과 같다.

우선, 데이터베이스를 스캔하여 최소 지지도를 기준으로 하고 있는 빈발항목집합을 내림차순으로 정렬한다. 그 후 "null"로 표시된 트리의 루트를 생성하고 난 다음, 데이터베이스를 스캔해서 각 트랜잭션의 정렬된 항목들 순서대로 가지로 생성한다. 이 때, 이미 생성된 경로와 공통되는 접두부를 가지고 있을 경우, 공통 접두부에 속하는 각 노드의 카운트를 1만큼씩 증가시키고 접두부 다음에 나타나는 항목에 대해서 새로운 노드를 생성한다. 트리 탐색을 위해 항목 헤더 테이블을 만든 후에 각 항목을 링크 체인으로 FP-tree의 모든 노드들과 연결한다. 모든 트랜잭션을 스캔하고 난 후에, 빈발 패턴을 마이닝하기 위해서 FP-tree를 마이닝하면 된다.

[그림 1]의 (a)는 예시 데이터이고, [그림 1]의 (b)는 예시 데이터의 FP-tree이다.



[그림 1] FP-tree 알고리즘 예시데이터

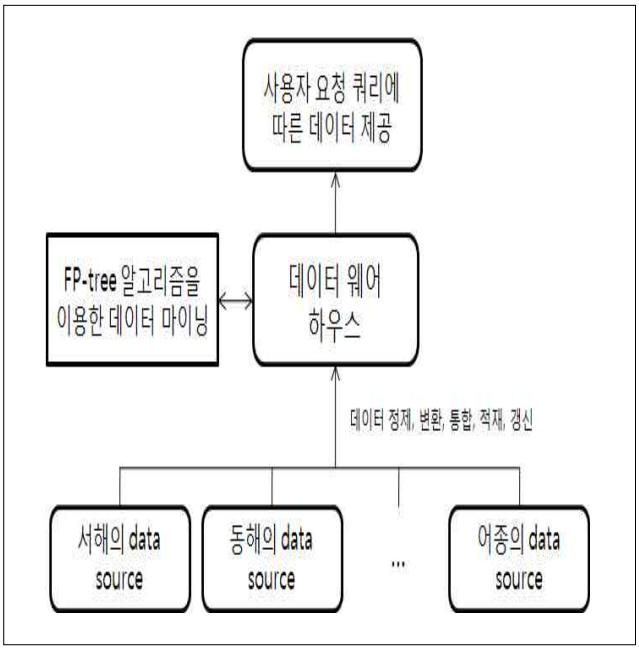
FP-tree 마이닝은 길이가 1인 빈발패턴에서 시작해서, 조건부 패턴 베이스를 생성하고, 이를 통해 조건부 FP-tree를 생성한다. 이 과정을 재귀적으로 수행하여 마이닝 해주고, 이를 통해서 빈발패턴을 얻어낸다.

3. 데이터 마이닝을 이용한 어종별 어장 예측

이 절에서는 첫 번째로, 제안되었던 어종별 어장 예측 시스템의 전체적인 구성과 전반적인 흐름에 대해서 설명하고 다음으로 실제 데이터의 예시를 통해 FP-tree를 통해 제안된 시스템의 마이닝을 소개와 사용자가 얻을 수 있는 빈발패턴에 대해서 설명한다.

3.1 제안된 어종별 어장 예측 시스템

데이터베이스에 현재까지 많은 양의 데이터가 저장되어 있다고 하자. 시스템 사용자는 자기가 잡고자 하는 어종을 입력하고, 이에 맞춰 FP-tree 알고리즘을 이용하여 이전의 사용자들이 가장 많이 이용했던 어장 정보를 얻을 수 있다. 얻은 정보를 이용하여 사용자는 객관성 있고 신뢰성 높은 어장을 예측할 수 있다. 제안하는 시스템은 다음과 같다.



[그림 2] 전체적인 시스템 흐름도

어장 예측 시스템은 FP-tree 알고리즘을 이용한 데이터 마이닝 과정을 통해 사용자가 얻고자 하는 쿼리에 대한 데이터를 제공해준다. 그리고 사용자에 의해 얻어진 데이터는 데이터베이스에 저장된 후, 정제 및 변환되어 이전 데이터들과 통합되고 데이터 웨어하우스에 적재, 갱신된다. 이 데이터들은 FP-tree 알고리즘을 이용해서 마이닝 되며 새로운 빈발 패턴들을 만들어 낸다. 그리고 이 새로운 빈발 패턴들은 다시 사용자의 쿼리에 대한 갱신된 데이터를 제공해주게 된다.

3.2 FP-tree를 이용한 제안된 시스템의 마이닝

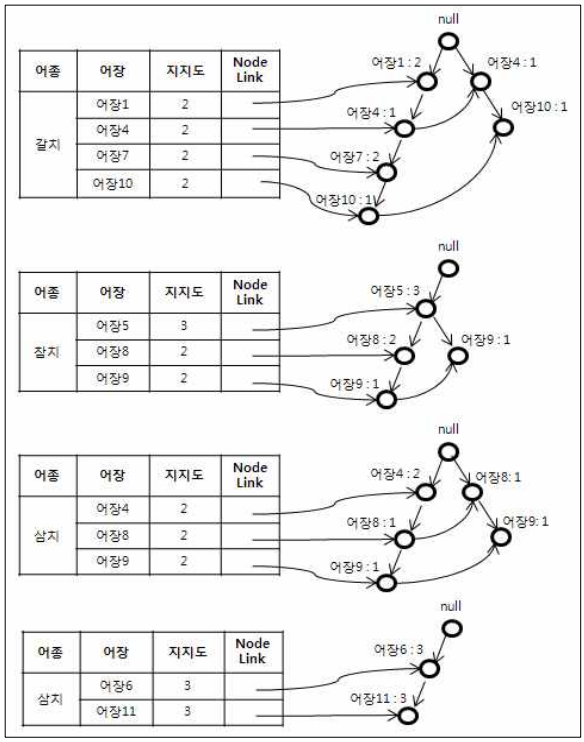
본 논문에서는 [표1]을 기반으로 FP-tree를 이용한 신뢰성 있는 어장 패턴 정보를 추출한다.

[표 1] 어종별 어장 정보를 담은 데이터

사용자 ID	어종	어장
100	갈치	어장1, 어장4, 어장7, 어장10
	참치	어장2, 어장5, 어장8, 어장9
200	삼치	어장3, 어장4, 어장7
	새우	어장2, 어장6, 어장11
300	참치	어장5, 어장6, 어장8
	새우	어장5, 어장6, 어장11
400	삼치	어장4, 어장8, 어장9
	갈치	어장1, 어장7, 어장8
500	참치	어장5, 어장9
	갈치	어장4, 어장5, 어장6, 어장10
600	새우	어장6, 어장11
	삼치	어장1, 어장8, 어장9

[표1]은 실제 마이닝에 사용되는 데이터 웨어 하우스에 저장되어 있는 데이터이다. 이 데이터에 FP-tree 알고리즘을 적용시킨다. 우선, 빈발항목들을 찾아내기 위해 각 항목들의 카운트를 확인한다. 이 때, 최소 지지도를 2로 한다. 그리고 이를 이용하여 FP-tree를 만든다.

다음의 [그림 3]은 그렇게 만들어진 FP-tree이다.



[그림 3] 전체적인 시스템 흐름도

위의 FP-tree를 이용하여 사용자는 지지도 2 이상의 객관성 있는 선별된 어종별 어장 데이터를 얻을 수

있다.

4. 결론 및 향후 연구과제

본 논문에서는 데이터 마이닝 기법을 이용하여 어종별 어장 예측을 할 수 있도록 했다. 본 논문에서 사용한 FP-tree 알고리즘은 어떠한 긴 트랜잭션도 끊어지지 않는 완전성과 원본 DB보다 커지지 않고, 부적절한 정보를 제거해주는 밀집성이라는 장점을 가지고 있다. 그러나 점진적인 데이터의 증가에도 기존 데이터의 재처리과정을 거친다는 단점을 가지고 있다. 본 논문의 향후 연구로는 어장 예측 시스템의 신뢰성을 더 높여줄 수 있는 다양한 종류의 패턴 분류 연구가 필요하고, 새로운 데이터 마이닝 기법을 통해 FP-tree 알고리즘의 단점을 보완해주는 방향으로 연구가 진행되어야 할 것이다.

감사의 글

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

참고문헌

[1] J. Han, M. Kamber "Data Mining : Concepts and Techniques" Acadamin Press 2000

[2] R. Agrawal, T. Imielinski and A. Swami "Mining association rules between sets of items in large database" In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data(SOGMOD'98), page 207-216, Washington, Dc, May 1993

[3] M.H. Dunham , Y. Xio, L.Greenwald, Z.Hossain "ASurvey of association rules" Dallas, Texas

[4] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. ACM-SIGMOD Int'l Conf. Management of Data, pp 1-12, May 2000

[5] J. Han, J. Pei, Y. Yin and R. Mao, "Mining Frequent Patterns without Candidate Generation : A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, vol. 8, no.1, pp. 53-87, 2004