Expert Systems with Applications 41 (2014) 1928–1936

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Extending market basket analysis with graph mining techniques: A real case

Ivan F. Videla-Cavieres [*], Sebastián A. Ríos

University of Chile, Department of Industrial Engineering, Business Intelligence Research Center (CEINE), Santiago, Chile

ARTICLE INFO

Keywords:
Market basket analysis
Graph mining
Retail
Product network
Big data
Overlap communities

ABSTRACT

A common problem for many companies, like retail stores, it is to find sets of products that are sold together. The only source of information available is the history of sales transactional data. Common techniques of market basket analysis fail when processing huge amounts of scattered data, finding meaningless relationships. We developed a novel approach for market basket analysis based on graph mining techniques, able to process millions of scattered transactions. We demonstrate the effectiveness of our approach in a wholesale supermarket chain and a retail supermarket chain, processing around 238,000,000 and 128,000,000 transactions respectively compared to classical approach.

1. Introduction

Over decades retail chains and department stores have been selling their products without using the transactional data generated by their sales as a source of knowledge. Recently – in the last two decades – companies started to use this data to discover information. In the 90's limited computational capabilities made the extraction of knowledge from millions of daily transactions unfeasible, and only analysis with simple models and reduced datasets were possible. In 1993, Agrawal (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994) showed that many organizations were getting bigger databases with transactional data, consumer data, sales records, etc. Therefore, they proposed the Apriori algorithm (Agrawal & Srikant, 1994) for a large data set for those years.

Today, computational systems have evolved – both hardware and software – and have been implemented in all areas of companies (CRMs, ERPs, MRPs, Data Marts, Data Warehouses, ad hoc systems, etc.), allowing the storage and processing of huge amounts of data. Similarly, it is possible to develop complex models and algorithms to gather knowledge from such huge databases.

A classical approach to getting information from data in retail and department stores is through market basket analysis (MBA), frequent item set discovery and clustering techniques such as K-means (Hartigan & Wong, 1979), SOM (Kohonen, 1990). The main idea behind this is to discover purchasing patterns from transactional data. However, when we used these techniques to process real supermarket chain data, the results obtained were of very poor quality. For example, with K-means techniques only one cluster grouped 93% of transactions and the 7% remaining is not meaningful. Therefore, poor quality information was generated disabling decisions such as finding customers profiles, discount offers generation, supermarket products layout, etc. Thus, we developed a novel approach to perform MBA based on graph mining techniques; specifically using overlap communities, that allows to generate highly related products to each other within the community. We benchmarked our method using several traditional approaches applied over millions of transactional data. The results of our evaluation show that our approach out–performs the traditional methods.

2. Definitions and related work

This work is focused on generating frequent item sets of products based on transactional data generated by a retail chain. The main idea is to obtain sets of meaningful products so we can generate customer profiles, product layout and recommendations from related products.

In the following sections we will explain the datasets over which we apply our methods; the classical approach and the state-of-art techniques based on graph mining over transactional data.

2.1. Data

We have data from two retail chains in Chile. One is a wholesale supermarket oriented to supply products to grocery store owners, hereafter, referred to as *Retail A*. The second is member of one of the biggest retail holdings in Chile called *Retail B*.

[*] Corresponding author.
E-mail addresses: ividela@dcc.uchile.cl (I.F. Videla-Cavieres), srios@dii.uchile.cl (S.A. Ríos).
URLs: http://www.ceine.cl (I.F. Videla-Cavieres), http://www.ceine.cl (S.A. Ríos).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.088

Our data was gathered within a period of thirty months, around 238 million transactions, approximately 160 thousand clients and over 11 thousand SKUs[1] in the case of *Retail A* chain. For *Retail B*, the gathered period was two months, with 128 million transactions, almost 2 million customers and 31 thousand different SKU.

2.2. Transactional data

We have a set of products and transactions. Products are defined formally as $P = \{p_1, p_2, \ldots, p_n\}$ where each p_i represents a specific SKU available. Indeed $|P| = number\ of\ distinct\ SKUs$. A transaction T is defined according to (Agrawal & Srikant, 1994) as a set of items (products in this case) purchased in the same buying opportunity, such that $T \subseteq P$.

In our datasets, products are organized in a three hierarchical level structure. Each level belongs to its predecessor based on an ad–hoc developed taxonomy by each retailer. Fig. 1 shows a subset of one of our taxonomy and Table 1 shows an example of product information with its hierarchy. *Retail A* has 23 product families, 150 lines of products and 415 sublines of products. *Retail B* has 50 product families, 287 lines and 1032 sublines of products.

This big amount of data are stored in a column oriented database because a classical relational database has a very low performance and the response time for every query took several hours or days, which is not acceptable.

Each transaction is identified by a unique number. An example of a transaction set is shown in Table 2 where we see that 925 is a transaction composed of three products: P1, P2 and P4. These products were bought by customer 10021 on the date May 7th, 2009. Suppose SKU of P1 is 13231. On Table 1, that would mean that the product is a Milk named "The Happy Cow" which belongs to Dairy Family, to Yogurt & Milk Line and to Liquid Milk Sub-line. On the other hand, transaction 926 has a *customer id* equal to -1 which means that retail does not have that customer registered or that the customer does not want to give their identifier.

Table 2 presents the set of data available and how that information is stored. Another way to store that information is by the one expressed in Table 3 which is a matrix whose rows are vectors of purchases. Each vector is composed by transactions and the set of products available. The first column stored the transactional id and in the following columns stored a number 1 or 0 which represents whether the product was purchased or not in that particular transaction.

2.3. Market basket analysis

This is one of the most applied techniques over transactional data. It is part of the vast family of Data Mining Techniques. The purpose of market basket analysis is to get a customer to spend more money based on two different principles: the first one is *Up-Selling*, which consists in buying a large quantity of the same product, or adding new features or warranties. The second way is *Cross-Selling*, which consists in adding more products from different categories.

The main purpose to discover frequent item sets. Also known as the discovery of if-then rules called *Association rules* (Agrawal et al., 1993; Agrawal & Srikant, 1994). The form of an association rule is $I \rightarrow j$ where I is a set of items (products) and j is a particular item. The process consist of finding sets of products (items) presents in a large number of transactions (basket).

2.4. Frequent item sets

Frequent item sets are formally defined, according to (Rajaraman, 2012), as follows:

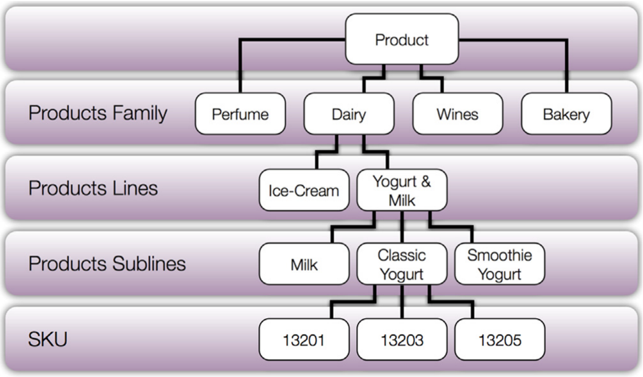

Fig. 1. Hierarchy of products.

Table 1
Products characterization available.

SKU	Product name	Product family	Product line	Product sub-line
13231	Milk "The Happy Cow"	Dairy	Yogurt & Milk	Milk
13201	Yogurt "Fancy Yogurt"	Dairy	Yogurt & Milk	Classic Yogurt
13245	Yogurt "Smoothiest"	Dairy	Yogurt & Milk	Smoothie Yogurt

Table 2
Example of a transaction set.

Transaction ID	Date	SKU	Customer ID	Quantity	Price	Total Price
925	05-07-2009	P1	10021	1	350	350
925	05-07-2009	P2	10021	3	500	1500
925	05-07-2009	P4	10021	2	500	1000
926	05-07-2009	P3	−1	4	600	2400
926	05-07-2009	P4	−1	9	500	4500
927	05-07-2009	P1	1308	4	350	1400
927	05-07-2009	P3	1308	7	600	4200

Table 3
Example of a transaction set as a vector of purchase.

Transaction ID	P1	P2	P3	P4
925	1	1	0	1
926	1	0	1	1
927	1	0	1	0

Let I be a set of items. Define *support* s as the number of transactions for which I is a subset. We will say I is frequent if its support s is bigger than a certain s' called support threshold.

Another important definition related to *association rules* is the *confidence* of a rule $I \rightarrow j$ which is defined as $\frac{support(I \cup j)}{support(I)}$. (In other words the fraction of the baskets with all of I that also contain j). Confidence can be interpreted as the probability of finding the right–hand–side of the rule (in this case j) under the condition that these transactions also contain the left–hand–side of the rule (in this case I).

We performed an experiment using this technique and found very poor results, obtaining a lot of meaningless rules or rules that apply only to a certain group of customers. For example, one of the rules found in the data of *Retail A* is *coke* → *rum*, with a high support and confidence despite the small values obtained in general (less than 0.15% of the transactions). This rule can be seen as a very good rule, but it is an expected rule because in Chile, a common drink named *ron-cola* is made from a base of mixed rum and coke.

[1] SKU: Stock Keeping Unit.