Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 730 – 738

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Generating groups of products using graph mining techniques

Sebastián A. Ríos[a,b,*], Ivan F. Videla–Cavieres[b]

[a]Department of Industrial Engineering, Universidad de Chile, Av. República 701, Santiago, Chile
[b]Business Intelligence Research Center (CEINE), University of Chile, Av. Domeyko 2369, Santiago, Chile

Abstract

Retail industry has evolved. Nowadays, companies around the world need a better and deeper understanding of their customers. In order to enhance store layout, generate customers groups, offers and personalized recommendations, among others. To accomplish these objectives, it is very important to know which products are related to each other.

Classical approaches for clustering products, such as K-means or SOFM, do not work when exist scattered and large amounts of data. Even association rules give results that are difficult to interpret. These facts motivate us to use a novel approach that generates communities of products. One of the main advantages of these communities is that are meaningful and easily interpretable by retail analysts. This approach allows the processing of billions of transaction records within a reasonable time, according to the needs of companies.

Peer-review under responsibility of KES International.

Keywords: Market Basket Analysis; Overlap Community Detection; Big Data; Graph Mining; Transactional Data;

1. Introduction

Large amounts of data are stored by companies for various reasons, such as increasing retention rates thanks to Customer Relationship Management (CRM) [14,19]; for recommender systems [8]; for understanding of consumer behavior and generating customer profiles [2] from transactional and factual information. In fact, [4], explains that profiling customers usually generates a description of the behavior and often, will lead us to a good explanation for it.

The retail industry generates huge amounts of transactional information. Indeed, this work is focused generating valuable customer information to complement existing information.

We will generate groups of products based on the novel extension of market basket analysis proposed by Videla–Cavieres and Ríos [18] and then, we studied the stability of the communities found over different time windows. Videla–Cavieres and Ríos [18] introduced the concept of overlapped communities of products, allowing a product to belong to more than one community. This approach has proved to be useful to retail analysts, generating groups of products purchased together that are easily interpreted.

* Corresponding author. Tel.: +56 2 2978 0545
E-mail address: srios@dii.uchile.cl ; ividela@dcc.uchile.cl

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.155

2. Definitions and related work

One of the uses classical of the graph mining techniques is in Social Network Analysis (SNA). We can find some examples in [21,23,22] where the main purpose it is to understand the underlying structure and content inside it..

Videla–Cavieres and Ríos [18] work is focused on generating frequent item sets of products based on transactional data generated by a retail chain. Their main idea is to obtain sets of meaningful products that are interpretable by analysts. Now, we will generate groups of products following their approach and then we will study the stability of the communities over time.

In the following sections we will explain the datasets over which we apply our method; the approach proposed by Videla–Cavieres and Ríos [18] in detail and how we finally studied the stability of communities of products.

2.1. Data

We have transactional records from a retail chain in Chile. In terms of volume, we have around half billion records gathered within a period of twenty months, approximately $2,200,000$ customers and over $42,000$ SKUs[1] globally.

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 an available specific SKU available. Indeed $|P|$ = *number of distinct SKUs*. A transaction T is defined according to Agrawal and Srikant [3] 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 three hierarchical level structures. Each level belongs to its predecessor based on an ad–hoc developed taxonomy by the retailer. Figure 1 shows a subset of one of our taxonomy and table 1 shows an example of product information with its hierarchy. In total we have 73 product families, 487 lines of products and $1,447$ sublines 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

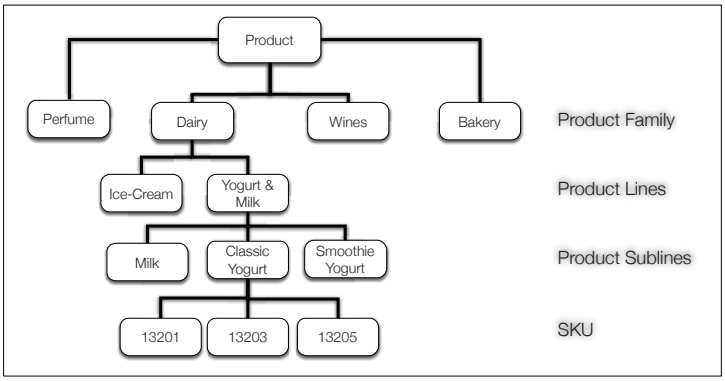

Fig. 1. Hierarchy of products.

[1] SKU : Stock Keeping Unit