Mining Large-Scale Retail Data

Vincent Leroy
Univ. Grenoble Alpes – LIG – CNRS

1 Introduction

Understanding customer buying patterns is of great interest in the retail industry. Applications include targeted advertising, optimized product placement, and cross-promotions. Association rules, expressed as $A \rightarrow B$ (**if** A **then** B) are a common and easily understandable ways to represent buying patterns. While the problem of mining such rules has received considerable attention over the past years, most of the approaches proposed have only be evaluated on relatively small datasets, and struggle at large scale.

In the context of the Datalyse project [1], Intermarché, our industrial partner, has given us access to 2 years of sales data: 3.5B sales records, 300M tickets, 9M customers, 200k products. This constitutes an opportunity to re-visit the problem association rules mining in the context of "big data". In the remainder of this paper, I will first give an overview of our work on designing mining algorithms adapted to long-tailed datasets. Then, I will describe our evaluation of quality measures for ranking association rules. Finally, I will present the systems architecture deployed to apply mining in production at Intermarché.

Collaborators This work was done in collaboration with members of the SLIDE research group Sihem Amer-Yahia, Martin Kirchgessner and Shashwat Mishra, as well as our partners from Intermarché STIME group.

2 Mining

Over the past 20 years, pattern mining has been applied successfully on a variety of datasets to uncover hidden associations. The use of these algorithms was popularized by the famous "beer and diapers" association observed in a supermarket [1]. With the availability of much larger datasets and powerful computers, one would expect to find new surprising and entertaining results. On the contrary, traditional frequent itemsets mining (FIM) algorithms today fail to output any itemset for the majority of products, and instead produce billions of itemsets combining the very few (typically less than 5%) most frequent products. We face a paradox: the more data is available, the less interesting the results are.

This issues comes from the characteristics of the data, as well as the core definition of FIM. Large datasets often exhibit a long-tail distribution. Few items are extremely popular (head), and the vas majority have a low frequency (tail). FIM focuses on mining the most frequent itemsets, and has an execution time linear in the number of results. Thus, to output results involving tail items, FIM algorithms first have to be able to compute the frequent itemsets from the head. As this set grows exponentially with the number of items involved, a larger dataset means in practice a smaller proportion of items present in the results.

We refer to the results of traditional FIM as narrow and deep. Few items are represented, but millions of itemsets are returned for each of them. Conversely, shallow and wide mining would ensure that each item is represented in the results, and involved in a few of itemsets. We developed TopPI, a new mining algorithm designed to achieve this goal. Rather than exhaustively extracting the most frequent itemsets, TopPI ensure that all items are covered by the results. TopPI follows a top-k approach, and computes for each item in the dataset the k most frequent itemsets that contain it. An analyst may request $k = 10$ results per item, to obtain an overview of the dataset over a wide range of products, while $k = 1{,}000$ may be used for automatic analysis or post-processing. We showed experimentally that TopPI scales to millions of items and transactions by focusing the mining efforts to a limited but valuable set of results and avoiding redundancy.

3 Ranking

By adopting a top-k approach and ensuring items coverage, TopPI, our mining algorithm, extracts interesting results in a reasonable time. Nevertheless, browsing hundreds of itemsets is cumbersome, and leads data analysts missing important results. Itemsets are generally presented as association rules $A \rightarrow B$, where A is the antecedent and B the consequent. The value of an association rules is expressed as a combination of *recall* and *precision*, which reflect how often a rule can be

[1] http://www.datalyse.fr/

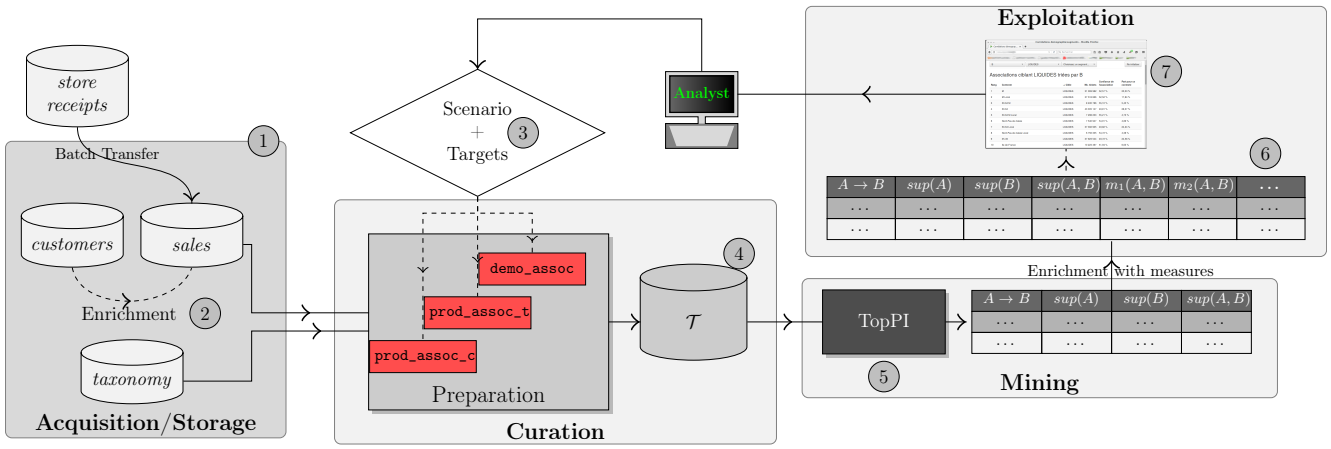

Figure 1: Overview of the architecture

used and how reliable it is respectively. Over 30 different interestingness measures [2], expressing different recall/precision trade-offs, were proposed to assign scores to association rules and rank them. Unfortunately, very little effort was done to assess which of these measures actually coincide with the opinion of the analysts.

We developed a two-stage approach to find the most appropriate ranking function in the context of the retail industry. We first performed an exhaustive comparative analysis by computing the ranking correlation of these measures over multiple lists of rules. We relied on different correlation functions, to study both the overall ranking agreement (Spearman) and the agreement on the first results (Overlap@k). Then, using hierarchical clustering, we identified 6 groups of quality measures. This result showed that, despite differences in their formulation and mathematical properties, many quality measures actually generate extremely similar rankings. Thus, we selected one representative measure for each group and performed a user study with the collaboration of Intermarché experts.

Our evaluation of association rules ranking for retail showed that analysis tend to favor precision over recall. Recall mostly constitutes a threshold constraint, as association rules must be actionnable and impact a sufficiently large fraction of a product's sales. But once recall reaches a sufficient level, analysts always preferred extra precision. This study also allowed us to confirm the suitability of TopPI as a mining algorithm, since it provides this recall threshold by focusing on the k most frequent itemsets of each product, while preserving high precision results.

4 Architecture

We developed an end-to-end data mining pipeline in collaboration with Intermarché. Figure 1 provides an overview of the different modules. We deployed this pipeline on a Hadoop Yarn platform.

The first module is **acquisition and storage**. Sales records are produced locally at each store, and are loaded daily into data center (1). Records are stored in a *sales* table on HBase, and are augmented with customer segments coming from the *customers* table (2). The **curation** module is used to build a transactions dataset. The analysts selects a scenario and a set of input targets (3), which are used to generate the appropriate collection of transactions $\mathcal{T}$ (4). Our system is flexible and allows the analysis of products associations, but also associations of customer segments and product categories, with the support of external knowledge such as product taxonomies. The analyst also selects the scope of her study, by choosing to group either by receipt (single visit to the store), or customer (long term associations). The **mining** component relies on TopPI to compute a set of association rules matching the input targets (5). TopPI is both multi-threaded and distributed using MapReduce to achieve low response times on billions of transactions. The **exploitation** component computes the quality of produced rules according to each interestingness measure (6), and loads them into a database. Results are presented to the analyst through a web application (7).

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proc. SIGMOD '93*, pages 207–216, 1993.

[2] L. Geng and H. J. Hamilton. Interestingness Measures for Data Mining: A Survey. *ACM Comput. Surv.*, 38(3), 2006.