Correlating entities based on DSL specifications of entity group relations

Erik Brännström
Chalmers University of Technology

1 Relevant keywords

data correlation, data mining, mining association rules, dsl

2 Related works

Useful terminology is defined by IBM [2006] in the field of autonomic computing. An autonomic manager is a component that collects data from a system and, based on this data, performs actions with the purpose of improving the system. This control loop is divided into four sub-tasks called monitor (collect system information), analyze (correlate and model data), plan (design behaviour required to reach goal) and execute (run the planned actions), sometimes refered to as MAPE.

An interesting area in data mining is association rule learning. Agrawal et al. [1993] describe how, given a large set of transactions containing any number of different items, rules can be identified between these items. In the case of commerce, such rules would describe how the purchase of one product would, with a probability above a certain threshold, also infer the purchase of another product.

Srikant et al. [1997] expands the above research by applying contraints on the items in the resulting association rules. These constraints can either be expressed using specific items or with taxonomy rules, e.g. if an item is a descendant or ancestor of another item.

Chen et al. [1996] give an overview of the field of data mining where they mention the aspect of multi-level data mining, which states that correlations may not commonly exist on the lowest level of granularity, but instead by forming groups of related items. An example given would be that a specific brand of milk does not necessarily imply the purchase of a specific brand of bread, however purchasing milk of any kind may still be correlated to the purchase of bread irrespective of brand.

Another area in data mining is that of classifiers, and more specifically decision trees. Quinlan [1986] decribe how decision trees can be created from a training set and how well it handles the problems of unknown attributes values and noise in the data.

3 Foundations

The use of item contraints in association rule mining [Srikant et al., 1997] can most likely be used for this project. The relation between target groups and ads can, based on existing metrics, be restructured into a transactional table where each transaction is an impression (in the case of the impression per click metrics) so that each entry would have an attribute stating whether or not the impression yielded a click. The item constraint would then be used to only formulate association rules that lead to clicks.

Considering that the original market basket problem assumes binary attributes and that relational data can contain both quantative and categorical data, [Srikant and Agrawal, 1996] gives a useful description on how such values can be mapped onto the binary (or boolean) problem space.

Hahsler et al. [2007] presents a package for the R software environment used in statistical computing, which implements a base for transaction databases as well as integration with two of the most common mining algorithms, Apriori and Eclat. One of the clever implementation choices is the use of the sparse matrix data structure, which greatly reduces the memory load.

A thorough literature study on domain-specific languages is covered by van Deursen et al. [2000], and the phases of development in the creation of a new DSL is described in detail by Mernik et al. [2005]. The latter provides insight into the relevant choices that need be made in the creation process as well as a useful terminology for discussing common DSL topics.

4 Suggested problems

- Define a DSL for transforming a relational data structure with metrics into a transaction-like table, perhaps also with taxonomies for generalizing values.
- Modify an existing association rule mining algorithm to work directly with non-transaction-like tables.
- Integrate existing transaction metrics to reduce load on rule mining algorithm.

5 Questions

- Which parts of the data should be input simultanously? Including a new campaign in a data set of long-running campaigns will most likely lead to a very low support for any identified rule. The problem with applying the algorithm on only one campaign at a time is that broader rules might not be identified.

6 Project plan

The proposed workflow of this thesis is inspired by the agile practices that perhaps most notably have become common in the software development industry. The most important methods that should be applied to this work are iterative

project development and frequent "customer collaboration", in this case thesis suporvisors from both Chalmers University of Technology and Duego Technologies AB.

The project is expected to run for about 20 weeks. By dividing this relatively long period of time into iterations of 1–3 weeks, the project will be easier both to manage from my point of view as well as provide ample opportunity for the supervisors to have their feedback incorporated.

The initial iteration will most likely require a heavy focus on research and studies of the related fields. The following iterations will all include work on the three areas mentioned in section 8.4 on page 4. Once the research has progressed to an extent that the problem space as well as the approach of the solution are well defined, more effort can be put into the development of the final application that is to be used by Duego.

7 Thesis outline

Todo

8 Original proposal

8.1 Introduction

Duego was founded in 2010 with the idea of creating a social networking site that would target people who wanted to meet new people. This stand in contrast to the existing networks that either manage people who the user already know (e.g. Facebook) or where the user try to find a life partner (i.e. dating sites). Online advertisement is leveraged as a way to promote the site to new users. This is done both using targeted ads that are adapted based on the target demographics on social networks, as well as more conventional ads on websites. Today, managing these campaigns is a manual process that is outsourced to another company. The wish of Duego is to automate this process using software. Such a system would be required to analyze the existing campaigns with regard to ads and target groups along with campaign metrics such as clicks per impression, conversion ratio, etcetera, and use this data to suggest, or even automatically add, new campaign ads.

8.2 Problem

The problem statement given in the background section is described from the view of Duego, however it can be generalized into a much more general problem in the sphere of software engineering.

Firstly, the terms entity, entity group and metrics need to be defined. An entity is something that has distinct properties that enables it to be compared to other entities. As such, entities do not have to be unique, and two separate entities are further defined to be equal if they share the same properties of which the corresponding values are equal.

An entity group is the set of all entities that share one or more defined property values as given by the group definition. In object-oriented terms, an entity group would be a class where as entities are instances of classes (i.e.

objects). Metrics is by definition a specific entity group, however for the sake of clarity I have chosen not to refer to it as such, as it is considered such an important concept in the problem statement.

By analyzing entities that belong to separate entity groups, I wish to identify correlations between unique properties of these with the goal of maximizing the appropriate metrics. Metrics are in turn measured for each combination of two entities that are not of the same type. The expected input to the resulting system are entities and their metrics which should yield an output of new combinations consisting of either already existing entities or suggested new entities.

The most prominent problem to be solved is how to design an algorithm that can efficiently calculate these correlations for large amounts of data. The data set that will be used in production consists of tens of millions of these entities, which means hundreds of millions of properties that should be correlated. Such data sizes can generally not be handled using ordinary brute-force-type algorithms, but requires some form of heuristics to increase efficiency while decreasing accuracy. Another expected requirement is to not have to run all previous data through the system each time, but rather caching results and improving them incrementally.

8.3 Relevance

The requirements discussed in the problem description are closely related to both the field of machine learning as well as that of algorithms and algorithm design, meaning that the final product will most likely have to incorporate strategies from both fields.

8.4 Approach

The general problem can be divided into three subtasks that need to be completed in order to create a system that fulfill the defined requirements.

- Create a domain specific language (DSL) for specifying the desired entity groups, their relations and which properties to optimize
- Design algorithm for optimizing the defined task which can be proven to provide output of high quality
- Extend system with machine learning methods to increase data handling capabilities without causing output quality to drop below reasonable thresholds

Since the general solution to the problem will simultaneously be applied to a specific problem area (described in the background), the definitions of quality are based on customer expectations. As such, it is recommended that the thesis work is performed using an agile approach to maximize the impact of input from supervisors both at Duego as well as Chalmers University of Technology.

8.5 Verification

The result of the thesis work would be verified by applying the results to the problem suggested by Duego in a real-world setting. The suggested advertise-

ment and target groups from the system will be added to Facebook which will allow us to analyze the effectiveness of the software. This can be done continuously over the course of the development in addition to more traditional software unit and system testing.

References

Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining Association Rules between Sets of Items in Large Databases. In *SIGMOD Record*, volume 22, page 207, 1993.

Ming-Syan Chen, Jiawei Han, and Philip S. Yu. Data Mining: An Overview from a Database Perspective. In *IEEE Transactions on Knowledge and Data Engineering*, volume 8, pages 866–883, 1996.

Michael Hahsler, Bettina Grün, and Kurt Hornik. Introduction to arules: Mining association rules and frequent item sets. *SIGKDD Explorations*, 2:4, 2007.

IBM. An architectural blueprint for autonomic computing. *Quality*, 36(June): 34, 2006.

Marjan Mernik, Jan Heering, and Anthony M. Sloane. When and How to Develop Domain-Specific Languages. 37(4):316–344, 2005.

J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986. ISSN 0885-6125.

Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. *SIGMOD Rec.*, 25(2):1–12, 1996.

Ramakrishnan Srikant, Quoc Vu, and Rakesh Agrawal. Mining association rules with item constraints. In *Proceedings of the Third International Conference of Knowledge Discovery and Data Mining*, pages 67–73. AAAI Press, 1997.

Arie van Deursen, Paul Klint, and Joost Visser. Domain-specific languages: An annotated bibliography. 35(6):26–36, 2000.