LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR INFORMATIK
LEHR- UND FORSCHUNGSEINHEIT FÜR
PROGRAMMIER- UND MODELLIERUNGSSPRACHEN

Institut
für
Informatik

Interactive Association Rule Exploration

Daniel Bauersachs

Bachelorarbeit

Beginn der Arbeit: 06.02.2017
Abgabe der Arbeit: 28.04.2017
Betreuer: Prof. Dr. François Bry
Dr. Clemens Schefels

Erklärung

Hiermit versichere ich, dass ich diese Bachelorarbeit selbständig verfasst habe. Ich habe dazu keine anderen als die angegebenen Quellen und Hilfsmittel verwendet.

München, den 28.04.2017 Daniel Bauersachs

Zusammenfassung

Das Erkennen von Zusammenhängen von Datensätzen einer Datenbank ist das sogenannte „Data Mining“. Dieses unterteilt sich in verschiedene Gebiete, von denen im Rahmen dieser Bachelorarbeit die Assoziationsregelgenerierung und besonders die Analyse der erstellten Regeln der Schwerpunkt sein soll. Dafür werden Implikationen aufgestellt, die von Merkmalen eines Datensatzes auf andere Merkmale schließen. Die Implikationen werden Assoziationsregeln genannt.
Das Ziel dieser Arbeit ist, die so erhaltenen Regeln im ARTigo Analytics Center, einer speziell auf dem ARTigo-Projekt aufbauenden Plattform, interaktiv bearbeiten zu können. Dazu werden verschiedene Möglichkeiten bereitgestellt, die auch ohne Hintergrundkenntnisse erlauben, aus den generierten Regeln Erkenntnisse zu ziehen. Nach Vorstellung der diversen Möglichkeiten werden in der Arbeit außerdem die so gewonnenen Daten evaluiert und auf ihre Bedeutung untersucht.
Die hier verwendeten Daten sind dem ARTigo-Projekt entnommen, einer Citizen Science-Plattform, auf der Nutzer sowohl über Stichworte Kunstwerke suchen als auch spielerisch Bildern Beschreibungen zuordnen können. Dies geschieht über sogenannte „Games with a purpose“, also Spielen mit dem Zweck, Informationen zu erhalten, die von Menschen generiert werden.
Um die Assoziationsregeln zu generieren und zu bewerten, wird das quelloffene Datenanalysetool ELKI benutzt, das hierzu bereits verschiedene Bewertungsmaße für Assoziationsregeln bietet. Erweitert um eine Kombination verschiedener solcher Bewertungsmaße zur Stabilisierung der erhaltenen Werte, werden so aus den vorhandenen Daten die Assoziationsregeln erstellt und deren Interessantheit bewertet.

Abstract

Data Mining is the process of finding relations between data records in a database. Divided in different subcategories, the one concentrated on in this work is the subject of generating association rules and in particular the analysis of those. For this purpose implications from some features of a data record on other ones are created and evaluated. These implications are called association rules.

The goal of this work is to permit the interactive modification of the created rules. This is done via the ARTigo Analytics Center, a platform based on the ARTigo project. Different possibilities, which can be used without background knowledge, are supplied to gain insight into the generated rules. After the presentation of the implemented features the gained data is evaluated and their meaning is analysed.

In this work, the used data is taken from the ARTigo project, which is a citizen science-platform enabling users to search artwork by keywords as well as playfully adding tags to the images. This is done by playing "games with a purpose", i.e. games with the purpose to get information generated by human input.

To generate and evaluate the association rules, the open-source tool ELKI is used, which already contains different interestingness measures. Expanded by a combination of different measures to stabilise the generated values, the association rules are created from the existing data and their interestingness is rated.

Danksagung

Ich danke meinem Betreuer Dr. Clemens Schefels, der mir mit viel Ratschlag zur Seite stand und so manchen guten Tipp gegeben hat. Außerdem Prof. Dr. François Bry, der mich durch sein Interesse an meiner Arbeit sehr motiviert und mit seinen Vorschlägen unterstützt hat. Dr. Arthur Zimek und Dr. Erich Schubert haben mir durch ihre Fachkentnisse im Bereich Data Mining sehr geholfen und viele Fragen beantwortet. Danke euch dafür.

Contents

1. Introduction

It was and is a goal of mankind to understand structures and rules in any kind of information. This is often due to the desire to predict patterns and to make some assumptions about the future. A well-known method, the Bayes' theorem [BP63] even dates back to the 18th century. Nowadays, there exists a whole field of research called "Data Mining" or more accurate "Knowledge Discovery in Databases", which comprises a whole lot of methods used for pre-processing input data, finding similarities, outliers and lots more. A well-known application, for example, is the market basket analysis [AIS93] as used by lots of online stores to predict the buying behaviour of customers.
The practical foundation of this work is the ARTigo project [Wie14][1], which is an image tagging project developed by several contributors including the Institute of Informatics and the Institute of Art History of the LMU. It consists of several so-called "Games with a purpose" (GWAP) and search functionality. Essentially, image describing tags are gained while playing the games, which are then added to the already gained set of tags. With an increasing amount of tags per image the overall quality of image description is also rising, which leads to a better description of the images.
As the database already consists of over 60.000 images, it lays a great foundation for experimenting with data mining algorithms. As later shown, there already exist many approaches to find out about underlying structures and patterns in the image database, to which this thesis adds a new one. For this approach I will use the ELKI framework [SKE$^+$15], which is an open source data mining software also used by Frederic Sautter in [Sau16], who implemented the association rule mining as described in the cited bachelor thesis.
The aim of this work is to implement an additional component for the ARTigo Analytics Center [Hoi14]. This component will provide a user-friendly approach of showing and modifying association rules generated from the ARTigo data. As the ARTigo Analytics Center is based on the ARTigo project, it allows to extract the needed data directly from the ARTigo database. Thus, it is not only possible to evaluate the tags gathered per picture, but also to analyse the single game rounds played by the users of the GWAP. This permits to recognise differences and similarities between both of the data sources.
The thesis is structured as follows: Chapter 2 shows related work. In Chapter 3, the theory this work is based on is looked at and in Chapter 4 the existing modules of the ARTigo Analytics Center are outlined.
Chapter 5 covers the implemented module for the ARTigo Analytics Center and its usage. Evaluation and analysis of the results is done in Chapter 6. Finally, in Chapter 7, possibilities for future work are highlighted and a conclusion is given.

[1] `http://www.artigo.org`, accessed 22 March 2017

2. Related Work

As the prediction of connections between features of data records is an important field of Data Science, the generation of association rules is a widely researched topic. The most well-known research is done in the field of market basket analysis, as this is also the original goal of generating association rules. Afterwards, we want to measure the interestingness of the rules. This helps to sort out rules which are not of importance.

2.1. Associations between Data

Generating association rules is originally done by Agrawal et al. [AIS93]. The aim of this paper is to create an algorithm which is capable of generating association rules with given constraints while also being efficient in aspects of memory and calculation time. Furthermore, the algorithm presented allows to sort rules by their confidence, i.e. the relative amount of transactions containing the implying tags that the rule is true for, or their support, i.e. the relative amount of transactions satisfying the rule.
Association rules are mined in very different aspects of research. As Koperski and Han [KH95] show, it is also possible to make assumptions about spatial data. They introduce an algorithm, which enables to find association rules concerning geographic data. This permits to draw conclusions from the relative location of an object to other geographical information, like the distance to rivers.
These approaches allow us to see the validity but not the value of an association rule, as the more trivial rules often are the most confident.

2.2. Rating Association Rules

To be able to measure the informative value of an association rule, the so-called interestingness measures are introduced. A comparison of some of these measures is done for example by Frederic Sautter [Sau16]. In this bachelor thesis, which also allows the proposed interestingness measures to be used by ELKI [SKE+15], a comparison of ten measures is done. Based on the results of this work, a combination of three of the measures is implemented in ELKI. This combination is used as the interestingness measure valuing the rules generated by the ARTigo Analytics Center.
Interestingness measures are also used to classify association rules generated from medical data, as done by Ohsaki et al. [OSYY02]. In this paper, the knowledge of a medical expert is used to value association rules generated from medical data about hepatitis patients. The results of the study show that it is important to let the rules be valued by an expert. As in the case of the hepatitis study a rule was generated and be found as interesting, but was refused to be true by the medical expert. Also the expert claimed the rules to be worth analysing, as the rules often were different from the experience made by the expert. This also is a point to memorise, as some of the rules generated in the process of this work are not as expected as others and still valued more interesting.

3. Theory

In this chapter, the theory of association rules is outlined, as it is necessary to understand some of the basic principles and terminology to concentrate on the features of the interactive rule exploration later on. At first, the origin as well as some basic concepts of association rules are explained. After that, the way from the raw transaction data to the rules rated by interestingness is shown. This process is split into three parts: At first, so-called frequent itemsets are generated. Later on, the association rules are mined from these itemsets. Afterwards, the rules are rated by their interestingness. At last, a new approach is discussed, which is the combination of interestingness measures.

3.1. Association Rules

As mentioned in Section 2, the first approach to generate association rules was done by Agrawal et al. [AIS93] in the early nineties. Their work aimed at making predictions about consumer behaviour. Therefore, the transactions made by the customers were collected and written to a database. The goal was then to create implications of what a customer would buy if he had already bought a specific set of products. As this process, also called market basket analysis, originates in the analysis of the buying behaviour, a lot of the terminology is adopted from the corresponding supermarket expressions.
The mentioned database consists of sets of products bought. The products are referred to as *items*, whereas a set of products is called *transaction*. An association rule is an implication of the form $A \Rightarrow B$, where A and B are *transactions* and have to be disjoint. This is due to the fact, that there would be no sense in implying someone will buy a product A, if he already has bought it before. A is now called *antecedent* and B as known as *consequent*. Notice that there is one restriction to the association rules, as there is no possibility to quantify an occurring *item*.
Now, for example, an association rule could be $\{Bread\} \Rightarrow \{Butter, Marmalade\}$. This would imply that there is a certain probability that if somebody buys bread, he will also buy butter and marmalade. This is useful information for the shopkeeper, as it allows him or her to place butter and marmalade so, that the customer will be held in the shop for longer or is otherwise stimulated to buy a larger amount of products.
The use of association rules is nowadays widespread as for example in analysis of medical data [OSYY02] or even to find out about geographical events and connections as done by Koperski and Han [KH95].

3.2. Frequent Itemset Mining

Before generating the association rules, it is necessary to find only those itemsets with a certain relevance. For this approach we follow some formal definitions made by Charu C. Aggarwal [Agg15]. At first, we assume that the database $\mathcal{T}$, containing a set of n transactions $T_1, \ldots, T_n$, is the database generated from the transactions we want to analyse. As mentioned the transactions are sets of items. These items are then grouped to the universe $\mathcal{U}$.
To find out about the relevance of an itemset we have to compare the number of its oc-

Identifier	**Transaction**
T_1	{Bread, Marmalade}
T_2	{Bread, Butter}
T_3	{Milk, Banana, Marmalade}
T_4	{Butter, Milk}
T_5	{Bread, Butter, Apple}
T_6	{Bread, Butter, Milk}

Table 1: Example consumer transactions. The universe $\mathcal{U}$ of the items is the set $\{Bread, Butter, Milk, Banana, Apple, Marmalade\}$
.

currences in the database $\mathcal{T}$ with the total number of transactions n. The amount of transactions containing an itemset is called its *support*, whereas the amount of transactions containing the itemset divided by n is called its *relative support*. We can now set a value *minsup* as the threshold the support of a frequent itemset has to be greater than or equal.

In the example of the market basket analysis, the universe $\mathcal{U}$ would be the set of all products sold by the shop, whereas the database $\mathcal{T}$ is the database in which all customer transactions are saved. Let Table 1 be a sample database containing some food related transactions. To calculate the support for the itemset $\{Bread, Milk\}$, we have to count the transactions containing the specified itemset. In this case T_1 and T_6 are supersets of the itemset $\{Bread, Milk\}$, so the support of the itemset would be 2. Consider a *minsup* of 2 or a relative *minsup* of $\frac{1}{3}$. In this example, the frequent itemsets would be $\{Bread, Butter\}$, $\{Butter, Milk\}$ and also the singletons $\{Bread\}$, $\{Butter\}$, $\{Marmalade\}$ and $\{Milk\}$.
There are several automated approaches to generate frequent itemsets. They greatly vary in terms of runtime and memory usage. As this bachelor thesis relies on the implementation of combined association rule mining and subsequent interestingness measuring by Frederic Sautter [Sau16] in ELKI [SKE+15], the FPGrowth algorithm is used. This algorithm excels in terms of generating candidates for frequent itemsets and also in the size of the resulting data structure. Thus, the relatively costly process of generating frequent itemsets is optimised.

3.3. Association Rule Mining

As we now have generated the frequent itemsets, the second part is the generation of association rules from these. Again, a lot of candidates will be created, but only those relevant enough are to be considered. To measure the relevance of an association rule, it has to satisfy two criteria [Agg15]. First, it has to occur in a relevant amount of transactions, as this excludes rules only found very sparsely. Second, if the antecedent occurs in a transaction, the consequent has to appear with a certain probability, too. This is called the confidence, see Definition 1.

Definition 1. *Given two itemsets X and Y. The confidence $conf(X \Rightarrow Y)$ is defined as follows:*

$$conf(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)}$$

Similar to the approach of sorting out non-frequent itemsets below the definition of the support, we can make sure to only mine relevant rules. As before one we now set two thresholds to limit the generated rules. These are called *minsup* as lower boundary of the support and *minconf* as lower boundary of the confidence. Now, we can formally define how to validate an association rule, see Definition 2.

Definition 2. *Given two itemsets X and Y. The implication $X \Rightarrow Y$ is now a valid association rule for a minimal support of minsup and a minimal confidence of minconf, if it satisfies the following criteria:*

1. $sup(X \cup Y) \geq minsup$
2. $conf(X \cup Y) \geq minconf$

Now it is clear that frequent itemset mining is a sub-problem of association rule generation. Thus, to generate association rules, at first, all frequent itemsets are generated. A simple way of mining association rules would be the following.
Assume the set of resulting frequent itemsets $\mathcal{F}$. For each itemset $I \in \mathcal{F}$ all possible combinations of itemsets X and Y with $X = I \setminus Y$ are created. These can now be analysed for their confidence and treated respectively. This approach of course can be optimised, as this is just a simple brute-force algorithm. At this point we have generated a set of association rules validated by both minimum support and minimum confidence.

3.4. Interestingness Measures

As we have now generated a set of association rules, one would further want to discriminate between those. Often the amount of generated rules is very high and cannot be sorted by pure human interaction. Thus, there is the need to weight the rules in some term of interestingness. This shall be done computational and automatised.
In this case three different interestingness measures are introduced, as these three are – as shown later – combined. They hold different ranges in the positive real numbers, where 0 always stands for an uninteresting rule, and higher values mean higher interestingness. The measure definitions are taken from Frederic Sautter's work [Sau16], as they are defined differently in various works, but implemented as the selected definitions used by him.

3.4.1. Conviction

Definition 3. *Let X and Y be itemsets and $sup(X)$ the relative support of X. Then the conviction is defined as:*

$$conviction(X \Rightarrow Y) = \frac{1 - sup(Y)}{1 - conf(X \Rightarrow Y)}$$

Conviction is a measure counting in both the uncommonness of the consequent and the confidence of the rule itself. As [ZJ14] mentions, it measures the expected error of a rule.

The rule would err in the event that a transaction contains X, but not Y. The conviction of an association rule can range from 0 to infinite, where 1 indicates that the rules are independent. Infinite is reached when the confidence is 1.

3.4.2. Gini Index

Definition 4. *Let X and Y be itemsets and $sup(X)$ the relative support of X. Then the Gini index is defined as:*

$$\begin{aligned} gini(X \Rightarrow Y) = sup(X) \times (conf(X \Rightarrow Y)^2 + conf(X \Rightarrow \overline{Y})^2) \\ +sup(\overline{X}) \times (conf(\overline{X} \Rightarrow Y)^2 + conf(\overline{X} \Rightarrow \overline{Y})^2) \\ -sup(Y)^2 - sup(\overline{Y})^2 \end{aligned}$$

The Gini index measures the impurity of a rule. Its range stretches from 0 to 0.5. Again, a rule rated with 0 is to be treated as uninteresting, whereas 0.5 implies that the rule is very interesting.

3.4.3. Jaccard Coefficient

Definition 5. *Let X and Y be itemsets. Then the Jaccard coefficient is defined as:*

$$jaccard(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X) + sup(Y) - sup(X \cup Y)}$$

The Jaccard coefficient measures the similarity between antecedent and consequent. Its values range from 0 to 1, where greater values indicate greater similarity [ZJ14].

3.5. Combination of Interestingness Measures

We have now gathered the tools to be able to create association rules measured by interestingness. At first, we generate the frequent itemsets from the transaction database, then we mine relevant association rules, and lastly we measure them by their interestingness. The problem about the last part is to choose the right measure depending on the type of the input data.

To choose one of the measures would be the first thought. However, it is also possible to combine measures to achieve more stability and avoid errors. This is commonly done in the fields of clustering and outlier detection, as to mention the work of Jaskowiak et al. [JMF+16], which explains the use of an ensemble of clustering methods. As Jaskowiak et al. describe, in their test the worst ranking combination of clustering algorithms ranked higher than the best ranking single algorithm. This is due to the fact, that different clustering methods achieve better in parting different structures of data. The combination of two algorithms will then provide a better overall result. As a conclusion, it was decided to also combine different interestingness measures to create a more stable function to value the rules.

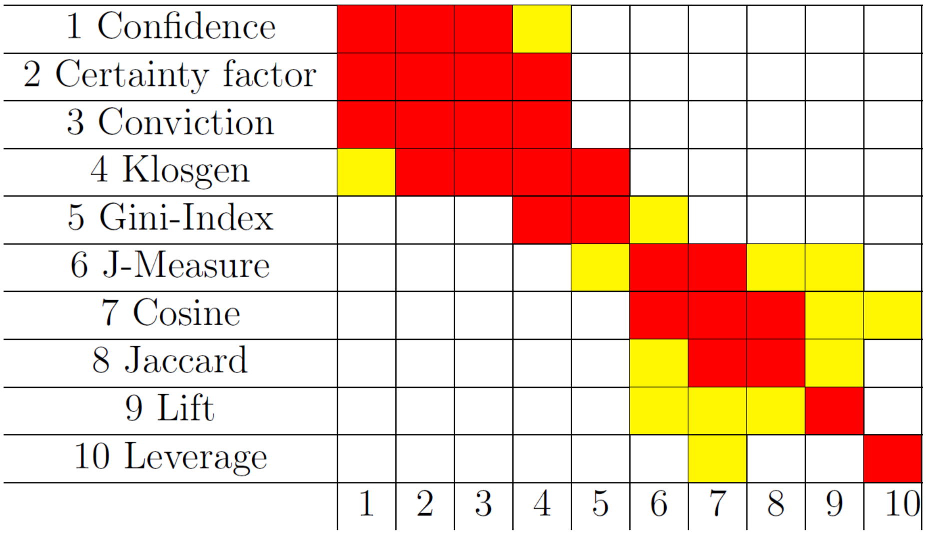


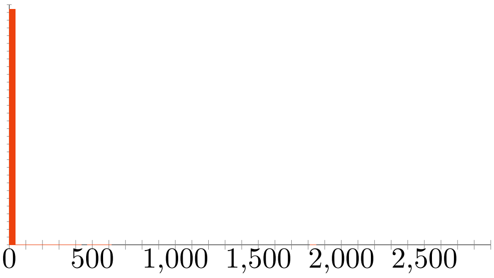
Figure 1: Correlation matrix between the analysed interestingness measures. Yellow fields indicate $\rho(A, B) \geq 0.85$, whereas red fields indicate $\rho(A, B) \geq 0.95$. [Sau16, S. 32]

A great part towards an applicable solution is already taken by the experiments set by Frederic Sautter [Sau16], as in his work he not only compared the measures by their underlying properties and attributes, but also showed the correlations between the measures. This was done by creating a single vector containing the rules sorted by interestingness for each measure. These vectors were then compared and the correlation computed. This allowed visualising the correlation values in a correlation matrix as shown in figure 1.
If we have a look at the correlation matrix, we can spot two different clusters of measures. Additionally, there are two hardly correlating measures, Leverage and the Gini Index. To get a more diverse combination of measures, the three measures chosen were taken one from each cluster and one more outstanding. Namely, they are the above mentioned measures: Conviction, Gini Index and the Jaccard Coefficient.

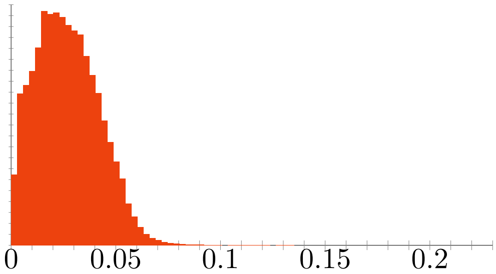
To research the behaviour of the three measures, the German artwork data of the ARTigo project is used. As described later on in Chapter 4, roughly 60.000 images together with their taggings are analysed. At last, there has to be dealt with the different ranges and also the different rule distribution throughout the measures. To illustrate the results of the three different rules, they are visualised via histograms in figure 2. It is now recognisable, that where the Jaccard Coefficient is somewhat widely distributed, the Gini Index is only set between 0 and 0.1. Conviction rates a lot of rules with 0 to 2, but some go up to 2800. Due to these facts, there have to be made some transformations, before assembling a combination of the three rules.

At first, we will have a look at measure Conviction. We will apply the transformation $\frac{x}{x+1}$, as values from 0 to infinity are now scaled down to 0 to 1. The Gini Index is transformed by taking the square root of every value. The Jaccard Coefficient will now be untouched, as it is already distributed somewhat smoothly. Now the histograms are depicted again in figure 3.

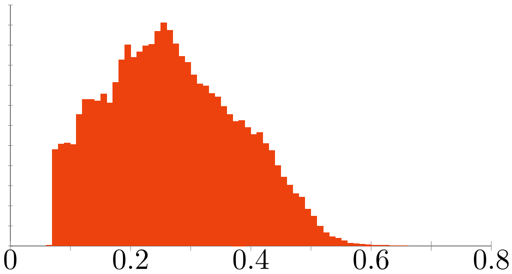
To finally round up the process of scaling the interestingness measures, we now subtract the means of the generated rules from the rating and divide it by their standard deviation.



(a) Conviction



(b) Gini Index



(c) Jaccard Coeffizient

Figure 2: Histograms of the three interestingness measures before any scaling

This allows us to roughly weight all measures the same. This is depicted in figure 4 together with the histogram of the resulting combination of the three rules.

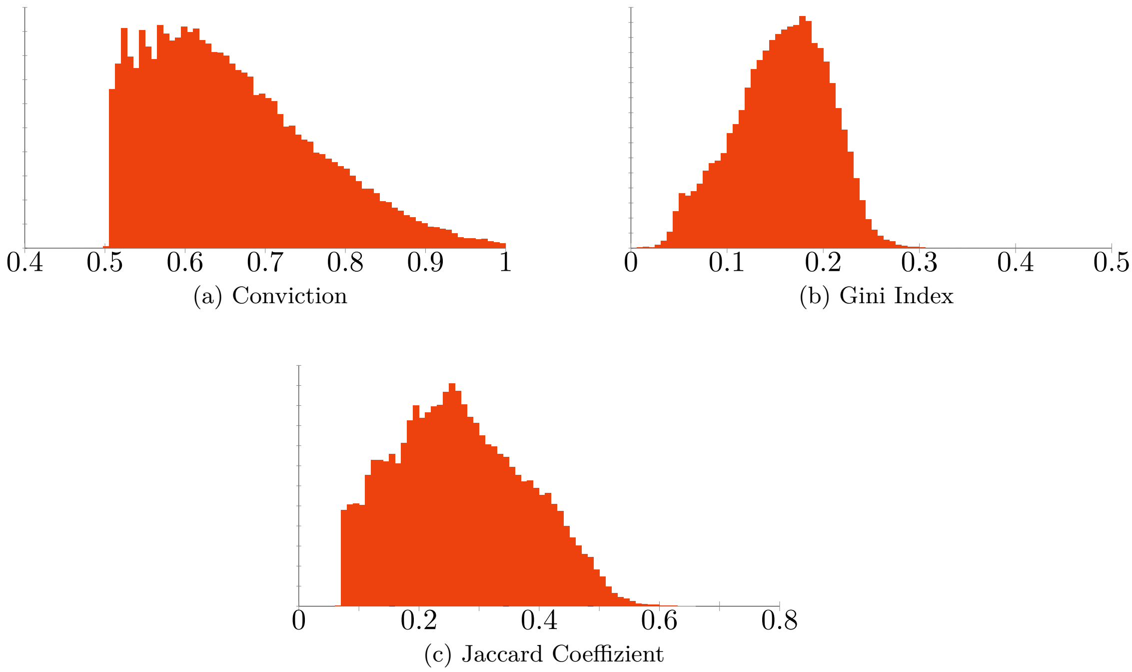


Figure 3: Histograms of the three interestingness measures with transformations done

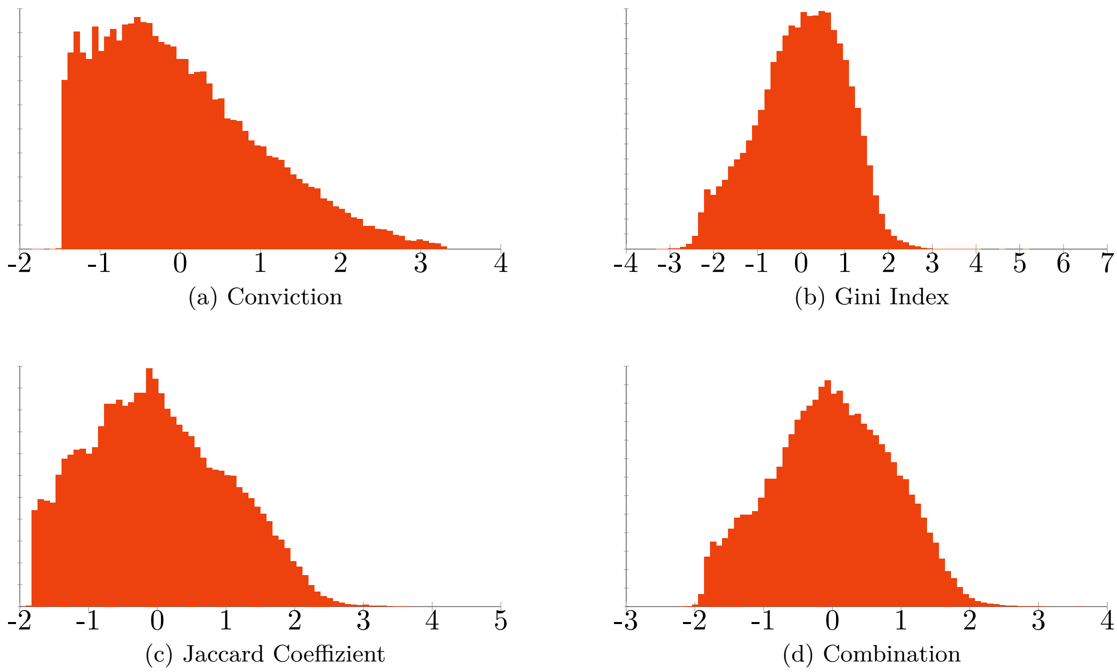


Figure 4: Final histograms of the three interestingness measures and the resulting combination

4. Data and Tools

4.1. ARTigo

As mentioned in the introduction, the data used for the experiments in this work is taken from the ARTigo project [Wie14]. The ARTigo project is a collaboration between the "Institut für Kunstgeschichte" and the "Institut für Informatik", which are both departments of the "Ludwig-Maximilians-Universität München". It offers the possibility to gather information and search for about 60.000 artworks produced in the long nineteenth century. As it is also a project to gather metadata, which are keywords describing the artwork, the user is offered to play a variety of so-called "Games With A Purpose" (GWAP). The ARTigo game for example lets the user play against other users, while tagging a picture with afore mentioned tags. The players are rewarded for tags, which are input by both players. Some points are also given for already confirmed tags, while new keywords are saved to a database. If an image is repeatedly tagged with a new tag, it will also be confirmed. This contributes to the database containing the overall taggings.

As mentioned before, it is also possible to search for certain keywords. The search is also

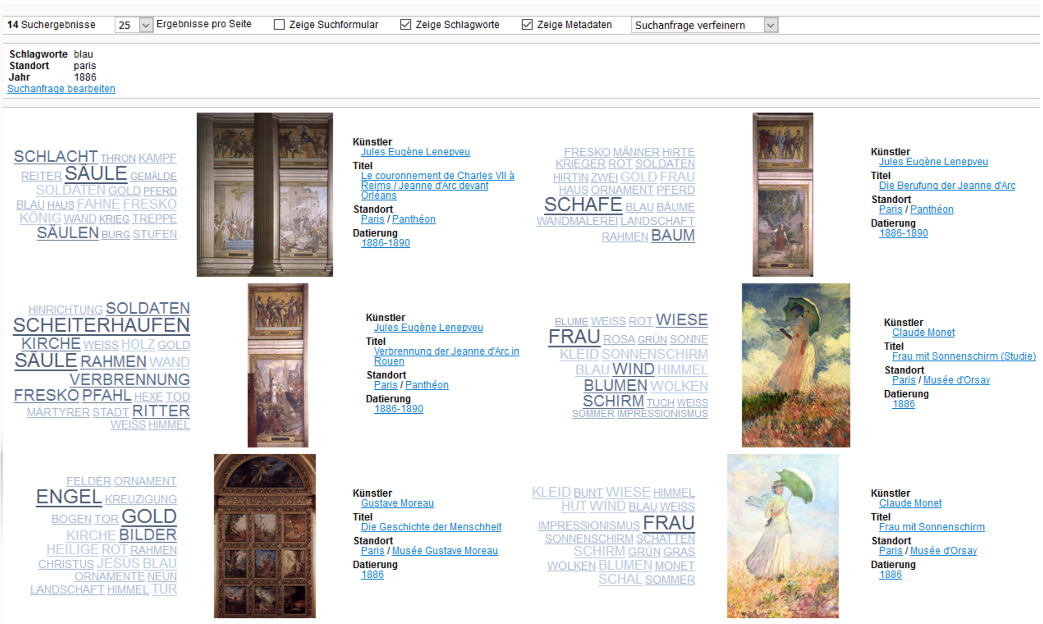


Figure 5: ARTigo search for artwork produced in 1886, tagged with "BLUE" (meaning blue) and placed in Paris

augmented by the possibility to add additional filters such as time period, location, title, artist and also tags the image must not contain. This combination of user functionality in form of a detailed search with the part of human computation to gather metadata, which would otherwise be different to acquire, leads to a platform both entertaining and informational, as well as having a scientifical purpose. An example of the search functionality is given in Figure 5.

4.2. ELKI

The **E**nvironment for Deve**l**oping **K**DD-Applications Supported by **I**ndex-Structures, short ELKI [SKE+15], is the tool that will be used to mine the association rules, rate them by interestingness and also create histograms for visualisation. As it is an open source framework written in Java, it allows simple modification for own purposes. It is mainly developed and supported by the Institute of Informatics of the Ludwig-Maximilians-Universität München. ELKI offers a wide range of different algorithms throughout different categories of data mining. It enables its users to conduct clustering, outlier detection, itemset mining, visualisation and lots more. It is mainly used to research new algorithms and their efficiency. This is possible due to the strict separation of algorithms and data management tasks.
As afore mentioned, Frederic Sautter extended ELKI to also create association rules and measure their interestingness [Sau16]. This now allows us to use ELKI as the underlying data mining tool for the newly implemented component. In this bachelor thesis the official git repository is forked, and version 0.7.2 is compiled, adding a combination of the three interestingness measures additionally implemented as described above.

4.3. ARTigo Analytics Center

While already providing the tools to filter for certain artwork, the ARTigo project's purpose is not supplying data mining tools in particular. As there is a lot of metadata gained throughout the game rounds, the idea of creating a project based on the ARTigo data and providing the necessary instruments to mine information out of it was a consequential one. The ARTigo Analytics Center was then implemented by Florian Hoidn in 2014 [Hoi14]. It provides different tools to conduct a variety of aspects of data mining on the ARTigo data. The ARTigo Analytics Center provides a web server based on the JBoss Seam framework. It directly connects to the PostgreSQL-database via Hibernate, which allows a very direct and relatively fast access to the ARTigo data. As implemented by Florian Hoidn, it consists of the following five user components:

1. The Frequency Query Component (Häufigkeitsgraphen) allows the user to set one or more keywords, as well as a start and end date. The relative frequencies of the tags are then plotted as a graph. This gives a certain overview of the tag's relevance in different time periods. An example is shown in figure 6.
2. The value of tags queried by the Poisson Overdispersion Query Component determines how specific they are. This permits the user to get a certain sense of the relevance of a tag.
3. The Association Rule Query Component lets the user compute the confidence of a self defined association rule. This allows the user for example to confirm or decline an assumption made on the basis of the artwork data. It is not capable of generating association rules or rating them by their interestingness, which is what this thesis wants to implement.
4. The most similar tags can be found with the Nearest Neighbour Ranking Component. A ranking composed of the "nearest" tags in terms of cosine or euclidian distance is generated depending on the user input tags.

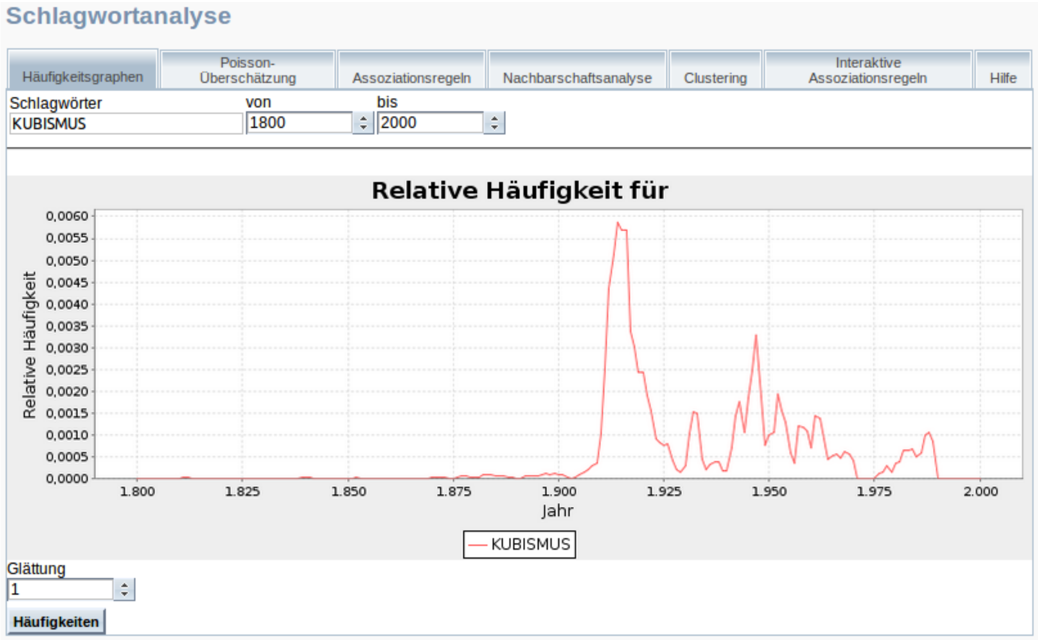


Figure 6: Frequency Query Component of the ARTigo Analytics Center; the keyword queried is “KUBISMUS”, which is cubism in German. The time period is limited to 1800 until 2000 and the graph is slightly smoothed.

5. Finally, it is possible to use the DBSCAN Cluster Expansion Component, which allows, similar to the Nearest Neighbour Ranking Component, to find a group of near keywords. However, using the DBSCAN algorithm, a cluster is expanded around the initial input.

The ARTigo Analytics Center is configured to allow easy modification and thus expansion of the initial tools. In this bachelor thesis a new component will be added, which focuses on the interactive exploration of association rules.

5. The Interactive Association Rule Component

As mentioned in the last chapter, it is possible to modify the ARTigo Analytics Center without greater obstacles. In this chapter I want to introduce a component newly implemented to the ARTigo Analytics Center, which allows an interactive analysis of association rules generated from the ARTigo artwork data. As we will see, it enables us to get a deeper understanding of the underlying artwork descriptions and the overall user tagging behaviour as well. In the following sections the functionality will be explained and, as to guide through the use of the component, some sample usage will be given.

5.1. Features of the Module

The Interactive Association Rule Component includes different features to help understanding the generated rules as easy as possible. The default view is shown in figure 7. At the

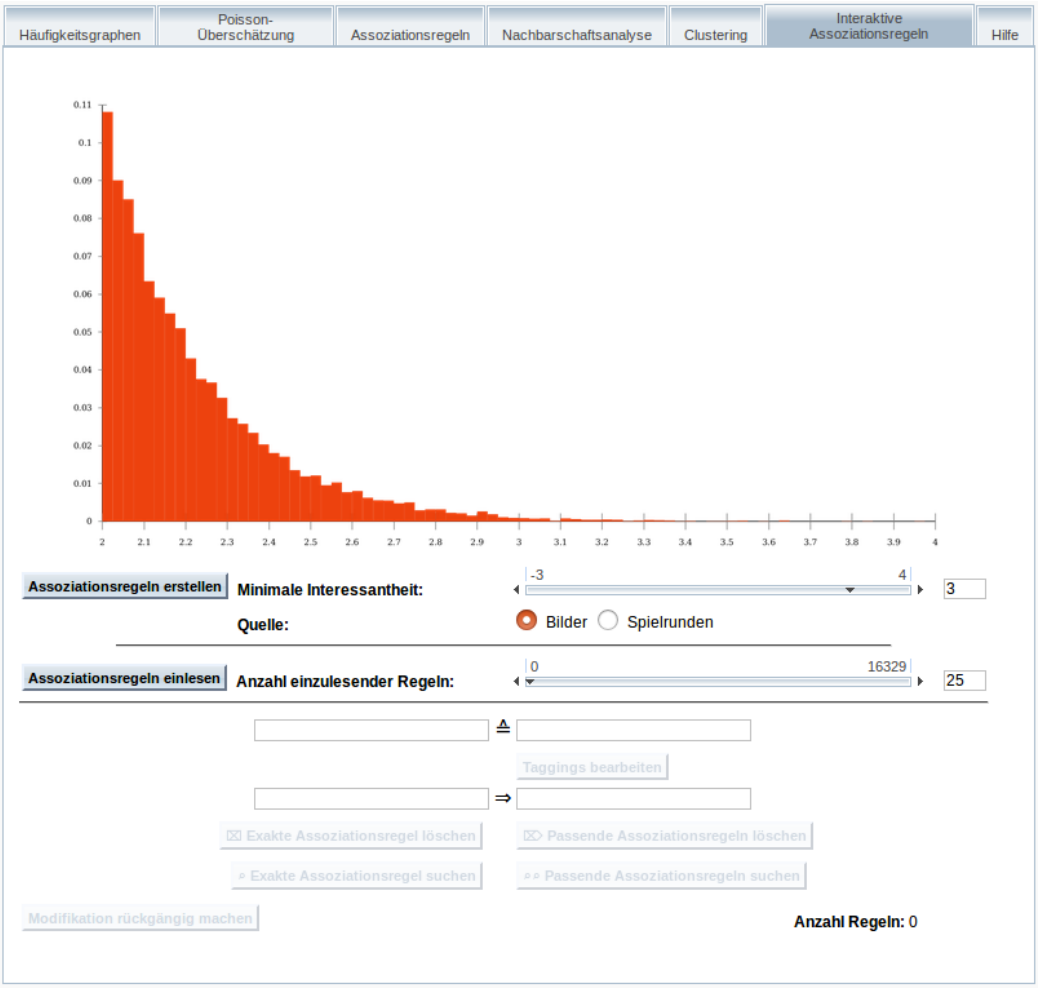


Figure 7: Overview of the Interactive Association Rule Component

server start the transactions - in this case the tags per image - are updated from the AR-

Tigo artwork database and some default rules are generated. Frequent itemsets are always mined with a minimum support of 0.035 as suggested by Frederic Sautter [Sau16]. These default rules are generated with a minimum interestingness measure of 2 and therefore just a bit of the interesting rules are generated. The distribution of these rules is depicted in a histogram, which is shown at the top of the component's page. As one can see, there are now two different possibilities to explore the data set:
First, it is possible to generate one's own association rules. For this you can set a minimum interestingness measure with the top slider. The two radio buttons enable to choose between the taggings per image ("Quelle: Bilder") as transactions and the taggings per gameround ("Quelle: Spielrunden") as transactions. This allows to set the focus from the overall tagging to a more single person centred approach. To generate the rules, simply press the button called "Assoziationsregeln erstellen". This will make ELKI read in the respective transactions from the text file extracted from the database at server start. The rules are then put in a single folder per session to allow multiple users work with the component at the same time.
Second, you can choose the number of rules, which were read into your working set with the second slider. The rules are always read from the most interesting rule downwards to the least interesting one. The button called "Assoziationsregeln einlesen" will parse either the default rules generated at server start or the custom rules generated by the user.
As depicted in Figure 8, the rules are now shown in an expandable tree structure. Each rule is shown with its interestingness in the first line and the implication itself below. The nodes are inserted in the order parsed from the ELKI generated file. A rule will be recursively inserted as a child node, if its antecedent and consequent contain the respective sets of the parent node. Otherwise it will be inserted at top level. This leads to rules being child nodes of another rule always being less interesting and more specific. As shown in figure 8, below the buttons at the right side, the number of rules ("Anzahl Regeln") contained in the tree is shown. Also some buttons, which were greyed out before, are now accessible. These allow to modify, remove and filter the rules as following:

- The top two input fields together with the button called "Taggings bearbeiten" allow to modify tags, which is also the standard action when pressing the enter key. This is helpful for example to merge tags semantically equal. It is possible to put a comma-separated list of tags in the fields. In case of the first input panel, it is checked if all tags are contained in the rule to modify. If so, the input tags are replaced by the tags contained in the second field. A sanity check is always made to remove tags contained in both antecedent and consequent from the consequent.
- The two input fields below are used for all four command buttons below. The first one always contains the antecedent and the second one the consequent of the rules to filter or remove. As before, it is possible to define comma-separated lists of tags.
  - The top buttons ("Assoziationsregel(n) löschen") allow the user to remove association rules depending on the above defined itemsets. Whilst the first button only removes a rule exactly matching, the second button will remove all rules with antecedent and consequent being supersets of the input tags.
  - Similar to the buttons above, the lower buttons ("Assoziationsregel(n) suchen") will also remove a certain set of rules. In this case of the first button, all rules

not matching the exact input will be removed. In case of the second button, it suffices for a rule to contain the input tags, to be retained. In other words, these two buttons enable the user to filter only the rules matching the input.

- The button greyed out in figure 8 ("Modifikation rückgängig machen") will undo the last modification made. This is useful in case of wrongly made modifications or to try out something. It is greyed out to indicate that no modifications are made at the moment. To completely reset the working set, it is always a possibility to just read the rules in again.

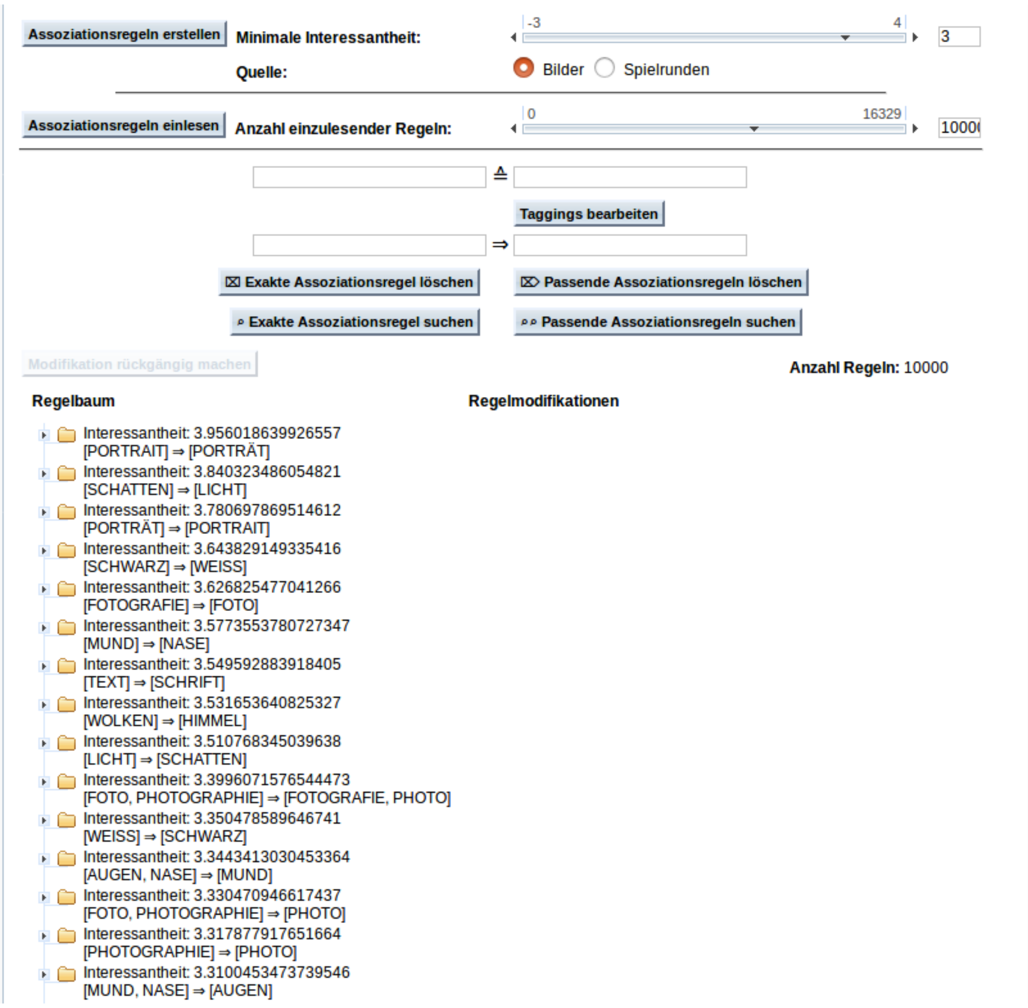


Figure 8: Rules read into the working set, shown in a tree structure

It is also some sanitisation done to remove some useless cases of modification. For example, tags will only be replaced if the itemset to be replaced does not equal the replacing one, the itemset to be replaced is not empty, and there are rules containing the tags to be replaced. Also a rule removal will only be conducted, if there are matching rules to remove.
Modifications done are shown on the right side of the rule tree. This allows for a simple overview of the order of alterations done to the tree. To round up the process of data

mining, user modifications are also saved to the ARTigo database for later information retrieval. Along with a automatically generated ID the modifications are saved in the form of an integer depending on the type of modification, the two arguments, and lastly an unique session identifier.

5.2. Example Usage

In this section, a short example on how to use this component shall be made. The situation depicted in Figure 9 results from the following steps made:

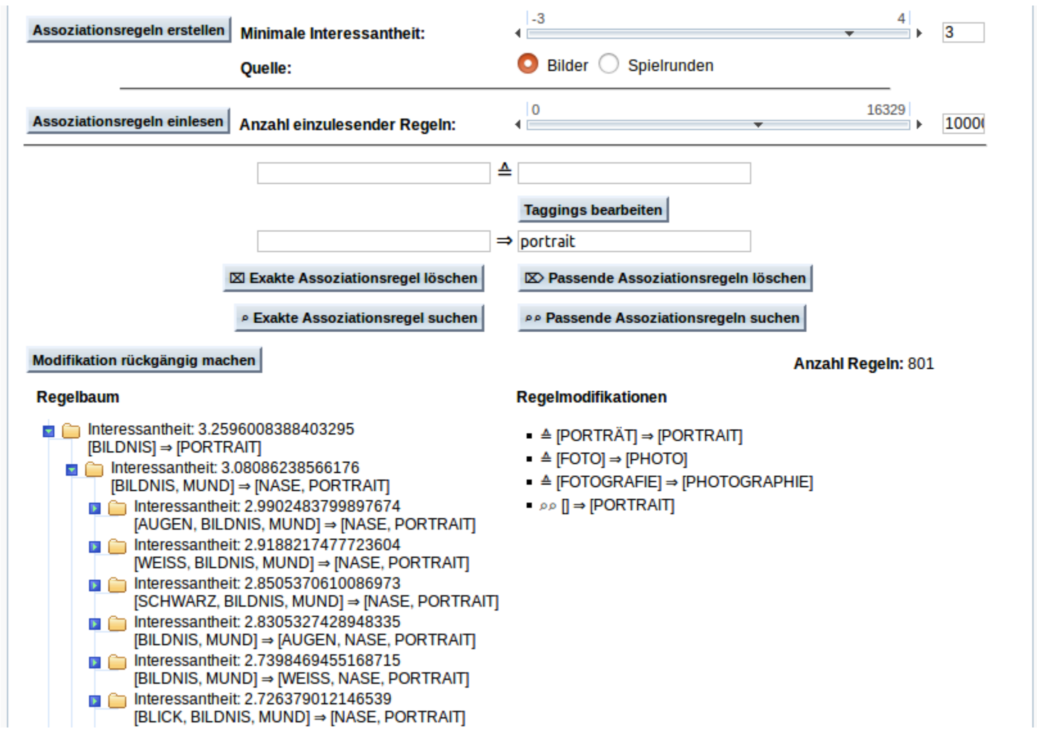


Figure 9: Modifications done to the working set of association rules

1. At first the number of rules to parse was set to 10.000 and the rules were read into the working set.
2. As shown in the lower right half of the image, three tags were modified to remove some synonyms.
3. At last, the rules were filtered for only the rules containing “PORTRAIT” in their antecedent. This results in 713 rules retained.

It is now also possible to revert modifications as indicated by the no more greyed out button right above the rule tree.

5.3. Finalisation

As the ARTigo Analytics Center is easily modifiable, there already exist some other additional components. Within the scope of this work, together with the Interactive Association Rule Component, the Outlier Explanation Component [FB16] and the ELKI Clustering Component [Ngu16] are combined to one build, so a concurrent use of them is possible.

6. Results and Evaluation

In this chapter, the limitations of the implemented component, possible combination with other components, and lastly the gained results are discussed. As later shown, some interesting insights in tagging behaviour, as well as some conclusions are gained.

6.1. Limitations of the Module

It is worth mentioning, that the modification mechanisms of this tool, as shown in Chapter 5, cannot substitute data preparation. Sometimes it is preferable to retain synonyms, so some modifications have to be handmade. As shown, for example, in Figure 9, some synonyms were replaced. This, of course, does not influence the interestingness rating of the retained rules. As mentioned earlier in the last chapter, the tool does merge duplicate rules. This is implemented by simply removing the lower rated rule. As always, it is preferable to conduct as much pre-processing as possible, before performing data mining on the data.
It is important to keep in mind, that the meaning of the interestingness is not always easy to interpret. Especially, as the used measure is a combination of three different measures, it is very difficult to explain a certain rating. In particular, this applies to the confidence of the rules. Rules rated as to be very interesting, may or may not be as confident as others. As now explained, there is a way to check rules more detailed with the ARTigo Analytics Center.

6.2. Verifying Results with ARTigo and the ARTigo Analytics Center

As mentioned in Chapter 4, the ARTigo Analytics Center [Hoi14] implements a method to calculate the confidence of an association rule. This can be very helpful to validate the significance of a rule. To calculate the confidence, one has to put in antecedent and consequent of the rule and simply press the calculate button. As seen in Figure 10, the interestingness of an association rule is not necessarily bound to its confidence. For this chart, the five top ranking rules and five randomly chosen rules were picked. Their confidence was then checked with the Association Rule Component of the ARTigo Analytics Center. Due to the calculation of the combined measure, the confidence of an association rule has a certain amount of influence on the interestingness. This, of course, leads to the fact that more interesting rules are also more often above a certain level of confidence. To tell the real significance of a rule, human oversight is indispensable.
Other tools to investigate the relevance or distribution of a rule are, for example, the image search of ARTigo [Wie14] and the Frequency Plot Component described in Chapter 4. The first one allows to get an overview of the amount of pictures tagged with a certain tagging. This can then be compared to the amount of rules generated which contain the tagging. The latter one can be used to find out about the distribution of taggings throughout time. This is helpful for recognising certain trends and correlations between the taggings.

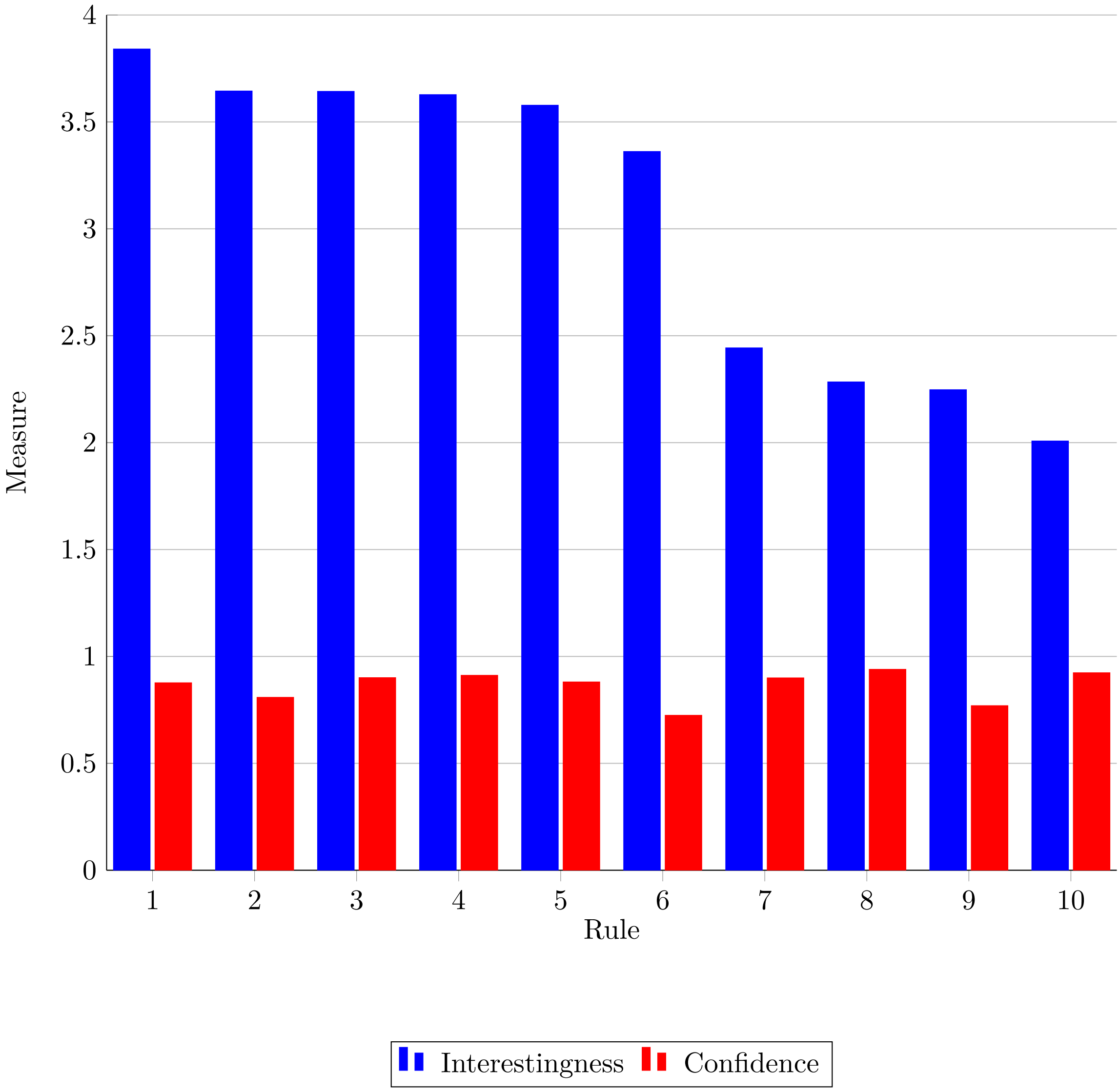


Figure 10: Comparison of ten different association rules with their corresponding interestingness and confidence. Note, that interestingness and confidence are not comparable to each other, but rather just use the same scale without units. The five higher ranking rules are the highest ranking rules as shown in Table 3. The five others were randomly chosen and are: $\{SCHATTEN\} \Rightarrow \{LICHT, WEISS\}$, $\{\ddot{O}L\} \Rightarrow \{GEM\ddot{A}LDE\}$, $\{AKT\} \Rightarrow \{LICHT, NACKT\}$, $\{SCHATTEN, PHOTO\} \Rightarrow \{LICHT, PHOTOGRAPHIE\}$, and $\{HIMMEL, H\ddot{U}GEL\} \Rightarrow \{LANDSCHAFT\}$.

6.3. Evaluation

In this section, the rules generated from the ARTigo [Wie14] data are analysed and some assumptions are made. The rules are generated from images as transactions, as this yields much more resulting rules, compared to rounds as transactions. As the rules are generated from human input, the results point out different aspects of observations made by users. This has to be remembered when evaluating the association rules, as the taggings are often very specific.
At first, we will have a look at the general structure of tags contained in the rules. Most of the rules contain tags concerning portraits. A much smaller part of the rules contain tags describing landscapes, like "HIMMEL" and "TREE" (meaning sky and tree). There are also some rules describing the overall content or kind of picture, as for example "PHOTOGRAPHIE" and "ÖL" (meaning photograph and oil). Searching for pictures tagged with "PORTRAIT" or "PORTRÄT" (meaning portrait) delivers around 6.000 pictures, which is about fifteen percent of the pictures to be found at total. When looking at the most interesting rules – the 10.000 most interesting rules, with different spellings sorted out –, about three of four rules contain portrait as a tag. This leads to some interesting conclusions.
An explanation would be the specific nature of a portrait. While there are about 7.000 pictures tagged with "LANDSCHAFT" (meaning landscape), when searching with ARTigo, there are much less rules rated as interesting containing this tag, compared to the rules concerning portraits. This may be because of the focus lying on the contents of the image. Whilst these can be very diverse in both cases, a portrait may be identified as one easier than a photography or painting of a landscape. The lack of rules containing more specific tags describing landscapes, such as "BAUM" or "HIMMEL", whilst these are still very common tags amongst the artwork, speaks against this theory.
Instead, it may be, that people are better in recognising faces than other objects. The relevance of human face recognition is shown for example in [HPJ02]. A lot of the rules containing portrait as antecedent or consequent and being rated interesting, also contain tags like "BILDNIS", "AUGE", "NASE", and "BLICK" (meaning image, eye, nose and look). This correlates with the distribution of the tags as explored for example with the Frequency Plot Component described in Chapter 4. It is clear, that the semantics of this rules often makes sense, as these are the things depicted in portraits. Some of these rules also contain tags like "WEISS" and "SCHWARZ" (meaning white and black). This is more of a trivial phenomenon, as a lot of different artwork is tagged with colours, especially with black and white, which is due to a lot of images being black and white.
The rules rated interesting rarely contain any tags describing art historical insights or conclusions of what is shown in the picture. The abstract nature of the rules is explained by the manner of rule generation. Due to ARTigo being a project relying on games and citizen science, most of the contributions are made by non-professionals and, therefore, describe the contents of the picture well, but do not go into historical details too much. Thus, there are only a small amount of taggings concerning moral or statements made by the artist. This leads to most rules being implications made of simple tags.
When comparing the results of the single measures, as done in [Sau16], it stands out, that Conviction has a lot of top rated rules containing "PORTRAIT", while these are much longer, and therefore more specific, than the ones rated high by the combination of the three measures. On the other hand, measuring with the Gini Index or the Jaccard Coeffi-

cient, a lot of colour and light based rules are rated high. Insofar, the combination of the rules seems to have a preferable outcome: Short and diverse rules are rated high. Of course, this is true for the data used in this thesis, and is no guarantee to be the perfect solution for other data sets.

Together with the theoretical basis in Chapter 3, it is shown that the combination of different interestingness measures is a perfectly applicable approach to improve both the quality of the measurement and the quality of top ranked rules. Of course, the exact selection of measures and the practical implementation has to be adjusted depending on the use case. The generated rules do not deliver complex statements about historical information, but rather simple implications. Also, there exists a clearly recognisable bias among the interesting rules towards portraits and their more detailed descriptions. This may be an effect of the human ability of quickly recognising faces.

7. Conclusion and Future Work

7.1. Conclusion

In this thesis the possibility is made to explore and examine association rules generated from the ARTigo data with the ARTigo Analytics Center. Therefore, at first, we had a look at the underlying theory and the rating of association rules by their interestingness. With a foundation lain by Frederic Sautter [Sau16], a combination of the three measures Gini Index, Jaccard Coefficient and Conviction was then outlined. As the measures differ greatly in aspects of value set and distribution, some transformations had to be applied to weight them correctly. The combination of these three measures provides a stable and reliable result.
With the implementation of a combined measure in ELKI and the infrastructure given by the ARTigo Analytics Center, a new module was added to the ARTigo Analytics Center. This new module comprises different methods of modifying and filtering the association rules, as well as a tree model to view the rules in a sorted manner. Together with already existing components of the ARTigo Analytics Center, this module enables interested people to analyse associations between different taggings belonging to the artwork of the ARTigo project. This can be done without deeper art historical knowledge or skills in the field of Informatics. Nevertheless, there are still a lot of possibilities to improve or expand the implemented module or the ARTigo Analytics Center itself. Some of these are presented in the following subsection.

7.2. Future Work

Data preparation still holds various possibilities to improve the input data. The taggings are read in from a spell-checked variant of the ARTigo database. As there are still different spellings of the same word, the association rules often imply from one spelling to the other. This is due to the fact, that the different spellings often occur together, as they are bound to the same semantics. Also, there are various kinds of synonyms, which are more difficult to remove, while still not losing different meanings. An intelligent way to prepare the data without losing information would not only reduce computational effort, but also increase the validity of the interestingness.
The implemented combination in ELKI relies on the pre-computed values of expected value and standard deviation of the single rule distributions generated from the latest ARTigo data. While the general structure of the rules will presumably not change completely, some discrepancy to the normalisation may occur. A long-term way to avoid this problem would be the possibility to input these values as parameters in ELKI.
The possibility to generate the rules from rounds played in the ARTigo game is momentarily a very RAM expensive process. This is due to the fact, that the amount of gamerounds is much higher than the amount of images. Nevertheless, the amount of rules generated is very small compared to the amount of rules compared with images as transactions. This derives from the itemsets only being very small, compared to the itemsets when looking at the artwork taggings. Thus, frequent itemsets are much harder to be found, and the

resulting amount of rules is much smaller. Here, a different approach in generating itemsets may be needed.
At last, a step most interesting would be to combine different modules of the ARTigo Analytics Center to one. The theses of Duc Nguyen [Ngu16] or Alexander Fischer-Brandies [FB16] would make an excellent example. While the first implements the possibility to cluster the artwork via ELKI much more detailed than the original component, the latter provides a way to find an explanation for a picture being an outlier, which means, it differs from the other pictures in an extraordinary manner. With the possibility to only examine a cluster of pictures, the amount of rules would be greatly decreased and simultaneously limited to artwork belonging together. This could deliver more detailed insights to this group of pictures, which can also be visualised with this module. The possibility to find out about outlying artwork, together with its tagging and the explanation for it being an outlier, could provide ideas of filtering and removing certain rules.

Appendices

A. Experimental results

Interestingness	Rule
3.95602	PORTRAIT ⇒ PORTRÄT
3.84032	SCHATTEN ⇒ LICHT
3.78070	PORTRÄT ⇒ PORTRAIT
3.64383	SCHWARZ ⇒ WEISS
3.64209	SCHATTEN, WEISS ⇒ LICHT
3.63785	PORTRAIT, WEISS ⇒ PORTRÄT
3.62683	FOTOGRAFIE ⇒ FOTO
3.57736	MUND ⇒ NASE
3.54959	TEXT ⇒ SCHRIFT
3.54301	AUGEN, MUND ⇒ NASE
3.53165	WOLKEN ⇒ HIMMEL
3.51077	LICHT ⇒ SCHATTEN
3.50413	PORTRÄT, WEISS ⇒ PORTRAIT
3.48434	PORTRAIT, SCHWARZ ⇒ PORTRÄT
3.47056	LICHT, WEISS ⇒ SCHATTEN

Table 2: The fifteen highest ranked rules created from images as transactions
.

Interestingness	Rule
3.84032	SCHATTEN ⇒ LICHT
3.64383	SCHWARZ ⇒ WEISS
3.64209	SCHATTEN, WEISS ⇒ LICHT
3.62683	PHOTOGRAFIE ⇒ PHOTO
3.57736	MUND ⇒ NASE
3.54959	TEXT ⇒ SCHRIFT
3.54301	AUGEN, MUND ⇒ NASE
3.53165	WOLKEN ⇒ HIMMEL
3.51077	LICHT ⇒ SCHATTEN
3.47056	LICHT, WEISS ⇒ SCHATTEN
3.41595	PORTRAIT, MUND ⇒ NASE
3.36085	SCHATTEN ⇒ LICHT, WEISS
3.35048	WEISS ⇒ SCHWARZ
3.34434	AUGEN, NASE ⇒ MUND
3.34041	PORTRAIT, SCHWARZ ⇒ WEISS

Table 3: The fifteen highest ranked rules after sorting out synonyms
.

Interestingness	**Rule**
1.43163	SCHWARZ ⇒ WEISS
0.93043	MUND ⇒ NASE
0.91008	WEISS ⇒ SCHWARZ
0.86272	WOLKEN ⇒ HIMMEL
0.57086	MUND ⇒ AUGEN
0.56374	AUGEN ⇒ NASE
0.50557	NASE ⇒ AUGEN
0.36464	NASE ⇒ MUND
0.22216	HIMMEL ⇒ WOLKEN
0.21790	GRAU, SCHWARZ ⇒ WEISS
0.12956	AUGEN ⇒ MUND
0.00242	BÄUME, WOLKEN ⇒ HIMMEL
-0.04939	GRAU, WEISS ⇒ SCHWARZ
-0.07638	LANDSCHAFT, WOLKEN ⇒ HIMMEL
-0.09164	FOTOGRAFIE ⇒ FOTO

Table 4: The fifteen highest ranked rules created from gamerounds as transactions. The minimum support had to be set to 0.015 to generate enough association rules.

.

B. Setup Manual

To install this component, the following steps are necessary. If the ARTigo Analytics Center is already configured, omit steps one to three.

1. At first install the PostgreSQL server as described in the “INSTALL.txt” delivered by the ARTigo git repository.
2. Follow the instructions found in the “README.md” provided by the ARTigo Analytics Center repository.
3. Replace the variable `JAVA-OPTS` in “jboss/bin/run.conf” with
   `JAVA_OPTS="-Xms1024m`
   `-Xmx4G -Dsun.rmi.dgc.client.gcInterval=3600000`
   `-Dsun.rmi.dgc.server.gcInterval=3600000"`.
4. Set the path to the folder containing the ARTigo Analytics Center source code in the “InteractiveAssociationRuleComponent.java” file found in “src/hot/analytics/center”. The variable to be set is called `projectFolder`.
5. If you want to use the ARTigo Analytics Center with other added components, like, for example, Duc Nguyen’s [Ngu16], refer to their work for additional steps. Often, there exists a readme or an explanation in the belonging thesis or source code folder. Some modules may also rely on external software.
6. Deploy the server via `ant deploy`.

References

[Agg15] Charu C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015.

[AIS93] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pages 207–216. ACM, 1993.

[BP63] Thomas Bayes and Richard Price. An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions*, 53:370–418, 1763.

[FB16] Alexander Fischer-Brandies. Explaining Outliers in ARTigo. Bachelor thesis, Institute of Computer Science, LMU, Munich, 2016.

[Hoi14] Florian Hoidn. The Analytics Center: Devising a Citizen Science Data Mining Tool for the ARTigo Image Tagging Project. Bachelor thesis, Institute of Computer Science, LMU, Munich, 2014.

[HPJ02] Michelle de Haan, Olivier Pascalis, and Mark H. Johnson. Specialization of Neural Mechanisms Underlying Face Recognition in Human Infants. *Journal of Cognitive Neuroscience*, 14(2):199–209, 2002.

[JMF+16] Pablo A. Jaskowiak, Davoud Moulavi, Antonio C. S. Furtado, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. On strategies for building effective ensembles of relative clustering validity criteria. *Knowledge and Information Systems*, 47(2):329–354, 2016.

[KH95] Krzysztof Koperski and Jiawei Han. *Discovery of spatial association rules in geographic information databases*, pages 47–66. Springer, 1995.

[Ngu16] Tien Duc Nguyen. Explorative Cluster analysis of artworks for the ARTigo Analytics Center. Bachelor thesis, Institute of Computer Science, LMU, Munich, 2016.

[OSYY02] Miho Ohsaki, Yoshinori Sato, Hideto Yokoi, and Takahira Yamaguchi. A Rule Discovery Support System for Sequential Medical Data. In *In the Case Study of a Chronic Hepatitis Dataset –. Int'l Workshop on Active Mining, IEEE Int'l Conf. on Data Mining (ICDM'02)*, pages 97–102, 2002.

[Sau16] Frederic Sautter. Association Rule Generation and Evaluation of Interestingness Measures for Artwork Tags. Bachelor thesis, Institute of Computer Science, LMU, Munich, 2016.

[SKE+15] Erich Schubert, Alexander Koos, Tobias Emrich, Andreas Züfle, Klaus Arthur Schmid, and Arthur Zimek. A Framework for Clustering Uncertain Data. *PVLDB*, 8(12):1976–1979, 2015.

[Wie14] Christoph Wieser. *Building a Semantic Search Engine with Games and Crowdsourcing*. Dissertation, Institute of Computer Science, LMU, Munich, 2014.

[ZJ14] Mohammed J. Zaki and Wagner Meira Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.