Die approbierte Originalversion dieser Diplom-/Masterarbeit ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (http://www.ub.tuwien.ac.at).

The approved original version of this diploma or master thesis is available at the main library of the Vienna University of Technology (http://www.ub.tuwien.ac.at/englweb/).

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Analysis of User-Generated Content in the Context of a Database of Artworks

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Michael Koutensky, BSc
Matrikelnummer 0625117

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung
Betreuer: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl
Mitwirkung: Projektass.(FWF) Dipl.-Ing. Max Arends

Wien, 17.10.2011

(Unterschrift Verfasser) (Unterschrift Betreuer)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Analysis of User-Generated Content in the Context of a Database of Artworks

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Michael Koutensky, BSc
Registration Number 0625117

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl
Assistance: Projektass.(FWF) Dipl.-Ing. Max Arends

Vienna, 17.10.2011

(Signature of Author)

(Signature of Advisor)

Erklärung zur Verfassung der Arbeit

Michael Koutensky, BSc
Bischoffgasse 1/8/21, 1120 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 17.10.2011 _____

(Unterschrift Verfasser)

Acknowledgements

I want to thank my advisor Dieter Merkl, who virtually invited me to write this thesis through his inspiring lectures on machine learning, information retrieval and cultural heritage informatics during my studies. His wide knowledge, his encouragement and his guidance have been of great value for me. I also want to thank Max Arends for his profound advice, continuous support and friendly help. It was really a great pleasure working with them.

For the best lifelong support anyone can think of, I am grateful to my parents Hilda and Alfred Koutensky, who made all this possible and always believed in me. I also want to thank Roxane Licandro for her loving support, her motivation and the inspiring discussions, which helped me a lot.

Abstract

In times of the Web 2.0, with nearly unlimited storage capacity and bandwidth, a lot of image collections are available on the Internet. Some of these collections are completely open for new material and contribution, some of them are completely closed for user input and some of them are a mixture, i.e. do not allow user uploads, but encourage user contribution through commenting, rating and annotating images, such as the explorARTorium[1]. The explorARTorium hosts a large collection of ∼20,000 digitized images of artworks, which can be explored along various dimensions such as time, region or theme. Users browsing the collection are able to tag as well as rate the artworks to express how much they like the picture. Through this practice of annotating content, a folksonomy (a system of classification based on user collaboration) is created. It is in the operator's interest to keep the users intrigued using the multimedia platform and tagging artworks, because untagged artworks do not contain the desired user input which is important for the operator, as it helps improve the folksonomy and create connections between artworks. Unfortunately, this goal is not easy to achieve, because tagging is a time-consuming task and without any incentive or help, the users' motivation to tag will decrease over time.

In the first part of the thesis, an extensive analysis of the explorARTorium's folksonomy related to art history is conducted, exploring the relationship between users and their tags. Firstly, it is confirmed that the users' tagging behavior can be set into relation to their liking of artworks. Secondly, the users' vocabulary is qualitatively and lexically analyzed discovering great differences between themes (e.g. portraits are described with different parts of speech than religious artworks). Thirdly, the role of the user regarding activity and learning effects is examined showing that the users' vocabulary gets more specific over time. Finally, the decrease of the users' motivation to tag is confirmed over time.

In order to give the users of the artwork collection an incentive to tag pictures and thus to prevent the users' tagging motivation from declining, a framework for system-generated suggestions for appropriate tags (based on tags extracted from the folksonomy) is developed and presented in the second part of this master's thesis, which offers the user an easier way to describe the artwork. The implementation of the framework makes heavy use of business intelligence techniques like Frequent Itemset Mining and Association Rule Mining for discovering interesting relations between variables in the database to provide reliable decision criteria for the recommender system. The quality and precision of the implemented Tag Recommendation Framework are evaluated revealing that the framework yields astonishing results for certain stereotypes of artworks and on average also performs better than a naive algorithm.

[1] `http://www.explorARTorium.info`; [accessed 04-October-2011]

Kurzfassung

In Zeiten des Web 2.0 mit beinahe unbegrenzter Speicherkapazität und Bandbreite, sind eine Fülle von Bildersammlungen im Internet verfügbar. Einige dieser Sammlungen sind völlig offen für neues Material und Beiträge, einige sind komplett geschlossen für Benutzerbeiträge, andere hingegen stellen eine Mischform dar, wie zum Beispiel das explorARTorium[2], d.h. sie erlauben zwar keine Uploads von neuen Bildern, ermöglichen jedoch nutzergenerierte Inhalte in Form von Kommentaren, Bewertungen und Annotieren von Bildern. Das explorARTorium beherbergt eine umfangreiche Sammlung von ca. 20.000 digitalisierten Bildern von Kunstwerken, die nach bestimmten Kriterien, wie z.B. Zeit, Region oder Thema erkundet werden können. Benutzern, die durch die Sammlung navigieren, steht die Möglichkeit offen, die Kunstwerke zu annotieren (taggen) bzw. zu bewerten. Durch diese Praxis, nämlich den Inhalt mit Anmerkungen zu versehen, wird eine Folksonomy (d.h. ein Klassifikationssystem basierend auf Benutzermitwirkung) geschaffen. Es liegt im Interesse des Betreibers, die Faszination der Benutzer an der multimedialen Plattform und am Taggen von Kunstwerken zu erhalten, da die nutzergenerierten Inhalte für den Betreiber von großer Wichtigkeit sind, weil sie dazu beitragen, die Folksonomy zu verbessern und Verbindungen zwischen den Kunstwerken herzustellen. Leider ist dieses Ziel nur schwer zu erreichen, denn das Taggen stellt eine zeitaufwändige Aufgabe dar, und ohne Anreiz oder Hilfestellung wird die Motivation des Benutzers dies zu tun im Laufe der Zeit abnehmen.

Im ersten Teil der Diplomarbeit wird eine umfassende Analyse der Folksonomy des explorARToriums im Zusammenhang mit Kunstgeschichte durchgeführt, indem das Verhältnis zwischen Benutzern und ihren Tags untersucht wird. Erstens wird bestätigt, dass das Taggingverhalten der Benutzer in Relation zu ihrem Bewertungsverhalten zu setzen ist. Zweitens bringt eine qualitative und lexikalische Analyse des Vokabulars der Benutzer große Unterschiede zwischen den Themen zutage (z.B. werden Porträts mit anderen Sprachmitteln beschrieben als religiöse Kunstwerke). Drittens wird die Rolle der Benutzer hinsichtlich Aktivität und Lerneffekten untersucht und aufgezeigt, dass das Vokabular der Benutzer im Laufe der Zeit spezieller wird. Letztlich wird die mit der Zeit abnehmende Motivation der Benutzer Kunstwerke zu taggen bestätigt.

Um den Benutzern von Kunstsammlungen einen Anreiz zum Taggen der Bilder zu bieten und damit das Absinken ihrer Motivation zu verhindern, wird ein Framework für systemgenerierte Vorschläge von passenden Tags entwickelt und im zweiten Teil dieser Diplomarbeit präsentiert. Dieses Framework erleichtert es dem Benutzer, die Kunstwerke zu beschreiben. Die Implementierung des Frameworks bedient sich Business Intelligence Techniken wie z.B. Fre-

[2] `http://www.explorARTorium.info`; [abgerufen am 4. Oktober 2011]

quent Itemset Mining und Association Rule Mining, um interessante Verbindungen zwischen Variablen in der Datengrundlage zu entdecken und damit Entscheidungskriterien für das Recommender System zur Verfügung zu stellen. Die Qualität und die Genauigkeit des implementierten Tag Recommendation Framework werden einer Evaluation unterzogen, die aufzeigt, dass das Framework erstaunliche Ergebnisse für bestimmte Stereotypen von Kunstwerken liefert und auch im Durchschnitt bessere Leistungen erzielt als ein naiver Algorithmus.

Contents

CHAPTER 1

Introduction

In this introductory chapter, an overview of this master's thesis is given. The motivation and the need for this thesis (Section 1.1) is described in the context of the problem statement (Section 1.2) along with the aim of this work (Section 1.3). Afterwards, the methodological approach is presented in Section 1.4 and the structure of the work is explained in Section 1.5.

1.1 Motivation

In times of the Web 2.0, with nearly unlimited storage capacity and bandwidth, a lot of image collections are available on the Internet. Some of these collections were founded based on the idea of web-based Art education with the goal to publicly provide access to huge collections of Art along with additional information. The Web Gallery of Art (WGA)[1] is a virtual museum and searchable database of European paintings and sculptures containing over 27,600 images. The Web portal Europeana[2] is a multi-lingual online collection offering more than 15 millions of digitized items from European museums, libraries, archives and multi-media collections. The Google Art Project[3] features a collaboration with some popular art museums to enable people to discover and view more than a thousand artworks online in high resolution via *Street View* technology through 360-degree street-level imagery and allows users to create their own artwork collection with personalized annotations and the possibility to share them with other users. By using a crowdsourcing approach (sourcing tasks to a community of undefined people instead of specific individuals), the Flickr Commons[4] project animates users to tag (i.e. annotate) images provided by organizations like the Smithsonian Institute[5] or the Library of Congress[6].

[1] `http://www.wga.hu`; [accessed 04-October-2011]
[2] `http://www.europeana.eu`; [accessed 04-October-2011]
[3] `http://www.googleartproject.com`; [accessed 04-October-2011]
[4] `http://www.flickr.com/commons`; [accessed 04-October-2011]
[5] `http://www.si.edu`; [accessed 04-October-2011]
[6] `http://www.loc.gov`; [accessed 04-October-2011]

Some of these image collections are completely open for new material and contribution (i.e. users can upload new pictures and comment on existing ones), e.g. Flickr[7]. Some of them are completely closed for user input (i.e. no uploading or commenting is possible), e.g. the Web Gallery of Art (WGA)[8]; and some of them are a mixture, i.e. do not allow user uploads, but encourage user contribution through commenting, rating and annotating images, e.g. the explorARTorium[9], part of The Virtual 3D Social Experience Museum (VSEM)[10].

The operators of these collections face different problems maintaining the platforms and improving their quality, which are discussed in the next section.

1.2 Problem Statement

The explorARTorium hosts a large collection of ∼20,000 digitized images of artworks, along with additional information about the artist, topic, time, theme, and region. Users exploring the collection are able to tag the images, i.e. comment on the paintings and share their thoughts about the artwork with other users as well as rate the artworks on a scale from 0 to 5 to express how much they like the picture. The collection can be divided into two parts: paintings that have already been tagged by users and therefore harbor valuable user-generated content, and paintings that have not received any tags. For both parts, the operator of the collection faces different challenges.

The phenomenon of "tagging" can be attributed to a set of motivations users experience to annotate images (Ames and Naaman, 2007). Reasons to tag include the wish for organizing the content for the general public (i.e. for photo pools, search, self-promotion), for self-organization (purpose of tagging to improve later retrieval) and social communication (adding context for family members, friends and the public).

It is in the operator's interest that eventually all artworks of the collection are tagged, i.e. that users are busy tagging artworks, because untagged artworks do not contain the desired user input, which is important for the operator, as it helps improve the folksonomy and create connections between artworks. Unfortunately, this goal is not easy to achieve, because tagging is a time-consuming task and without any incentive or help, the users' motivation to tag will decrease over time.

1.3 Aim of the Work

In order to be able to propose suitable strategies to meet this challenge, an extensive analysis of the explorARTorium's folksonomy related to art history has to be conducted, exploring the relationship between users and their tags. Questions regarding the relation of the users' tagging behavior to their liking of artworks are investigated, the users' vocabulary is qualitatively and lexically analyzed and the role of the user regarding tagging activity and learning effects is examined.

[7] `http://www.flickr.com`; [accessed 04-October-2011]

[8] `http://www.wga.hu`; [accessed 04-October-2011]

[9] `http://www.explorARTorium.info`; [accessed 04-October-2011]

[10] `http://vsem.ec.tuwien.ac.at`; [accessed 04-October-2011]

By giving the users of the artwork collection an incentive to tag pictures and therefore to prevent the users' tagging motivation from declining, this master's thesis provides a framework for system-generated suggestions for appropriate tags (based on tags extracted from similar artworks), which offers the user an easier way to describe the artwork. When the tag suggestion is presented to the user, he or she can easily accept or decline the suggested tag. Through this methodology not only the folksonomy of the explorARTorium is enriched, but also the user is invited to take a closer look at the artwork to find and verify the suggested tags in the artwork. Thereby he or she might explore previously unseen elements in the artwork, which might subsequently lead to new tags. Another application of the tag recommendations is the ability to present the user additional artworks in the same context if he or she wants to further explore artworks assigned with a particular tag. By giving the user the possibility to decline the recommended tag, the operator of the collection also gets useful information about what the user regards as an *in*appropriate description of the artwork. This information is of high value for the data mining process behind the suggestion engine, because "negative" tags are generally not available. With the help of business intelligence techniques like Frequent Itemset Mining and Association Rule Mining for discovering interesting relations between variables in the database, reliable decision criteria for the recommender system are provided.

1.4 Methodological Approach

The methodological approach of this master's thesis consists of three parts:

1. **Literature survey**: A literature research and comprehensive analysis of the state-of-the-art in the analysis of user-generated content, recommender systems and data mining are conducted.

2. **Experimentation**: This part of this master's thesis deals with the problem of untagged artworks. With the help of Frequent Itemset Mining and Association Rule Mining, suggestions for appropriate tags are generated. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases (Agrawal et al., 1993). The Tag Recommendation Framework for the explorARTorium is implemented in Java using the Waikato Environment for Knowledge Analysis (Weka[11], respectively the Weka Java Library) for data mining tasks.

3. **Evaluation**: The implemented Tag Recommendation Framework is evaluated and an extensive analysis of user-generated content of a folksonomy related to art history is conducted.

1.5 Structure of the Work

This thesis consists of five further chapters and starts with an introduction to business intelligence techniques in Chapter 2. Important data mining methods like Frequent Itemset Mining

[11] `http://www.cs.waikato.ac.nz/ml/weka/`; [accessed 04-October-2011]

and Association Rules Mining as well as Recommender Systems are discussed to provide the theoretical background of this thesis. Furthermore, applications of these techniques in the context of image databases and folksonomies are presented. An overview of related work in the fields of user-generated content, recommender systems and data mining is given.

Chapter 3 gives an overview of the explorARTorium project and analyzes the data set of a specific snapshot of the explorARTorium database, which serves as an underlying data basis for the evaluation of the implementation of the Tag Recommendation Engine in the following chapter.

User-generated content of a folksonomy related to art history is analyzed in Chapter 4, exploring the relationship between users and their tags.

In Chapter 5 the practical part of this thesis, i.e. the implementation of the Tag Recommendation Framework, is presented and the outcome is evaluated.

Finally, this thesis concludes with a summary and gives an outlook on future work in Chapter 6.

CHAPTER 2

Data Mining and Recommender Systems

The following sections cover the theoretical background and the technological aspects of this master's thesis. Based on introductions and definitions of data mining (especially association rules mining) in Section 2.1 and recommender systems in Section 2.2, this chapter draws the connection from the theory behind these techniques to the possibilities and opportunities their applications offer in the context of image databases in Section 2.3. Afterwards, related work regarding these topics is presented in Section 2.4.

2.1 Data Mining

Data Mining is an interdisciplinary subject in the field of computer science and can be seen as the process of discovering new patterns from large data sets by combining the use of statistical, artificial intelligence and database management techniques. Because of the interdisciplinarity and the wide range covered by analysis methods referred to as data mining techniques, many different definitions of data mining exist (Han et al., 2011). This section gives a general introduction to data mining presenting popular data mining techniques, their potential and applications laying special focus on association rules mining.

2.1.1 Introduction to Data Mining

According to Gupta (2006), data mining may be defined as follows:

> Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so that they may be used in an enterprise's decision making process.

Another definition of data mining is given by Fayyad et al. (1996):

> Data mining is a step in the KDD [Knowledge Discovery in Databases; note from the author] process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.

According to Witten and Frank (2005), data mining is about solving problems by analyzing data already present in databases and is defined as the process of discovering patterns in data. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data in order to extract these previously unknown interesting patterns.

In this context, the term "data warehouse" is often named. A widely accepted definition by Inmon (1992) is "an integrated subject-oriented and time-variant repository of information in support of management's decision making process".

A typical data mining process consists of the following six steps (Gupta, 2006):

1. **Requirements analysis**: First of all, the requirements are analyzed and the goals for the data mining process to achieve are formulated. It is of high importance that the goals are clearly defined and measurable, in order to evaluate the outcome of the data mining process.

2. **Data selection and collection**: Not every data source is suitable for a data mining process. Depending on the requirements and goals, the best available data sources and databases are selected in this step.

3. **Cleaning and preparing data**: Once appropriate data sources have been selected, the data is cleaned and prepared for the actual data mining task in this step. Often data mining tools require the data to be in a certain form with specific requirements (e.g. in a data warehouse, therefore this step is concerned with the adjustment or elimination of missing or corrupt data, conflicts and ambiguities to avoid incorrect results. This can be carried out through a particular ETL (extraction, transformation and loading) process, where "further data transformations deal with schema/data translation and integration, and with filtering and aggregating data to be stored in the warehouse" (Rahm and Do, 2000).

4. **Data mining exploration and validation**: In this step, a data mining model is constructed and different data mining exploration techniques and tools are applied to the data in an iterative process. Based on the evaluation of the results, a set of suitable techniques is selected and the data mining model is refined.

5. **Implementing, evaluating and monitoring**: This step can be seen as the actual implementation of the selected data mining model, where software may be developed for the visualization of results, the generation of reports and the explanation of the outcome of the data mining task. The results are analyzed and evaluated against the defined requirements and goals defined in the first step of the process by measuring accuracy and effectiveness of the implementation.

6. **Results visualization**: In this important last step of the process, the results are visualized and explained to the decision makers. Most of the common data mining tools include features to facilitate these tasks.

Of course there exist also models which suggest slightly different phases of a data mining process, e.g. Fayyad et al. (1996) group the KDD process into five stages, being "data mining" one of them:

1. Selection
2. Preprocessing
3. Transformation
4. Data Mining
5. Interpretation/Evaluation

Another example is the *Cross Industry Standard Process for Data Mining (CRISP-DM)* (see Wikipedia, 2011b) which is divided into six phases and shown in Figure 2.1.

Data Mining is being used for a wide variety of applications, including the following (see Gupta, 2006; Wikipedia, 2011c):

- **Business**: For businesses, several use cases for the application of data mining techniques have evolved:
  - **Relationship marketing**: Businesses are not interested in just a single sale, but in keeping their customers' loyalty to the company. Data mining can help improve the customer relationship management (CRM) to attract new customers, to retain the ones the company already has and to get former customers to return by improving the company's interaction with customers and identify reasons for client loyalty.
  - **Customer profiling**: By creating profiles of their customers, companies are able to improve their knowledge about their customers (e.g. their clients' interest in the products or services of the company), to determine their most valuable customers or to propose personalized offers to their customers.
  - **Customer segmentation**: By assigning customers into segments based on their profile, needs and status, businesses can target certain groups of customers directly, e.g. instead of sending an offer to all customers, the offer is only sent to people who are determined likely to respond. Gupta (2006) names also the promotion of cross-selling of services as an example of customer segmentation.
  - **Outliers identification and fraud detection**: Through discovering anomalies in large data sets, outliers, fraud or unusual cases can be identified. Allowing reductions in cost and risk, the detection of change and deviation is crucial for businesses.

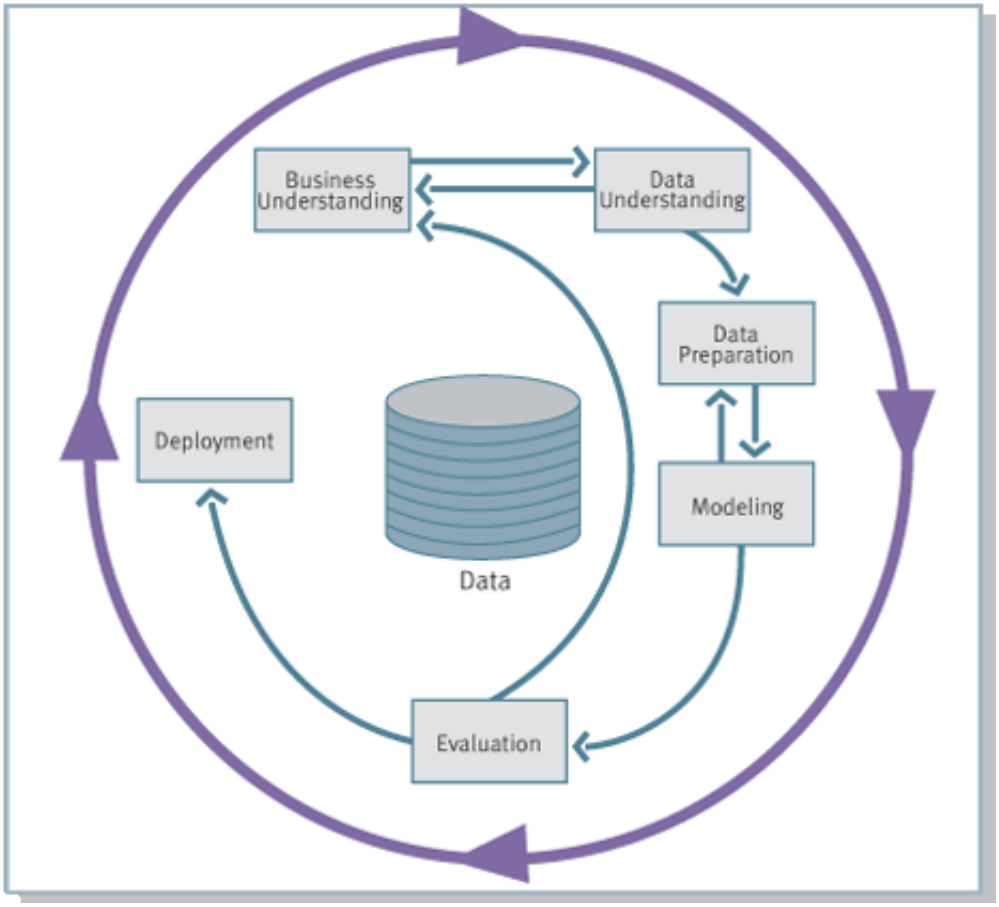

Figure 2.1: CRISP Data Mining Process Model (Gupta, 2006).

- **Spatial data mining**: The term *spatial data mining* refers to data mining techniques applied to spatial data with the aim to find geographic patterns in the data. Due to the immense explosion in geographically referenced data during the last years, the inclusion of Geographic Information Systems (GIS) into the data mining process yields enormous possibilities, e.g. health services searching for explanations of disease clusters - Google Flu Trends[1] being a popular example showing that certain aggregated search terms are good indicators of flu activity.

- **Surveillance**: The National Research Council[2] distinguishes between two different approaches of data mining techniques - pattern-based and subject-based - and gives the following definitions: "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related

[1] `http://www.google.org/flutrends/`; [accessed 04-October-2011]
[2] `http://www.nationalacademies.org/nrc/`; [accessed 04-October-2011]

to that initiating datum. Pattern-based data mining looks for patterns (including anomalous data patterns) that might be associated with terrorist activity - these patterns might be regarded as small signals in a large ocean of noise." (National Research Council, 2008, chap. 1.4)

- **Science and engineering**: In the fields of science and engineering, data mining is widely used in many areas, e.g. in medical science and bioinformatics for the study of human genetics (Zhu and Davidson, 2007, chap. 2) or in transportation for the analysis of traffic using self-organizing maps (SOM) (Chen et al., 2006).

- **Games**: For certain combinatorial games, the use of data mining provides insight into gameplay patterns by extracting human-usable strategies from oracle machines (abstract machines used to study decision problems) (Wikipedia, 2011c).

According to Fayyad et al. (1996), two primary goals of data mining techniques can be identified: description and prediction. Although the distinction between these two goal might sometimes be hard to make because the boundaries seem to be blurred, it can be stated that with description one aims to explore previously unknown patterns in data that help to explain certain cases, whereas with prediction, which is closely related to uncertainty, the focus lies on finding appropriate values for new instances of data in the future. Common classes of tasks for prediction and description based on data mining techniques involve the following as stated in Fayyad et al. (1996):

- **Dependency modeling**: The goal of dependency modeling is to build a model that describes significant relationships between variables. A common approach is the use of association rules mining (sometimes referred to as market basket analysis) to discover relationships between items in a large scale database, e.g. the sales data of supermarkets (Agrawal et al., 1993), and use this information for marketing purposes. Association rules are used in this master's thesis as data mining technique for the implementation of the Tag Recommendation Framework in Chapter 5 and therefore discussed in detail in Section 2.1.2.

- **Classification**: The aim of this supervised machine learning technique is to predict a class that an item of the data set is likely to belong. Classification is used if the classes are already known and some training data (items with characteristic attributes and known classes) are available in order to "train" the algorithm. The decision tree technique (see Quinlan, 1986) and the Naive Bayes method (see Wikipedia, 2011d) are some of the most widely used classification methods.

- **Clustering**: As a method of unsupervised learning, the aim of cluster analysis is similar to classification: the grouping of similar items in the data set. The methodology is, however, different: due to the fact that no training data are available and that the classes in the data are not already known, cluster analysis uses algorithms to explore the underlying structure of the data and to find clusters, in which the items of the data set can be classified. Popular cluster analysis techniques are hierarchical clustering (agglomerative; divisive) or partitional clustering (e.g. *k*-means clustering (MacQueen, 1967)).

- **Change and deviation detection**: "Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior"(Chandola et al., 2009). Applications of anomaly detection mechanisms include fraud detection for credit cards, insurance, or health care, intrusion detection for computer systems, fault detection in performance or safety critical systems, etc.

- **Regression**: "Regression is learning a function that maps a data item to a real-valued prediction variable" (Fayyad et al., 1996), i.e. regression analysis includes techniques for modeling several variables and analyzing the impact of changes on the dependent variable while varying the independent variables.

- **Summarization**: Summarization aims at finding a more compact representation of a data set and is often utilized in interactive exploratory data analysis and automated report generation.

2.1.2 Association Rules Mining

As noted earlier, association rules mining (or association rule learning) is a data mining technique for discovering interesting relations between variables in large databases. In this section, association rules are formally defined and several quality measures are presented. Afterwards, popular association rule mining algorithms are compared.

Businesses hold an enormous amount of enterprise data because of the growth in data due to online transaction processing (OLTP) data, credit and loyalty cards, the web and other sources and at the same time the growth in data storage capacity and decrease of processing costs. With the use of association rules mining, enterprises are able to identify patterns in their databases regarding customers and habits which help them to improve their customer relationship management. For example, the rule $\{\mathrm{X}, \mathrm{Y}\} \Rightarrow \{\mathrm{Z}\}$ indicates that if a customer buys the goods X and Y, he or she is also likely to buy good Z.

According to Agrawal et al. (1993, chap. 2), the formal model of association rules is defined as follows: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called "items". Let $T = \{t_1, t_2, \ldots, t_m\}$ be a set of N transactions called the "database". Each transaction in T has a unique identifier (TID) and contains a subset of the items in I (possibly a small subset). Let each transaction of m items be $\{i_1, i_2, \ldots, i_m\}$ with $m \leq n$. A "rule" is defined as an implication of the form $X \Rightarrow Y$ meaning whenever X appears, Y is also likely to appear, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short "itemsets") X and Y are called "antecedent" (left-hand-side or LHS) and "consequent" (right-hand-side or RHS) of the rule respectively.

Several measurements have been developed to indicate whether the discovered rules seem to be representative (i.e. if the rule is backed by enough transactions in the data set or if the consequence of the rule is likely to be true) and therefore are the most interesting rules from the (possibly very large) set of all possible rules.

The following measurements are discussed by Gupta (2006):

- **Support**: The support of X ($supp(X)$) is defined as the proportion of transactions in the

data set which contain X (see Equations (2.1) and (2.2)).

$$\text{supp}(X) = \frac{\text{Number of times } X \text{ appears}}{N} = P(X) \tag{2.1}$$

$$\text{supp}(XY) = \frac{\text{Number of times X and } Y \text{ appear together}}{N} = P(X \cap Y) \tag{2.2}$$

A high level of support is a good indicator that the rule might be of interest due to the high frequency. When choosing the minimum support threshold in an association rules mining process, one has to pay attetion to the *rare item problem*. Items that occur very infrequently in the data set are pruned according to the minimum support threshold, although they might still provide interesting and potentially valuable rules.

- **Confidence**: is defined in Equation (2.3) as the ratio of the items covered by the antecedent of the rule that are also covered by the consequence of the rule, in other words the proportion of the support for X and Y together to the support for X. Confidence can be interpreted as an estimate of the probability $P(Y|X)$, i.e. the probability of the consequence under the condition that the transactions contain the antecedent.

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(XY)}{\text{supp}(X)} = \frac{P(X \cap Y)}{P(X)} = P(Y|X) \tag{2.3}$$

- **Lift**: Originally called interest, lift is a measure of the importance of the association given in Equation (2.4), that is independent of the support, i.e. for the example of a shop "lift essentially indicates how much more likely an item Y is to be purchased if the customer has bought the item X that has been identified as having an association with the first item Y, compared to the likelihood of Y being purchased without the other item being purchased" (Gupta, 2006).

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)} = \frac{\text{supp}(XY)}{\text{supp}(Y) \times \text{supp}(X)} = \frac{P(X \cap Y)}{P(X) \times P(Y)} = \frac{P(Y|X)}{P(Y)} \tag{2.4}$$

- **Conviction**: The measure of conviction was introduced by Brin et al. (1997) and is defined as the proportion of the probability that X occurs without Y if X and Y are independent compared to the actual frequency of the occurrence of X without Y. In contrast to lift, conviction is a measure of implication because it is directional and maximal for perfect implications as shown in Equation (2.5).

$$\text{conv}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)} = \frac{P(X)P(\neg Y)}{P(X \cap \neg Y)} \tag{2.5}$$

- **Leverage**: measures the difference of X and Y appearing together in the data set to what would be expected if X and Y were statistically independent of each other. The definition of leverage was introduced by Piatetsky-Shapiro (1991) and is given in Equation (2.6).

$$\text{lev}(X \Rightarrow Y) = P(X \cap Y) - (P(X)P(Y)) \tag{2.6}$$

In the following chapters, the focus lies on constraints of minimum thresholds on support and confidence for finding interesting rules.

Each of the algorithms that are presented in the following sections can be divided into two parts:

1. In the first part, *frequent itemsets* according to the specified minimum support threshold are mined.
2. In the second part, association rules are generated out of the itemsets according to the specified minimum confidence threshold.

Whereas the second part is relatively straightforward and the same for all the different algorithms, the implementation of the first part to find frequent itemsets differs from algorithm to algorithm.

Naive Algorithm

The simplest possible algorithm for association rule mining is the naive brute force algorithm, which explores all the possible combinations of items in the database and calculates their frequencies. Then the algorithm checks the frequencies of the itemsets against the minimum support threshold. Those which pass the requirements are taken into account for the derivation of rules. Since this naive approach needs huge amounts of memory ($2^n - 1$ with n items excluding the null combination which is not a valid itemset) and time (every possible combination of items is evaluated, even if the combination does not occur at all in the transactions database), the naive algorithm is not suitable for larger problems and other algorithms are used instead.

Apriori Algorithm

The Apriori algorithm was proposed by Agrawal and Srikant (1994) and is the best-known algorithm to mine association rules. Apriori uses a "bottom up" approach, where frequent subsets are extended by one item at a time (candidate generation) and groups of candidates are tested against the data. It uses a breadth-first search strategy and a tree structure to count candidate itemsets efficiently. The algorithm by Agrawal and Srikant (1994) is discussed in detail below, using the following notation according to Gupta (2006):

- A k-itemset is a set of k items.
- The set C_k is a set of potentially frequent candidate k-itemsets.
- The set L_k, which is a subset of C_k, is the set of k-itemsets that are frequent.

1. **First Part - Finding Frequent Itemsets**: For the first part of the Apriori algorithm, a set of transactions is used as input and a value of $p\%$ has to be chosen for the minimum support threshold. The search for frequent itemsets starts with sets containing only one item and is continued iteratively with k-itemsets until no itemsets meeting the specified support criterium are found. Therefore candidate sets are generated in every iteration and checked against the minimum support threshold $min\_support$.

Step 1 In the first step, every transaction is examined and all frequent items (with $support > min\_support$) are found. The collection of these frequent items is called L_1 ($k = 1$).

Step 2 This step is also called the Apriori-gen function, which increments k and takes L_{k-1} as input parameter and returns a set of all candidate k-itemsets. This is done by "building potential sets of k items from L_{k-1} by using pairs of itemsets in L_{k-1} such that each pair has the first $k-2$ items in common. Now the $k-2$ common items and the one remaining item from each of the two itemsets are combined to form a k-itemset" (Gupta, 2006). These potentially frequent k-itemsets are called the candidate set C_k. Instead of calculating the support for all itemsets, the Apriori-gen delivers a subset of already found frequent itemsets to reduce the amount of support calculations.

Step 3 In this third step, the k-itemsets in C_k that are frequent, i.e. have $support > min\_support$, are found and the other k-itemsets are pruned. The resulting set is named L_k.

Steps 2 and 3 are repeated until no further frequent itemsets are found.

The result after this first part contains all frequent itemsets.

2. **Second Part - Forming the Rules**: To find the association rules from the frequent itemsets computed in the first part, the idea behind the Apriori-gen function is used again. The algorithm generates association rules for every found frequent itemset by starting with conclusions containing one item, which are magnified iteratively and checked against the specified minimum confidence threshold $min\_confidence$:

   a) Generate association rules $X \rightarrow Y$ with $\mid Y \mid = 1$ and $X = Z - Y$ with $\text{Confidence}(X \rightarrow Y) > min\_confidence$ for every itemset Z.

   b) Generate H_1 with each itemset containing one conclusion.

   c) Generate H_k by using Apriori-gen.

   d) For each conclusion $h_k \in H_k$ check if $min\_confidence < \text{Confidence}(Z - h_k) \rightarrow h_k$. If the inequation evaluates to false, h_k is removed from H_k.

   e) Terminate if H_k is empty.

   f) Repeat steps c) to e) for every k, then return $\bigcup H_k$.

   All of the generated association rules comply with the minimum support and confidence thresholds.

For large sets of transactions having large sets of frequent items, the Apriori algorithm is very resource intensive, due to the following reasons: huge candidate sets are derived because the number of candidate itemsets expands quickly; many scans of the database are required; redundant rules are generated and the Apriori algorithm is inefficient for dense data. To overcome these drawbacks, several improvements have been considered useful, resulting in variations of the Apriori algorithm (e.g. Apriori-TID) or new techniques (e.g. FP-growth), which are discussed in the following sections.

Apriori-TID Algorithm

Like the Apriori algorithm, the Apriori-TID algorithm also uses the Apriori-gen function to determine the candidate itemsets but the database is not used for counting the support after the first pass. Instead, Apriori-TID generates a separate (usually smaller) database by encoding the candidate itemsets used in the previous pass (T_k). Only transactions containing candidate k-itemsets are maintained in T_k, with the result that the size of T_k is decreasing with k being incremented.

The exact mode of operation is outlined in Gupta (2006):

1. Scan the entire database to derive T_1 along with the transaction ID (TID) for every item of the itemsets.
2. Calculate L_1 on the basis of T_1.
3. Compute C_2 with Apriori-gen function.
4. Calculate the support for C_2 with the help of T_1.
5. Compute T_2.
6. Generate L_2 from C_2 and C_3 from L_2.
7. Compute T_3 on the basis of T_2 and C_3.
8. Repeat this algorithm until the k-itemset contains no more items.

As noted earlier, both Apriori as well as Apriori-TID use the Apriori-gen function to generate candidate sets. The Apriori algorithm shows better performance during the start-up phase (when k is still small), because entries in T_k (used by Apriori-TID) may be larger than the entries in the database (used by Apriori) accordingly, whereas Apriori-TID proves more suitable in later stages with higher k, due to its ability to filter unnecessary candidate itemsets.

DHP Algorithm

The direct hashing and pruning (DHP) algorithm shows similarities to the Apriori algorithm, but in contrast, uses a hash-based approach in order to reduce the number of candidate k-itemsets computed in the first pass. According to Gupta (2006), the DHP algorithm generates frequent itemsets efficiently and tries to trim the database by pruning transactions which do no have to be scanned in subsequent passes because the necessary minimum support threshold is not met. The algorithm can be divided into three parts and works as follows:

- **Part 1**: In the first part of the DHP algorithm, each transaction in the database is scanned (as the algorithms discussed earlier do), with the difference that at the same time of the scan, all the possible 2-itemsets are hashed to a hash table. A bit vector assigned to the hash table keeps track of the number of items in each bucket of the hash table, signaling 1 if the minimum support threshold has been reached and 0 otherwise.

- **Part 2**: The second part of this algorithm consists of two phases. In the first phase, the hash table is used to reduce the number of candidate itemsets when C_k is generated. As mentioned before, the bit vector reveals if an item is included in C_k. Due to the fact that collision may occur during the hashing process, there is no guarantee that the itemset is in fact frequent, but nevertheless the size of C_k is reduced. Each itemset of C_k is inserted into a hash tree, which comes into use during the second phase of this part, where unnecessary itemsets are pruned, reducing the number of itemsets as well as transactions, which do not contain frequent itemsets anymore.

- **Part 3**: In the third and last part of the DHP algorithm, the pruned database is taken as basis for the computation of the support for each itemset. For the itemsets that have been classified as frequent, candidate itemsets are generated and this part of the algorithm is repeated until no more candidate itemsets are found.

It is worth noting that the DHP algorithm shows better performance than the Apriori algorithm during the early stages (e.g. by generating L_2). Due to the fact, that pruning of the transaction database is carried out at every pass, the efficiency of the DHP algorithm is improved.

FP-growth Algorithm

The frequent pattern-growth algorithm, introduced in Han et al. (2000), depicts a method which avoids candidate generation-and-test and uses a new data structure to reduce the cost in frequent-pattern mining. This compact data structure, the FP-tree, which is "an extended prefix-tree structure storing crucial, quantitative information about frequent patterns" (Han et al., 2004), is constructed during the early stages of the algorithm. "To ensure that the tree structure is compact and informative, only frequent length-1 items will have nodes in the tree, and the tree nodes are arranged in such a way that more frequently occurring nodes will have better chances of node sharing than less frequently occurring ones." (Han et al., 2004). FP-growth implements a divide-and-conquer approach to split the mining as well as the database containing the transactions and itemsets into smaller pieces and subtasks in contrast to the Apriori algorithm.

The algorithm for the first part, the generation of an FP-tree as described in Gupta (2006), is as follows:

1. Similar to the other algorithms presented earlier, the whole database is scanned to find all frequent items, which are sorted by their support in descending order.

2. The FP-tree is initialized with an empty root.

3. The following step is repeated for every transaction in the database:

    - All non-frequent items are removed from the current transaction. The remaining already-sorted items are inserted into the FP-tree in that particular order along with their frequency, each node representing a frequent item. If the item already exists in the tree, the item count is increased.

When the algorithm terminates, the resulting FP-tree has the following properties:

- The nodes near the root of the tree are more frequent than the nodes near the leaves.
- The height of the tree is equal to the maximum number of items in a frequent itemset minus 1 (the root node).
- Identical itemsets are represented only once in the tree.

The second part of the algorithm is concerned with the mining of the previously generated FP-tree for frequent items, called FP-growth. By exploring the conditional pattern base for each item, i.e. finding all patterns (paths in the tree) leading to the particular item, a new conditional frequent pattern tree is generated, containing only frequent items. To find all possible combinations of frequent itemsets, it is necessary to start this process with the least frequent items due to the fact that these are stored in the leaves of the tree.

The advantages of the FP-tree algorithm are the complete avoidance of the costly candidate generation and the fact that the database has to be scanned only twice. The FP-growth algorithm can prove its superiority against the Apriori-algorithm especially in situations where the minimum support threshold is very low, because a low support threshold results in large candidate sets, with which the FP-tree does not have to deal with.

2.2 Recommender Systems

In this section, an introduction to recommender systems is given. Firstly, the term *recommender system* is defined. Secondly, different functions of recommender systems are identified and discussed. Finally, popular recommendation techniques are explored and their characteristics are analyzed.

2.2.1 Introduction to Recommender Systems

Recommender systems are tools that support users in their decision making process and aim to provide recommendations of high quality via easy accessibility for a large user community (Jannach et al., 2010). Ricci et al. (2010) define recommender systems as follows:

> "Recommender Systems (RSs) are software tools and techniques providing suggestions for items to be of use to a user. The suggestions relate to various decision-making processes, such as what items to buy, what music to listen to, or what online news to read.
>
> "Item" is the general term used to denote what the system recommends to users. A RS normally focuses on a specific type of item (e.g., CDs, or news) and accordingly its design, its graphical user interface, and the core recommendation technique used to generate the recommendations are all customized to provide useful and effective suggestions for that specific type of item."

According to Wikipedia (2011f), the following characteristics of a recommender system can be added to the previous definition:

> "Typically, a recommender system compares a user profile to some reference characteristics, and seeks to predict the 'rating' or 'preference' that a user would give to an item they had not yet considered. These characteristics may be from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach)."

2.2.2 Recommender Systems Function

It is stated in Ricci et al. (2010), that the "recommendation problem can be defined as estimating the response of a user for new items, based on historical information stored in the system, and suggesting to this user *novel* and *original* items for which the predicted response is *high*." Based on this definition, several functions, features, use cases and reasons why recommender systems are used by *businesses* or *service providers*, can be derived:

- **Increase the number of items sold**: This is one of the most popular goals a service provider tries to achieve with a recommender system. Through the recommendations of possible useful items to a user (e.g. on basis of his or her profile), her or she might consider purchasing the recommended item(s), resulting in an increase of sold items for the service provider.

- **Sell more diverse items**: In cases where a service provider is interested in not just selling popular items but also niche products, a recommender system can be used to present personalized recommendations for unpopular items the user might be interested in according to his or her profile or previous buying history.

- **Increase the user satisfaction**: The system provider is able to increase the user's satisfaction by improving the user experience and the human-computer interaction with a well-designed system. This will in turn motivate the user to use the system more often and interact with the recommendations, i.e. take a closer look at them, and therefore increase the probability that the user accepts the recommendations.

- **Increase user fidelity**: The ability of a recommender system to recognize a returning user and to steadily update the profile of the user by improving the accuracy of preferences, the loyalty of the user to the website the recommender system works for will grow, resulting in longer interactions with the website, the system and the recommendations.

- **Better understand what the user wants**: Through the knowledge the recommender system gathers about the user regarding his or her preferences and through the information the system collects about its users, which is represented in models or profiles, the service provider is able to improve other parts of his business, e.g. customer relationship management or warehouse management.

In contrast to the previous listing, where goals of recommender systems, which businesses, service providers, marketers and other stakeholders pursue, are listed, Herlocker et al. (2004) identify goals and tasks for *end-users* a recommender system might fulfill. For the evaluation

of a recommender system, it is important to analyze which of these tasks are targeted with the particular recommender. Whereas the first two goals, *Annotation in Context* and *Find Good Items* are the most popular tasks, several other tasks emerged over time, some of them also aiming at the rating functions of a recommender system, like *Improve Profile* or *Express Self*.

- **Annotation in Context**: In situations, where the user views an ordered sequence of items (e.g. messages, news, TV program, etc.), the aim of *Annotation in Context* is to recommend the user some of these ordered items through highlighting or annotating but retaining the order and context of items.

- **Find Good Items**: The user is presented a ranked list of recommendations, along with predictions showing the probability that the recommendations are accurate and fit the user's needs.

- **Find All Good Items**: In some cases it is not enough to present the user just some of the probably useful items (as noted previously with *Find Good Items*), but to give the user a complete list of all "good items" the recommender system generates (e.g. for medical, financial or legal applications).

- **Recommend Sequence**: To accomplish this task, the recommender system generates suggestions not only for single items, but for items that can be consumed in a particular sequence, e.g. TV shows or recommendations for books which should be read in a particular order.

- **Just Browsing**: Herlocker et al. (2004) found out that many users just like to browse recommendations and enjoy using the recommender system without having purchase intentions. Nevertheless, if the "just browsing" goal is well implemented in the recommender system (e.g. well-thought-out user interface and interaction design), it is likely to increase the user's satisfaction.

- **Improve Profile**: By improving their profile (i.e. their representational model of preferences visible to the recommender system), users aim to improve the quality of recommendations, because they are aware of the fact that an up-to-date profile may result in more personalized recommendations.

- **Express Self**: Some users may only wish to express themselves by contributing their ratings to the system. For instance, this user behavior can be observed at Amazon[3], where users rate different products to share their opinion with other users.

- **Help Others**: With the intention to help other users by rating items, some users feel that they can help the community and that other users benefit from their contribution. Similar to this aspect of rating items for the benefit of the community, the social phenomenon of tagging can be observed regarding the Web 2.0, as noted earlier in Section 1.2.

[3] `http://www.amazon.com`; [accessed 04-October-2011]

- **Influence Others**: According to Herlocker et al. (2004), there also exists the risk that the recommender system is being manipulated with fake ratings by users whose intention is to influence other users according to their point of view.

2.2.3 Recommendation Techniques

The core competence of a recommender system is the ability to identify the usefulness of an item for the user. Therefore, the utility of an item has to be computed in form of a prediction, in order to be able to compare the utility of a set of items to determine the most valuable recommendations. According to Ricci et al. (2010), the degree of utility of the user c for an item i is modeled through the function $R(c, i)$. The job of the recommender system is to predict the value of R over pairs of users and items; this estimation is denoted by $\hat{R}(c, i)$. Once the values of $\hat{R}$ have been computed for a set of items $i_1 \dots i_N$ in relation to a specific user c, the system filters the items and recommends $i_{j1} \dots i_{jK}$ with the highest computed utility. This is formally specified by Adomavicius and Tuzhilin (2005) in Equation (2.7): for each user $c \in C$, the item $i' \in I$ which maximizes the user's utility function $u(c, i)$ is chosen.

$$\forall c \in C,\ i'_c = \arg\max_{i \in I} u(c, i) \tag{2.7}$$

For the analysis of a recommender system, different types of systems can be distinguished depending on the applied techniques to generate recommendations. Since the mid-1990s, recommender systems have become an important research area, and according to Jannach et al. (2010); Ricci et al. (2010), the following approaches emerged:

- Content-based recommendation
- Collaborative recommendation
- Demographic recommendation
- Knowledge-based recommendation
- Community-based recommendation
- Hybrid approaches

These approaches are now discussed in detail in the following sections.

Content-based recommendation

Systems implementing a content-based recommendation approach learn to recommend items to the user that correspond to the items the user liked in the past by analyzing a set of items which have been rated or purchased (in electronic commerce scenarios) previously by the user. With this information, a model (or profile) of the user can be built, representing the user's preferences. Attributes or features of the items are compared to determine the similarity. During the recommendation process, the attributes of the user profile are compared and matched against the attributes of other objects. The resulting recommendations reflect the characteristics of the

user's profile. The better the preferences of the user are captured in his or her profile, the higher the effectiveness of the recommender system will be. For example, if the user has rated or purchased CDs of a particular music genre in the past, the content-based system will recommend other CDs matching that style of music to the user.

Collaborative recommendation

In contrast to content-based recommendation techniques, where recommendations of items for a user are generated based on the history of similar items the user liked/rated/purchased in the past, collaborative filtering methods recommend items to the user that other users with similar tastes liked in the past. To determine similar tastes, the histories of the users are compared and analyzed for correlations, e.g. with the help of neighborhood-based methods for collaborative filtering (see Sarwar et al., 2001).

Demographic recommendation

With this approach, recommendations for items are generated by the recommender system based on the available demographic data of the user. The idea behind demographic-based recommendations is that it might be useful to determine the recommendations not only based on the preferences contained in the user profile, but also based on the language, geographic location, age or gender of the user, resulting in different suggestions for different demographic properties. Typical use cases are recommendations customized to the age of the user or personalized websites based on the user's language or country.

Knowledge-based recommendation

Recommender systems applying a knowledge-based approach use knowledge about users and products to generate a recommendation, reasoning about what products meet the user's requirements (Burke, 2000). Whereas other recommendation approaches like the collaborative filtering technique experience problems at the start-up period of the recommender, e.g. the collaborative filtering system has to be filled with a large amount of ratings, a knowledge-based recommender system avoids this, since its recommendations do not depend on a base of user ratings. The knowledge-based system does not have to collect information on particular users because the algorithm is independent of individual tastes (see Burke, 2000).

Community-based recommendation

The idea behind a community-based recommendation approach is based on the assumption that users tend to rely more on personal suggestions for appropriate items (e.g. by family members, friends, or personally-known persons in general) than on suggestions provided by other anonymous users. Recommender systems using this technique, sometimes also referred to as social recommender systems, explore the relationship between the active user and his or her friends on that platform. Items are recommended to the user based on the ratings derived from the user's friends. Since the increasing popularity of social networks, a lot of research is currently

ongoing concerning community-based recommender systems (see Siersdorfer and Sizov, 2009; Shepitsen et al., 2008; Golbeck, 2006; Guy et al., 2009).

Hybrid approaches

Hybrid recommender systems combine two or more of the above listed approaches and try to compensate the drawbacks of a particular technique through the advantages the other technique(s) offer(s). Ricci et al. (2010) take the *new-item problem* as example of the useful deployment of a hybrid recommender system: collaborative filtering methods have to deal with the *new-item problem*, i.e. an item that has not been rated by users is not included in the recommendation process and therefore not suggested to other users. By combining the collaborative filtering technique with the content-based technique to a hybrid recommender systems, this problem is overcome, since this approach generates recommendations based on the descriptions or features of an item. In Burke (2002) it is confirmed that semantic ratings generated by a knowledge-based recommender system can help to enhance the effectiveness of collaborative filtering system. A study comparing several hybrid web recommender systems is presented in Burke (2007).

2.3 Applications in the context of image databases / folksonomies

In this section, applications of the data mining and recommender systems techniques discussed earlier in this chapter are presented in the context of image databases, folksonomies and social tagging tools in general.

2.3.1 Introduction to Social Tagging and Folksonomies

Social tagging tools are rapidly emerging on the Web. According to Wikipedia (2011g), the term *tag* in online computer systems terminology is defined as "a non-hierarchical keyword or term assigned to a piece of information (such as an Internet bookmark, digital image, or computer file). This kind of metadata helps describe an item and allows it to be found again by browsing or searching. Tags are generally chosen informally and personally by the item's creator or by its viewer, depending on the system". It is also stated that tagging has become popular by websites associated with Web 2.0 and can be seen as an important feature of many Web 2.0 services. The collection of all the assigned tags for a specific user is called *personomy*, whereas the combination of all personomies results in a *folksonomy*[4], a term coined by Thomas Vander Wal, a portmanteau word combining *folks* and *taxonomy*. A lot of research has been done lately analyzing the possibilities of data mining technologies in social tagging systems (Schmitz et al., 2006; Sigurbjörnsson and van Zwol, 2008; Hotho, 2010).

Schmitz et al. (2006) discuss in their work the mining of association rules in folksonomies, where the following formal definition of a folksonomy is given:

Definition 1. A *folksonomy* is a tuple $\mathbb{F} := (U, T, R, Y, \prec)$ where

[4] `http://vanderwal.net/folksonomy.html`; [accessed 04-October-2011]

- U, T and R are finite sets, representing the elements of *users*, *tags* and *resources*
- Y is a ternary relation between them, i.e. $Y \subseteq U \times T \times R$, called assignments
- $\prec$ is a user-specific *subtag/supertag-relation*, i.e. $\prec \subseteq U \times ((T \times T)/\{(t,t)|t \in T\})$.

The personomy $\mathbb{B}_u$ of a specific user $u \in U$ is the restriction of $\mathbb{F}$ to u, i.e. $\mathbb{P}_u := (T_u, R_u, I_u, \prec_u)$ with $I_u := \{(t,r) \in T \times R|(u,t,r) \in Y\}$, $T_u := \pi_1(I_u)$, $R_u := \pi_2(I_u)$ and $\prec_u := \{(t_1,t_2) \in T \times T|(u,t_1,t_2) \in \prec\}$.

2.3.2 Types of resource sharing systems

The social resource sharing systems can be divided into different categories, depending on the type of content that is shared, the most popular being the following:

- **Images and photos**, e.g. Flickr[5]
- **Bookmarks**, e.g. del.icio.us[6]
- **Bibliographic references**, e.g. CiteULike[7] or Zotero[8] (for instance, Zotero is used to collect, organize and manage the research sources for this master's thesis)
- **Personal goals** in private life, e.g. 43 Things[9]

With the success of these systems, the amount of information maintained by them is increasing steadily. In order to be able to cope with the growth of information and the need to organize the resources, improvements concerning the structuring of the content are necessary. According to Schmitz et al. (2006), the first step towards more structure is to "discover knowledge that is already implicitly present by the way different users assign tags to resources". By projecting the three-dimensional data set of the folksonomy (users, tags and resources) to a two-dimensional one (items and transactions), the use of association rule mining allows to generate a hierarchy of the already existing tags as well as the recommendation of additional tags to resources.

2.3.3 Examples of applications

This concept of applying data mining on folksonomies is also discussed by Hotho (2010), where two aspects approving the application of mining techniques on folksonomies are identified:

- For the process of ontology learning, folksonomies can be a useful source due to their rich source of data.
- The analysis of folksonomy data by using data mining techniques can be seen as Semantic Web Mining.

[5] `http://www.flickr.com`; [accessed 04-October-2011]
[6] `http://del.icio.us`; [accessed 04-October-2011]
[7] `http://www.citeulike.org`; [accessed 04-October-2011]
[8] `http://www.zotero.org`; [accessed 04-October-2011]
[9] `http://www.43things.com`; [accessed 04-October-2011]

The aim of the application of data mining techniques lies therefore in the effort to bridge the gap between folksonomies and the Semantic Web by extracting hidden information in the data sources and improving the understanding of the hidden semantics. In the following, concrete applications and use cases of these techniques in the context of folksonomies are discussed (Hotho, 2010; Sigurbjörnsson and van Zwol, 2008):

Spam Detection

Similar to the problems search engines have to face with web spam (i.e. manipulation of search results via fake web pages or manipulation of the popularity and therefore the ranking of web pages), social tagging systems show similar vulnerabilities, which may by exploited by spammers, e.g. a social bookmarking system, where spammers are able to easily create entries and bookmark web pages they want to promote. In terms of data mining techniques, this problem can be understood as a binary classification task (cf. Section 2.1). The classification model is trained to distinguish between two types of bookmarks (or posts), which are inserted as new content into the system: *spam* and *non-spam*. This distinction is made by the classifier based on the idea that spammers reveal their identity by using a similar vocabulary and similar resources. Together with additional information such as the IP address, the classification model can be trained according to these "features" and predict the probability of spam for new instances inserted into the system.

Ranking in Folksonomies

Due to the recent growth of folksonomies, a ranking of items can improve the user experience by identifying popular topics or trending posts, similar to the Trending Topics[10] on Twitter[11]. In Hotho (2010), an algorithm named *FolkRank* is presented, which reflects the idea of the *PageRank* algorithm introduced in Brin and Page (1998), but takes the special structures of folksonomies into account. The underlying principle of the algorithm is as follows: the importance of a resource is affected by the importance of the tags which are assigned to it and the importance of the users who tag it (similar to PageRank where the importance of a web page is determined by the number of hyper links pointing to it and the importance of the referring pages). By computing a topic-specific ranking FolkRank overcomes the problem that the rankings only mirror the overall trends of the folksonomy and do not respond to the preferences of the user. The FolkRank algorithm is implemented in BibSonomy[12], a social bookmark and publication sharing system.

Tag Recommendation

The application of recommending tags to users constitutes a combination of the techniques presented in Section 2.1 (data mining) and Section 2.2 (recommender systems). Recommender

[10] `https://support.twitter.com/articles/101125-about-trending-topics`; [accessed 04-October-2011]

[11] `http://www.twitter.com`; [accessed 04-October-2011]

[12] `http://www.bibsonomy.org/`; [accessed 04-October-2011]

systems can be used in folksonomies to recommend similar users, interesting resources or help the user by providing suitable keywords to describe the resource. This is for instance implemented in the social bookmarking system del.icio.us and the online photo service Flickr. In Sigurbjörnsson and van Zwol (2008), tag recommendation based on collective knowledge for images on Flickr is discussed. With the analysis of a Flickr data set containing 52 million publicly available photos with annotations, it is confirmed that the tag frequency distribution follows a power law and the majority of photos is tagged with only a few keywords. Users assign tags over a broad semantic spectrum, i.e. locations, persons, things, time or simply impressions of the photo. Based on these findings, a tag recommendation strategy is developed, incorporating ideas of tag co-occurrence and tag aggregation and promotion. By using tag co-occurrence, i.e. measuring the number of photos where two tags are used in the same annotation, a list of candidate tags for each user-defined tag is generated. With the help of tag aggregation, these lists are merged into a single ranking. Along with a "promotion function", the most descriptive tags are promoted for recommendation incorporating the ranking of tags. A total of four different recommendation strategies is presented, including "Vote" and "Sum". The evaluation of the tag recommendation systems shows that the algorithm is particularly good at recommending locations, artifacts and objects. Due to the fact that the system is based on the statistical patterns of the Flickr data set, a change in the vocabulary of the users (which is likely to occur because social tagging systems evolve continuously) does not pose a great challenge. However, the presented tag recommendation system is unable to recommend tags for photos which have not been tagged so far, due to the underlying principle of tag co-occurrence. It is one of this master's thesis goals to overcome this shortcoming and to generate tag recommendations for untagged images.

2.4 Related Work

In this section, an overview of related work to the topics of this master's thesis is presented.

A number of machine learning frameworks has been proposed to address the problem of automatic tag recommendation for both text and digital data on the web (Li and Wang, 2006; Song et al., 2008b), showing that a single computer processor can suggest tags in real-time with good accuracy by applying training algorithms for semantic concepts. Chirita et al. (2007) suggested a method named P-TAG for automatically generating personalized tags by extracting keywords from similar documents for recommendation. Applications of this approach include personalized web search, web recommendations for desktop tasks and ontology learning. An alternative keyword-oriented approach finds the co-occurrence of terms in different documents and recommends the remaining tags from similar documents (Song et al., 2008a) with the help of multi-class sparse Gaussian process classification. Advanced algorithms for the aggregation of preferences in recommender systems, offering more flexibility and adaptability than standard functions such as arithmetic mean or minimum/maximum functions are analyzed by Ricci et al. (2010).

The phenomenon of tagging is discussed in Ames and Naaman (2007), giving explanations for possible motivations for annotation in mobile and online media and concluding that in particular, social incentives for tagging appear to be surprisingly important in motivating users to

tag their photographs.

Since its introduction by Agrawal et al. (1993), the research field of association rule mining has brought up a lot of different techniques (e.g. Apriori (Agrawal and Srikant, 1994), Dynamic Itemset Count (DIC), Partition, FP-growth). Today there are several efficient algorithms that cope with the popular and computationally expensive task of association rule mining (Hipp et al., 2000). Several measures to evaluate the interestingness of association rules exist, but each of them is useful for some applications, but not for others. Tan et al. (2004) discuss several key properties which should be examined in order to select the right objective measure for a given application.

According to Yang et al. (2010), automatic media tagging plays a crucial role in modern tag-based media retrieval systems. Existing tagging schemes mostly perform tag assignment based on community contributed media resources, where the tags are provided by users interactively. However, such social resources usually contain dirty and incomplete tags, which severely limit the performance of these tagging methods. Yang et al. (2010) propose an automatic image tagging method aiming to automatically discover more complete tags associated with information importance. Because of the sensitivity of parameters used in their approach due to the used parametric data mining techniques, it is difficult to extend it to general cases.

Wang et al. (2009) explore the correlation between classification and annotation and develop a probabilistic model for jointly modeling the image, its class label, and its annotations, guided by the intuition that classification and annotation are related.

Active learning is a supervised machine learning technique that learns a model in an interactive way. The learning algorithm can actively query the user for labeling data and is able to select the most representative data. Through this iterative process active learning techniques are capable to reduce human annotation effort or to achieve better results with the same effort (Wang and Hua, 2011).

A completely different approach to search for digital images in large databases constitutes content-based image retrieval (CBIR). With the help of computer vision techniques the actual content of the image is analyzed in contrast to other methods where the metadata of the image is taken into account (Smeulders et al., 2000). A comprehensive survey of different image retrieval techniques is conducted in Datta et al. (2008).

CHAPTER 3

explorARTorium

This chapter introduces the explorARTorium, a multimedia platform for user-generated textual annotations to artworks, in Section 3.1. In Section 3.2, a specific data set extracted from the database of the explorARTorium is analyzed, which serves as a data basis for the following Chapter 5 (implementation of the Tag Recommendation Framework).

3.1 explorARTorium - the Project

The explorARTorium is an interactive environment that allows users to navigate through art history. Various levels of information concerning artworks are placed in context, such as title, artist, theme, time, geographical area in which the artwork was created, etc. VSEM (2011) gives an overview of the platform, its design and goals. The explorARTorium is maintained as part of the research project "The Virtual 3D Social Experience Museum" (VSEM)[1] of the Electronic Commerce Group[2] at the Institute of Software Technology and Interactive Systems[3] at the Vienna University of Technology[4] and is funded by the FWF (Fonds zur Förderung der wissenschaftlichen Forschung / Austrian Science Fund), Project No. L602.

Historically speaking, the explorARTorium has its roots in the Tagging-tool[5], where users are able to annotate textual information to artworks completely unbiased, because no information about the artwork (artist, title, etc.) is provided and the previously assigned tags for the artwork by other users are only displayed if the user opts for it. A screenshot of the Tagging-tool is shown in Figure 3.1. The artwork itself is placed in the center of the page, whereas assigned tags appear on the right side of the artwork. At the time of taking this screenshot, the following tags were assigned to the artwork (given in the original language, with the English translation in brackets if necessary): *sense* (scythe), *engel* (angel), *angst* (fear), *maske* (mask), *mask*, *lion*, *earth*, *kugel*

[1] `http://vsem.ec.tuwien.ac.at`; [accessed 04-October-2011]
[2] `http://www.ec.tuwien.ac.at`; [accessed 04-October-2011]
[3] `http://www.isis.tuwien.ac.at`; [accessed 04-October-2011]
[4] `http://www.tuwien.ac.at`; [accessed 04-October-2011]
[5] `http://vsem.ec.tuwien.ac.at/taggingtool/`; [accessed 04-October-2011]

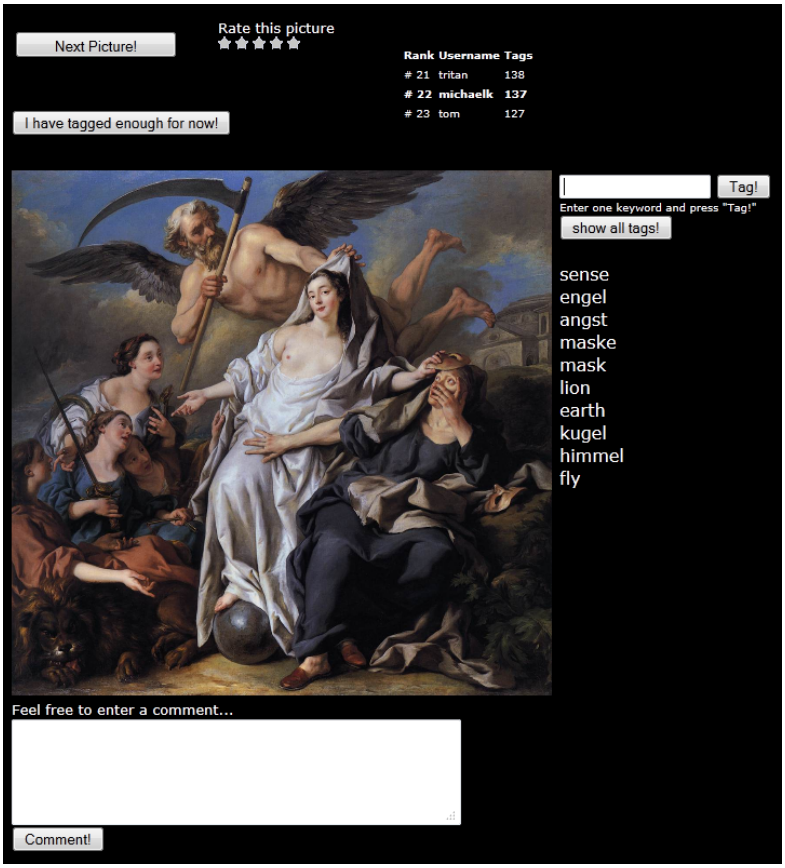

Figure 3.1: Screenshot of the Tagging-tool.

(globe, sphere), *himmel* (sky, heaven) and *fly*. This example of both German and English tags also illustrates the multilingualism of the explorARTorium and the subsequent advantage of allowing users to explore and search the collection in different languages. According to Arends et al. (2011), users of the Tagging-tool assigned 80,000 tags to the artworks between October 2010 and January 2011, which was a huge success. But within the term, the users' motivation to tag artworks decreased, so the team behind VSEM decided to take the Tagging-tool to the next level in order to give the users feedback and a reason to return again to the platform: the idea of the explorARTorium was born.

The explorARTorium enables the user to discover an artwork not only isolated (as with the Tagging-tool), but embedded in a greater context within art history, which encourages the user to scrutinize different artworks and thus comprehend that each artwork is part of a larger environment (Arends et al., 2011). The possibility for visitors to tag artworks, i.e. assign descriptive keywords to the artworks, is the key feature of the explorARTorium being used in this master's thesis.

Visitors of the explorARTorium are able to explore about 20,000 paintings. Figure 3.2 shows a screenshot of the explorARTorium. An artwork (in this case *The Last Supper* by *Leonardo Da*

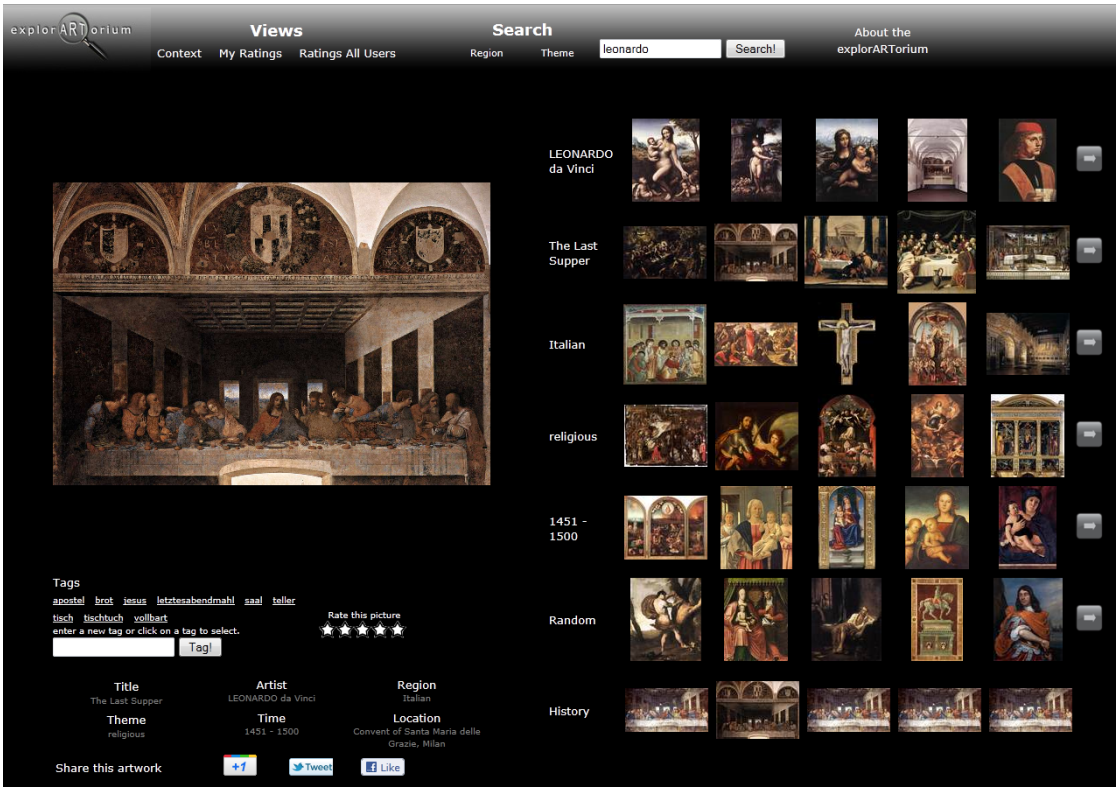

Figure 3.2: Screenshot of the explorARTorium.

Vinci) is notably presented in the left half of the gallery. Further information such as the artist name, title, etc is available in a fold-out menu below the artwork, so that visitors can choose if they want to see additional information or just examine and possibly tag the artwork without previous knowledge. On the right half of the gallery, other artworks which are connected to the current artwork are shown as thumbnails to invite the users to further explore the image database. Overall, the explorARTorium uses five predefined and two dynamic dimensions of contextualization (VSEM, 2011):

Predefined dimensions:

- **Artist**: In the first dimension, the context of the artist is used to give the visitor an impression of the work of the artist by displaying some of his or her other artworks.

- **Title**: In the second dimension, the contextualization targets the subject of the artwork. By presenting other artworks with the same title (regardless of other dimensions like artist or time), the user can explore the iconographic program of the image subject.

- **Region**: The third dimension relates to the geographic region in which the artwork was created, respectively the school of the artist (e.g. Italian, Dutch, German, etc). With this

dimension, the visitor is able to gain insight into the characteristics of painting in that particular region.

- **Theme**: The fourth dimension is dedicated to the theme of the artwork (e.g. *portrait*, *religious*, *mythological*, etc). Through the examination of this dimension, the visitor gets a feeling of the different realizations of this theme over time.

- **Time**: The fifth and last of the predefined dimensions of contextualization relates to the time period in which the artwork was created and shows the visitor other artworks of the same time period (in steps of 50 years), so the visitor gets an overview of the style of painting and typical themes at that particular time.

Dynamic dimensions:

- The sixth dimension presents five randomly chosen paintings from the collection which might attract the visitor's attention and set a completely new context. If the user clicks on a tag below the current artwork, this dimension is used to show artworks associated with the same tag.

- Finally, the seventh and last dimension shows the visitor's history of artworks in order to be able to return to a certain artwork and its context.

To keep the interface clearly arranged, at most five artworks for each dimension are presented as thumbnails to the visitor. Through the search interface on top of the web page, it is also possible to search for artworks by tags, titles, names of artists, regions or themes to present the user the artworks he or she is wants to explore.

In order to fill the explorARTorium with valuable content, images and additional information (e.g. the current location of the artwork, available through one of the fold-out menus as seen in Figure 3.2), the Web Gallery of Art (WGA) was chosen as data source. For more information on the WGA please refer to the next section. In Arends et al. (2011) the process of the transformation of the information from the WGA into the CIDOC Conceptual Reference Model (CRM)[6] is discussed, which "provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation"(CIDOC, 2007) and is the international standard (ISO 21127:2006) for the controlled exchange of cultural heritage information.

Between October 2010 and August 2011 more than 94,000 tags have been assigned to more than 10.000 images of the collection by more than 120 different users. Table 3.1 contains more relevant facts and numbers of the explorARTorium (derived from the explorARTorium database as of August 22, 2011). The user-generated input is normalized according to Hsu and Chen (2008), i.e. spaces and special characters are removed, tags are converted into lower case letters and concatenated to one single word. German umlauts are converted into combinations of vowels (e.g. "ä" turns to "ae").

[6] `http://cidoc.mediahost.org/standard_crm%28en%29%28E1%29.xml`; [accessed 04-October-2011]

Artworks	20313	
Tagged Artworks	10942	54%
Untagged Artworks	9370	46%
Users	125	
Active Users during the last 3 months	10	8%
Tags	94710	
Distinct Tags	14496	
Average number of tags per tagged artwork	8.66	

Table 3.1: Statistical overview of the explorARTorium (August 22, 2011).

Social media play an important role in today's development of the world wide web. To be part of that social interaction on the web, the explorARTorium enables the user to share the artwork he or she is currently exploring with his or her friends on Facebook[7], Twitter[8] and Google+[9] and to comment on the artwork over these social media channels.

3.1.1 Web Gallery of Art

The Web Gallery of Art (WGA) is a virtual museum and searchable database of European painting and sculpture from 11th to mid-19th centuries and was chosen as data source for the explorARTorium. According to Web Gallery of Art (2011a), "it was started in 1996 as a topical site of the Renaissance art, originated in the Italian city-states of the 14th century and spread to other countries in the 15th and 16th centuries. Intending to present Renaissance art as comprehensively as possible, the scope of the collection was later extended to show its Medieval roots as well as its evolution to Baroque and Rococo via Mannerism. More recently the periods of Neoclassicism, Romanticism and Realism were also included."

The WGA offers a collection containing over 27,600 reproductions of over 3,000 artists. In addition to the images of the artworks, some curatorial information is provided for each artwork, like biographical information about the artist, size and location of the artwork, the school of the painting, etc.

Most of the artworks in the gallery are no longer under copyright, but for reproductions the copyright situation within some legal systems remains unclear. However, the following copyright statement is given under Web Gallery of Art (2011b): "The Web Gallery of Art is copyrighted as a database. Images and documents downloaded from this database can only be used for educational and personal purposes. Distribution of the images in any form is prohibited without the authorization of their legal owner."

[7] `http://www.facebook.com`; [accessed 04-October-2011]
[8] `http://www.twitter.com`; [accessed 04-October-2011]
[9] `http://plus.google.com`; [accessed 04-October-2011]

3.2 Data Set

This section is devoted to a specific snapshot extracted from the database of the explorARTorium, which serves as a data basis for the proposed Tag Recommendation Framework in Chapter 5. A fixed and stable data set is needed in order to be able to objectively analyze and compare the results of the Tag Recommendation Framework.

The snapshot was taken on June 1, 2011 and contains about 7,000 artworks and 2,800 distinct tags. The following two conditions were applied during the extraction of the snapshot from the database:

1. At least one tag has to be assigned to the artwork to include the artwork into the data set.
2. A tag has to be assigned to at least three different artworks to be included into the data set.

Table 3.2 gives a statistical overview of this data set with relevant key figures like the average number of assigned tags per artwork, variance, standard deviation, etc.

Artworks	6726
Tags	53701
Distinct tags	2821
Average number of tags per artwork	7.68
Variance of tags per artwork	32.958
Standard deviation of tags per artwork	5.74
Minimum of tags per artwork	1
Maximum of tags per artwork	44
Least frequently used tag	3
Most frequently used tag	1148

Table 3.2: Statistical overview of the extracted data set (June 1, 2011).

3.2.1 Artist

Table 3.3 shows the Top 15 artists with the most artworks in the data set. For each artist, the school of the artist together with the number of artworks for each theme (*genre*, *historical*, etc.) is listed. For each of the Top 5 artists (*Pieter Pauwel RUBENS*, *El GRECO*, *TIZIANO Vecellio*, *GIOTTO di Bondone* and *TINTORETTO*) there exist more than 80 artworks in the data set. It is interesting to see that the data set contains artworks of almost all Top 15 artists for the three most frequent themes (*religious*, *portrait* and *mythological*), whereas the fourth frequent theme *landscape*, is almost exclusively "in the hand" of one artist (*Canaletto*).

Artist	School	Genre	Historical	Interior	Landscape	Mythological	Other	Portrait	Religious	Still-life	Study	Total
Pieter Pauwel RUBENS	Flemish	1	5		6	28	2	23	43		9	117
El GRECO	Spanish				1	1		19	92			113
TIZIANO Vecellio	Italian		1	1		20		29	49			100
GIOTTO di Bondone	Italian			3			1		89			93
TINTORETTO	Italian		5	4		13		13	47			82
REMBRANDT van Rijn	Dutch	3	1		5	7		34	13			63
MICHELANGELO	Italian			3			1		57			61
Hans MEMLING	Flemish					2		14	44	1		61
CANALETTO	Italian				59							59
Frans HALS	Dutch	6						50	2			58
Francisco de GOYA	Spanish	5	1		2	2	9	33	4	1		57
Paolo VERONESE	Italian		2	9		13	2	3	28			57
Lucas the Elder CRANACH	German	2	1		1	16		13	22			55
RAFFAELLO Sanzio	Italian			6		5		10	32			53
Fra ANGELICO	Italian								49			49
Total		**17**	**16**	**26**	**74**	**107**	**15**	**241**	**571**	**2**	**9**	**1078**

Table 3.3: Top 15 artists with their corresponding number of artworks sub-divided by theme in the data set.

3.2.2 Title

In Table 3.4 the 20 most frequent titles of artworks in the data set are listed along with their corresponding themes. In this context it is worth noting, that these 20 top titles only account for a total of 618 artworks, which represent only 9% of all the artworks in the data set. Overall, 4505 different titles of artworks exist in the data set and only 519 titles are used more than once as a description of an artwork in the data set. It will be very important to keep this strong diversification of titles in mind for the evaluation of the Tag Recommendation Framework in Chapter 5. It is also interesting to see that only four different themes (*portrait*, *still-life*, *religious* and *landscape*) from a total of 10 themes occur in this Top 20 list with *religious* clearly being the dominant theme. Another interesting fact is that the most frequent title for the theme *portrait* is *Self-Portrait* (i.e. the artist portrayed himself/herself), followed by *Portrait of a Man* far ahead of *Portrait of a Woman*.

Title	Theme	Count
Self-Portrait	portrait	69
Still-Life	still-life	51
Portrait of a Man	portrait	50
Annunciation	religious	48
Madonna and Child	religious	41
Crucifixion	religious	39
Virgin and Child	religious	38
Adoration of the Magi	religious	35
Adoration of the Shepherds	religious	28
The Annunciation	religious	27
St Jerome	religious	25
Nativity	religious	23
Portrait of a Woman	portrait	22
Portrait of a Young Man	portrait	20
St Sebastian	religious	20
The Adoration of the Magi	religious	18
Pietá	religious	17
The Holy Family	religious	17
The Last Supper	religious	15
Landscape	landscape	15
Total		**618**

Table 3.4: Top 20 titles of artworks in the data set.

3.2.3 Theme

Table 3.5 gives an overview of the distribution of the 10 different themes of artworks in the data set. By looking at the numbers and percentages and the chart in Figure 3.3, it becomes obvious

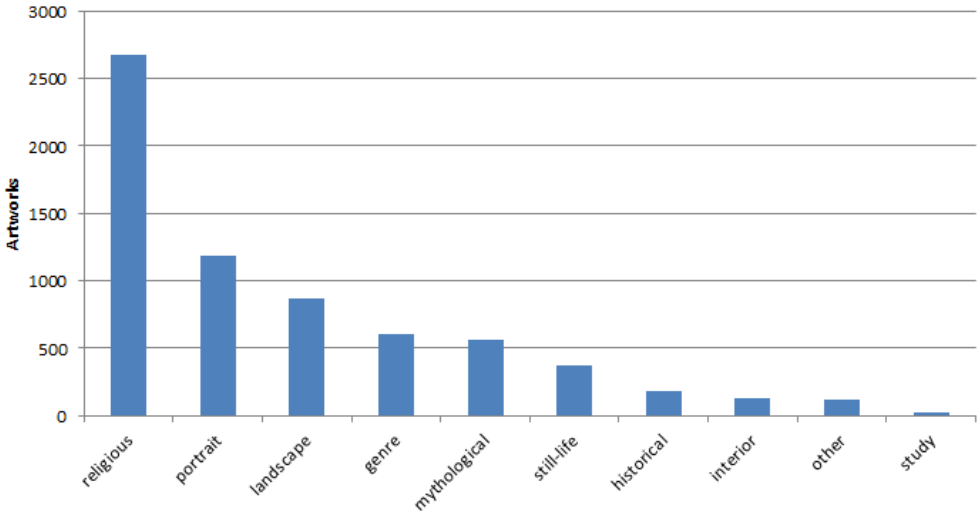

Figure 3.3: Distribution of artworks on the level of themes.

that the themes are very inhomogeneously distributed (with the Top 3 themes (*religious*, *portrait* and *landscape*) amounting up to 71%. Therefore it does not come as a surprise that regarding themes the distribution of artworks has a mean of 672.6 with a high standard deviation of 793.91.

Theme	Count	Percent
Religious	2672	40%
Portrait	1185	18%
Landscape	873	13%
Genre	609	9%
Mythological	563	8%
Still-life	372	6%
Historical	183	3%
Interior	131	2%
Other	115	2%
Study	23	0%
Total	**6726**	**100%**

Table 3.5: Distribution of artworks on the level of themes.

3.2.4 Region

Table 3.6 gives an overview of the distribution of the 25 different regions where the artworks contained in the data set were created. The six most frequent regions (*Italian*, *Dutch*, *Flemish*, *French*, *Spanish* and *German*) cover 94% of all the artworks in the data set, whereas the remaining 19 regions only play a secondary role in the context of regions, which is illustrated in

Figure 3.4. Similarly to the previously discussed category, the theme of the artworks, variance and standard deviation regarding regions compute to high values: the distribution of artworks has a mean of 269.04 with a standard deviation of 578.72.

Region	Count	Percent
Italian	2599	39%
Dutch	1137	17%
Flemish	849	13%
French	811	12%
Spanish	469	7%
German	410	6%
English	125	2%
Netherlandish	104	2%
Austrian	64	1%
Swiss	27	0%
American	25	0%
Hungarian	23	0%
Scottish	15	0%
Danish	15	0%
Belgian	12	0%
Russian	12	0%
Swedish	9	0%
Catalan	7	0%
Greek	4	0%
Portuguese	2	0%
Irish	2	0%
Norwegian	2	0%
Bohemian	1	0%
Other	1	0%
Polish	1	0%
Total	**6726**	**100%**

Table 3.6: Distribution of artworks on the level of regions.

3.2.5 Timeframe

The particular points in time, in which each artwork of the data set was created, are grouped together in periods of 50 years. Table 3.7 lists the frequency of artworks in these timeframes in the data set. Figure 3.5 allows an analysis of the chronological development of artworks in the data set. The oldest artworks in the data set date back to the 13^{th} century, staying at a low quantitative level until 1450. The number of artworks in the data set increases dramatically during the second half of the 15^{th} century and reaches the peak of 1462 artworks between 1601

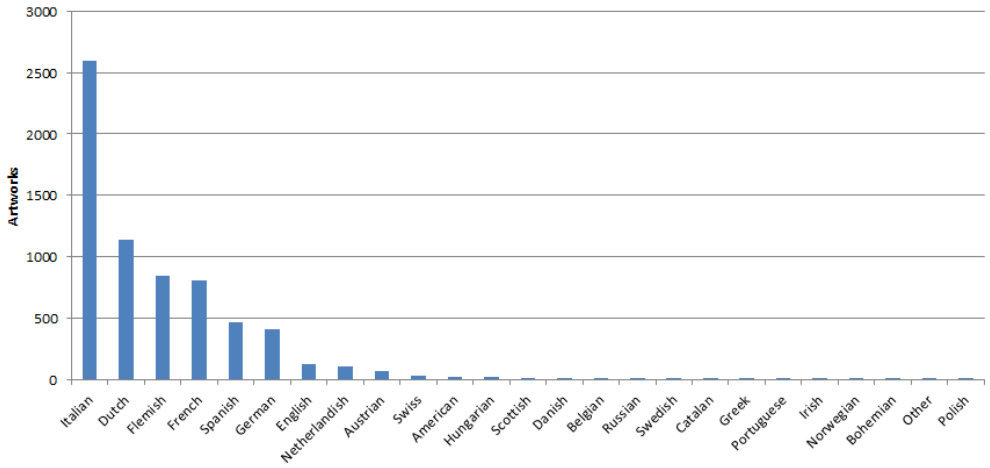

Figure 3.4: Distribution of artworks on the level of regions.

and 1650. From then on, the number of artworks starts to decline rapidly. The most recent artworks in the data set date back to the end of the 19th century. Unsurprisingly, variance and standard deviation in relation to the timeframes of the artworks compute to high values, due to the inhomogeneous distribution of the collection shown in the explorARTorium: the distribution of artworks has a mean of 517.38 with a standard deviation of 425.45.

Timeframe	Count	Percent
1251-1300	8	0%
1301-1350	198	3%
1351-1400	35	1%
1401-1450	268	4%
1451-1500	720	11%
1501-1550	1013	15%
1551-1600	633	9%
1601-1650	1462	22%
1651-1700	872	13%
1701-1750	559	8%
1751-1800	361	5%
1801-1850	493	7%
1851-1900	104	2%
Total	**6726**	**100%**

Table 3.7: Distribution of artworks on the level of time.

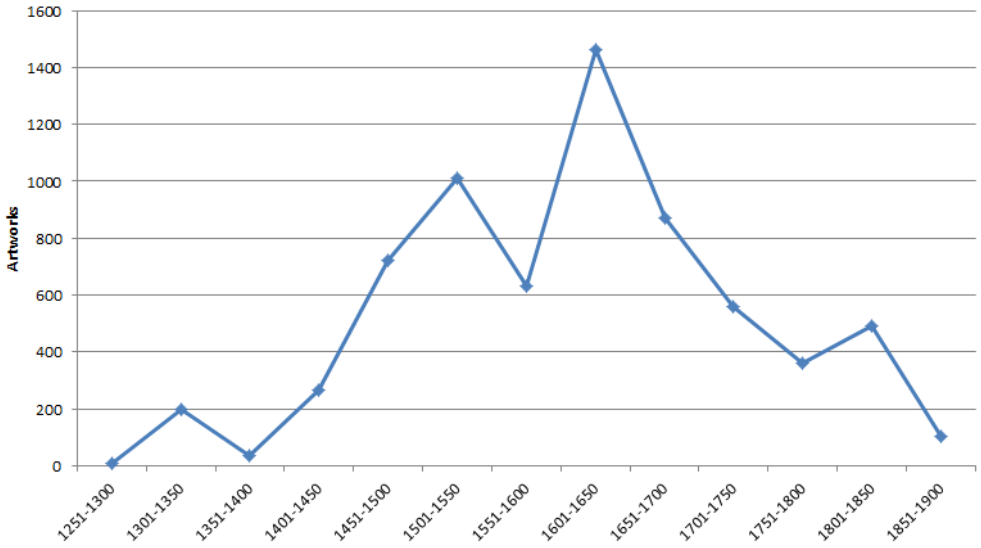

Figure 3.5: Distribution of artworks on the level of time.

3.2.6 Tags

In this section, the focus lies on the tags contained in the data set and their relation to the artworks. Table 3.8 lists the total number of tags and the average number of tags assigned to an artwork for each theme. The average number of tags per artwork ranges from 8.69 (*landscape*) to 5.52 (*study*) and shows a low variance of 1.24. It is interesting to see that the themes with the most artworks (*religious*, *portrait*, etc. see Table 3.5) are not necessarily the same themes which have the highest *tags per artwork* value (e.g. the theme *landscape*).

Theme	Tags	Tags per artwork
Landscape	7586	8.69
Genre	5241	8.61
Religious	22808	8.54
Historical	1529	8.36
Mythological	4339	7.71
Interior	992	7.57
Still-life	2800	7.53
Portrait	7513	6.34
Other	726	6.31
Study	127	5.52

Table 3.8: Distribution of tags on the level of themes.

Similarly to the previous table 3.8 showing the relation of theme and tags, Table 3.9 gives an impression of the total number of tags and the average number of tags per artwork for each

region. It can be stated that artworks from regions which harbor fewer artworks (e.g. *Catalan*, *Irish* or *Scottish*) received more tags on average than other regions. The average number of tags per artwork ranges from 10.57 (*landscape*) to 2.00 (*study*) and shows a variance of 4.04. This can be attributed to the fact that there exist a lot of regions in the data set which contain only a very small number of artworks compared to the total amount of artworks in the data set and due to this small number of artworks in these regions it may be considered as coincidence if these artworks received lots of tags (or not).

Region	Tags	Tags per artwork
Catalan	74	10.57
Irish	19	9.50
Scottish	140	9.33
Netherlandish	943	9.07
Belgian	108	9.00
Flemish	7434	8.76
Dutch	9707	8.54
German	3286	8.01
Italian	20041	7.71
French	6226	7.68
Austrian	486	7.59
English	939	7.51
American	185	7.40
Swiss	199	7.37
Spanish	3421	7.29
Russian	87	7.25
Danish	106	7.07
Swedish	63	7.00
Polish	7	7.00
Hungarian	150	6.52
Portuguese	12	6.00
Bohemian	5	5.00
Greek	15	3.75
Norwegian	6	3.00
Other	2	2.00

Table 3.9: Distribution of tags on the level of regions.

The 25 most frequent tags assigned to artworks in the data set are listed in Table 3.10 to give the reader an overview of the semantic context of the data set. If the original tag given in the first column is in German language, the English translation is provided in the second column.

A histogram of the tags in the data set is depicted in Figure 3.6 showing the inhomogeneous distribution of the frequency of assigned tags to artworks, i.e. graphically answering the question: how many distinct tags were how many times assigned? Because Sturges' formula

Tag	Translation	Occurrence
wolken	clouds	1148
himmel	sky, heaven	1040
engel	angel	804
frau	woman	822
mann	man	792
heiligenschein	halo	688
maria		592
portrait		546
baeume	trees	547
jesus		508
buch	book	427
kind	child	406
hund	dog	403
maenner	men	396
fluegel	wing(s)	374
frauen	women	351
tisch	table	335
hut	hat	310
pferd	horse	299
baum	tree	296
kreuz	cross	291
landschaft	landscape	285
saeulen	pillars	282
felsen	cliff, rock	261
jesukind	infant Jesus	268

Table 3.10: Top 25 most frequently assigned tags.

for the optimal number of bins for histograms (see Sturges, 1926) given in Equation (3.1) does not account for statistical dispersion, the formula for the optimal bin width by Scott (1979) was used to calculate the optimal bin width. According to Equation (3.2), h computes to 14.14 with $\sigma = 57.13$ and $n = 2801$. Due to the characteristics of the present distribution, it was chosen to change the bin width to $h = 5$ for the range from 0 to 100 to achieve a more significant chart (Figure 3.6).

$$k = 1 + \log_2 n = 1 + 3,3 \cdot \log_{10} n \tag{3.1}$$

$$h = \frac{3{,}49 \cdot \sigma}{\sqrt[3]{n}} \tag{3.2}$$

It becomes obvious that the number of tags assigned to different artworks decreases rapidly. Contrary to numerous tags assigned to less than 5, 10 or 15 artworks, there are only a few tags

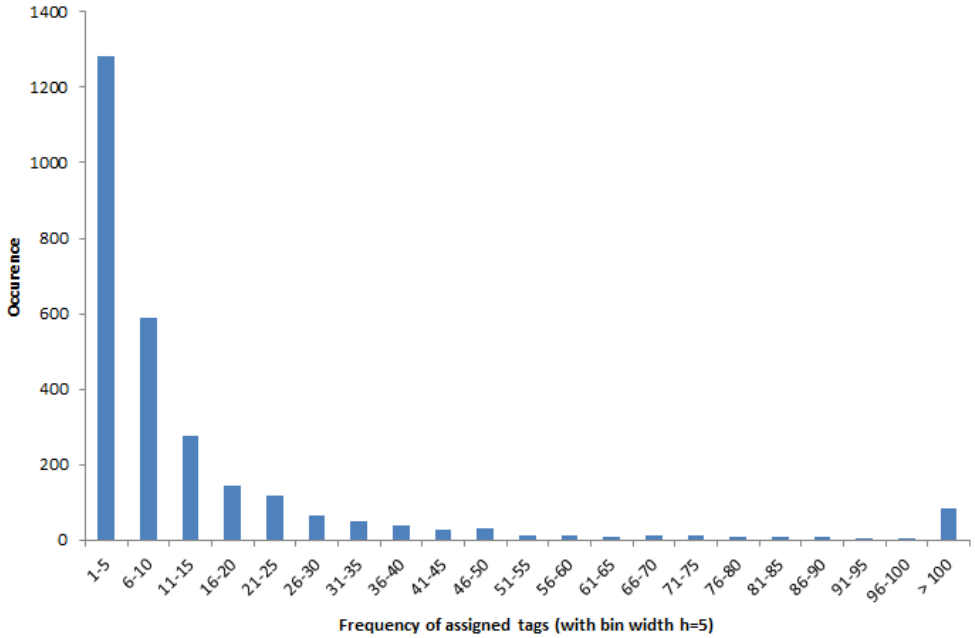

Figure 3.6: Histogram of the frequency of assigned tags.

which are assigned to artworks more than 100 times.

Table 3.11 lists the most tagged artwork for each theme along with its title, artist, the number of tags the artwork received and the reference to the image in the Appendix of this thesis.

This specifically fixed and stable data set extracted from the database of the explorARTorium serves as a data basis for the proposed Tag Recommendation Framework in Chapter 5.

Theme	Title	Artist	Tags	Image ref.
Religious	Sacred Allegory	Giovanni BELLINI	66	Figure A.1
Landscape	Gloomy Day (detail)	Pieter the Elder BRUEGEL	46	Figure A.2
Genre	In Luxury, Look Out	Jan STEEN	45	Figure A.3
Historical	Napoleon Bonaparte on the Battlefield of Eylau, 1807	Antoine-Jean GROS	45	Figure A.4
Interior	Duet	Frans van MIERIS, the Elder	45	Figure A.5
Mythological	Perseus Frees Andromeda (detail)	PIERO DI COSIMO	42	Figure A.6
Other	Dinner	Thomas ROWLANDSON	40	Figure A.7
Still-life	Still-Life of Flowers and Fruits	Jean-Baptiste MONNOYER	39	Figure A.8
Study	The Adoration of the Wise Man	Albrecht DÜRER	37	Figure A.9
portrait	Portrait of the Saltykov Family	Johann Friedrich August TISCHBEIN	36	Figure A.10

Table 3.11: Artworks with the most assigned tags.

CHAPTER 4

Analysis of User-Generated Content of a Folksonomy

In this chapter, user-generated content of a folksonomy related to art history is analyzed: it is explored if the users' tagging behavior is related to their liking of artworks in Section 4.1, the users' vocabulary is qualitatively and lexically analyzed in Section 4.2, and finally the role of the users regarding activity and learning effects is examined in Section 4.3.

4.1 Rated Artworks

With the rating-function of the explorARTorium, users are able to rate artworks based on a scale of five stars, with the user being able to define his or her own "meaning" of a star. Until March 02, 2011, it was possible to rate artworks in steps of "half stars" including zero stars (i.e. the accepted values in the database were 0, 0.5, 1, 1.5, . . . , 4.5, 5), but since then the rating-function has only allowed the assignment of "full stars" (i.e. the accepted values in the database are 1, 2, 3, 4 or 5). Two questions are formulated concerning the rating of artworks, which are discussed in the following sections:

1. Is the users' rating activity changing over time?
2. Are users more likely to tag artworks they like?

4.1.1 Is the users' rating activity changing over time?

The chronological sequence of ratings is presented in Figure 4.1, which is backed by the values in Table 4.1, with time (in months) plotted on the x-axis and the number of ratings users assigned to artworks on the y-axis. It is interesting to see that the rating activity shows no continuity: during the last three months of 2010, users rated four times as many artworks as during the first nine months of 2011. The curve illustrates that the users' rating activity has decreased dramatically over time since the peak of activity in November 2010.

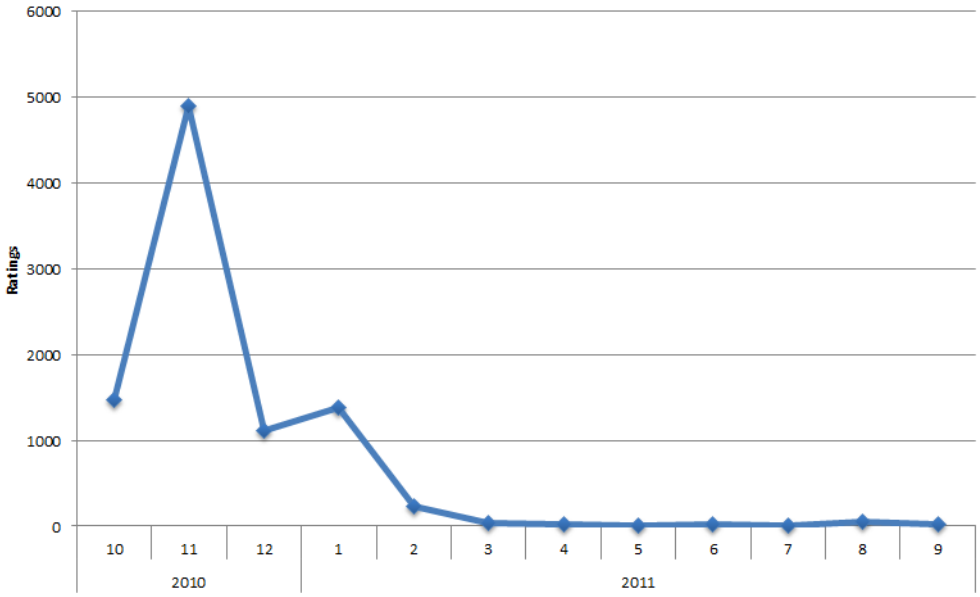

Figure 4.1: Chronological Sequence of Ratings.

4.1.2 Are users more likely to tag artworks they like?

In order to be able to answer this question, different measures are computed for each rating category (0, 0.5, 1, 1.5, ..., 4.5, 5 stars) shown in Table 4.2. Firstly, numbers not necessarily corresponding to the *same* user who rated *and* tagged an artwork are calculated:

- The number of artworks rated in this category (column $Artworks$)
- The number of tags these rated artworks received in this category by *all* users (column $Tags_1$)
- The mean of assigned tags per artwork in this category (column $Mean_1$)
- The deviation of $Mean_1$ for this category from the overall mean (over all categories) (column $Dev._1$)

It is worth noting that these numbers include tags for rated artworks assigned by *all* users and not just the tags allotted by users who actually rated *and* tagged the same artworks. But since this measure is the most exact one to answer the initial question, additional values for tags that were assigned to artworks by users who also rated these artworks are computed (again shown in the three columns on the right of Table 4.2):

- The number of tags artworks received which were rated by the same user (column $Tags_2$)
- The mean of assigned tags per artwork which was rated by the same user (column $Mean_2$)

Month	Count
2010	**7463**
10	1469
11	4889
12	1105
2011	**1807**
1	1385
2	234
3	34
4	28
5	10
6	28
7	13
8	59
9	16
Total	**9270**

Table 4.1: Chronological rating activity.

- The deviation of $Mean_2$ for this category from the overall mean (over all categories) (column $Dev._2$)

The most interesting column in Table 4.2 is the last one ($Dev._2$), which is visualized in Figure 4.2, answering the question if users are more motivated to tag artworks they like. The number of assigned stars is plotted on the abscissa, whereas the deviation of the mean of assigned tags per artwork in that particular category (the number of assigned stars) in percent from the overall mean is plotted on the ordinate. The chart shows clearly that users of the explorARTorium tend to assign up to 89% more tags to artworks they find appealing and up to 47% fewer tags to artworks they do not like. It is also interesting to see that users are likely to assign tags above average to artworks they rated with zero stars. Due to the limited possibilities to draw conclusions from this particular analysis, as the intentions of the users to tag artworks are unknown, no direct statement can be made, but it might be speculated that the rejection of the artwork stimulates the user to assign it with more than average tags.

Furthermore, the following interesting statistics are calculated:

- Only 0.64% of all ratings are assigned to artworks which received *no* tags.
- The overwhelming majority of all ratings (99.36%) is assigned to tagged artworks.
- The average tag count for users who tag and rate the same artwork computes to 8.19, whereas users who do not rate the artworks assign only 1.33 tags on average.

	Artworks		$Tags_1$		$Mean_1$	$Dev._1$	$Tags_2$		$Mean_2$	$Dev._2$
5 stars	152	2%	2042	2%	13.43	26%	1672	2%	11	34%
4.5 stars	177	2%	3106	3%	17.55	64%	2745	4%	15.51	89%
4 stars	522	6%	7747	8%	14.84	39%	6553	9%	12.55	53%
3.5 stars	1053	11%	14524	15%	13.79	29%	12115	16%	11.51	40%
3 stars	2389	26%	26882	27%	11.25	5%	21178	28%	8.86	8%
2.5 stars	2125	23%	21319	22%	10.03	-6%	16042	21%	7.55	-8%
2 stars	990	11%	8294	8%	8.38	-21%	5672	7%	5.73	-30%
1.5 stars	882	10%	7387	7%	8.38	-21%	5139	7%	5.83	-29%
1 star	484	5%	3576	4%	7.39	-31%	2096	3%	4.33	-47%
0.5 stars	393	4%	2874	3%	7.31	-31%	1770	2%	4.50	-45%
0 stars	103	1%	1144	1%	11.11	4%	956	1%	9.28	13%
Total	**9270**	**100%**	**98895**	**100%**	**10.67**		**75938**	**100%**	**8.19**	

Table 4.2: Parameters of the users' tagging activity based on the rating of the artworks.

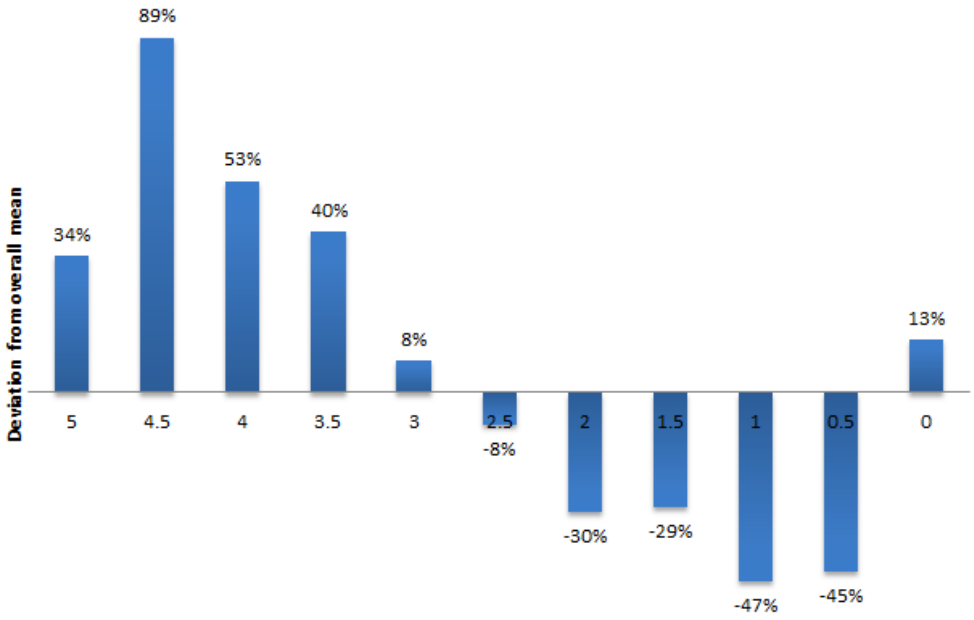

Figure 4.2: Deviation of the average tag count based on the user's rating of the artworks.

4.2 Tagged Artworks

With the tagging-function of the explorARTorium, users are able to annotate descriptive keywords (tags) to artworks. For the operator of the explorARTorium it is important to have knowledge about the habits in tagging of the users in order to be able to analyze and improve the platform. Four questions are formulated concerning the tagging of artworks, which are discussed in the following sections:

1. Is the users' tagging activity changing over time?

2. How are parts of speech distributed through the users' vocabulary?

3. Is there a part of speech bias in different themes?

4. Are users more likely to identify historical persons and places in an artwork or the creator of the artwork?

5. Does the quality of adjectives for different themes differ?

4.2.1 Is the users' tagging activity changing over time?

Similar to the question concerning rated artworks, the question arises if the users' tagging activity remains at a constant level or changes over time. Figure 4.3 is backed by the values in Table 4.3 and shows the chronological tagging sequence, plotting time (in months) on the x-axis and the number of assigned tags to artworks on the y-axis. The diagram reveals that the development of the tagging activity over time bears a striking resemblance to the rating activity over time (depicted earlier in Figure 4.1) and also shows a continuous downward trend after the peak in November 2010.

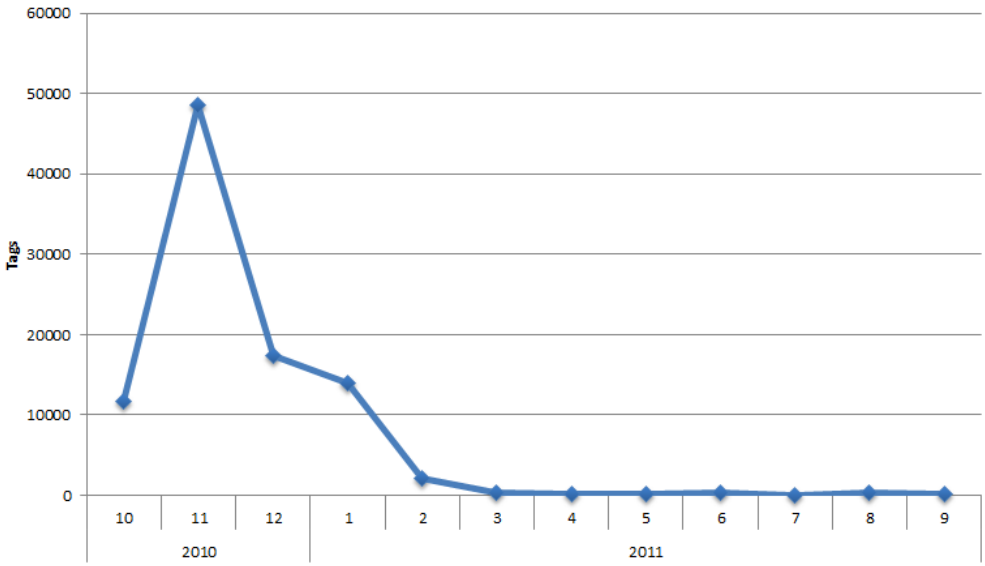

Figure 4.3: Chronological Sequence of Tags.

Month	Count
2010	**77581**
10	11702
11	48495
12	17384
2011	**17355**
1	13955
2	2131
3	257
4	204
5	87
6	253
7	50
8	293
9	125
Total	**94936**

Table 4.3: Chronological tagging activity.

4.2.2 How are parts of speech distributed through the users' vocabulary?

Till September 07, 2011, users of the explorARTorium have assigned a total of 94,936 tags to artworks. In order to be able to answer the question of the deviation behind the parts of speech[1] of the vocabulary of the explorARTorium, each tag has to be analyzed and assigned to a lexical class. For this task, so-called part of speech taggers have been developed (see Brill, 1992; Cutting et al., 1992; Schmid, 1994; Brants, 2000), which are able to determine the lexical class of a word: "In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context - i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph"(Wikipedia, 2001). In this particular case of the explorARTorium, the words to be analyzed are not embedded in a context (e.g. a phrase or a sentence), so the POS tagger can only determine the lexical class of the word based on its definition.

For this task, the data is loaded into the POS tagger of the Apache OpenNLP[2] project, a machine learning based toolkit for natural language processing (NLP), which incorporates the feature to mark tokens (words) with their corresponding word type. Although the OpenNLP tagger provides an easy-to-use application programming interface (API) and delivers good results for a small training set of tags, the achieved results for the whole data set turn out to be unfeasible, due to the following circumstances:

[1] "In grammar, a part of speech (also a word class, a lexical class, or a lexical category) is a linguistic category of words (or more precisely lexical items), which is generally defined by the syntactic or morphological behavior of the lexical item in question. Common linguistic categories include noun and verb, among others."(Wikipedia, 2011e)

[2] `http://incubator.apache.org/opennlp/index.html`; [accessed 04-October-2011]

- In the German language, nouns are usually written with an initial upper case letter. Since the majority of the words in the data set is in German, but unfortunately just available in lower case letters, the POS tagger misclassifies most of the nouns.
- In the data set, German umlauts are converted into combinations of vowels (e.g. “ä” turns to “ae”), with which the POS tagger is unfamiliar with.
- There exist lots of composite words in the data set (e.g. *ringamzeigefinger* (meaning *ring on the forefinger*), sometimes even containing proper nouns (e.g. *hlchristophorus* (meaning Saint Christopher) which the POS tagger is unable to annotate with the correct lexical class.

It is possible to solve these problems by reconverting the combination of vowels into umlauts, applying a word tokenizer to separate composite words, etc, but since this task is more complex and time-consuming than a manual classification, the part of speech for 3750 distinct tags, which have been assigned at least three times to an artwork, is annotated manually. On the basis of the content of the database, five different categories for parts of speech are distinguished (with the most frequent representative in brackets):

- Nouns (e.g. *wolken* (clouds))
- Verbs (e.g. *beten* (to pray))
- Adjectives (e.g. *nackt* (naked))
- Proper nouns (persons and places, e.g. *maria* (Mary), *venedig* (Venice) or *elgreco* (El Greco))
- Others (every other part of speech, including multiple tags written as one word, e.g. *blauerhimmel* (blue sky))

Figure 4.4 presents the distribution of parts of speech for these tags. The bar chart depicts that users of the explorARTorium tend to assign nouns as tags to the artworks, being the category $noun$ the predominant part of speech with a proportion of 86%, followed by $adjectives$ with a 6% and $proper\ nouns$ with a 5% share. This allows the conclusion that the users of the explorARTorium rather like to name things or persons they see in the artworks (with *nouns*), than to describe them (with *adjectives*) or to convey an action (with *verbs*).

4.2.3 Is there a part of speech bias in different themes?

In Figure 4.4 the overall distribution of the parts of speech is presented. To answer the question if there is part of speech bias in different themes, i.e. if different parts of speech are used more often for certain themes than others, the distribution of parts of speech is analyzed on the level of themes. As noted earlier, the vast majority of tags are nouns which show an even distribution over all themes. Therefore nouns are omitted and the focus lies on the other four parts of speech, which are visualized in Figure 4.5. The ten different themes are plotted on the x-axis and the proportion of the share of the parts of speech in relation to the number of artworks for each

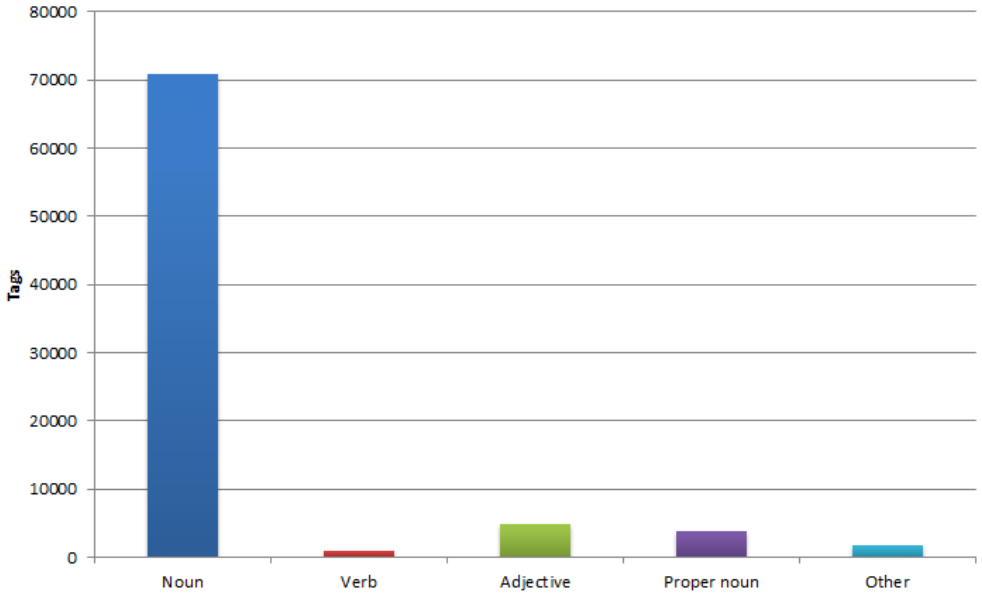

Figure 4.4: Distribution of parts of speech.

theme is given on the y-axis with the part of speech "noun" being left out. Based on the chart, the following observations are worth noting: The ratio of adjectives for the themes *mythological*, *other*, *portrait* and *study* is almost twice as high as for the rest of themes, whereas the number of verbs is at an extremely low level for every theme (especially for *interior*, *portrait* and *still-life*, categories of artworks traditionally barely conveying actions) with *genre* being the highest verb-ranked theme of artworks (possibly due to frequent portrayals of everyday life). It is also very interesting to see that the proportion of proper nouns for the theme *religious* is three times higher than for any other theme. This allows the conclusion that users of the explorARTorium are more likely to identify religious figures (e.g. Jesus, Mother Mary, Saint Joseph, etc.) than historical figures in portraits (e.g. Napoleon, Maria Theresa, Rembrandt, etc). It can also be concluded that the taggers use different parts of speech based on the theme of the artwork they are describing. Table 4.4 lists the Top 5 tags per theme for the parts of speech *noun*, *verb* and *adjective*.

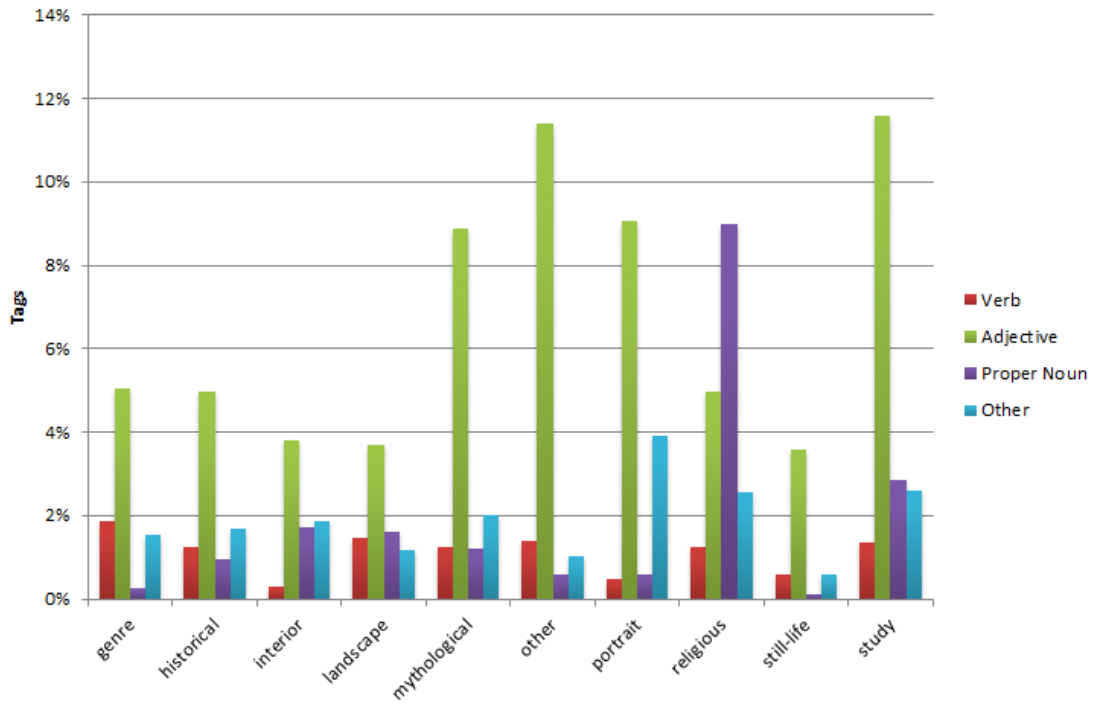

Figure 4.5: Distribution of parts of speech based on the theme of the artwork.

	Noun		Verb		Adjective	
Genre	frau	166	trinken	15	fest	16
	tisch	143	essen	13	pelzverbraemt	14
	hut	123	lesen	7	heiter	10
	mann	121	braten	4	nackt	9
	hund	112	fliegen	4	baeuerlich	7
Historical	wolken	47	essen	3	nackt	10
	pferd	41	bitten	3	barbusig	3
	himmel	38	servieren	2	besiegt	3
	pferde	31	zeigen	2	dunkel	3
	maenner	29	beten	1	fest	3
Interior	kirche	38	promenieren	1	museal	8
	saeulen	31	saufen	1	gold	4
	hund	30	schreiben	1	knieend	2
	altar	21	servieren	1	schwarz-weiss	2
	fenster	19			ueberladen	2
Landscape	wolken	426	fliegen	14	steil	17
	himmel	342	weiden	10	ruhig	15
	landschaft	246	eislaufen	8	schwarz-weiss	14
	baeume	217	trinken	7	beruhigend	8
	fluss	168	grasen	6	friedlich	8
Mythological	frau	229	fliegen	5	nackt	193
	engel	161	trinken	5	naked	24
	wolken	159	essen	4	langhaarig	17
	himmel	141	baden	3	barbusig	16
	mann	133	blasen	3	schwarz-weiss	11
Other	frau	54	fliegen	2	nackt	30
	mann	32	hunt	2	schwarz-weiss	21
	himmel	29	saeugen	2	blau	6
	wolken	29	trinken	2	gold	6
	pferd	28	angeln	1	golden	6
Portrait	portrait	627	falten	8	langhaarig	36
	mann	521	schreiben	6	braunaeugig	20
	frau	288	beten	2	dunkelhaarig	20
	hut	159	essen	2	freundlich	19
	bart	153	lachen	2	pelzverbraemt	19
Religious	heiligenschein	1040	beten	55	nackt	131
	engel	1036	strahlen	39	knieend	62
	himmel	574	fliegen	23	gold	56
	wolken	564	weinen	16	langhaarig	37
	frau	462	essen	13	schwarz-weiss	37
Still-life	stillleben	96	essen	7	orange	16
	stilleben	94	kochen	4	bunt	14
	blaetter	86	blasen	1	gestreift	11
	tisch	77	fliegen	1	erlegt	7
	blumen	71	ranken	1	aufgeblueht	6
Study	zeichnung	53	saeugen	2	schwarz-weiss	27
	mann	35	zeichnen	2	nackt	22
	skizze	35	beten	1	blackwhite	5
	frau	33	laecheln	1	naked	4
	studie	27	zeigen	1	unleserlich	4

Table 4.4: Top 5 tags per part of speech for each theme.

4.2.4 Are users more likely to identify historical persons and places in an artwork or the creator of the artwork?

In Figures 4.4 and 4.5 it is not differentiated for the part of speech *proper noun* between proper nouns describing the characters and locations depicted in the artwork (e.g. Jesus, Mary, etc.) and proper nouns referencing the artist (e.g. El Greco, Michelangelo, Caravaggio, etc.). Since it is an interesting question if users of the explorARTorium rather identify historical figures than the creator of the artwork, this distinction is undertaken and the part of speech *proper noun* is divided into two subcategories:

- Proper nouns describing the characters and locations depicted in the artwork
- Proper nouns referencing the artist

The outcome of this analysis is presented in Figure 4.6, showing the distribution of identified characters and locations and identified artists in percent of overall assigned tags for each theme. It becomes obvious that the users are far more likely to recognize persons and places than the creator of the artwork, who is identified on average in only 0.2% of all artworks. This finding applies to every theme except for portraits, where the distribution between these two categories of proper nouns is at an even level. Based on the diagram, it can be concluded that the users rather tend to recognize the artist of artworks with the theme *study* (e.g. Michelangelo, Leonardo da Vinci) than artworks showing *mythological* motives (e.g. Caravaggio, Velazquez).

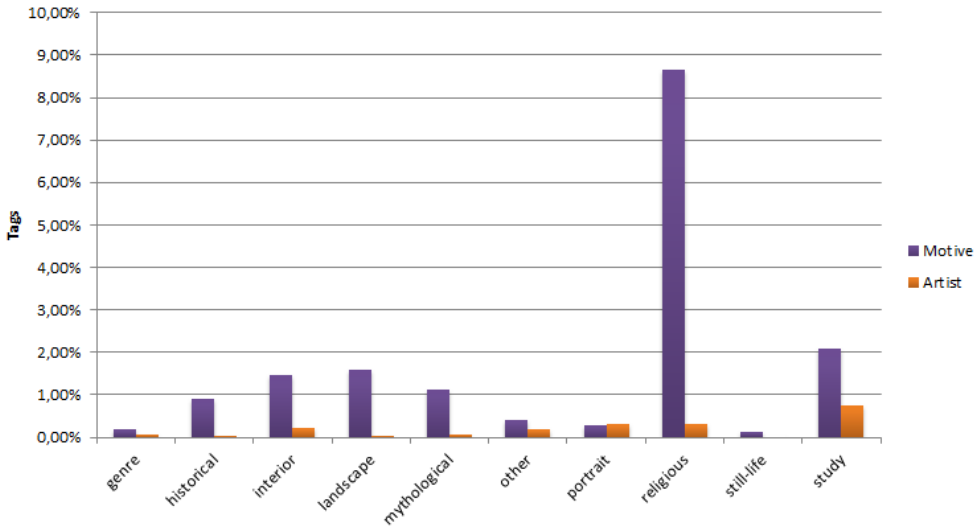

Figure 4.6: Distribution of proper nouns differentiated between motive and artist based on the theme of the artwork.

In contrast to Figure 4.6, which presents the overall distribution of proper nouns in the context of all assigned tags, Figure 4.7 focuses on the distribution of *distinct* proper nouns (i.e. identified motives and artists) for each theme, showing that the users are identifying more different

motives than different artists, resulting in a similar distribution as in Figure 4.6. The numbers backing Figure 4.7 are given in Table 4.5. The fact that both the *total* number of identified motives as well as the number of *different* identified motives (i.e. distinct proper nouns) is higher for each theme than the corresponding numbers for identified artists, confirms the previously stated conclusion derived from Figure 4.6 that the users of the explorARTorium rather tend to recognize persons and places than the creator of the artwork. In Table 4.6 the Top 5 tags identifying motives and artists for each theme are presented. Most of the Top 5 tags are quite expectable (e.g. *venedig* (Venice) for the theme *landscape*, *venus* for *mythological* or *maria* for *religious* artworks), but the fact that tags like *jesukind* (Infant Jesus), *madonna* and *heiligefamilie* (Holy Family) are assigned to *still-life* artworks is quite surprising and allows to suspect an error or misclassification of the artwork in the database. But in fact, both the classification of the artwork as well as the assigned tags are correct (cf. Figure 4.8).

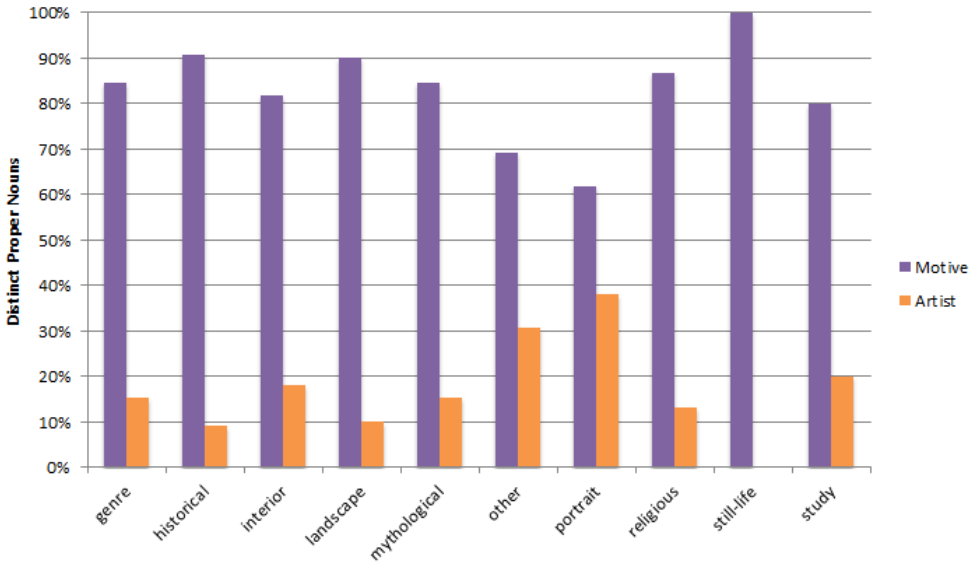

Figure 4.7: Distinct proper nouns differentiated between motive and artist based on the theme of the artwork.

Figure 4.8: Still-life: *Holy Family* by *Jan van KESSEL, junior*.

	Motive	Artist	Total
Genre	11	2	13
Historical	10	1	11
Interior	9	2	11
Landscape	9	1	10
Mythological	22	4	26
Other	9	4	13
Portrait	13	8	21
Religious	86	13	99
Still-life	3		3
Study	16	4	20
Total	**188**	**39**	**227**

Table 4.5: Distribution of *distinct* proper nouns based on the theme of the artwork.

	Motive		Artist	
Genre	venedig	3	caravaggio	3
	maria	1	breughel	1
	jesus	1		
	jesukind	1		
	holofernes	1		
Historical	venedig	5	cranach	1
	napoleon	4		
	david	3		
	maria	2		
	venus	2		
Interior	maria	4	michelangelo	2
	jesukind	3	giotto	1
	christus	3		
	sixtinischekapelle	3		
	jesus	2		
Landscape	venedig	101	elgreco	3
	venice	26		
	sanmarco	8		
	canalegrande	7		
	rom	5		
Mythological	venus	21	elgreco	1
	zeus	8	caravaggio	1
	bacchus	7	cranach	1
	jupiter	5	velazquez	1
	merkur	5		
Other	adam	3	goya	2
	jesus	1	elgreco	1
	venedig	1	breughel	1
	eva	1	albrechtduerer	1
	napoleon	1		
Portrait	napoleon	7	rembrandt	11
	venedig	3	elgreco	9
	raphael	3	leonardodavinci	3
	luther	3	cranach	3
	david	2	velazquez	3
Religious	maria	801	elgreco	60
	jesus	716	michelangelo	19
	jesukind	356	caravaggio	10
	madonna	248	giotto	9
	christus	247	michaelangelo	4
Still-life	jesukind	2		
	madonna	1		
	heiligefamilie	1		
Study	maria	4	davinci	4
	jesus	4	leonardodavinci	3
	madonna	4	leonardo	2
	jesukind	3	michelangelo	1
	rubens	2		

Table 4.6: Top 5 proper nouns for each theme.

4.2.5 Does the quality of adjectives for different themes differ?

Since adjectives have been determined as the second most frequent part of speech of the assigned tags in the explorARTorium earlier, the question opens up if different types of adjectives are used and, if so, if different themes show different ratios of these types. In this context, two types of adjectives are distinguished:

- **Type 1**: Adjectives describing emotions, feelings or impressions (e.g. friendly, comforting, calming, silent, . . .)
- **Type 2**: Adjectives describing non-emotional facts or things (e.g. long-haired, black-and-white, naked, . . .).

For five themes (*landscape*, *mythological*, *portrait*, *religious* and *study*) the most frequently used adjectives are compiled and visualized through tag clouds. According to Halvey and Keane (2007), tag clouds (word clouds, or weighted lists in visual design) are "visual presentations of a set of words, typically a set of tags, in which attributes of the text such as size, weight or color can be used to represent features (e.g., frequency) of the associated terms."

For this master's thesis, the web application Wordle[3] is used to create artistic word clouds from the weighted adjectives for each theme. Figures 4.9, 4.10, 4.11, 4.12 and 4.13 show tag clouds containing the 150 most popular adjectives for the themes *landscape*, *mythological*, *portrait*, *religious* and *study*. The importance of each tag is shown with the font size, whereas the color of the tag depends on the assigned adjective-type of the tag (Type 1 in red, Type 2 in blue).

Figure 4.9: Weighted tag cloud with adjectives used to describe *landscape* artworks.

The analysis of the tag clouds reveals that there are strong differences between the "emotional quality" of adjectives for different themes. Whereas artworks showing landscapes and portraits are described with both types of adjectives, artworks with mythological, religious or study themes almost completely lack adjectives of Type 1. Therefore, based on the tag clouds, it can be inferred that users are not likely to use adjectives describing emotions or feelings to tag mythological, religious or study artworks and use non-emotional adjectives like *nackt* (naked),

[3] `http://www.wordle.net`; [accessed 04-October-2011]

Figure 4.10: Weighted tag cloud with adjectives used to describe *mythological* artworks.

Figure 4.11: Weighted tag cloud with adjectives used to describe *portrait* artworks.

schwarz-weiss (black-and-white), *langhaarig* (long-haired) or *steil* (cliffy/steep) instead. In contrast, users respond also emotionally to landscapes and portraits, assigned adjectives like *ruhig* (calm/quiet), *freundlich* (friendly) or *stuermisch* (stormy) to artworks of these themes.

In order to confirm the conclusions drawn from the analysis of the tag clouds, the total number of Type 1-adjectives is compared to the total number of Type 2-adjectives for each theme (in proportion to the number of tagged artworks for that particular theme). Figure 4.14 shows a graphical representation of this comparison, revealing that the previously stated difference between the themes landscape/portrait and mythological/religious/study is not as high as expected. This can be attributed to the fact that even mythological/religious/study themes are described with "emotional" adjectives to some extent, but with many different tags which occur very infrequently and are therefore either printed extremely small or ignored completely in the visualization of the tag clouds.

4.3 Users

In the previous two sections, interesting aspects of the rating and the tagging function of the explorARTorium are evaluated. But since the driving force behind all this generated content and

Figure 4.12: Weighted tag cloud with adjectives used to describe *religious* artworks.

Figure 4.13: Weighted tag cloud with adjectives used to describe *study* artworks.

interaction with the online museum are the visitors, emphasis is laid on the evaluation of the users in this section. Three questions concerning the users of the explorARTorium are formulated and will be discussed in the following sections:

1. What does the user landscape look like?
2. Are there different types of users regarding their activity over time?
3. Is the users' vocabulary getting more specific over time?

4.3.1 What does the user landscape look like?

For every operator of an interactive platform on the Internet it is important to have knowledge of the user landscape of the platform in order to be able to improve the platform based on the conclusions of the users' statistics. As for the explorARTorium, Figure 4.15 presents a bar chart showing the number of new users (in green color), active users (i.e. the sum of new and returning users; in red color) and the total number of users accumulated over time (in blue color), plotting time (in months) on the abscissa and the number of users on the ordinate. In this particular context, the term *user* refers to a visitor who has rated or tagged at least one artwork. Figure 4.15, which is backed by the values of Table 4.7, shows that the explorARTorium was able to increase its total number of users in November 2010 (the second month after the start) by 61%, but has failed to repeat this success since then by only attracting 1%-5% new users each

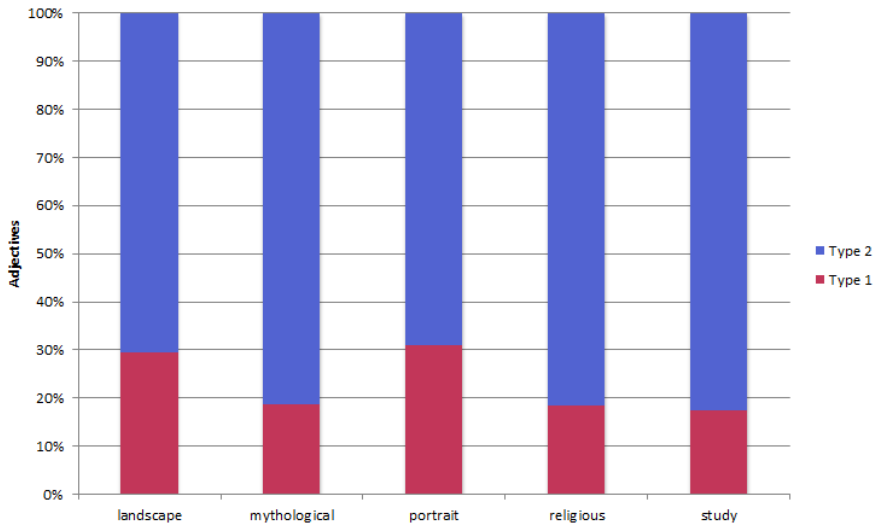

Figure 4.14: Distribution of "emotional" and "non-emotional" adjectives for different themes.

month. Moreover, it is interesting to see that disregarding the first two months, the number of active users of the explorARTorium remains at an almost constant level (average of 9 users per month).

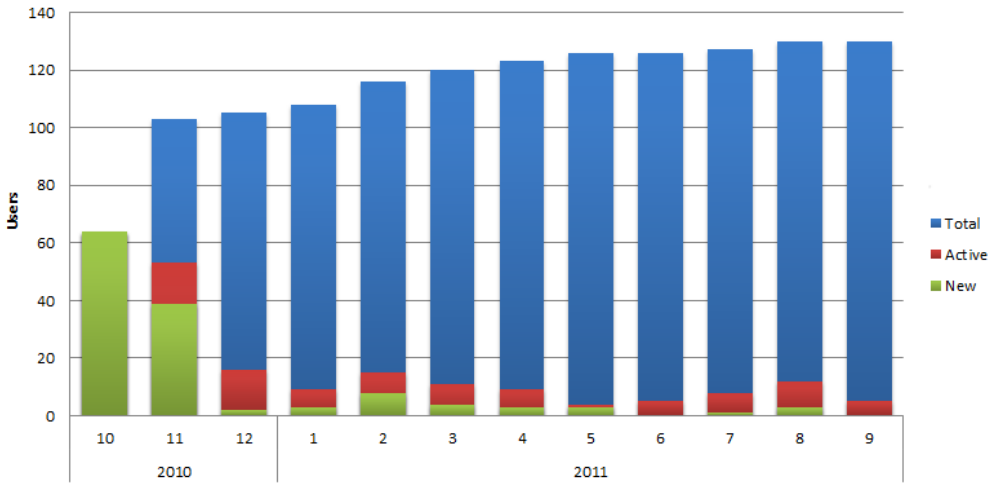

Figure 4.15: User activity chart.

Month	Total	Active	New
2010			
10	64	64	64
11	103	53	39
12	105	16	2
2011			
1	108	9	3
2	116	15	8
3	120	11	4
4	123	9	3
5	126	4	3
6	126	5	0
7	127	8	1
8	130	12	3
9	130	5	0

Table 4.7: User activity statistics.

4.3.2 Are there different types of users regarding their activity over time?

Figure 4.16 presents the tagging activity at the explorARTorium from a more user-centric perspective than in the previous question, analyzing the tagging trend for each user. For this particular analysis, only users who provided at least a total of 5 tags were taken into account. In the diagram on the left, the number of tags each user assigned during *his* or *her* first, second, third, ... month of usage of the explorARTorium is plotted as dots in the chart with the x-axis outlining the months of usage (i.e. *not* the calendar months as in Figure 4.3) and the y-axis the number of tags per user. The diagram on the right shows the same data but with a logarithmically scaled y-axis to better integrate the outliers of the second, third and fourth month into the plot. Due to an evaluation of the chart in Figure 4.16 in combination with Figure 4.15, two types (or groups) of users can be identified:

- Type 1: It can be observed that some of the users are active taggers during their first five months visiting the explorARTorium, but then lose interest in actively participating at the explorARTorium and either start tagging much less than before or even completely quit tagging artworks (this is implicated by the fact that the density of the dots in Figure 4.16 is getting sparse in the course of time).
- Type 2: This user type shows constant interest in annotating artworks and returns to the explorARTorium on a regular basis.

4.3.3 Is the users' vocabulary getting more specific over time?

One of the most interesting questions regarding the explorARTorium is concerned with the learning effect of the users. It may be hypothesized that users who spend time at the explorARTorium

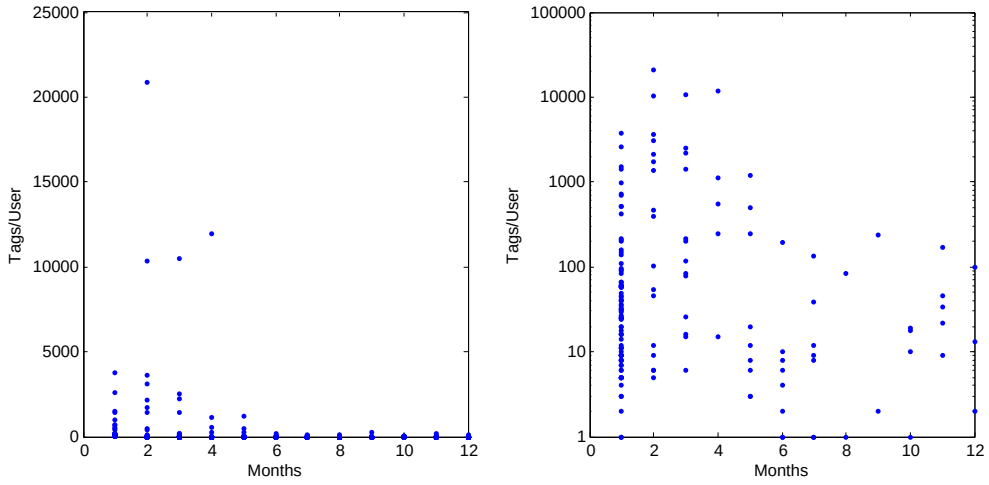

Figure 4.16: User-centric tagging activity.

discovering artworks and the tags of other users for these artworks, may broaden their art historical knowledge. Since it is a very subjective process to determine if someone has actually learned something and the available data set does not contain sufficient parameters to do so, the emphasis in this section lies on the specificity of the users' vocabulary with the aim to conclude if users tend to use more specific tags to describe the artworks over time.

In order to be able to compare the "quality" of tags, an algorithm to determine the specificity has to be chosen. The only way to achieve an automated measure of the tags is to use an algorithm based on the occurrence and frequency of the tags. Popular ways in information retrieval and text mining to reach these goals are the following (Salton and McGill, 1983):

- **TF (term frequency)**: is based on the idea the more often a term occurs in a document, the more important it is in describing that document.
- **IDF (inverse document frequency)**: measures the relevance of the term with regard to the whole document collection.
- **TF-IDF (term frequency – inverse document frequency)**: combination of TF and IDF, computed by multiplication of these two values.

By taking a look at the data of the explorARTorium, one can see that each tag occurs at most once for each painting, so the TF algorithm is unfeasible for this purpose. Therefore also the combination of TF and IDF (TF-IDF) does not fit the needs. However, the IDF algorithm is able to measure the specificity of the tags, because the frequency of one tag in the whole collection is set into relation with the size of the artwork collection. The IDF is calculated for a tag according to Equation (4.1), where $N = |D|$ is the number of artworks in the collection and n_i the number of artworks which are tagged with term i.

$$idf_i = \log \frac{N}{n_i} \tag{4.1}$$

In a first step, the IDF is calculated for every tag which appears at least twice in the data set in order to eliminate typing errors. The IDF values range from 0.88 to 3.74. In the second step, a special data set is created, containing additional information for each tag, namely the username of the user who assigned the tag, the timestamp of the tag and the IDF for the tag calculated in the first step. Afterwards, a matrix is computed showing the mean of the assigned tags for each user over time (in intervals of months). As this matrix now contains the desired values to answer the question if the vocabulary of users gets more specific over time, different measures to test the hypothesis are computed:

- Comparison between first half and second half (C_HALF): for each user who was active tagging artworks in at least two different time intervals (months), the two means of the IDF of the tags assigned to artworks during his or her first and second half of active participation at the explorARTorium are calculated.

- Comparison between first quarter and last quarter ($C\_QUARTER$): the same approach is taken as before, but only the tags the user assigned during the first quarter and last quarter of his or her active period are taken into account.

In both cases, the two calculated means are compared: if the mean of the second time period is of a higher value than the mean of the first period, it can be concluded that the vocabulary of the user got more specific over time using the explorARTorium. If the IDF means do not differ from each other by more than a certain value (0.3), it can be inferred that the tag specificity for that particular user remained unchanged.

Figure 4.17 presents the comparison of the means, i.e. evaluates how many users enhanced their vocabulary with more specific terms. It becomes obvious that for both measures (C_HALF and $C\_QUARTER$) the number of users whose vocabulary got more specific is higher than the number of users whose vocabulary got less specific, while the number of users whose tag specificity remained unchanged is quite similar for both measures. By means of the $C\_QUARTER$ measure, which eminently analyzed the tags at the very beginning and the very end of the active period of the users, it can be concluded that actually 40% of users enrich their vocabulary with more specific terms over time using the explorARTorium. A possible explanation for this phenomenon might be that if users tag an artwork that has already been tagged previously with less specific tags like *man* or *woman*, they naturally have to assign more specific tags and thus the specificity of the vocabulary increases.

4.4 Conclusion

The analysis of the explorARTorium's folksonomy conducted in this chapter clearly shows that the users' motivation to tag artworks drastically decreases over time. Therefore, the explorARTorium is in need to provide an incentive to its users not to lose interest and to start tagging artworks again. Sigurbjörnsson and van Zwol (2008) showed that suggestions for tags help users to annotate images and that the number of tags increases by providing tag recommendations. In order to use this opportunity and to give the users of the artwork collection an incentive to tag pictures a Tag Recommendation Framework for system-generated suggestions for appropriate

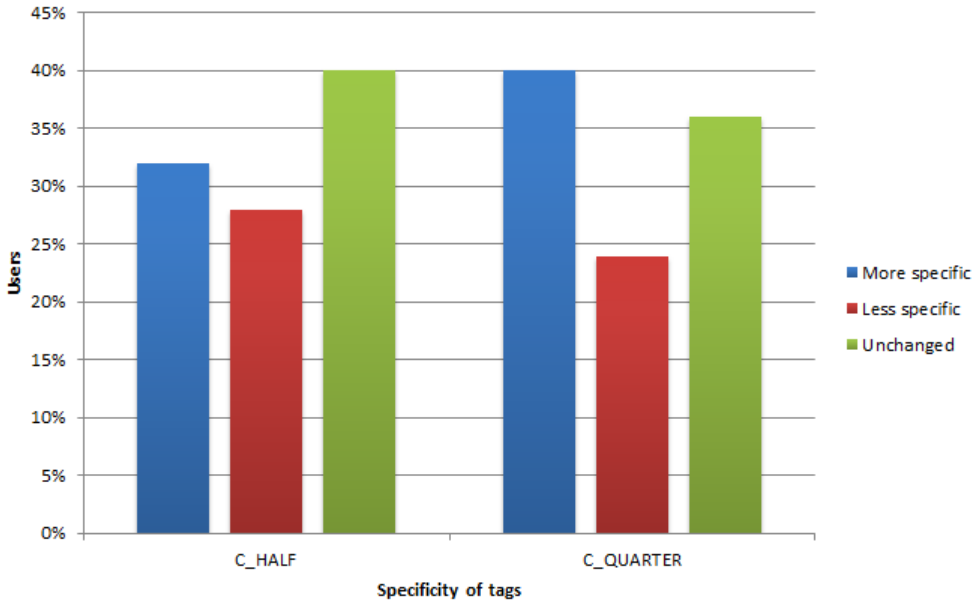

Figure 4.17: Comparison of specificity of the users' vocabulary over time.

tags (based on tags extracted from the folksonomy) is developed and presented in Chapter 5. As noted earlier in Chapter 1, this approach invites the user to start looking for the suggested tags in the artwork in order to verify them. In the following, this might lead to new tags, because the user might discover previously unnoticed elements in the artwork. The tag recommendations are also used to present the user additional artworks in the same context if he or she wants to further explore artworks assigned with a particular tag.

CHAPTER 5

Tag Recommendation

In this chapter, the essence of this master's thesis, the Tag Recommendation Framework for the explorARTorium, is presented. Firstly, a general overview of the model of the framework is given in Section 5.1 and the different phases of the framework are inspected in detail. In Section 5.2, an evaluation of the framework is conducted and the results are discussed.

5.1 Tag Recommendation Framework

The framework developed for this master's thesis constitutes an automatic tag recommending system for untagged artworks of the explorARTorium by combining the data mining and recommender system techniques as described in Chapter 2.

The life cycle of the framework consists of the following four chronological phases:

1. **Import and data preparation**: Read the tag database, i.e. import the whole artwork collection and their assigned tags for artworks that have already been tagged by users.

2. **Data mining**: Mining of frequent itemsets and association rules in the subset of tagged artworks of the imported data.

3. **Recommendation engine**: With the help of calculated frequent itemsets and association rules, the recommender system assigns tag recommendations to the untagged artworks.

4. **Result visualization**: The output of the previous phase, the generated tag recommendations, are inserted in the database of the artwork collection to be presented to the user.

These phases form the Tag Recommendation Process, which is fully automated and intended to be scheduled on a regular basis, so that new user-generated content can be incorporated in the next run of the process. Therefore, a system covering every aspect of the life cycle has to be designed.

Figure 5.1 illustrates the model of the framework and the four phases as described above.

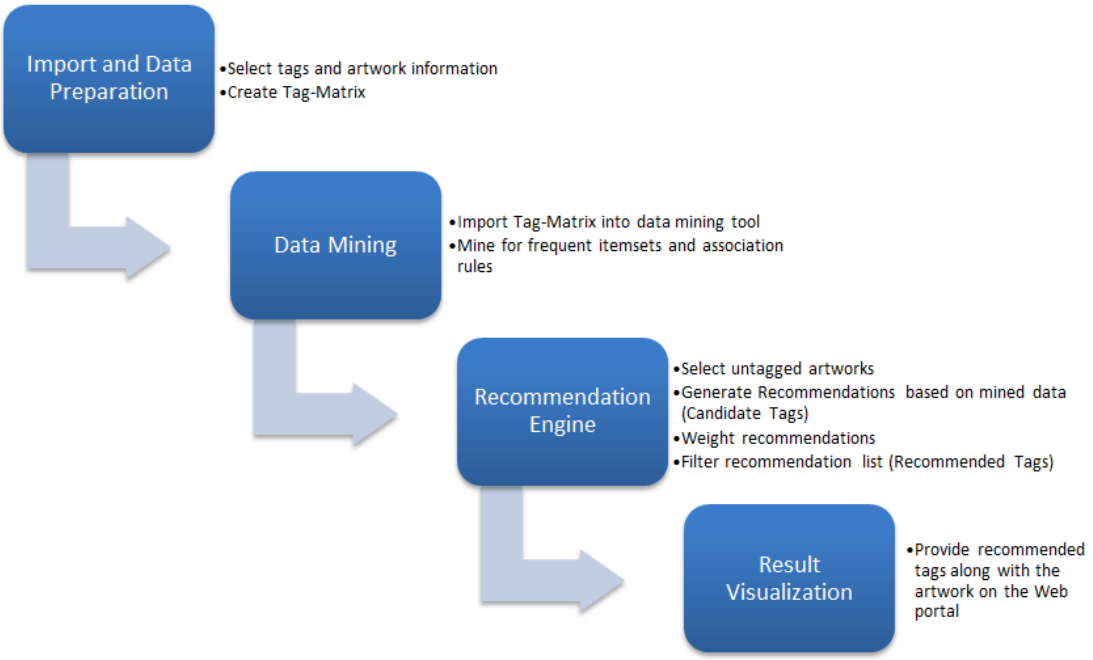

Figure 5.1: Schematic model of the Tag Recommendation Framework.

5.1.1 Import and Data Preparation

The *Import and Data Preparation* phase is the first step in the tag recommendation process. With the help of the Structured Query Language (SQL), information about the tagged artworks and their assigned tags is selected out of the explorARTorium database and combined into a tag matrix. This matrix contains descriptions of the artworks along with their attributes and their corresponding tags and is exported as a comma-separated value (CSV) file based on a defined structure. Every line of the file (except the first line which contains header information) represents an artwork and contains the following information:

- **PICID**: The ID of the artwork in the database.
- **ARTISTNAME**: The artist name of the artwork.
- **TITLE**: The title of the artwork.
- **THEME**: The theme or motive of the artwork (e.g. *portrait*, *religious*, *mythological*, etc).
- **REGION**: The region in which the artist lived and was influenced or the school the artist is associated with respectively (e.g. *Italian*, *Dutch*, *German*, etc).
- **TIMEFRAME**: The time frame in which the artwork was created (in steps of 50 years).

These attributes of the artworks represent their categorization, always appear in the same order and are followed by all distinct tags that exist in the database. There are 2 possible values:

- **0**: The tag is not assigned to the artwork.
- **1**: The tag is assigned to the artwork.

The tags are ordered decreasingly according to their overall count.
Table 5.1 shows a small excerpt of the tag matrix.

PICID	ARTISTNAME	TITLE	THEME	REGION	TIMEFRAME	wolken	himmel	engel	frau	mann	heiligenschein	maria	portrait	baeume	...
...															
12759	MICHELANGELO	Medallion	religious	Italian	1501-1550	0	0	0	0	0	0	0	0	0	
12761	MICHELANGELO	Last Judgment	religious	Italian	1501-1550	0	0	0	0	0	0	0	0	0	
12762	MICHELANGELO	Last Judgment	religious	Italian	1501-1550	0	0	0	0	0	0	0	0	0	
12856	MICHELE DA VERONA	Crucifixion	religious	Italian	1501-1550	0	0	0	0	0	0	0	0	0	
12870	JAN MIEL	Carnival Time in Rome	genre	Flemish	1601-1650	0	0	0	0	0	0	0	0	0	
12871	JAN MIEL	Genre Scene	genre	Flemish	1601-1650	1	1	0	1	0	0	0	0	1	
12872	JAN MIEL	Hunters at Rest	genre	Flemish	1601-1650	0	0	0	0	0	0	0	0	0	
12873	Hans MIELICH	High Altar	religious	German	1501-1550	0	0	0	0	0	0	0	0	0	
12875	MIEREVELD	Anatomy Lesson of Dr. Willem van der Meer	genre	Dutch	1601-1650	0	0	0	0	0	0	0	0	0	
12876	MIEREVELD	Portrait of Frederick Hendrick Prince of Orange-Nassau	portrait	Dutch	1601-1650	0	0	0	0	1	0	0	1	0	
12877	MIEREVELD	Portrait of Lubert Gerritsz.	portrait	Dutch	1601-1650	0	0	0	0	0	0	0	1	0	
12878	MIEREVELD	Prince Maurits Stadhouder	portrait	Dutch	1601-1650	0	0	0	0	1	0	0	0	0	
12879	MIEREVELD	Portrait of Maurits Orince of Orange-Nassau	portrait	Dutch	1601-1650	0	0	0	0	0	0	0	0	0	
12880	MIEREVELD	Portrait of a Woman	portrait	Dutch	1601-1650	0	0	0	1	0	0	0	1	0	
12884	the Elder MIERIS	Boy Blowing Bubbles	genre	Dutch	1651-1700	0	0	0	0	0	0	0	0	0	
12885	the Elder MIERIS	Carousing Couple	genre	Dutch	1651-1700	0	0	0	0	0	0	0	0	0	
12886	the Elder MIERIS	The Cloth Shop	genre	Dutch	1651-1700	0	0	0	0	0	0	0	0	0	
12887	the Elder MIERIS	Duet	interior	Dutch	1651-1700	0	0	0	1	1	0	0	0	0	
12889	the Elder MIERIS	A Meal of Oysters	genre	Dutch	1651-1700	0	0	0	1	1	0	0	0	0	
12890	the Elder MIERIS	Young Woman in the Morning	genre	Dutch	1651-1700	0	0	0	1	0	0	0	0	0	
12895	the Elder MIERIS	Portrait of a Young Man	portrait	Dutch	1651-1700	0	0	0	0	0	0	0	0	0	
12896	the Elder MIERIS	The Painters Studio	genre	Dutch	1651-1700	0	0	0	0	0	0	0	0	0	
12899	the Elder MIERIS	Woman Writing a Letter	genre	Dutch	1651-1700	0	0	0	1	1	0	0	0	0	
12900	the younger MIERIS	Old Peasant Holding a Jug	genre	Dutch	1701-1750	0	0	0	0	0	0	0	1	0	
12901	Willem van MIERIS	The Escape Bird	genre	Dutch	1701-1750	0	0	0	0	0	0	0	0	0	
12903	Willem van MIERIS	The Greengrocer	genre	Dutch	1701-1750	0	0	0	0	0	0	0	0	0	
12904	Willem van MIERIS	The Lute Player	genre	Dutch	1701-1750	0	0	0	0	0	0	0	0	0	
12905	Willem van MIERIS	Man with Pipe	genre	Dutch	1701-1750	0	0	0	0	0	0	0	1	0	
12909	Willem van MIERIS	The Spinner	genre	Dutch	1701-1750	0	0	0	1	1	0	0	0	0	
12910	Giovanni MIGLIARA	Venetian View	landscape	Italian	1801-1850	1	1	0	0	0	0	0	0	0	
12911	Nicolas MIGNARD	Virgin and Child	religious	French	1651-1700	1	0	0	0	0	0	0	0	0	
12912	Piere MIGNARD	Clio	mythological	French	1651-1700	0	0	0	1	0	0	0	0	0	
12913	Piere MIGNARD	Girl Blowing Soap Bubbles	genre	French	1651-1700	0	0	0	0	0	0	0	0	0	
12914	Piere MIGNARD	Mystic Marriage of St Catherine	religious	French	1651-1700	0	0	1	0	0	0	1	0	0	
12914	Piere MIGNARD	The Presentation of the Virgin in the Temple	religious	French	1651-1700	0	0	0	0	0	0	0	0	0	
...															

Table 5.1: Small excerpt of the tag matrix.

For the next step of the tag recommendation process, this CSV file containing the tag matrix is imported into Weka.

5.1.2 Data Mining

In this second phase, the imported tag matrix is read by using the Weka Java Library with the help of the class `weka.core.converters.ConverterUtils.DataSource`, which converts the CSV data with the `getDataSet` method into Weka instances (`weka.core.Instances`). Now the data is mined for frequent itemsets and association rules according to the specified preferences with the Apriori algorithm (`weka.associations.Apriori`).

The Weka implementation of the Apriori algorithm allows the following parameters to be defined (Community documentation for Weka, 2011):

- **car**: Boolean value; if set to $true$, class association rules are mined instead of (general) association rules.
- **classIndex**: Integer value; depicts the index of the class attribute.
- **delta**: Double value; the support of the mined rules is iteratively decreased by this delta factor until the minimum support is reached or the required number of rules (*numRules*) has been generated.
- **lowerBoundMinSupport**: Double value; the lower bound for minimum support.
- **upperBoundMinSupport**: Double value; the upper bound for minimum support.
- **metricType**: Sets the type of the 4 available metrics (confidence, leverage, lift and conviction) by which the mined rules are ranked.
- **minMetric**: Double value; only rules with higher scores than this minimum metric score value are considered.
- **numRules**: Integer value; the number of rules to find.
- **outputItemSets**: Boolean value; if set to $true$, the itemsets are added to the output of the associator.
- **removeAllMissingCols**: Boolean value; if set to $true$, columns with all missing values are removed.
- **significanceLevel**: Double value; sets the significance level, if *confidence* is chosen as metric.

The setting of these parameters is carried out by invoking the corresponding setter methods of the `weka.associations.Apriori` class. The actual association rule mining task is started with the `buildAssociations` method, which takes the previously created `Instances` as input. The output of the associator (list of large itemsets and association rules) is written into a text file, because Weka lacks the possibility to access the mined frequent itemsets and

association rules in a way that is suitable for the tag recommendation process. All this functionality is encapsulated in the classes of the package `da.mining`. This file is used amongst others as input for the recommendation engine in the next phase of the process.

5.1.3 Recommendation Engine

In this third phase of the tag recommendation process, the actual recommendation of tags for untagged artworks takes place. Therefore, several assisting steps are necessary: Firstly, the previously mined frequent itemsets and association rules have to be read and imported into a separate suitable data structure. Secondly, information about the untagged artworks are imported into the data structure as well. This is done by using the parsers and database-importers in the package `da.readers`:

- `ItemSetReader`: imports the frequent itemsets into the data structure.
- `AssociationRulesReader`: imports the association rules into the data structure.
- `UntaggedPictureReader`: imports information on the untagged artworks, for which recommendations are to be generated, into the data structure.

Whereas the `ItemSetReader` and `AssociationRulesReader` import the data from the generated text file in the prior phase, the `UntaggedPictureReader` can either read from a CSV file or can access the explorARTorium database directly via SQL.

The `ItemSetReader` takes the size of the large itemsets, which is to be read, as input and then starts parsing the relevant itemsets. Only itemsets which contain at least one of the inherent categorizations of an artwork (title, theme, region, timeframe) are accepted by the parser. The same restriction applies to the `AssociationRulesReader`.

After these required preparations, the `Recommender` (located in the package `da.engine`) is ready to start generating recommended tags. The general and simplified mode of operation of the recommender is as follows:

- The recommender tries to find frequent itemsets and association rules containing the same attribute values which the artwork to receive tag recommendations has, i.e. find matches in the data structure.
- Every match is weighted depending on the category the matching attribute value belongs to and the tag of the matching itemset/rule is added to a list along with computed quality criteria measuring the accuracy of fit for the tag.
- This list of tag recommendations is sorted by the accuracy and trimmed by pruning tags with low relevance according to the quality measures, to finally contain the best possible tag recommendations for the given untagged artwork.

This conceptual design of the Recommendation Engine also incorporates the following considerations, which are indirectly addressed in the previous itemization:

- **Selection of significant attributes**: The following attributes are available for the matching process of the Recommendation Engine: $artist$, $title$, $theme$, $region$ and $timeframe$. All of these attributes are considered valuable for the matching process except for the attribute $artist$, because with the four attributes it is possible to identify image perception stereotypes for the artworks (artworks with the same $title/theme/region/timeframe$ share similarities and are therefore suitable for the matching process), whereas artworks of the same artist do not necessarily show similar properties concerning these stereotypes. Therefore the attribute $artist$ is not included in the algorithm of the Recommendation Engine.

- **Weighting of matches**: As noted earlier, every recommended tag is weighted according to the matching attribute category. The weighting scheme is developed according to Equation (5.1); an example with values proven useful is depicted in Table 5.2.

$$weight(title) > weight(theme) > weight(region) > weight(timeframe) \quad (5.1)$$

  If more than one attribute matches the association rule (or itemset), the weights of the matching categories are accumulated. For some special cases, the weighting is altered in order to optimize the result of the recommendation engine: if the theme of the artwork is either *portrait*, *landscape* or *religious*, the weight of $region$ is doubled (in case the rule/itemset also contains the $region$ attribute); for the themes *portrait* and *religious* the weight of $timeframe$ is tripled (in case the rule/itemset also contains the $timeframe$ attribute). These weighting parameters have been derived heuristically after reviewing the first results of the recommender system. The rationale behind the increased weighting in *region* for *landscape*, for example, was driven by the expectation that a painter shows landscapes representative of his or her region or origin. Of course, this might be wrong for the specific case, in general, however, it seems that it might hold true.

- **Computing of quality criteria**: To give an estimate of the accuracy of fit for the recommended tags and therefore to be able to rank them, a measure named $recorating$ is computed, combining the following two values by multiplication:

  - Confidence of the association rule (or logarithm of the frequency in the case of frequent itemsets)
  - Weighting of the matching attribute category: see previous bullet point.

  Through the combination of these two values, it can be ensured that the matching attribute category as well as the "strength" of the association rule (or itemset) is weighted accordingly. In order not to overburden the users of the explorARTorium with too many recommended tags, only the statistically most valuable recommendations are displayed to the user. To determine the number of tags to be displayed, the algorithm looks for significant "breaks" (i.e. gaps) in the $recorating$ between the tags in the list of recommended tags. Subsequently, this finding of "breaks" is discussed in detail along with the algorithm.

Due to the modular design of the framework, it is very easy to alter all of the addressed values, properties and weighting parameters without touching the implementation of the recommendation engine through adjusting the desired variables in the properties file `reco.properties`, if there may be the need to change these values in the future. In the implementation of the framework, information about objects like artworks, itemsets, association rules or recommended tags is encapsulated in corresponding classes, e.g. an artwork is represented by the class `da.models.Picture` holding information on the artist, theme, region, timeframe, etc., whereas a tag recommendation is represented by the class `da.models.RecoTag` encapsulating information about the tag, the weighting, the strength and the $recorating$ and implements the `java.lang.Comparable`-Interface in order to allow easy comparison of tag recommendations based on their $recoratings$.

Matching category	Weight	Comment
Title	20	
Theme	5	
Region	3	in case of *portrait*, *landscape* or *religious* theme, the weight of $region$ is doubled
Timeframe	1	in case of *portrait* or *religious* theme, the weight of $timeframe$ is tripled

Table 5.2: Weighting scheme for the recommendation engine.

The recommender works according to the following algorithmic procedure:

1. The recommendation engine is started and given an artwork (along with attributes like $title$, $theme$, $region$ and $timeframe$) to recommend tags.

2. For every mined association rule, the itemsets of these rules are compared to the attributes of the given artwork:

   - For every item in the itemset that equals one of the given artwork's attributes, the according weight is added to a variable i. In special cases (such as portraits, landscapes or religious paintings as described earlier) some weights are changed.
   - If at least one weight was assigned, i.e. $i > 0$, it is checked for every item in the itemset if it is already included in the recommendation list. If there is a match, two cases are distinguished:
     a) If the current item has a higher $recorating$ than the one already in the list, the item with the lower $recorating$ is replaced with the higher one.
     b) If the current item has a lower $recorating$ than the one already in the list, it is pruned.

     If the current item was not in the recommendation list before, it is added to the list.

3. After all possible tag recommendations are identified for the artwork, the list of recommendations is sorted descending by the value of $recorating$.

4. The list is trimmed to the maximum number of allowed recommendations specified in the properties file.

5. It is checked if a "break" in $recorating$ between two tags can be found. Therefore a specified percentage of the difference between the $recorating$ values of the first and the last recommendation is taken (named as $breakvalue$) and checked for every recommendation if the difference in $recorating$ of two adjacent recommendations is larger than the $breakvalue$. If this is the case, all the following recommendations after the identified break are pruned.

6. If no break can be identified, it is checked if every recommendation complies with the specified minimum $recorating$-threshold.

7. After all the previous steps are completed, the recommendation engine outputs the tag recommendation list.

Algorithm 5.1 shows the first part of this algorithm (steps 1-2) in pseudocode which is applied to every untagged artwork and generates a list of tag recommendations for it. Algorithm 5.2 shows the second part of the algorithm (steps 3-7), sorting the list of tag recommendations and finding breaks.

The generated tag recommendations calculated in this phase are now passed to the subprocess of the next phase.

5.1.4 Result Visualization

This fourth and last phase of the tag recommendation process deals with the output and result visualization of the tags generated in the previous phase. The functionality is encapsulated by the `RecommendedPictureWriter` in the package `da.writers`. The tag recommendations for the untagged pictures are written into a text file and inserted into the explorARTorium database table $ent\_tag\_recommendations$. The table structure is shown in Table 5.3. The column id contains the primary key (PK) for this table, with $ent\_tag\_recommendations\_id\_seq$ being a simple sequence which returns an incremented value for the ID of the $ent\_tag\_recommendations$ table whenever a data set is inserted. The field $ent\_document\_id$ is used to reference the ID of the artwork via a foreign key (FK) to the field id of the table $ent\_documents$. The field tag contains the tag to be recommended. The field $date$ harbors the date and time when the tag was recommended via the Tag Recommendation Framework. The fields $value$, $confidence$ and $recorating$ contain the computed quality measure values for the recommended tag (according to their definitions discussed in the previous phase).

Users of the explorARTorium can now experience the benefit of recommendations of possible tags when viewing untagged artworks. Figure 5.2 shows a screenshot of the explorARTorium where tag recommendations generated by the Tag Recommendation Framework for an untagged artwork (*Young Woman Drinking* by *Pieter de HOOCH*) below the image are provided (*frau* (woman), *tisch* (desk), *hut* (hat), *krug* (jar, mug)).

input : An artwork pic with attributes $picid, title, theme, region, timeframe$
output : A list of tag recommendations $reco\_list$ for this artwork
variables: $weight\_title, weight\_theme, weight\_region, weight\_timeframe$

```
1  foreach itemset 'iset' in the list of itemsets 'ilist' do
2      i ← 0;
3      reset weighting values to default;
4      if pic.title = iset.title then
5          i = i + weight_title
6      end
7      if pic.theme = iset.theme then
8          i = i + weight_theme;
9          if pic.theme = "portrait" ∨ pic.theme = "religious" then
10             weight_region = weight_region * 2;
11             weight_timeframe = weight_timeframe * 3;
12         end
13         if pic.theme = "landscape" then
14             weight_region = weight_region * 2;
15         end
16     end
17     if pic.region = iset.region then
18         i = i + weight_region
19     end
20     if pic.timeframe = iset.timeframe then
21         i = i + weight_timeframe
22     end
23     if i > 0 then
24         foreach Tag 'tag1' in the list of tags 'iset.tags' do
25             foreach Tag 'tag2' in the list of tags 'reco_list' do
26                 if tag1 = tag2 ∧ tag1.recorating > tag2.recorating then
27                     Replace tag2 with tag1 in reco_list
28                 end
29             end
30             if tag1 is not in reco_list then
31                 Add tag1 to reco_list
32             end
33         end
34     end
35 end
36 return reco_list
```

Algorithm 5.1: Tag recommendation engine algorithm part 1 in pseudocode.

input : A list of tag recommendations $reco\_list$ for this artwork
output : A sorted and trimmed list of tag recommendations $reco\_list$ for this artwork
variables: $max\_number\_reco, diff factor, reco\_min\_threshold$

```
1  Sort reco_list by recorating descending;
2  Trim reco_list to max_number_reco size;
3  diff = reco_list.FirstElement.recorating − reco_list.LastElement.recorating;
4  breakvalue = diff ∗ diff factor;
5  foreach Tag 'tag' in the list of tags 'reco_list' do
6  |  if tag.recorating − reco_list.NextTag.recorating > breakvalue then
7  |  |  Dump all elements of reco_list after tag
8  |  end
9  end
10 if No break was found then
11 |  Dump all elements of reco_list with recorating < reco_min_threshold
12 end
13 return reco_list
```

Algorithm 5.2: Tag recommendation engine algorithm part 2 in pseudocode.

Column	Type	Default value	Constraints
id	integer	nextval('ent_tag_recommendations_id_seq')	PK
ent_document_id	bigint		FK: ent_documents(id)
tag	character varying		
date	timestamp without time zone	now()	
value	integer		
confidence	double precision		
recorating	double precision		

Table 5.3: Structure of the explorARTorium database table $ent\_tag\_recommendations$.

5.2 Evaluation

In this section the results from an evaluation of the Tag Recommendation Framework are presented. The first part gives an overview of the methodology of the evaluation process. In the second part the result of the evaluation is presented and discussed.

5.2.1 Evaluation methodology

In order to evaluate the quality and precision of the Tag Recommendation Framework, a group of users reviewed the accuracy of the generated tag recommendations. Therefore an evaluation

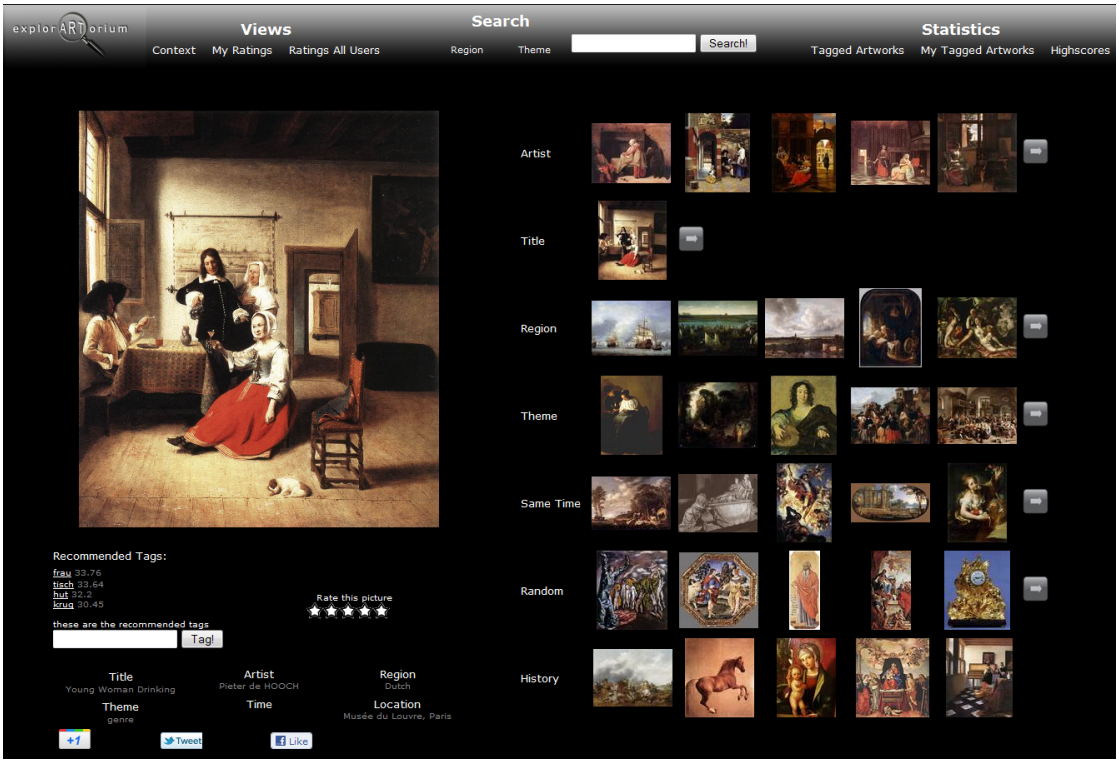

Figure 5.2: Screenshot of the explorARTorium showing tag recommendations for *Young Woman Drinking* by *Pieter de HOOCH*.

system called the Evaluatorium[1] was set up based on the Tagging-tool[2] (cf. Section 3.1).

The evaluation process of tag recommendations for a given artwork can be described as follows:

- The user views an artwork chosen by the evaluation system (according to some specific criteria, which are discussed in detail later).
- The tag recommendations are listed right next to the artwork.
- The user rates these tag recommendations with the help of thumbs-up and thumbs-down symbols right next to each tag, based on his or her opinion of the accuracy of the tag, i.e. if he or she thinks the tag is relevant for the artwork or not based on the following scale:
  - 2 thumbs up = tag fits well
  - 1 thumb up = tag fits approximately
  - 1 thumb down = tag fits barely

[1] `http://vsem.ec.tuwien.ac.at/evaluatorium/`; [accessed 13-October-2011]
[2] `http://vsem.ec.tuwien.ac.at/taggingtool/`; [accessed 04-October-2011]

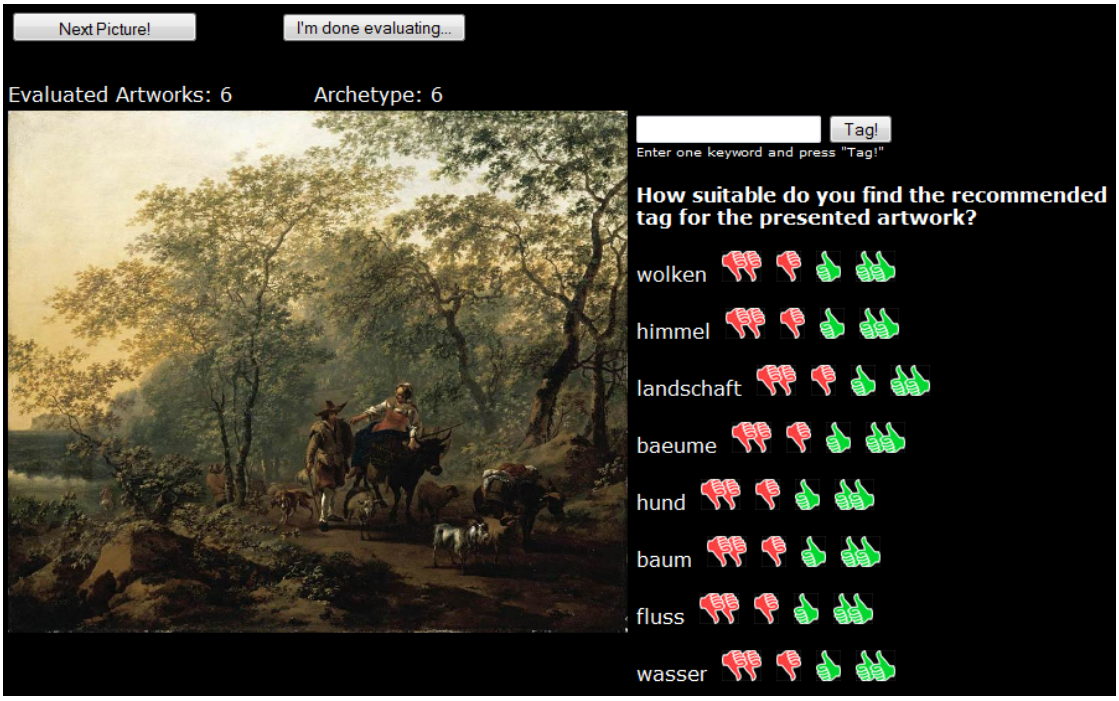

Figure 5.3: Screenshot of the evaluation system.

– 2 thumbs down = tag does not fit at all

Through this process the user is shown as many artworks as he or she likes to evaluate. Figure 5.3 shows a screenshot of the Evaluatorium.

For the selection of an artwork presented to the user to evaluate, the following considerations are incorporated in the selection algorithm:

- Due to the bias in the data set of the collection of artworks (as discovered during the analysis in Chapter 3) regarding the different attributes (themes, regions, timeframes, etc.), a random selection of artworks would result in a biased evaluation.
- To counteract this bias, different archetypes[3] of artworks (i.e. artworks which have certain attributes in common) in the data set are identified.
- For each of these archetypes, an artwork representing the specific archetype is chosen randomly to be evaluated.

With this approach it is assured that the distribution of evaluated artworks is even regarding the archetypes.

[3] "An archetype is a universally understood symbol or term or pattern of behavior, a prototype upon which others are copied, patterned, or emulated." (Wikipedia, 2011a)

The results of the users' evaluation are stored in the explorARTorium's database table $ent\_reco\_evaluation$ to be later used in this chapter for the analysis of the evaluation. The table structure is shown in Table 5.4. The column id contains the primary key (PK) for this table, with $ent\_reco\_evaluation\_id\_seq$ being a simple sequence which returns an incremented value for the ID of the $ent\_reco\_evaluation$ table whenever a data set is inserted. The field $ent\_tag\_recommendation\_id$ is used to reference the ID of the recommended tag via a foreign key (FK) to the field id of the table $ent\_tag\_recommendations$. The field $ent\_document\_id$ is used to reference the ID of the artwork via a foreign key (FK) to the field id of the table $ent\_documents$. The field $username$ contains the username of the evaluating user. The field $date$ harbors the date and time when the tag was evaluated. The field $evaluation$ contains the actual evaluation as integer value (2 thumbs up = 2; 1 thumb up = 1; 1 thumb down = -1; 2 thumbs down = -2).

For the whole evaluation process it is very important to always keep in mind that the archetypical perception is completely based on user-generated content and derived from a random, biased and naturally incomplete data set (the folksonomy of the explorARTorium), so the Tag Recommendation Framework and its suggestions can only be as "good" as the underlying folksonomy.

Column	Type	Default value	Constraints
id	integer	nextval('ent_reco_evaluation_id_seq')	PK
ent_tag_recommendation_id	integer		FK: ent_tag_recommendations(id)
ent_document_id	integer		FK: ent_documents(id)
username	character varying		
date	timestamp without time zone	now()	
evaluation	integer		

Table 5.4: Structure of the explorARTorium database table $ent\_reco\_evaluation$.

Archetypes

For the identification of the different archetypes contained in the data set of the explorARTorium the following approach using cluster analysis is developed:

- **Idea**: The set of artworks is divided into groups (clusters) so that the artworks in the same cluster are more similar (in terms of tags assigned by the Tag Recommendation Framework) to each other than to those in other clusters (cf. Section 2.1).

- With the help of the data mining tool *Weka* (introduced earlier in this chapter), the **clustering** algorithm *SimpleKMeans* is applied to the artworks along with their tags contained in the data set.

Figure 5.4: Archetype 1: *Still-Life* by *Frans SNYDERS*.

- The result of the cluster analysis are 12 clusters, representing the different archetypes of artworks in the dataset.

Table 5.5 provides the distribution of themes and regions over the different archetypes (regions with less than 10 artworks in total are omitted to improve the clarity of the table) as of October 11, 2011. The following paragraphs give an overview of the identified archetypes.

Archetype 1 All artworks corresponding to *Archetype 1* represent the theme *still-life*, with the majority of them created in *Flemish*, *French*, *Dutch* and *Spanish* regions. The most frequent tag recommendations for this archetype are depicted in Table B.1 in the Appendix. Figure 5.4 shows a typical example of an artwork for this archetype (*Still-Life* by *Frans SNYDERS*).

Archetype 2 *Archetype 2* harbors only *religious* artworks, with the majority of them created in *Flemish* and *Spanish* regions. The most frequent tag recommendations for this archetype are depicted in Table B.2 in the Appendix. Figure 5.5 shows a typical example of an artwork for this archetype (*Triptych of the Sedano Family* by *Gerard DAVID*).

Archetype 3 Nearly all of the artworks corresponding to *Archetype 3* represent the theme *religious*, with the majority of them created in *Italian* regions. The most frequent tag recommendations for this archetype are depicted in Table B.3 in the Appendix. Figure 5.6 shows a typical example of an artwork for this archetype (*Annunciation* by *Fra Filippo LIPPI*).

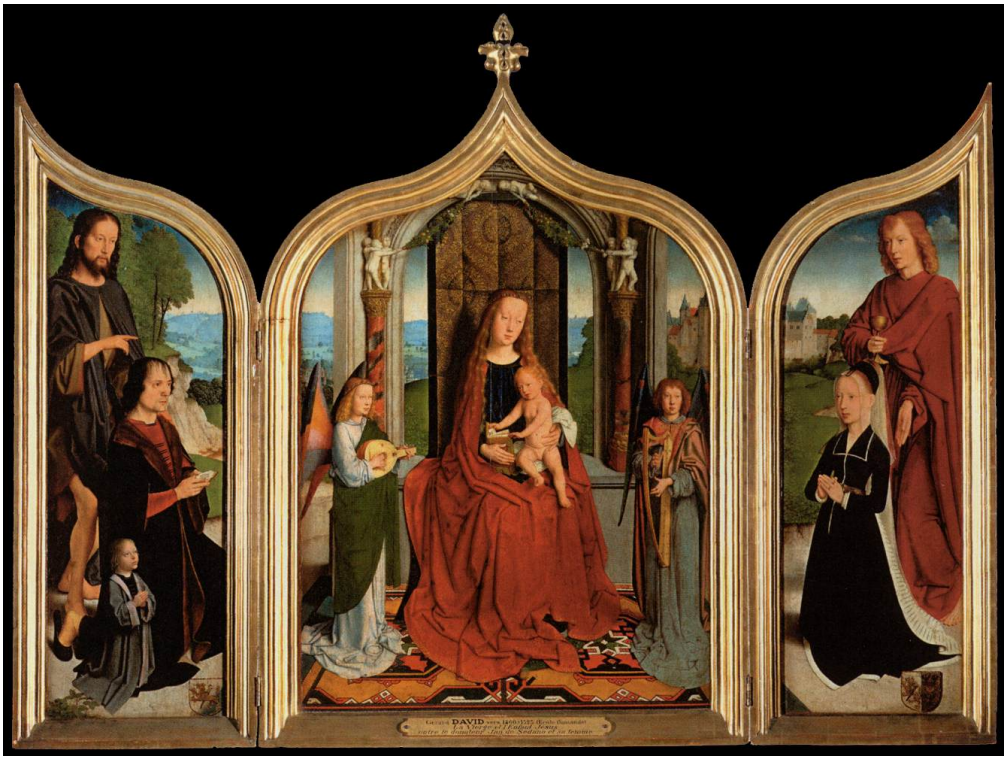

Figure 5.5: Archetype 2: *Triptych of the Sedano Family* by *Gerard DAVID*.

Archetype 4 Most of the artworks corresponding to *Archetype 4* illustrate the theme *religious*, with the majority of them created in *French*, *Dutch*, *German* and *Netherlandish* regions. The most frequent tag recommendations for this archetype are depicted in Table B.4 in the Appendix. Figure 5.7 shows a typical example of an artwork for this archetype (*The Angel Is Opening Christ's Tomb* by *Benjamin Gerritsz. CUYP*).

Archetype 5 All artworks corresponding to *Archetype 5* represent the theme *landscape*, with the majority of them created in *Dutch* and *Flemish* regions. The most frequent tag recommendations for this archetype are depicted in Table B.5 in the Appendix. Figure 5.8 shows a typical example of an artwork for this archetype (*Ferry-boat* by *Jan VICTORS*).

Archetype 6 Like *Archetype 5*, all artworks corresponding to *Archetype 6* represent the theme *landscape*, but with the majority of them created in *Italian* and *French* regions. The most frequent tag recommendations for this archetype are depicted in Table B.6 in the Appendix. Figure 5.9 shows a typical example of an artwork for this archetype (*Return of the Bucentoro to the Molo on Ascension Day* by *CANALETTO*).

Archetype 7 Most of the artworks corresponding to *Archetype 7* illustrate the themes *portrait* and *genre*, with the majority of them created in *Dutch*, *French* and *Spanish* regions. The most

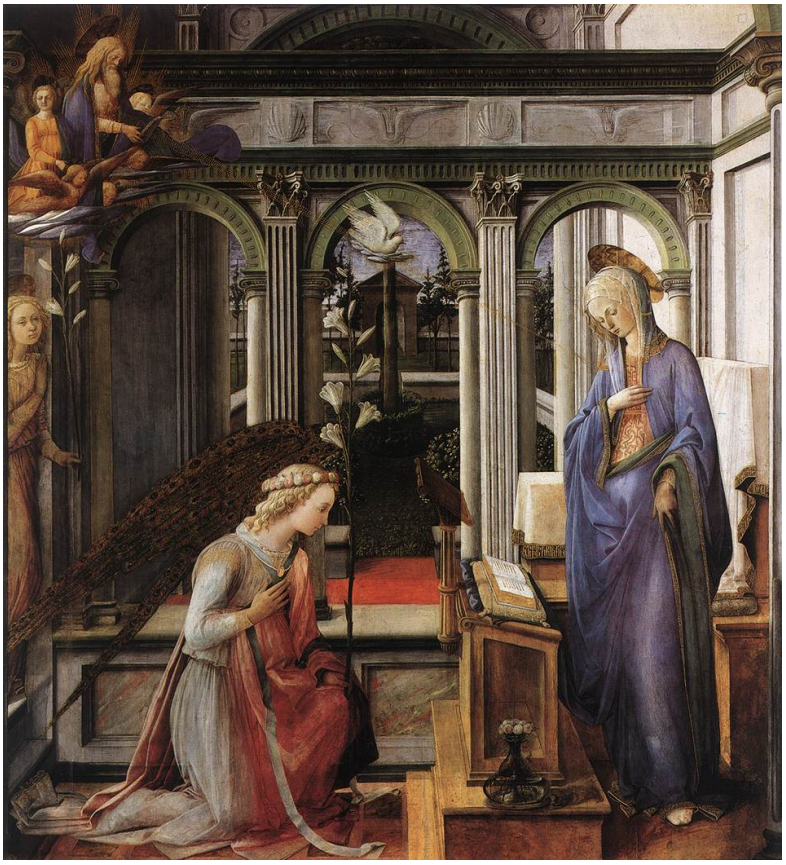

Figure 5.6: Archetype 3: *Annunciation* by *Fra Filippo LIPPI*.

frequent tag recommendations for this archetype are depicted in Table B.7 in the Appendix. Figures 5.10 (*Self-portrait* by *Pieter Jansz. van ASCH*), 5.11 (*The Procuress* by *Johannes VERMEER*) and 5.12 (*Young Woman Drinking* by *Pieter de HOOCH*) show typical examples of artworks for this archetype.

Archetype 8 All artworks corresponding to *Archetype 8* are *portraits* created in *Italy*. The most frequent tag recommendations for this archetype are depicted in Table B.8 in the Appendix. Figure 5.13 shows a typical example of an artwork for this archetype (*Man in Military Costume* by *TIZIANO Vecellio*).

Archetype 9 All artworks corresponding to *Archetype 9* are *portraits*, with the majority of them created in *Flemish* and *German* regions. The most frequent tag recommendations for this archetype are depicted in Table B.9 in the Appendix. Figure 5.14 shows a typical example of an artwork for this archetype (*Portrait of a Man* by *Frans the Elder POURBUS*).

Figure 5.7: Archetype 4: *The Angel Is Opening Christ's Tomb* by *Benjamin Gerritsz. CUYP*.

Archetype 10 Most of the artworks corresponding to *Archetype 10* represent the themes *genre* and *interior*, with the majority of them created in *French* and *Dutch* regions. The most frequent tag recommendations for this archetype are depicted in Table B.10 in the Appendix. Figure 5.15 shows a typical example of an artwork for this archetype (*The Progress of Love: The Lover Crowned* by *Jean-Honoré FRAGONARD*).

Archetype 11 Like *Archetype 10*, the artworks corresponding to *Archetype 11* represent the themes *genre* and *interior*, with the majority of them created in *Flemish* and *Dutch* regions. The most frequent tag recommendations for this archetype are depicted in Table B.11 in the Appendix. Figure 5.16 shows a typical example of an artwork for this archetype (*Tea Time* by *Jan Jozef II HOREMANS*).

Archetype 12 The majority of the artworks corresponding to *Archetype 12* are *mythological* artworks from *Italy*. The most frequent tag recommendations for this archetype are depicted in Table B.12 in the Appendix. Figure 5.17 shows a typical example of an artwork for this archetype (*Diana and Actaeon* by *Francesco ALBANI*).

Figure 5.8: Archetype 5: *Ferry-boat* by *Jan VICTORS*.

Figure 5.9: Archetype 6: *Return of the Bucentoro to the Molo on Ascension Day* by *CANALETTO*.

Figure 5.10: Archetype 7: *Self-portrait* by *Pieter Jansz. van ASCH*.

Figure 5.11: Archetype 7: *The Procuress* by *Johannes VERMEER*.

Figure 5.12: Archetype 7: *Young Woman Drinking* by *Pieter de HOOCH*.

Figure 5.13: Archetype 8: *Man in Military Costume* by *TIZIANO Vecellio*.

Figure 5.14: Archetype 9: *Portrait of a Man* by *Frans the Elder POURBUS*.

Figure 5.15: Archetype 10: *The Progress of Love: The Lover Crowned* by *Jean-Honoré FRAGONARD*.

Figure 5.16: Archetype 11: *Tea Time* by *Jan Jozef II HOREMANS*.

Figure 5.17: Archetype 12: *Diana and Actaeon* by *Francesco ALBANI*.

Archetype	Theme	American	Austrian	Dutch	English	Flemish	French	German	Hungarian	Italian	Netherlandish	Russian	Scottish	Spanish	Swiss	**Total**
1	still-life	1	2	28		46	34	13	2	5	4		2	28		**165**
2	religious					501								206		**707**
3	other										8					**8**
	religious			10			8	218		2618	37			14	1	**2906**
	study										1					**1**
4	historical							3			11			3	1	**18**
	other														1	**1**
	religious		39	129	11	5	144	96			91		1	2	9	**527**
	still-life									1						**1**
	study							1			17					**18**
5	landscape			305		116		4						5	2	**432**
6	landscape	6	2		44		106	76	4	171	13	3		4	7	**436**
7	genre			293				1			1			3		**298**
	historical			5				3						6		**14**
	interior			4												**4**
	other			7	9			10								**26**
	portrait	6	18	211	44	9	214	62	7	11	3	8	7	102	5	**707**
	study			3				3								**6**
8	portrait									313						**313**
9	portrait			6	7	148		97			1					**259**
10	genre					24	137									**161**
	historical				3		1					1	4			**9**
	interior		1	42								1				**44**
11	genre					33	1			1				3		**38**
	interior			18												**18**
12	genre									9						**9**
	historical	2				26	35					1				**64**
	interior								1							**1**
	mythological									478						**478**
	other	6	1		4	7	41								5	**64**
	study		1		1	8	12									**22**
Total		**21**	**64**	**1061**	**123**	**923**	**733**	**587**	**14**	**3607**	**187**	**14**	**14**	**376**	**31**	**7755**

Table 5.5: Distribution of themes and regions over the different archetypes (October 11, 2011).

5.2.2 Evaluation result

In this section, the result of the evaluation process is presented. Firstly, an overview of the evaluated data set and key data is given. Secondly, the outcome of the evaluation process is analyzed and discussed. All the numbers, charts and tables mentioned in the following sections are based on the evaluation data extracted from the evaluation database on October 17, 2011.

Overview of the evaluated data set

The key data of the evaluation are provided in Table 5.6. The data set for the evaluation contains 7,813 different artworks with 89,514 corresponding tag recommendations. Nine users conducted the evaluation and rated 5,891 tag recommendations for 652 different artworks according to the earlier presented scheme. On average, each user evaluated 72.44 artworks and 654.56 tags respectively.

Artworks with tag recommendations	7,813
Tag recommendations	89,514
Evaluated artworks	652
Evaluated recommendations	5,891
Evaluating users	9
Average number of evaluated artworks per user	72.44
Average number of evaluated tags per user	654.56

Table 5.6: Statistical overview of the evaluation (October 17, 2011).

At this point of the evaluation it has to be noted that the explorARTorium does not only contain images of paintings, but also images of ceramics, furniture, sculptures, etc. and also images showing a detailed view (i.e. only a small clipping) of some original artwork. Although these images (from now on called "detail-artworks") were shown in the explorARTorium for some time and therefore received tags, which are subsequently mined for frequent patterns in the Data Mining phase of the Tag Recommendation Framework, it can be expected that these "detail-artworks" will have a negative influence on the outcome of the evaluation. This problem can be illustrated by means of an example: Figure 5.18 shows *The Mystical Nativity* by *Sandro BOTTICELLI* on the left, depicting Bethlehem, the Infant Jesus, Mother Mary, Joseph, angels, the ox, the donkey, the creche, the stable, etc. On the right of Figure 5.18 only a particular detail of this artwork is shown, namely an angel and a character. Although the content of these two images is nearly totally different, both of them share the same attributes (*region*, *theme*, *title*, etc.). Therefore also the "detail-artwork" on the right receives the same tag recommendations as the complete artwork on the left (e.g. Mary, Jesus, ox, donkey, etc.), which will inevitably cause misleading evaluations of these tags. To overcome this bias of the evaluation, in the following tables and charts presenting the outcome of the evaluation, a distinction between a set of artworks containing *all* artworks (paintings, ceramics, furniture, sculptures, etc) and a set of artworks containing *only* paintings is made.

In Table 5.7 the numbers of evaluated artworks and tags (both with and without the "detail-artworks") are presented. The second column shows the number of evaluated artworks per

Figure 5.18: *The Mystical Nativity* by *Sandro BOTTICELLI*: original artwork on the left, detailed view on the right.

archetype, revealing that the users liked to evaluate artworks with *Archetypes 1-9* (with every archetype being evaluated with at least 50 different artworks), whereas artworks with *Archetypes 10-12* were not evaluated so often (with every archetype being evaluated with less than 50 different artworks). The third column excludes the "detail-artworks", but shows a similar distribution. In the fourth column in Table 5.7 the evaluated tag recommendations per archetype are provided, showing a similar pattern in general but with some minor exceptions (e.g. few tag evaluations for the *Archetypes 4* and *7*), possibly due to the fact that the Tag Recommendation Framework generated a smaller number of tag recommendations for artworks with these archetypes. The fifth column again excludes tags for "detail-artworks".

Result

Figure 5.19 shows the actual result of the evaluation by providing the distribution of the four possible ratings over the set of evaluated tag recommendations with the help of two pie charts. In the pie chart on the left, the evaluations of all artworks are taken into consideration, whereas the pie chart on the right only shows the ratings of non-"detail-artworks". More than 55% of the tag recommendations fit the artworks either exactly or approximately, whereas less than 45% of the tag recommendations fit barely or do not fit at all. The set of artworks excluding "detail-artworks" performs even better with more than 58% acceptance rate and only 42% rejection.

Figure 5.20 provides a more detailed result of the evaluation, showing the distribution of

Archetype	Artworks	Artworks excl. detail	Tags	Tags excl. details
1	62	54	595	516
2	57	36	626	405
3	67	37	722	424
4	52	27	277	160
5	59	52	792	705
6	62	43	559	388
7	59	42	367	252
8	54	43	600	482
9	55	43	429	340
10	45	40	343	306
11	34	32	333	319
12	46	26	248	144
Total	**652**	**475**	**5891**	**4441**

Table 5.7: Evaluated artworks and tag recommendations per archetype.

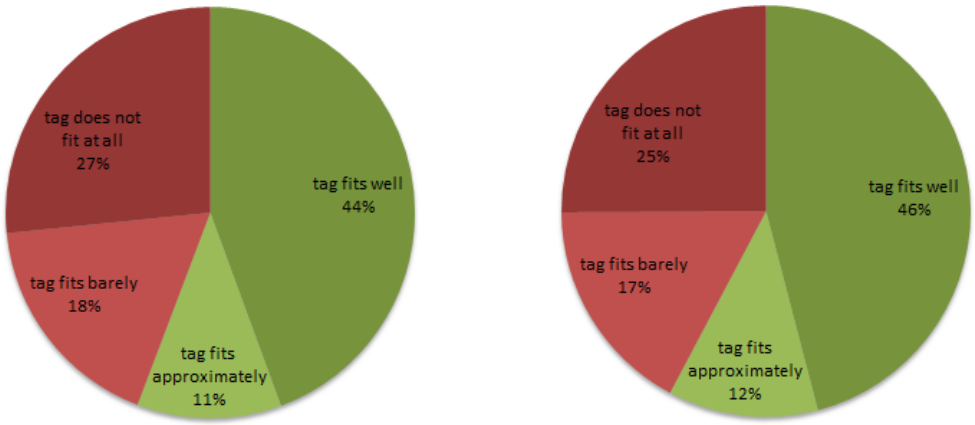

Figure 5.19: Distribution of ratings for the evaluated tag recommendations. On the left: for all artworks, on the right: only for non-"detail-artworks".

the four possible ratings (two thumbs up = 2; one thumb up = 1; one thumb down = -1; two thumbs down = -2) over the set of evaluated tag recommendations *for each* identified archetype. The bar chart reveals that tag recommendations for artworks corresponding to *Archetypes 1, 5, 6, 7, 11* and *12* perform extremely well. This allows the conclusion that the Tag Recommendation Framework provides adequate and suitable tag recommendations for artworks with the themes *still-life* and *landscape*, for *Dutch* and *Flemish genre*, *Dutch, French* and *Spanish portraits* and *Italian mythological* artworks. Artworks corresponding to *Archetypes 3* and *9* show average results (i.e. *Italian religious* artworks and *Flemish* and *German portraits*), whereas the evaluation reveals the *Archetypes 2, 4, 8* and *10* (artworks with *religious* theme except *Italian*

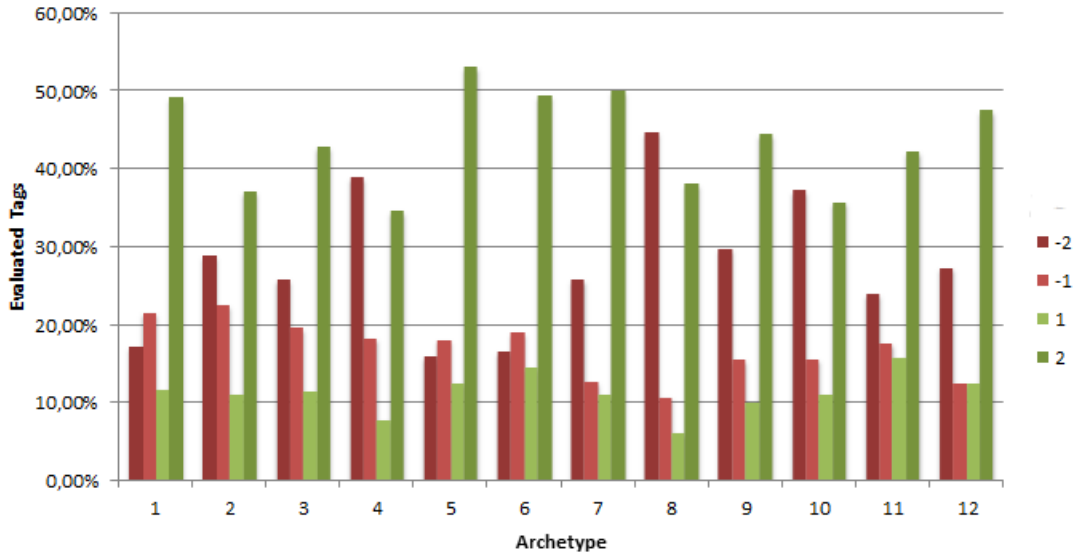

Figure 5.20: Distribution of ratings for the evaluated tag recommendations for each archetype.

artworks as well as *Italian portraits* and *French genre*) as soft spots of the Tag Recommendation Framework. Figure 5.21 also provides the distribution of the four possible ratings over the set of evaluated tag recommendations *for each* identified archetype, but without the "detail-artworks". As expected earlier, the Tag Recommendation Framework performed even better on this set of artworks concerning the evaluation, especially for the *Archetypes 4* and *12*.

The success of the Tag Recommendation Framework regarding the themes *still-life* and *landscape* might be attributed to the fact that artworks of these themes typically show stronger similarities than artworks with *religious* theme. A possible explanation for the performance of tag recommendations for *portraits* can be obtained by taking a look at the most frequently recommended tags for this theme: both *mann* (man) as well as *frau* (woman) are recommended for nearly every portrait. The fact that the vast majority of portraits in the data set of the explorARTorium shows either a man or a woman inevitably results in the "2 thumbs down" evaluation of either of them.

To finally conclude the evaluation of the Tag Recommendation Framework, an overview of the ten "best" (i.e. most often suitable) and ten "worst" (least often suitable) tag recommendations according to the evaluation is given in Table 5.8. It is quite surprising to see that both the ten "best" as well as ten "worst" tag recommendations mostly occur in the themes *religious, landscape* and *portrait*.

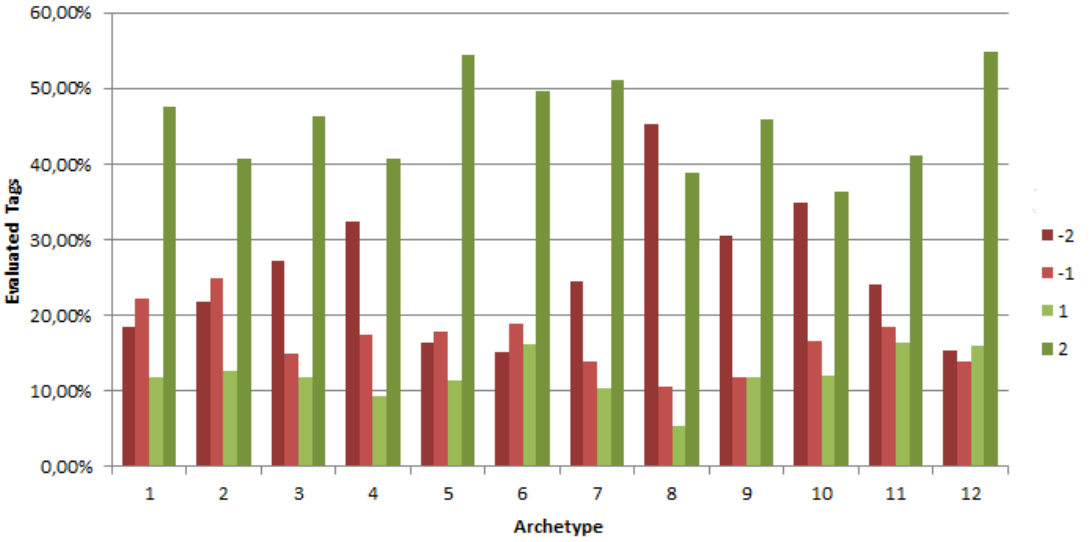

Figure 5.21: Distribution of ratings for the evaluated tag recommendations for each archetype without “detail-artworks”.

Best tags	Fitting	Worst tags	Fitting
esel	88.89%	christus	48.48%
josef	88.89%	man	48.12%
frauen	75.00%	laute	45.56%
verkuendigung	68.97%	muetze	44.33%
woman	68.18%	ruestung	44.33%
fenster	66.82%	schwert	44.33%
huegel	66.31%	vollbart	44.33%
baum	66.08%	genre	37.50%
schnurrbart	65.66%	baby	11.11%
pferd	65.66%	madonna	11.11%

Table 5.8: Top 10 best and worst tag recommendations.

CHAPTER 6

Conclusion and Future Work

In this final chapter, the summary and the conclusions of this master's thesis are presented in Section 6.1. Afterwards, an outlook on possible future work in context to this master's thesis is given in Section 6.2.

6.1 Conclusion

In this master's thesis, user-generated content in the context of a database of artworks (the explorARTorium) is analyzed. The explorARTorium hosts a large collection of ∼20,000 digitized images of artworks, which can be explored along various dimensions such as time, region or theme. Through the practice of annotating content, a folksonomy (a system of classification based on user collaboration) is created. It is in the operator's interest to keep the users intrigued using the multimedia platform and tagging artworks, which is not easy to achieve, because tagging is a time-consuming task and without any incentive or help, the users' motivation to tag will decrease over time.

In the first part of the thesis, an extensive analysis of the explorARTorium's folksonomy related to art history is conducted, exploring the relationship between users and their tags. It is confirmed that the users' tagging behavior can be set into relation to their liking of artworks. The users' vocabulary is qualitatively and lexically analyzed discovering an extremely unequal distribution of parts of speech within the users' vocabulary. Accordingly, great differences in the description of themes are revealed (i.e. a part of speech bias is identified, e.g. portraits are annotated with different parts of speech than religious artworks). Furthermore, it is confirmed that users are more likely to identify historical persons and places in an artwork than the creator of the artwork. The analysis also shows that there are striking differences between the "emotional quality" of adjectives used for different themes. The role of the user regarding activity and learning effects is examined revealing that there are different types of users regarding their activity over time and that the users' vocabulary gets more specific in the course of time. Finally, the gradual decrease of the users' motivation to tag is confirmed.

In order to give the users of the artwork collection an incentive to tag pictures and thus to prevent the users' tagging motivation from declining, a framework for system-generated suggestions for appropriate tags (based on tags extracted from the folksonomy) is developed and presented in the second part of this master's thesis, which offers the user an easier way to describe the artwork. Furthermore, this approach invites the user to start looking for the suggested tags in the artwork in order to verify them, which might subsequently lead to new tags. The tag recommendations are also used to present the user additional artworks in the same context of tags. The implementation of the framework makes heavy use of business intelligence techniques like Frequent Itemset Mining and Association Rule Mining for discovering interesting relations between variables in the database to provide reliable decision criteria for the recommender system.

The quality and precision of the implemented Tag Recommendation Framework are evaluated by users reviewing the accuracy of the generated tag recommendations. To counter the bias in the data set of the collection of artworks regarding the different attributes, distinct archetypes of artworks (i.e. artworks which have certain attributes in common) in the data set are identified by using cluster analysis techniques. The analysis of the evaluation with regard to these archetypes concludes that the Tag Recommendation Framework provides adequate and suitable tag recommendations for artworks with the themes *still-life*, *landscape*, *mythological* and also performs well for certain *genres* and *portraits*.

6.2 Future Work

Although many important questions regarding the folksonomy of the explorARTorium have been answered in this master's thesis, new ones have appeared and should be investigated in order to improve the understanding of the data set.

- **Specificity of tags**: An analysis regarding the question how the specificity of tags develops if new tags are assigned to an artwork would bring more insight into the users' tagging behavior, i.e. do users start to tag the artwork with less specific tags and tend to assign more specific tags only if less specific tags have already been assigned?
- **Change due to recommendation**: An exploration of the possible change in the users' tagging behavior due to the tag recommendations could be conducted, analyzing if the users favor the suggested tags or if they still like to assign their self-chosen tags to the artworks. It would be interesting to see if different types of users regarding the acceptance of the tag recommendations can be identified.
- **Influence by implicit recommendations**: As noted earlier, the tags suggested by the Tag Recommendation Framework can also be used to present the user untagged artworks in the context of the artwork which is currently shown to the user. Analyzing how users adopt this "hidden" influence by the operator might be an interesting task.

The Tag Recommendation Framework (TRF) proposed in this master's thesis could be enhanced to provide even more functionality, with the following extensions being the most interesting ones:

- **Parts of speech**: The analysis of the explorARTorium's folksonomy has shown the existence of a part of speech bias. Currently, the TRF pays no attention to the parts of speech of the tags it is recommending. In order to either represent the part of speech bias in the distribution of the tags suggested by the TRF or, on the contrary, to counter this bias, the recommender system can be extended to give the operator the possibility to influence the suggestion of tags by means of parts of speech.

- **Personalizing tags**: On the one hand, the personalization of tags, i.e. the suggestion of tags corresponding with the user's own vocabulary, yields an interesting possibility to further improve the user experience and satisfaction. On the other hand, this approach should not be relied on exclusively, due to the imminent risk to restrict the lexical spectrum of the user by suggesting only keywords which reflect the system's model of the user's vocabulary.

- **Incorporate results of the evaluation**: The TRF can be improved regarding the particular archetypes of artworks for which the evaluation of the tag recommendations yields results below average as shown in Section 5.2. A potential approach to refine the data basis on which the TRF relies on is to present the users artworks of these archetypes more often in the context of other artworks with the expectation to collect more tags describing these archetypes. Thereby the TRF can profit from an improved data basis and generate more accurate results.

APPENDIX A

Images

As noted earlier in Section 3.2.6, the most frequently tagged artwork for each theme of the explorARTorium's classification (*genre*, *landscape*, *portrait*, *religious*, etc.) according to Table 3.11 is presented in this chapter along with its title, artist and the number of tags the artwork received.

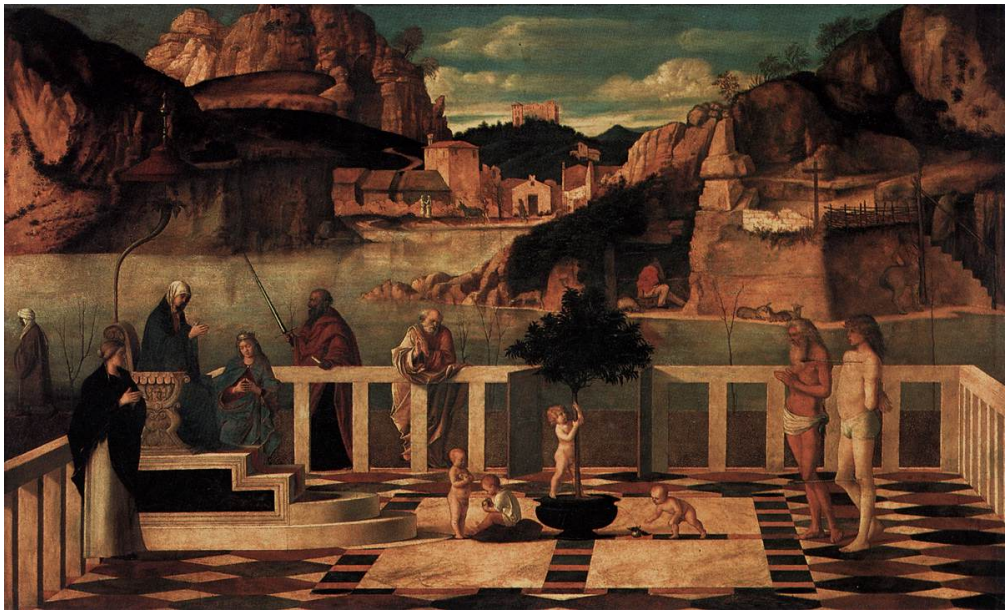

Figure A.1: Artwork with the most assigned tags for the theme *religious*: *Sacred Allegory* by *Giovanni BELLINI* with 66 assigned tags.

Figure A.2: Artwork with the most assigned tags for the theme *landscape*: *Gloomy Day (detail)* by *Pieter the Elder BRUEGEL* with 46 assigned tags.

Figure A.3: Artwork with the most assigned tags for the theme *genre*: *In Luxury, Look Out* by *Jan STEEN* with 45 assigned tags.

Figure A.4: Artwork with the most assigned tags for the theme *historical*: *Napoleon Bonaparte on the Battlefield of Eylau, 1807* by *Antoine-Jean GROS* with 45 assigned tags.

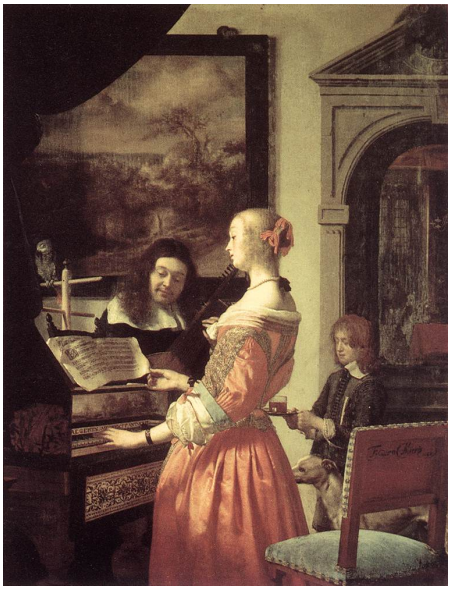

Figure A.5: Artwork with the most assigned tags for the theme *interior*: *Duet* by *Frans van MIERIS, the Elder* with 45 assigned tags.

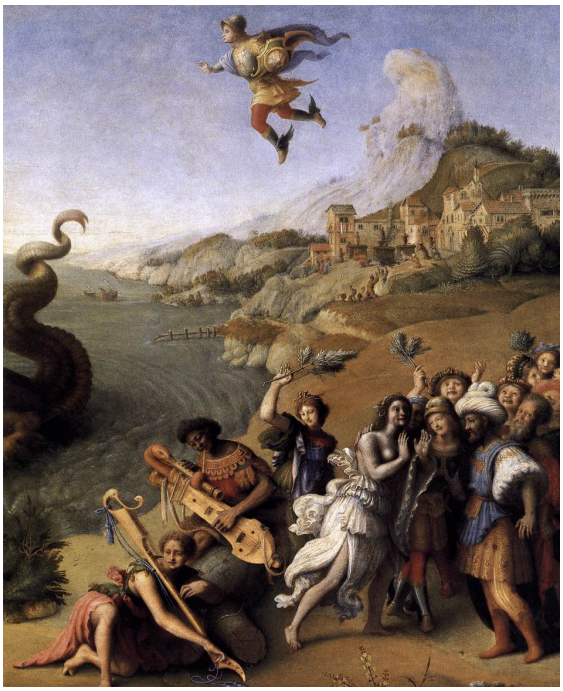

Figure A.6: Artwork with the most assigned tags for the theme *mythological*: *Perseus Frees Andromeda (detail)* by *PIERO DI COSIMO* with 42 assigned tags.

Figure A.7: Artwork with the most assigned tags for the theme *other*: *Dinner* by *Thomas ROWLANDSON* with 40 assigned tags.

Figure A.8: Artwork with the most assigned tags for the theme *still-life*: *Still-Life of Flowers and Fruits* by *Jean-Baptiste MONNOYER* with 39 assigned tags.

Figure A.9: Artwork with the most assigned tags for the theme *study*: *The Adoration of the Wise Man* by *Albrecht DÜRER* with 37 assigned tags.

Figure A.10: Artwork with the most assigned tags for the theme *portrait*: *Portrait of the Saltykov Family* by *Johann Friedrich August TISCHBEIN* with 36 assigned tags.

APPENDIX B

Tables

In this chapter the most frequent tag recommendations for each archetype are listed in the following tables. For each tag in the left column of the tables, the number of artworks which received this particular tag through the Tag Recommendations Framework is given in the right column of the tables.

Tag	Count
stillleben	168
blumen	164
stilleben	146
blaetter	139
tisch	130
glas	95
obst	89
messer	86
tulpen	79
vase	75
weintrauben	74
rosen	73
tischtuch	59
fruechte	54
schmetterling	42
teller	29
zitrone	28
krug	27
wolken	11
himmel	7
engel	5

Table B.1: The most frequent tag recommendations for Archetype 1 (*still-life*).

Tag	Count
himmel	706
fluegel	706
engel	706
christus	706
kreuz	706
buch	706
maria	706
heiligenschein	705
jesus	702
wolken	701
frau	691
jesukind	506
baeume	500
berge	495
felsen	494
moench	206
bart	206
elgreco	206
mann	206
verkuendigung	5

Table B.2: The most frequent tag recommendations for Archetype 2 (*Flemish* and *Spanish religious* artworks).

Tag	Count
kreuz	2915
jesus	2915
wolken	2915
himmel	2913
jesukind	2908
buch	2907
baeume	2906
fluegel	2904
maria	2902
frau	2900
engel	2898
heiligenschein	2898
kind	2895
mann	2859
maenner	2618
christus	266
kreuzigung	28
verkuendigung	25
elgreco	14
bart	14
portrait	7
esel	7
josef	7
heilige	2

Table B.3: The most frequent tag recommendations for Archetype 3 (*Italian religious* artworks).

Tag	Count
heiligenschein	563
engel	562
maria	513
wolken	495
jesus	472
himmel	265
frau	246
jesukind	183
kreuz	182
fluegel	180
buch	170
kind	161
mann	155
baeume	22
madonna	18
baby	17
maenner	11
christus	8
verkuendigung	7
kreuzigung	6
portrait	5
esel	4
josef	4
kette	4
heilige	3
elgreco	3
tisch	2
berge	2

Table B.4: The most frequent tag recommendations for Archetype 4 (*French*, *Dutch*, *German* and *Netherlandish religious* artworks).

Tag	Count
himmel	433
wolken	433
fluss	433
baeume	433
reiter	433
hund	433
meer	432
landschaft	425
kirche	421
pferd	421
baum	421
boot	317
wasser	306
haus	305
spiegelung	305
huegel	116

Table B.5: The most frequent tag recommendations for Archetype 5 (*Dutch* and *Flemish landscapes*).

Tag	Count
wolken	444
himmel	444
baeume	353
venedig	348
meer	281
felsen	273
huegel	273
bruecke	269
fluss	261
landschaft	192
wasser	189
haeuser	171
saeulen	171
kanal	171
reiter	167
menschen	98
berge	98
kirche	86

Table B.6: The most frequent tag recommendations for Archetype 6 (*Italian* and *French landscapes*).

Tag	Count
frau	827
portrait	771
mann	712
hut	668
buch	420
man	411
kragen	347
bart	301
tisch	298
krug	297
halskrause	290
locken	277
peruecke	226
wolken	208
schnurrbart	203
vorhang	201
spitzbart	198
himmel	128
orden	65
hund	62
kette	40
baeume	30
woman	21
inschrift	16
fluss	15
tischtuch	15
berge	12
landschaft	12
heiligenschein	12
ruestung	11
venedig	9
boot	8
maria	7
elgreco	6
christus	6
maenner	6
fenster	5
genre	5
mittelscheitel	5
kopfbedeckung	5
frauen	5
baum	4
kirche	4
jesus	4

Table B.7: The most frequent tag recommendations for Archetype 7 (*Dutch*, *French* and *Spanish portraits* and *genre* artworks).

Tag	Count
mann	313
vollbart	313
kette	313
ruestung	313
buch	313
portrait	313
bart	313
frau	313
hut	313
schwert	313
man	313
muetze	313

Table B.8: The most frequent tag recommendations for Archetype 8 (*Italian portraits*).

Tag	Count
portrait	259
mann	259
kette	259
hut	258
ring	252
halskrause	161
frau	161
wolken	155
ringe	104
buch	104
kragen	56
locken	50
spitzbart	50
woman	7
orden	6

Table B.9: The most frequent tag recommendations for Archetype 9 (*Flemish* and *German portraits*).

Tag	Count
himmel	215
sessel	214
baeume	214
tisch	204
mann	173
wolken	171
laute	161
krug	137
hund	69
hut	66
kirche	44
landschaft	43
frau	24
fluss	12
wald	11
berge	9

Table B.10: The most frequent tag recommendations for Archetype 10 (*French* and *Dutch genre* and *interior* artworks).

Tag	Count
frau	56
himmel	56
hund	56
tisch	55
tischtuch	54
kind	38
sessel	38
fenster	37
kirche	18
maenner	18
landschaft	18

Table B.11: The most frequent tag recommendations for Archetype 11 (*Flemish* and *Dutch genre* and *interior* artworks).

Tag	Count
himmel	640
wolken	638
frau	629
nackt	575
engel	510
hund	124
mann	103
portrait	84
baeume	83
peruecke	80
berge	79
landschaft	65
maenner	29
baum	29
pferd	18
meer	18
orden	18
buch	12
fluss	11
tisch	9
maria	9
fenster	9
hut	9
huegel	9
fluegel	8
kreuz	7
jesukind	5
kirche	4
bruecke	2

Table B.12: The most frequent tag recommendations for Archetype 12 (*Italian mythological* artworks).

Bibliography

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734– 749. (Cited on page 19.)

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22:207–216. ACM ID: 170072. (Cited on pages 3, 9, 10, and 25.)

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499. (Cited on pages 12 and 25.)

Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. *Proceedings of the SIGCHI conference on Human factors in computing systems*, page 971–980. ACM ID: 1240772. (Cited on pages 2 and 24.)

Arends, M., Froschauer, J., Goldfarb, D., and Merkl, D. (2011). Analysing user generated content related to art history. In *Proceedings of the Int'l Conference on Knowledge-Management and Knowledge Technologies*, Graz, Austria. (Cited on pages 28 and 30.)

Brants, T. (2000). TnT: a statistical part-of-speech tagger. In *ANLC '00 Proceedings of the sixth conference on Applied natural language processing*, pages 224–231. Association for Computational Linguistics. (Cited on page 48.)

Brill, E. (1992). A simple rule-based part of speech tagger. In *ANLC '92 Proceedings of the third conference on Applied natural language processing*, page 152. Association for Computational Linguistics. (Cited on page 48.)

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 255–264. ACM Press. (Cited on page 11.)

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117. (Cited on page 23.)

Burke, R. (2000). Knowledge-based recommender systems. *Encyclopedia of Library and Information Systems*, 69(Supplement 32):175–186. (Cited on page 20.)

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370. (Cited on page 21.)

Burke, R. (2007). Hybrid web recommender systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web*, volume 4321, pages 377–408. Springer Berlin Heidelberg, Berlin, Heidelberg. (Cited on page 21.)

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, 41(3):1–58. (Cited on page 10.)

Chen, Y., Hu, J., Zhang, Y., and Li, X. (2006). Traffic data analysis using kernel PCA and Self-Organizing map. In *2006 IEEE Intelligent Vehicles Symposium*, pages 472–477. IEEE. (Cited on page 9.)

Chirita, P. A., Costache, S., Nejdl, W., and Handschuh, S. (2007). P-TAG: large scale automatic generation of personalized annotation tags for the web. *Proceedings of the 16th international conference on World Wide Web*, page 845–854. ACM ID: 1242686. (Cited on page 24.)

CIDOC (2007). The CIDOC CRM. http://www.cidoc-crm.org/. [Online; accessed 04-October-2011]. (Cited on page 30.)

Community documentation for Weka (2011). Apriori - pentaho data mining - pentaho wiki. http://wiki.pentaho.com/display/DATAMINING/Apriori. [Online; accessed 04-October-2011]. (Cited on page 69.)

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. In *ANLC '92 Proceedings of the third conference on Applied natural language processing*, page 133. Association for Computational Linguistics. (Cited on page 48.)

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40:5:1–5:60. ACM ID: 1348248. (Cited on page 25.)

Fayyad, U. M., Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. In *Advances in knowledge discovery and data mining*. AAAI Press/The MIT Press. (Cited on pages 6, 7, 9, and 10.)

Golbeck, J. (2006). Generating predictive movie recommendations from trust in social networks. In Stølen, K., Winsborough, W. H., Martinelli, F., and Massacci, F., editors, *Trust Management*, volume 3986 of *Lecture Notes in Computer Science*, pages 93–104. Springer Berlin Heidelberg, Berlin, Heidelberg. (Cited on page 21.)

Gupta, G. K. (2006). *Introduction to Data Mining with Case Studies*. Prentice-Hall Of India Pvt. Ltd. (Cited on pages 5, 6, 7, 8, 10, 11, 12, 13, 14, and 15.)

Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., and Ofek-Koifman, S. (2009). Personalized recommendation of social software items based on social relations. In *Proceedings of the third ACM conference on Recommender systems*, page 53. ACM Press. (Cited on page 21.)

Halvey, M. J. and Keane, M. T. (2007). An assessment of tag presentation techniques. In *WWW '07 Proceedings of the 16th international conference on World Wide Web*, page 1313. ACM Press. (Cited on page 57.)

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier. (Cited on page 5.)

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 1–12. ACM Press. (Cited on page 15.)

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A Frequent-Pattern tree approach. *Data Mining and Knowledge Discovery*, 8:53–87. (Cited on page 15.)

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22:5–53. (Cited on pages 17, 18, and 19.)

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining - a general survey and comparison. *ACM SIGKDD Explorations Newsletter*, 2:58–64. ACM ID: 360421. (Cited on page 25.)

Hotho, A. (2010). Data mining on folksonomies. In Armano, G., Gemmis, M., Semeraro, G., and Vargiu, E., editors, *Intelligent Information Access*, volume 301 of *Studies in Computational Intelligence*, pages 57–82. Springer Berlin Heidelberg, Berlin, Heidelberg. (Cited on pages 21, 22, and 23.)

Hsu, M. and Chen, H. (2008). Tag normalization and prediction for effective social media retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08*, volume 1, pages 770–774. IEEE. (Cited on page 30.)

Inmon, W. H. (1992). *Building the data warehouse*. John Wiley. (Cited on page 6.)

Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender Systems: An Introduction*. Cambridge University Press. (Cited on pages 16 and 19.)

Li, J. and Wang, J. Z. (2006). Real-time computerized annotation of pictures. *Proceedings of the 14th annual ACM international conference on Multimedia*, page 911–920. ACM ID: 1180841. (Cited on page 24.)

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, page 281–297. University of California Press. Published: Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967). (Cited on page 9.)

National Research Council (2008). *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment*. National Academies Press, Washington, DC. (Cited on page 9.)

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, pages 229–238. (Cited on page 11.)

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106. (Cited on page 9.)

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23:2000. (Cited on page 6.)

Ricci, F., Rokach, L., and Kantor, P. B. (2010). *Recommender Systems Handbook*. Springer. (Cited on pages 16, 17, 19, 21, and 24.)

Salton, G. and McGill, M., editors (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill. (Cited on page 62.)

Sarwar, B., Karypis, G., Konstan, J., and Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM Press. (Cited on page 20.)

Schmid, H. (1994). Probabilistic Part-of-Speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12, pages 44–49. (Cited on page 48.)

Schmitz, C., Hotho, A., Jäschke, R., and Stumme, G. (2006). Mining association rules in folksonomies. In Batagelj, V., Bock, H., Ferligoj, A., and Žiberna, A., editors, *Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270. Springer Berlin Heidelberg. (Cited on pages 21 and 22.)

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605 –610. (Cited on page 40.)

Shepitsen, A., Gemmell, J., Mobasher, B., and Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, page 259. ACM Press. (Cited on page 21.)

Siersdorfer, S. and Sizov, S. (2009). Social recommender systems for web 2.0 folksonomies. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, page 261. ACM Press. (Cited on page 21.)

Sigurbjörnsson, B. and van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, page 327. ACM Press. (Cited on pages 21, 23, 24, and 63.)

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380. (Cited on page 25.)

Song, Y., Zhang, L., and Giles, C. L. (2008a). A sparse gaussian processes classification framework for fast tag suggestions. *Proceeding of the 17th ACM conference on Information and knowledge management*, page 93–102. ACM ID: 1458098. (Cited on page 24.)

Song, Y., Zhuang, Z., Li, H., Zhao, Q., Li, J., Lee, W., and Giles, C. L. (2008b). Real-time automatic tag recommendation. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, page 515–522. ACM ID: 1390423. (Cited on page 24.)

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21:65–66. (Cited on page 40.)

Tan, P., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313. (Cited on page 25.)

VSEM (2011). About the explorARTorium. http://explorartorium.info/about.html. [Online; accessed 04-October-2011]. (Cited on pages 27 and 29.)

Wang, C., Blei, D., and Li, F. (2009). Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. (Cited on page 25.)

Wang, M. and Hua, X. (2011). Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2:10:1–10:21. ACM ID: 1899414. (Cited on page 25.)

Web Gallery of Art (2011a). Information about the WGA. http://www.wga.hu/index1.html. [Online; accessed 04-October-2011]. (Cited on page 31.)

Web Gallery of Art (2011b). Legal information. http://www.wga.hu/legal.html. [Online; accessed 04-October-2011]. (Cited on page 31.)

Wikipedia (2001). Part-of-speech tagging. http://en.wikipedia.org/wiki/Part-of-speech_tagging. [Online; accessed 04-October-2011]. (Cited on page 48.)

Wikipedia (2011a). Archetype. http://en.wikipedia.org/wiki/Archetype. [Online; accessed 04-October-2011]. (Cited on page 77.)

Wikipedia (2011b). Cross industry standard process for data mining. http://en.wikipedia.org/wiki/CRISP-DM. [Online; accessed 04-October-2011]. (Cited on page 7.)

Wikipedia (2011c). Data mining. http://en.wikipedia.org/wiki/Data_mining. [Online; accessed 04-October-2011]. (Cited on pages 7 and 9.)

Wikipedia (2011d). Naive bayes classifier. http://en.wikipedia.org/wiki/Naive_Bayes_classifier. [Online; accessed 04-October-2011]. (Cited on page 9.)

Wikipedia (2011e). Part of speech. http://en.wikipedia.org/wiki/Part_of_speech. [Online; accessed 04-October-2011]. (Cited on page 48.)

Wikipedia (2011f). Recommender system. http://en.wikipedia.org/wiki/Recommender_system. [Online; accessed 04-October-2011]. (Cited on page 16.)

Wikipedia (2011g). Tag (metadata). http://en.wikipedia.org/wiki/Tag_(metadata). [Online; accessed 04-October-2011]. (Cited on page 21.)

Witten, I. H. and Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann. (Cited on page 6.)

Yang, Y., Huang, Z., Shen, H. T., and Zhou, X. (2010). Mining multi-tag association for image tagging. *World Wide Web*, 14(2):133–156. (Cited on page 25.)

Zhu, X. and Davidson, I. (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. Hershey, New York. (Cited on page 9.)