Pattern Research in the Digital Humanities: How Data Mining Techniques Support the Identification of Costume Patterns

Michael Falkenthal[1], Johanna Barzen[1], Uwe Breitenbücher[1],
Sascha Brügmann[2], Daniel Joos[2], Frank Leymann[1], Michael Wurster[2]

[1]Institute of Architecture of Application Systems,
University of Stuttgart, Germany
{falkenthal, barzen, breitenbuecher, leymann}@iaas.uni-stuttgart.de

[2]Herman-Hollerith Zentrum,
University of Applied Sciences Reutlingen, Germany
{michael1.wurster, daniel.joos, sascha.bruegmann}@student.reutlingen-university.de

BibTeX:

```
@article{Falkenthal2016,
   author = {Falkenthal, Michael and Barzen, Johanna and Breitenb{\"{u}}cher,
   Uwe and Br{\"{u}}gmann, Sascha and Joos, Daniel and Leymann, Frank and
   Wurster, Michael},
   doi = {10.1007/s00450-016-0331-6},
   journal = {Computer Science - Research and Development},
   number = {74},
   title = {{Pattern research in the digital humanities: how data mining
   techniques support the identification of costume patterns}},
   volume = {22},
   year = {2016}
}
```

Noname manuscript No.
(will be inserted by the editor)

Pattern Research in the Digital Humanities

How Data Mining Techniques Support the Identification of Costume Patterns

Michael Falkenthal · Johanna Barzen · Uwe Breitenbücher · Sascha Brügmann · Daniel Joos · Frank Leymann · Michael Wurster

Received: date / Accepted: date

Abstract Costumes are prominent in transporting a character's mood, a certain stereotype, or character trait in a film. The concept of patterns, applied to the domain of costumes in films, can help costume designers to improve their work by capturing knowledge and experience about proven solutions for recurring design problems. However, finding such Costume Patterns is a difficult and time-consuming task, because possibly hundreds of different costumes of a huge number of films have to be analyzed to find commonalities. In this paper, we present a Semi-Automated Costume Pattern Mining Method to discover indicators for Costume Patterns from a large data set of documented costumes using data mining and data warehouse techniques. We validate the presented approach by a prototypical implementation that builds upon the Apriori algorithm for mining association rules and standard data warehouse technologies.

Keywords Pattern Languages · Pattern Mining · Pattern Identification · Data Mining · Costume Languages · Costume Patterns · Digital Hummanities

1 Introduction

When watching a movie, the first impression of the characters is often not caused by how they move or what they say but by what they wear. Costumes are special types of clothes used by costume designers to support a certain character, his moods and transformations or to give the recipient hints on where the movie is set in geographical or historical terms. This communication, based on clothes, is called *vestimentary communication* (from the Latin term "vestimentum" meaning clothes) and is used by the costume designers to *communicate* certain stereotypes, character traits, professions, or a specific age of a certain character. Because the vestimentary communication is a nonverbal communication, mainly experienced unconsciously by the recipients, and interpreted based on their social and socioeconomic background, it is rather complex to gain insight in how this communication works. However, there are some

Michael Falkenthal
Institute of Architecture of Application Systems
University of Stuttgart
Germany Tel.: +49-711-68588482
E-mail: falkenthal@iaas.uni-stuttgart.de

Johanna Barzen
Institute of Architecture of Application Systems
University of Stuttgart
Germany Tel.: +49-711-68588487
E-mail: barzen@iaas.uni-stuttgart.de

Uwe Breitenbücher
Institute of Architecture of Application Systems
University of Stuttgart
Germany Tel.: +49-711-68588261
E-mail: breitenbuecher@iaas.uni-stuttgart.de

Sascha Brügmann
Herrmann-Hollerith Zentrum
University of Applied Sciences Reutlingen
Germany E-mail: sascha.bruegmann@student.reutlingen-university.de

Daniel Joos
Herrmann-Hollerith Zentrum
University of Applied Sciences Reutlingen
Germany E-mail: daniel.joos@student.reutlingen-university.de

Frank Leymann
Institute of Architecture of Application Systems
University of Stuttgart
Germany Tel.: +49-711-68588470
E-mail: leymann@iaas.uni-stuttgart.de

Michael Wurster
Herrmann-Hollerith Zentrum
University of Applied Sciences Reutlingen
Germany E-mail: michael1.wurster@student.reutlingen-university.de

rules that allow us to distinguish between a *villain* and a *hero* in a classic western movie. Whether we interpret characters as villains, because they wear black and dirty-looking costume elements, or as heroes (often represented as *sheriffs*) because they wear rather tidy costumes including a sheriff's star, is one of the questions we want to answer. A costume designer has the challenging job to create an appropriate costume for a specific character, which relies heavily on the designer's experience.

Systematically capturing of insights in which design conventions have established in films into patterns would strongly support the creative process of finding adequate textile expressions for specific design problems at hand (Schumm et al, 2012). This is because patterns and pattern languages, which originated by Alexander et al (1977) in the domain of architecture, aim to capture knowledge gathered from experience in order to provide proven solutions for frequently reoccurring problems.

According to Barzen and Leymann (2015) as well as Fehling et al (2015), patterns can be detected by analyzing existing, documented *concrete solutions* and abstracting the essence of detected commonalities into structured pattern documents. For investigating the vestimentary communication in films, concrete solutions correspond to concrete costumes worn in films. To provide machine-readable data about costumes, we built the MUSE-Repository[1] as a database that captures costumes and their relevant attributes (Barzen et al, 2015). This so-called *Costume Repository* contains (i) *general information* about the captured movies, e.g., title, year of publication, producer, and costume designer, as well as (ii) *specific information* in terms of the involved roles, such as gender, profession, age, main personality, and stereotype attributes. Further, each role is linked to a set of concrete costumes worn during the movie. Each costume consists of a set of *base elements*, e.g., trousers, shirts, and shoes, and *primitives*, e.g., sleeves, collar, and cuffs. Base elements and primitives are described by means of specific categorical properties, which are organized into taxonomies. At the time of writing this paper, the Costume Repository contained 25 movies, about 900 corresponding roles, 2,100 costumes, 10,360 base elements, and 20,660 primitives.

Although this Costume Repository is a first step to reduce efforts for investigating costumes in films in order to identify common design principles, the process of pattern identification still bases on manual and, thus, time-consuming work because of the lack of automation for discovering similarities in the documented costumes. Therefore, we introduced a first approach to support analyzing the captured costumes using *On-Line Analytical Processing (OLAP)* technologies (Falkenthal et al, 2015). Nevertheless, this approach is only capable of answering concrete questions and verifying assumptions by specifying them as multidimensional queries that are executed on the database, but the approach does not help to detect yet unknown coherences in the data set.

To overcome these issues, applying data mining techniques for identifying similarities, relations, and rules in the captured costume data is promising[2]. However, data mining, in fact, is hardly an easy exercise to accomplish and comprehensive conceptual as well as technical knowledge about database systems is inevitable. Detailed knowledge about data mining algorithms and how to apply them to the captured data set is required, while domain experts require knowledge on how the data is structured and how to interpret the mining results. This leads to a complex challenge in terms of how such data mining techniques can be applied to a concrete domain, in this paper, the domain of costumes in films. Furthermore, it is rarely the case that data mining is a straight process defining a concrete start and end. It is rather seen as an iteration-based process in which domain experts work on the results and the mining configuration incrementally based on gathered insights.

In this work, we contribute to the field of *Digital Humanities* by introducing a *Semi-Automated Costume Pattern Mining Method* that (i) leverages the capabilities of data mining to enable detecting indications regarding potential Costume Patterns from concrete costume documentations. The method (ii) can be partially automated to support analyzing the vast amount of captured costumes while (iii) it supports the iterative refinement of the data analysis. Thereby, we show how general data mining techniques for mining association rules can be applied to the domain of costumes in films. To prove the practical feasibility of the presented method, we conducted two case studies and present a prototypical implementation of a *Costume Pattern Mining Framework* that is based on standard IT-technologies. In summary, we show how IT can contribute to the domain of humanities and, thus, how it provides fundamentals for the new research discipline of the *Digital Humanities*. While the techniques and methods this work is predicated upon are well understood from the perspective of data analytics, their application in the domain of cos-

[1] As part of the MUSE project (last accessed on 25.02.2016): http://www.iaas.uni-stuttgart.de/forschung/projects/MUSE

[2] As the term *pattern* is ambigious and used besides the costume domain also in the domain of data mining (cf. (Bishop, 2006)) we clarify the different meanings at this point. While data mining is utilized to find patterns in large data sets in the form of similarities, relations, and rules, costume patterns follow the principles of the pattern approach by Alexander et al (1977).

tumes clearly fosters the endeavours to tackle research challenges in the humanities by means of IT.

The remainder of this paper is structured as follows: We discuss related work and approaches, which this work builds upon, in Section 2. Section 3 introduces the Semi-Automated Costume Pattern Mining Method that enables to find common design practices in a set of concrete, documented costumes. The challenges of applying data mining techniques to the domain of costumes are discussed in Section 4. Section 5 presents the case studies and prototypical implementation of the Costume Pattern Mining Framework. In Section 6, we conclude and give an outlook on future work.

2 Related Work

Patterns are commonly used to capture knowledge and experience about proven solutions for recurring problems (Reiners, 2013). In the past, patterns and pattern languages have been used in various different research domains (Alexander et al, 1977; Hohpe and Woolf, 2003). In literature, discovering patterns is described as a generative process and is referred as pattern mining (Dearden and Finlay, 2006; Appleton, 1997), which is a metaphor for discovering patterns from existing designs (Dearden and Finlay, 2006).

Reiners et al. propose different pattern mining methods (Reiners, 2013; Reiners et al, 2015). A pattern mining process in their perspective is a manual assessment of existing solutions with domain experts, e.g., in workshops, and relies heavily on the experts' experience. In addition to workshops, a community-based platform with online discussions, commenting, rating, and voting is used to share knowledge and to assess existing solutions.

Fehling et al (2015) propose a pattern research methodology where pattern candidates shall be identified from concrete solutions, which are then linked to the abstracted patterns (Falkenthal et al, 2014a,b). In another research, Fehling et al (2014) published a general pattern identification, authoring, and application process, which is applicable for several research domains. The iteration-based process consists of three phases: (i) pattern identification, (ii) pattern authoring, and (iii) pattern application. Each phase is broken down into a separate cycle that consists of multiple sub-activities. Our work applies to the phase pattern identification, which is the structuring, collection, and analysis of information in a domain in which patterns shall be identified. Following this method, we build upon a *Costume Repository* that contains a large number of documented concrete solutions. Moreover, this repository provides a machine-accessible interface that can be used for analyzing the contained data. We provide details about these approaches in Section 3.

Fayyad et al (1996) introduce the process of Knowledge Discovery in Databases (KDD). KDD refers to the overall process of discovering useful knowledge from data. This process incorporates the concepts of data mining and proposes a comprehensive approach to identify potential coherences in data. Our approach bases on the KDD process in order to analyze existing documented solutions for potential pattern indicators in the area of costumes in films. Data mining can be used to "discover hidden, previously unknown and usable information from a large amount of data" (ISO, 2006). Data mining techniques are used to gather knowledge from an underlying data set for a better understanding usually without any expectation on the outcome (ISO, 2006). The Apriori algorithm, as proposed by Agrawal and Srikant (1994), is one well-known algorithm in the area of data mining. It is used for discovering association rules between items in a database of sales transactions (Agrawal and Srikant, 1994). As a prominent example, consider the market basket analysis, helping retailers to find out, which of their offered products are typically sold in combination with other products. The resulting association rules can be used, for example, to optimize the store layout or to adapt the advertising strategy of the retailer.

3 Semi-Automated Costume Pattern Mining Method

In order to efficiently support the identification of common design principles hidden in the Costume Repository that captures concrete costume descriptions, we present a Costume Pattern Mining Method that extends KDD (cf. Section 2) and builds upon Barzen and Leymann (2016) to analyze existing costumes for indicators about patterns in the area of costumes in films. The process consists of three phases: (i) Data Preparation, (ii) Hypothesis Discovery, and (iii) Hypothesis Validation. An overview of the method is shown in Figure 1 with the respective phases depicted as a sequence of chevrons resulting in pattern indicators that represent frequent design principles contained in the analyzed data set. We also describe automation capabilities to handle the huge amount of data.

To provide an overview, we first summarize the method as follows and provide details about each phase in the following subsections. In the (i) Data Preparation phase, data is structured, prepared, and transformed to a domain-specific data set in order to be processed by data mining algorithms.

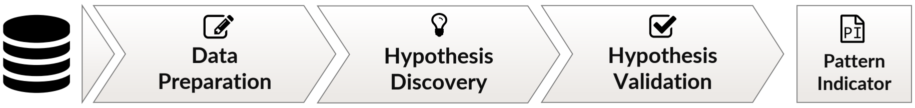


Fig. 1 Overview of the proposed Costume Pattern Mining Method

In the (ii) Hypothesis Discovery phase, analysis interests are first manually translated into specific configurations of a mining algorithm, to be precise, the costume data set is reduced to the attributes relevant for the analysis interests. Afterwards, the configured data mining algorithm gets automatically executed to find hypotheses about coherences in the investigated data set in the form of association rules. Finally, in the (iii) Hypothesis Validation phase, the discovered hypotheses are validated manually against the data set in the Costume Repository, i.e., domain experts interpret them and evaluate them based on OLAP techniques.

3.1 Data Preparation

In the first phase, the data of the Costume Repository needs to be prepared so that it can be used in the succeeding phases: On the one hand, this includes preparations required for the data mining algorithm to work. On the other hand, OLAP cube techniques used for validating hypotheses require the data of the Costume Repository, which are structured and stored in a relational schema, to be converted to a different data model. Moreover, additional data structures required for cube-based data analysis have to be created in the database. The data preparation phase follows the principles of an Extract-Transform-Load (ETL) process, typically used in the area of Data-Warehouse applications (Fayyad et al, 1996). For the analysis of the data set in the Costume Repository, we built upon an existing ETL process, described by Falkenthal et al (2015).

The data of the Costume Repository first gets extracted into a separate database. A reason for this is that the application of data mining algorithms can cause heavy load on the underlying database and might cause the Costume Repository to be slow or unavailable for parallel insertion of new costume data. As an additional advantage, working on a copy of the Costume Repository provides a certain level of data consistency and isolation as no new costumes are added and no data gets changed while applying data mining algorithms and executing the OLAP analysis.

After the data has been imported into a separate database (also called data staging area), automated steps for filtering, cleansing, and transformation of the data are executed (Fayyad et al, 1996). For instance, the Costume Repository contains a lot of screenshots, e.g., showing a costume in several scenes of a movie. This data is not used for analytical processing and is, therefore, filtered-out. Furthermore, the Costume Repository contains entries that are considered not valid, e.g., having NULL values or invalid strings. Those are filtered-out by the ETL process as well. Various minor transformation steps are applied to the data, for example, the applied implementation of the data mining algorithm could require non-composed primary keys. Therefore, the creation of surrogate keys would be required – a typical operation in ETL processes. The data are further denormalized into a star schema, while the hierarchical structure of the data describing the relevant parameters of a costume is preserved.

3.2 Hypothesis Discovery

Discovering common combinations in costume design that indicate a Costume Pattern is realized using data mining techniques as these enable to find similarities and associations in the data set of costumes. Data mining algorithms are applied to the data set of costumes resulting in hypotheses about similarities and associations. Thus, the second phase of the Costume Pattern Mining Method is called Hypothesis Discovery. It is refined in Figure 2, where it is broken down into four steps, described in the following.

The Hypothesis Discovery phase starts with defining the specific Interest we have about a certain area of the costume domain. As an example, it could be interesting to find out, if there is a relation between personality attributes of a character, wearing a certain costume, and the composition of the costume's base elements. Finding such relations in the costume data would support the identification of costume patterns for special character traits, meaning which base elements are mainly used to, e.g., communicate a certain character trait like "conceited" or "cool". Therefore, this step heavily relies on the input, given by an expert of the costume domain. As an output, this step clarifies the costume parameters that can be used to answer the question.

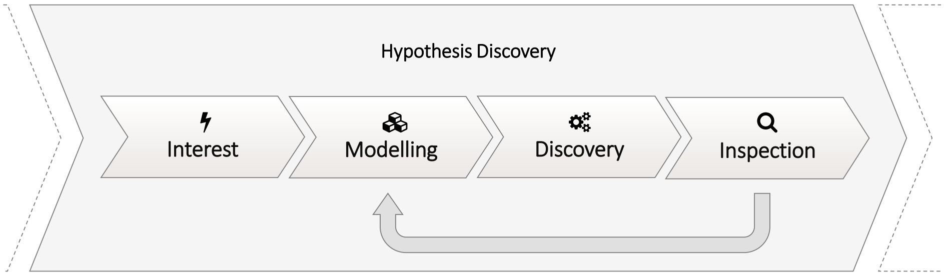


Fig. 2 Refinement of Hypothesis Discovery phase

So, after focusing on an Interest, the relevant costume parameters have to be modeled regarding their structure in order to get processed by the data mining algorithm. Using the above example, this could mean to use "character traits" as the input data set and the set of available "base elements" as the output data set. This step most likely requires support of a technical expert of the used data mining toolset.

The third step, the Discovery, involves the actual execution of the data mining algorithm, using the mining structures, created in the previous step. It also includes adaptations of filters and parameters specific to the algorithm. The form of the results of this step depends on the executed data mining algorithm, which could be, e.g., a set of association rules or clusters representing discovered similarities of the input data set. Independently from the actually run data mining algorithm the results can be stated as hypotheses about coherences in the input data set of costume descriptions.

The last step is called Inspection and focusses on presenting the results of the Discovery step to an expert. Additional filtering can be applied to the result to enable the expert to put focus on specific and interesting aspects. If no appropriate results were generated, the overall method can be continued back at the Modelling step. This allows to adapt the input data set or specific parameters of the data mining algorithm in order to refine the conducted analysis.

3.3 Hypothesis Validation

In this subsection we describe how identified hypotheses can be validated using the concept of OLAP cubes. A refinement of the Hypothesis Validation phase is shown in Figure 3 and the depicted steps are described in the following.

First, the data we are interested in have to be discovered in the OLAP cube. Analysis with the OLAP cube requires column- and row-filtering to be applied, focusing on a certain slice of the underlying data. Having found a hypothesis, the column- and row-filters of the OLAP cube need to be set to the dimensions (properties) that include the subset of costume parameters that are relevant for the hypothesis. This operation is often referred to as Slice and Dice as the view on the data focuses on a certain slice or sub-cube of the overall OLAP cube (Codd et al, 1993).

As second step, the data that lead to the inspected hypothesis needs to be identified in the cube. Therefore, search and filtering capabilities as well as drill-down and roll-up (cf. Codd et al (1993); Golfarelli et al (1998)) operations have to be applied on the dimensions of the OLAP cube.

In order to compare a hypothesis, the cube needs to count the appearance of costumes with given properties. So, the combination of relevant parameters of the hypothesis at hand can be compared with the number of appearances of other property combinations. If other property combinations are not considered significant in contrast to the examined combination, the discovered hypothesis can be considered validated. Then it can be grasped as an indicator for a pattern (cf. PI in Figure 3) because it represents a design principle, which contributes substantially to achieve the intended effects considered by the formerly defined Interests. Otherwise, the Hypothesis Discovery step starts all over again, either to focus on other Interests or to refine the input data set or parameters of the data mining algorithm.

4 Mining Association Rules from the Costume Repository

As explained in Section 3.2, data mining algorithms have to be applied to the data set of costumes. To understand the challenges of the Hypothesis Discovery phase and to grasp the required expertise to apply data mining techniques to the domain of costumes in films, we describe the application of the Apriori algorithm developed by Agrawal and Srikant (1994) to the data set of the Costume Repository.

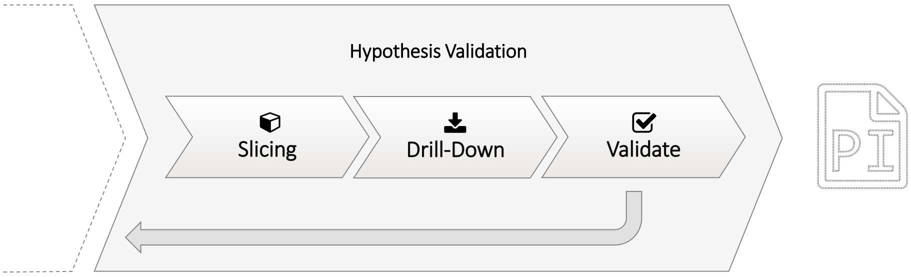


Fig. 3 Refinement of the Hypothesis Validation phase

The Apriori algorithm was originally designed to work with transactional sales data. As an example, consider a store selling multiple products. One of the store's customers bought three of those products together, so in this case, the transaction would include those three products. In order to apply this algorithm to find frequent associations between costume elements we have to transform the costume descriptions into corresponding concepts and data structures.

Looking at the domain of costumes in films, we consider a transaction to be a single costume that occurs in a specific movie. Barzen (2013) describes the relevant parameters of a costume, such as color, design, and material, as well as the base elements, which compose the costume. For each of those parameters, Barzen defines taxonomies, providing a well-defined and hierarchical set of parameter values. Fehling et al (2015) define a costume to consist of (i) "clothes" as a haptic basis, which itself is composed of base elements and (ii) an "intended effect". Through this intended effect, costumes communicate attributes about a character such as character traits, mood or social standing, as well as represented stereotypes to the recipient (Barzen and Leymann, 2015).

Applying the concept of transactions of the Apriori algorithm to the domain of costumes in films, the base elements of a costume correspond to the products bought by a customer. Costume parameters such as character traits, gender, stereotypical information of a costume as well as the information about the related movie correspond to attributes of a sales transaction, like time of day, store location, or product numbers. Let $P := P_1 \cup P_2 \cup \dots \cup P_n$ be the set of all available costume parameters, where $P_1 := \{b_1, b_2, \dots, b_m\}$ be the set of available base elements, $P_2 := \{s_1, s_2, \dots, s_k\}$ be the set of available stereotypical information, and so on. Then, the transaction representing a costume is defined as $T_{costume} := T_1 \cup T_2 \cup \dots \cup T_n$ such that $T_1 \subseteq P_1$, $T_2 \subseteq P_2$, and so on. Let $C := \{T_{costume_1}, T_{costume_2}, \dots, T_{costume_n}\}$ be the set of all available costumes in the database, which corresponds to the set of all transactions D of Apriori (cf. Agrawal and Srikant (1994)).

In general, an association rule $X \implies Y$ produced by Apriori is an implication from a set X to a set Y, both containing elements of a set of elements I, i.e., $X, Y \subset I$, and $X \cap Y = \emptyset$. In our case, I corresponds to P. An association rule has support s, which describes the number of transactions in D that contain $X \cup Y$. For the transactions in D that contain X the confidence c describes the percentage of transactions that also contain Y. For representing association rules between a specific set of costume parameters that have to be investigated, let $P_x \subset P$ be the set of permitted left hand parameters and $P_y \subset P$ be the set of permitted right hand parameters, $P_x \cap P_y = \emptyset$, then for every association rule $X \implies Y$ holds $X \subseteq P_x$ and $Y \subseteq P_y$.

The following list contains association rules that we have found during our analysis of the corpus of films. The set of character traits $P_{character} = \{$ *active*, *evil*, *good*, $\dots\}$ and the set of genders $P_{gender} = \{male, female\}$ are used as left hand parameters, i.e., $P_x = P_{character} \cup P_{gender}$, while the set of investigated available base elements $P_{base} = \{trousers, necklace, \dots\}$ is used as right hand parameters, i.e., $P_y = P_{base}$:

- $\{evil, male\} \implies \{trousers\} (confidence\ c = 87,5\%)$
- $\{evil, female\} \implies \{necklace\} (confidence\ c = 75,6\%)$
- $\{active, male\} \implies \{boots\} (confidence\ c = 59,4\%)$

(Algorithm parameters: minimum support $s \geq 10$)

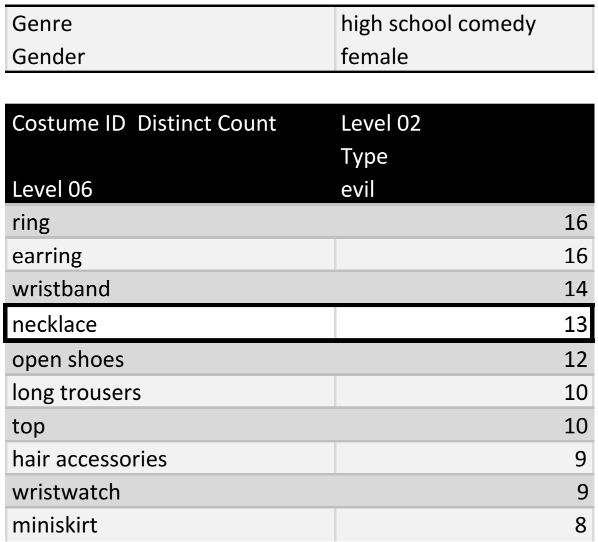
Looking at the second item in the example list, we identify that 75% of the costumes worn by evil female characters have a necklace available in their composition of base elements. We also know that this combination occurs in at least 10 costumes by looking at the configured minimum support threshold.

Rule quality can be derived from the confidence c of each association rule. Rules of high quality lead to a hypothesis about certain aspects of a costume. For example, the rule $\{evil, female\} \implies \{necklace\}$ leads to the hypothesis that evil female characters can be expressed by adding a necklace to the costume.

If no rules with appropriate confidence have been found in the Inspection step, the method can be continued back at the Modelling step, as described in Section 3.2. This allows to change parameters of the association rule discovery to find stronger rules. For example, additional model filters could be applied for focusing on a specific genre. Also parameters of the mining algorithm, such as support and confidence can be adapted to the requirements of a domain expert in order to properly investigate the data set. Specifically for mining design patterns from present concrete solutions, support and confidence have to be set to values that ensure the resulting rules to be the frequently occuring essence from many concrete solutions. In order to classify a detected solution the so called rule of three has emerged in the pattern community, which defines a detected solution to be relevant for formulating a pattern if it occurs at least three times in different concrete implementations (Coplien, 1996).

If rules with sufficient confidence have been mined, they can be validated using OLAP techniques. To understand the impact of the conceptually described capabilities in Section 3.3, the above stated example of the association rule that indicates that female characters, who are evil also wear a necklace is depicted as a pivot table in Figure 4. For this case, the filtering is configured to include only movies of the high school comedy genre and female characters. Rows present the base elements of a costume, drilled-down to the sixth level of the base element taxonomy. Columns show the character traits, drilled-down to the second level of the character traits taxonomy.

In this case, a filter is applied to only show the character trait "evil". The table holds the distinct count of costumes with the respective column/row properties. Clearly, the base element "necklace", highlighted by the bold border in Figure 4, is one of the top 5 base elements for evil female characters, so the discovered association rule of the above example can be considered as validated. Discovering association rules on the actual Costume Repository would probably also produce rules for the "ring", "earring" and "bracelet" base elements, as those are used in a similar number of costumes.



Genre	high school comedy
Gender	female

Costume ID Distinct Count	Level 02
	Type
Level 06	evil
ring	16
earring	16
wristband	14
necklace	13
open shoes	12
long trousers	10
top	10
hair accessories	9
wristwatch	9
miniskirt	8

Fig. 4 Example of Hypothesis Validation using Microsoft Excel Pivot Table

5 Prototypical Implementation and Case Studies

In this section, we describe our prototypical implementation of the introduced Costume Pattern Mining Method. In addition, we exercise the method based on two example case studies: the overall question is if we can find a typical costume for a villain or "Bad Guy" in the genres "western" and "high school comedy". At the time of writing, the database contained approximately 350 costumes from 23 western movies and approximately 2,200 costumes from 21 high school comedy movies.

As part of the Digital Humanities research, one of the core goals of the proposed prototype is to hide as much of the complexity as possible that domain experts in the area of humanities with a lower level of detailed knowledge in data mining techniques and algorithms can effectively and efficiently work with the proposed toolchain and immediately benefit from the results.

5.1 Prototype Implementation – Data Preparation

Starting with the data preparation step, as described in Section 3.1, we set up an ETL process using the Microsoft SQL Server Integration Services. In addition to the tasks already mentioned, the ETL process migrates the data from a MySQL database of the Costume Repository to a Microsoft SQL Server database staging area. The ETL process is scheduled to run once every night fully automated. This enables the analysis to be repeatable, as the daily database backup could be used to restore a specific state of the Costume Repository. The data of the Costume Repository changes rather slowly, so the interval of one day does not cause the data mining algorithm to work on outdated data.

5.2 Prototype Implementation – Hypothesis Discovery

In the Costume Repository, attributes to express a "villain" or a "Bad Guy" are represented by stereotype and

Table 1 Mining structures for discovering association rules for “villains”

Input Set		**Output Set**
Character traits, Gender	$\Longrightarrow$	Base element (appearance) Base element design Base element color Base element material Base element condition
Stereotype information, Gender	$\Longrightarrow$	Base element (appearance) Base element design Base element color Base element material Base element condition

character traits. Therefore, we can refine our overall interests into the following questions:

– Are there relations between stereotype, character traits and worn costume base elements?
– Are there relations between stereotype, character traits, design, color, material and condition of costume base elements?

Using Microsoft’s implementation of the Apriori algorithm (Microsoft Association Algorithm) the set of input and output attributes, P_x and P_y, as described in Section 3.2, have to be configured in a so called mining model. Multiple mining models are grouped in a mining structure. Having a set of mining models available allows to easily re-trigger the discovery of association rules with a defined set of parameters as the number of costumes in the costume database grows over time.

We transformed the questions above into ten mining structures. Each mining structure is designed to answer a specific part of the questions, e.g., there is one mining structure to represent the question about relations between personality attributes and base elements and there is a second mining structure representing the question about relations between stereotype information and base elements, and so on. The mining structures as depicted in Table 1 have been defined.

For each mining structure we created two mining models, one to answer the question in the perspective of western movies and another one to answer the question in the perspective of high school comedy movies. Due to the fact that we approximately have 2,200 costumes for the genre high school comedy we decided to limit the data to costumes worn by supporting roles and extra artists. Such characters have a rather short screen time and, therefore, need to communicate the stereotype and personality characteristics more efficiently than costumes with longer screen time.

The defined mining models were processed by the Microsoft SQL Server Analysis Services. It allows to apply a keyword-based filtering on the set of produced association rules. We filtered the rules by the keywords “evil” and “bad”. This gave a first impression on how a possible pattern indicator could look like. In addition, we put focus on rules including a gender property and having this property set to “male”. This resulted in the rules described in Table 2.

5.3 Prototype Implementation – Hypothesis Validation

The rules depicted in Table 2 give a first impression about pattern indicators for a villain in western movies and high school comedy movies. Scanning the rules of western movies, one can easily picture a typical costume of a bandit - wearing a black cowboy hat, a black scarf, a worn-out shirt and brown boots.

Nevertheless, the rules have to be considered as independent as one can only tell that a villain in a western movie often wears a cowboy hat and boots. In addition, villains often seem to wear costume elements made out of worn-out cotton in brown or black color. But, by now one cannot relate such base element attributes, like color and material, to specific worn base elements. Further, a domain expert has to analyze the mined rules in order to decide which rules provide meaningful information and, thus, are actually relevant for the hypothesis at hand.

Having found such a set of rules, thus, they have to be validated against the data set. We used an OLAP cube by Falkenthal et al (2015) to validate these pattern indicators. To access the provided OLAP cube we used Microsoft Excel and its pivot functionality. Regarding the domain of Digital Humanities, using Microsoft Excel provides the opportunity for users with only little IT background to easily access the functions provided. We applied filters to set genre, gender, stereotype and character traits, we are looking for.

We validated the rules by comparing two dimensions. We used the dimension base element as column values and the dimensions “base element design”, “base element color”, “base element material”, and “base element condition” as row values each combination in a separate pivot table. For the column values, we set the range of possible values, which are found by the association rule algorithm.

In case of high school comedy movies, we set the range to “Long Trousers”, “Tie”, “Business Shirt”, and “Wristwatch”. Having this setting, we were able to relate each base element to the base element attributes and validate if a discovered rule significantly expresses a state in the OLAP cube.

By applying this process, we identified that a villain in a high school comedy typically wears black trousers,

Table 2 Discovered rules for villains in genres "high school comedy" and "western"

Genre "high school comedy" (min. $s = 5$)				c	Genre "western" (min. $s = 10$)				c
1.	$\{evil, male\}$	$\Longrightarrow$	$\{long\ trousers\}$	67.6 %	16.	$\{villain, male\}$	$\Longrightarrow$	$\{long\ trousers\}$	75.4 %
2.	$\{villain, male\}$	$\Longrightarrow$	$\{tie\}$	43.8 %	17.	$\{villain, male\}$	$\Longrightarrow$	$\{boots\}$	55.4 %
3.	$\{villain, male\}$	$\Longrightarrow$	$\{business\ shirt\}$	50.0 %	18.	$\{evil, male\}$	$\Longrightarrow$	$\{revolver\}$	39.1 %
4.	$\{villain, male\}$	$\Longrightarrow$	$\{wristwatch\}$	56.3 %	19.	$\{evil, male\}$	$\Longrightarrow$	$\{cartridge\ belt\}$	26.2 %
5.	$\{villain, male\}$	$\Longrightarrow$	$\{striped\}$	50.0 %	20.	$\{villain, male\}$	$\Longrightarrow$	$\{casual\ shirt\}$	43.1 %
6.	$\{evil, male\}$	$\Longrightarrow$	$\{black\}$	71.8 %	21.	$\{villain, male\}$	$\Longrightarrow$	$\{jacket\}$	24.1 %
7.	$\{evil, male\}$	$\Longrightarrow$	$\{blue\ tones\}$	63.4 %	22.	$\{evil, male\}$	$\Longrightarrow$	$\{cowboy\ hat\}$	35.6 %
8.	$\{evil, male\}$	$\Longrightarrow$	$\{metallic\ colors\}$	73.2 %	23.	$\{villain, male\}$	$\Longrightarrow$	$\{scarf\}$	21.1 %
9.	$\{villain, male\}$	$\Longrightarrow$	$\{powerful\ color\}$	93.8 %	24.	$\{villain, male\}$	$\Longrightarrow$	$\{striped\}$	27.7 %
10.	$\{villain, male\}$	$\Longrightarrow$	$\{shiny\}$	87.5 %	25.	$\{evil, male\}$	$\Longrightarrow$	$\{checkered\}$	13.8 %
11.	$\{evil, male\}$	$\Longrightarrow$	$\{gold\}$	56.3 %	26.	$\{villain, male\}$	$\Longrightarrow$	$\{black\}$	78.5 %
12.	$\{villain, male\}$	$\Longrightarrow$	$\{cotton\}$	100.0 %	27.	$\{evil, male\}$	$\Longrightarrow$	$\{browntones\}$	66.7 %
13.	$\{evil, male\}$	$\Longrightarrow$	$\{clean\}$	73.2 %	28.	$\{evil, male\}$	$\Longrightarrow$	$\{cotton\}$	81.6 %
14.	$\{evil, male\}$	$\Longrightarrow$	$\{tidy\}$	18.3 %	29.	$\{villain, male\}$	$\Longrightarrow$	$\{leather\}$	80.0 %
15.	$\{villain, male\}$	$\Longrightarrow$	$\{ironed\}$	56.3 %	30.	$\{villain, male\}$	$\Longrightarrow$	$\{worn - out\}$	52.3 %

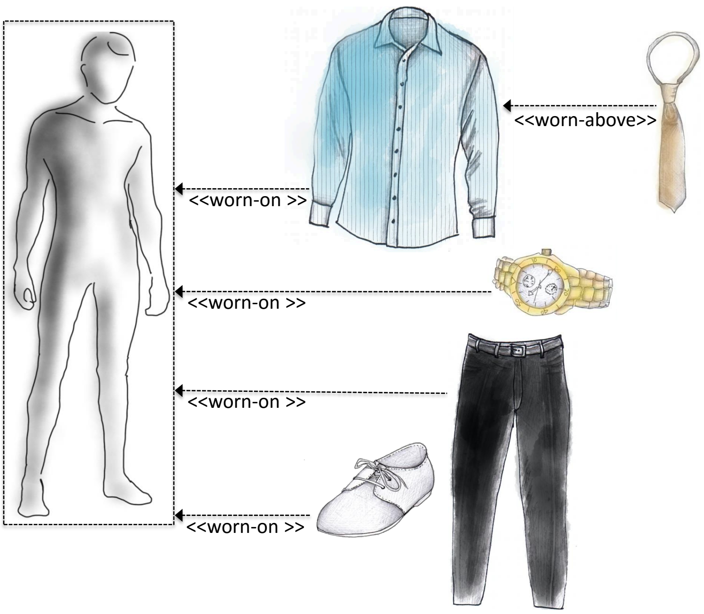


Fig. 5 Villain in a high school comedy movie

a blue-striped shirt, a brown tie and a golden watch, as depicted by the costume composition graph in Fig. 5. Looking at the profession of roles wearing such costumes, one can discover that mostly "teachers" and "attorneys" are expressed by such costumes and these roles often act as the counterpart of the protagonist, namely the "popular boy" or the "prom king" in high school comedy movies. In contrast, a villain in a western movie typically wears a black cowboy hat, a black scarf, a brownish, worn-out and checkered shirt, a black jacket and brown boots and pants as depicted in Fig. 6. If drilling-down to the type of role that is represented by such costumes, we can determine that such roles are often expressed as bandits.

These resulting costume composition graphs, show the feasibility of the presented approach. Applying the Costume Pattern Mining Method on our corpus of concrete costume descriptions, we were able to identify indicators for costume patterns. To optimally reuse these findings, a domain expert can build upon them to author costume patterns by refining the results to include even more relevant parameters and abstracting the essence into patterns.

6 Summary and Outlook

In this paper, we presented an approach to identify indicators for costume patterns in movies by using data mining techniques for finding coherences between cos-

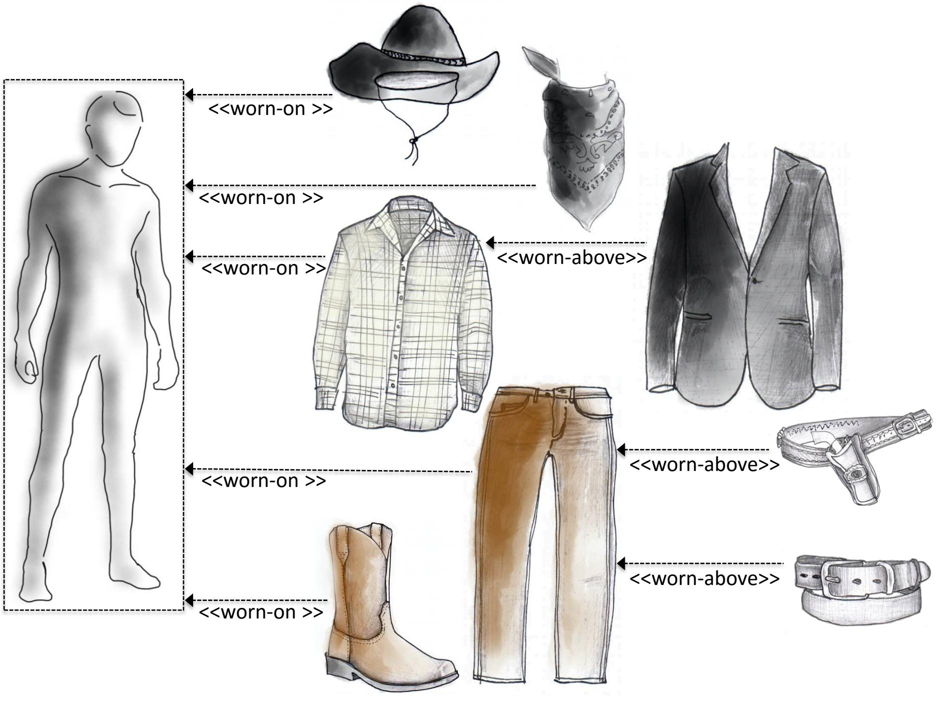


Fig. 6 Villain in a western movie

tumes. A prototype was presented that builds upon data mining techniques for basic validation of those. This can support the identification of costume patterns and the creation of a costume language, as described by Barzen and Leymann (2015) as well as Fehling et al (2015). Our work shows how IT can seize issues from the humanities and contributes approaches and ideas typically not utilized in this domain. Therefore, the presented approach is a motivating example on how the emergent field of Digital Humanities can be enabled and influenced building upon established methods and techniques from IT.

We were focusing on a specific area of costume parts and attributes for identifying costume patterns in specific genres, as the current film corpus has the best coverage on those genres. To enhance the results of the presented method in the future, the number of data mining structures has to be increased. This also includes the usage of additional data mining algorithms, such as clustering. Also the set of genres can be expanded, as the film corpus will grow. Executing the presented method with different movie genres can also give additional confirmation for the approach to work.

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., San Francisco, USA, VLDB '94, pp 487–499

Alexander C, Ishikawa S, Silverstein M (1977) A pattern language: towns, buildings, construction. Oxford University Press, New York

Appleton B (1997) Patterns and software: essential concepts and terminology. Object Magazine Online 3(5)

Barzen J (2013) Taxonomien kostümrelevanter Parameter: Annäherung an eine Ontologisierung der Domäne des Filmkostüms. Technical Report 2013/04, University of Stuttgart, Faculty of Computer Science, Electrical Engineering and Information Technology, Germany

Barzen J, Leymann F (2015) Costume languages as pattern languages. In: Baumgartner P, Sickinger R (eds) Proceedings of PURPLSOC (Pursuit of Pattern Languages for Societal Change). The Workshop 2014., epubli GmbH, pp 88–117

Barzen J, Leymann F (2016) Patterns as Formulas: Applying the Scientific Method to the Humanities. Technical Report 2016/01, University of Stuttgart, Faculty of Computer Science, Electrical Engineering and Information Technology, Germany, University of Stuttgart, Institute of Architectur of Application Systems

Barzen J, Falkenthal M, Hentschel F, Leymann F (2015) Musterforschung in den Geisteswissenschaften: Werkzeugumgebung zur Musterextraktion aus Filmkostümen. In: Extended Abstract Digital Humanities im deutschsprachigen Raum (DHd 2015), DHd 2015, Graz

Bishop C (2006) Pattern Recognition and Machine Learning. Springer, New York

Codd EF, Codd SB, Salley CT (1993) Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. E. F. Codd and Associates

Coplien J (1996) Software Patterns. SIGS

Dearden A, Finlay J (2006) Pattern Languages in HCI: A Critical Review. Human-Comp Interaction 21(1):49–102

Falkenthal M, Barzen J, Breitenbücher U, Fehling C, Leymann F (2014a) Efficient Pattern Application: Validating the Concept of Solution Implementations in Different Domains. International Journal On Advances in Software 7(3&4):710–726

Falkenthal M, Barzen J, Breitenbücher U, Fehling C, Leymann F (2014b) From pattern languages to solution implementations. In: Proceedings of the 6th International Conferences on Pervasive Patterns and Applications (PATTERNS), pp 12–21

Falkenthal M, Barzen J, Dörner S, Elkind V, Fauser J, Leymann F, Strehl T (2015) Datenanalyse in den Digital Humanities – Eine Annäherung an Kostümmuster mittels OLAP Cubes. In: Datenbanksysteme für Business, Technologie und Web (BTW), 16. Fachtagung des GI-Fachbereichs “Datenbanken und Informationssysteme” (DBIS), Lecture Notes in Informatics

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39(11):27–34

Fehling C, Barzen J, Breitenbücher U, Leymann F (2014) A process for pattern identification, authoring, and application. In: Proceedings of the 19th European Conference on Pattern Languages of Programs – EuroPLoP '14, Association for Computing Machinery (ACM)

Fehling C, Barzen J, Falkenthal M, Leymann F (2015) PatternPedia - Collaborative Pattern Identification and Authoring. In: Proceedings of PURPLSOC (Pursuit of Pattern Languages for Societal Change). The Workshop 2014., epubli GmbH, pp 252–284

Golfarelli M, Maio D, Rizzi S (1998) The Dimensional Fact Model: A Conceptual Model For Data Warehouses. International Journal of Cooperative Information Systems 7:215–247

Hohpe G, Woolf B (2003) Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions. Addison-Wesley Longman Publishing Co., Inc.

ISO (2006) ISO/IEC 13249-6:2006 Information technology – Database languages – SQL Multimedia and Application Packages – Part 6: Data Mining

Reiners R (2013) An Evolving Pattern Library for Collaborative Project Documentation. Phd thesis, RWTH Aachen University

Reiners R, Falkenthal M, Jugel D, Zimmermann A (2015) Requirements for a collaborative formulation process of evolutionary patterns. In: Proceedings of the 18th European Conference on Pattern Languages of Program – EuroPLoP '13, Association for Computing Machinery (ACM)

Schumm D, Barzen J, Leymann F, Ellrich L (2012) A pattern language for costumes in films. In: Proceedings of the 17th European Conference on Pattern Languages of Programs – EuroPLoP '12, Association for Computing Machinery (ACM)