# Knowledge Discovery and Knowledge Visualization

Katharina Morik,
Michael Wurst

University of Dortmund,
Artificial Intelligence Dept.,
D-44221 Dortmund, Germany
Tel. 0231 - 755 5101
wurst@ls8.cs.uni-dortmund.de

## Abstract

Past experiences have shown that there is a strong connection between Knowledge Discovery in Databases and Knowledge Visualization. This connection is twofold. On the one hand, Visualization can serve as a powerful tool for identifying interesting patterns and relationships in data in an intuitive way. On the other hand, large amounts of data can often not be visualized directly, as the resulting graphical representation gets much to complex to be captured by human user. Therefore, intelligent methods are needed to extract and aggregate the essential information contained in the database first. The resulting structures can then be visualized, to give the user an intuitive access to the whole data set. Following this observation, we argue that in many domains only the combination of modern Knowledge Discovery and Knowledge Visualization techniques will lead to successful solutions. As to exemplify and support this thesis, several state-of-the-art approaches in the field of Knowledge Discovery in large databases will be presented. These approaches derive from practical applications and therefore offer an interesting insight in "real life" Knowledge Discovery.
Additional to this more general overview of the connection of Knowledge Discovery and Visualization, the special relevance of this connection to large databases of documents will be pointed out, as this is a central issue in the development of digital archives and portals.

## 1. Introduction

To analyze the relationship between Knowledge Discovery and Visualization, we should first take a look at two common definitions for these terms.

According to U. Fayyad, Knowledge Discovery is

"[..] the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"[1].

A typical definition for (scientific) Visualization is the following:

"[Scientific] Visualization is concerned with exploring data and information is such a way as to gain understanding and insight into the data"[16]

According to this, it seems, that Knowledge Discovery and Visualization are two alternative methods for a common goal: making large amounts of data comprehensible and utilizable for human users.

Though, as will be argued, this oversimplifies the situation. Instead of thinking of them as alternatives, Knowledge Discovery and Visualization should be seen as two steps in one common process, both with their individual strengths and limitations.

In section two, this relationship between Knowledge Discovery and Visualization will be discussed in more detail.

Section three discusses possibilities, of how to go the other way around and utilize user actions gained from interactive Visualization to enrich the sources of information for Knowledge Discovery.

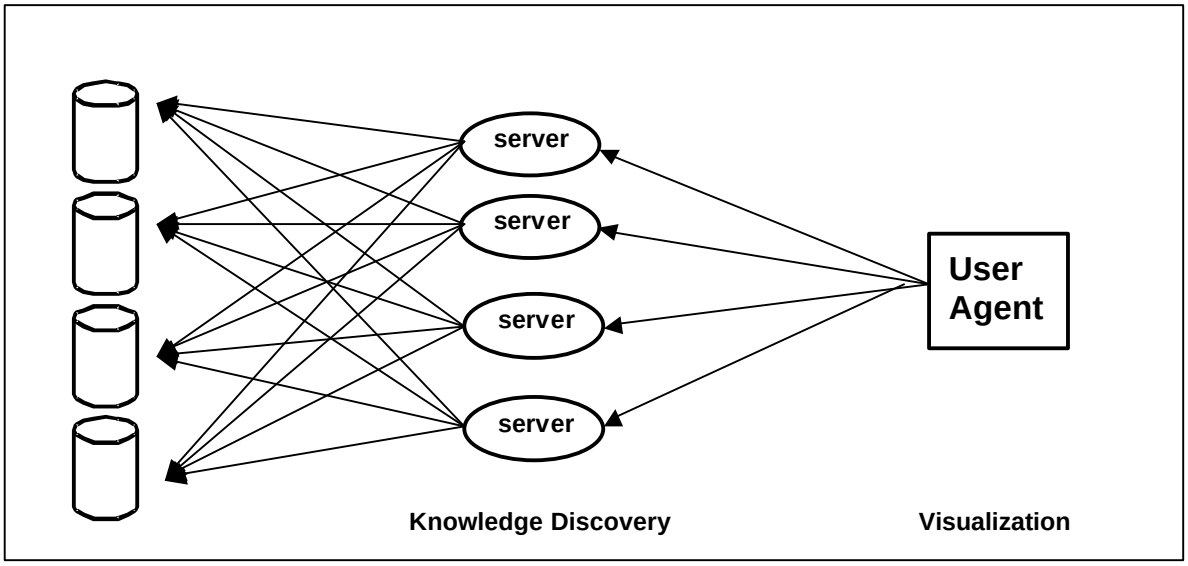


**Figure 1**

The fourth section provides a brief summary of the whole paper.

## 2. The relationship between Knowledge Discovery and Knowledge Visualization

Knowledge Discovery and Knowledge Visualization often share a common goal: to make large amounts of data comprehensible and utilizable for human users. The fact, that the access to large amounts of data is one of the basic problems in a modern information society, has been pointed out many times (e.g. [1]). Some of the problems are obvious while others are more subtle. One obvious problem is, that the number of available data sets (e.g. research documents) has increased in such a way, that human users cannot go through all the data sets sequentially as this would simply exceed their cognitive abilities. More subtle though not less important is the fact, that also the information available about individual datasets as well as the interrelations between data sets have increased dramatically. Even in cases, in which it is possible to go through all the data sets by hand, there is still the problem that relevant information can be revealed only by putting several pieces of information together. In a large space of relations and attributes, this is not feasible even for expert users.

Both, Knowledge Discovery and Knowledge Visualization provide methods to deal with these problems. Though they are more than just two alternative approaches to solve a common problem. Rather they can be seen as different steps in a process that leads from raw data to information relevant and useful for a human user

This relationship is depicted in figure 1, which serves as a typical (but simplified) model of the whole process of discovering and visualizing information.

At one end of this process, we have data, possibly stored in different sources and distributed among a network. At the other end, there is a user with specific information needs, preferences and constraints. Knowledge Discovery and Visualization are located in between. While Knowledge Discovery mostly takes place at server side, Visualization is located at a user client.

This separation is motivated by the individual contributions of both in the overall process.

The aim of Knowledge Discovery is to summarize and index the underlying information, in order to reveal relevant patterns describing aspects of the original data in an aggregated form. This requires intensive access to the connected databases as well as a considerable amount of computation. In large financial databases for example, processing times

of several weeks are quite common. Obviously, it is not reasonable to perform such kind of processing on the client side. Another reason why to locate Knowledge Discovery at the server side is that the discovered structures are often of interest for many different users. Performing this process anew for every user, would lead to redundant computation.

After Knowledge Discovery has been performed, user agents can efficiently access the aggregated information provided by the server. These agents offer users the possibility to choose, which information is relevant to them and to present it in an adequate way. This includes also the adaptation to user preferences and constraints, which arise e.g. from the use of mobile devices.

Although the depicted process is quite typical, there will be certainly situations in which the underlying data can be reasonably visualized just as it is, without any intermediate pre-processing necessary. On the other hand, Knowledge Discovery results are sometimes simply printed to a terminal including no Visualization at all. Finally, in some cases parts of the task assigned to the user agent are actually located at the server.

Nevertheless, paradigmatically, the presented separation of Knowledge Discovery and Visualization can be very helpful as it offers insights in the individual strengths and contributions of both of them to the overall process.

In the remainder of this section, these contributions will be discussed in more detail.

### 2.1 Knowledge Discovery as preprocessing for Visualization

#### 2.1.1.Connecting heterogeneous sources

Information relevant for a given domain is most often not just simply stored in one central database, but rather distributed among different sources of diverse types. To find information about movies, for example, we may have a database with some meta information about these movies, such as who directed a film. Moreover we have individual internet pages that contain additional facts, newsgroups with film reviews, movie recommender systems providing ratings for movies and so on.

This situation is the source of severe problems concerning the access as well as the automatic processing of this information. Without any preprocessing, the only way to access the data is through the native interfaces of the individual information sources. This is very unsatisfactory for the user, as she cannot access all the information in a unified way. For automatic processing this is even worse, as relationships and connections between data sets from different sources cannot be established without further processing.

There are two common solutions of this problem. Firstly, the heterogeneous sources can be mapped manually to a common schema. This solution is quite simple and straightforward. Though the problem is, that a knowledge engineer has to maintain the system in order to incorporate new data sources or as a reaction to changes. As this is a non-trivial task, the costs as well as the delay of maintenance are serious disadvantages of this approach.

Alternatively the mapping can be performed automatically or semi-automatically. Although this is much more complicated, it offers some important advantages. First of all, the work of the knowledge engineer is simplified considerably, reducing cost and time for maintenance. Secondly, also information stored in semi-structured or even natural language sources can be extracted, which is mostly beyond the possibilities of manual mapping of data sources.

Currently there are three major research areas dealing with problems related to the automated integration of different data sources aiming at different types of underlying data: intelligent schema matching, wrapper induction and information extraction.

Intelligent schema matching is used, to map information between databases, which describe objects or relations in different ways. For example a first movie data base might contain a field

"name" for actors, while another one contains two fields "name" and "surname". Intelligent schema matching can be used to find corresponding attributes (semi-) automatically. A comprehensive overview on current problems and approaches in this area can be found in [3].

Wrapper induction aims at information, which is structured in different ways, but unlike databases, not only the attributes, but also, the format in which the information is stored can differ. A typical example is product information at the internet pages of different online vendors. The format of the information provided by one vendor is the same for all products at this site, but differs from the format used by other vendors. To get a common view on all sources, they have to be wrapped by a mechanism translating the proprietary format to a common format. There are several approaches how to "learn", to generate, such a mapping mechanism automatically (see e.g. [4]).

Finally, often information is contained in natural language texts, such as news articles. There is a wide variety of approaches trying to extract information from such sources. This is usually much more difficult than schema matching or wrapper induction, but is of great relevance, as many information is stored as natural text only. Information extraction has a long history in Artificial Intelligence and Natural Language Processing. For a brief overview refer to [2].

### 2.1.2. Dealing with massive amounts of data

After several data sources are connected, there is still the problem of processing the underlying data. The methods needed to process huge amounts of data often differ completely from methods to handle smaller amounts of data as they have to be much more efficient. This is even more significant, if not all of the processed data can be stored in main memory. In such cases sophisticated swapping strategies have to be used, to make the whole process sufficiently efficient.

Just to give an impression, of how much data maybe involved, the following list contains statistics of the amount of data sets contained in the Internet Movie Data Base, a comprehensive source of movie information on the internet:

| | |
|---|---:|
| Movies | 300.000 |
| Persons | 1.000.000 |
| Relations | 8.000.000 |
| Ratings | 18.000.000 |

Beside the development of Modern Database Technology, research in the domain of Knowledge Discovery has resulted in many algorithms specialized in dealing with very large amounts of data. A classical example for such an algorithm is Apriori([12]) used to find association rules in large data sets.

### 2.1.3 Dealing with data distributed among different relations

The majority of Knowledge Discovery, as well as Visualization models and methods (e.g. the popular SOM[7]) expect the underlying data to be represented in attribute-value format. This means that all of the datasets are described by the same set of attributes. Though this is in contrast to the way how data is stored in many real world applications. Relational databases still represent the most wide spread form of storing large amounts of data. In these databases, information is distributed among different relations.

In the Internet Movie Database ([15]) for example, there is a relation for persons, a relation for movies and many relations as actor or director, to describe the relationship between both.

Again, there are several possibilities to deal with this problem. On is to convert the data to attribute-value form, a process that is called "propositionalization". This process can be performed manually or automatically ([14]). Please note that either way, this is far from being trivial. To describe movies, for example, we are certainly interested in the actors appearing in a movie. Though how to find a common set of attributes for this purpose? One possibility would be to have

attribute fields with “Actor #1” “Actor #2“, etc. This is problematic as the order of actors can be misleading. Another possibility would be to provide a binary attribute for each actor, denoting whether he is appearing in a given movie. Though this would result (in case of IMDB) in over 500.000 attributes per movie.

Regarding this problems, there are some (though quite few) methods, which operate directly on relational data. One example for such a method is the “Rule Discovery Tool”(RDT, [13]), which is able to find rules in a subset of first-order predicate logic in relational databases.

For a broader view on approaches connected with Knowledge Discovery in relational databases refer to [14].

### 2.1.4. Global vs. Local patterns

The aim of Knowledge Discovery is usually to find useful patterns in data. We can now distinguish between two types of patterns: global ones and local ones. A global pattern is the observable result of an underlying dominant process. Exceptions to this global pattern (data sets which do not fit into the pattern) can be either noise or interesting and useful local patterns. In the first case, these datasets are simply random deviations. In the latter case, they maybe exactly what the user was looking for.

While global patterns are often “obvious” and easy to observe directly by just looking at the datasets through an adequate Visualization technique, local patterns are most often hidden behind the global patterns. So, while Visualization can serve as a powerful tool to reveal global, it is not suitable at all for finding hidden local patterns.

At this point, the use of Knowledge Discovery can help to find even such hidden patterns and to make them accessible to Visualization.

The following example should make the distinction between global and local patterns clearer (see [5] for details).

The Internet Movie Database([15]) contains a list of the one hundred top ranking movies. Now the question is, what characterizes a movie ranking among the top movies. To answer this question, we use the RDT method to generate a set of rules, describing these movies. In a first step, we try to find the global patterns, the “obvious” rules. The result is depicted in figure 2. The first rule states for example, that if a movie is from the US and a drama, then it ranges among the top movies. A cineaste, or at least someone knowing some of the films in the top-ranking list, would have found this rule by simply taking a quick glance at the list of films.

$h_1$ : usa(X) Ù drama(X) ® top(X)
$h_2$ : director(X,Y) Ù topDir(Y) ® top(X)
$h_3$ : actor(X,Y) Ù topActor(X) ® top(X)
$h_4$ : director(X,Y) Ù notbotDir(Y) ® top(X)
$h_5$ : actor(X,Y) Ù notbotActor(Y) ® top(X)

**Figure 2**

But although the rules in figure 2 are already quite accurate they do not cover all movies correctly. There are two types of mistakes such a set of rule can produce. Firstly, they can predict that a film ranges among the top ranking movies, while it doesn’t. The rule set above contains 14 mistakes of this kind. Secondly, they may not cover a movie, although it actually belongs to the top one hundred movies. This happened 15 times.

As stated above, these mistakes can be either random noise, or they may contain interesting local patterns. To find out, whether there actually are local patterns, RDT is applied to the misclassified movies only, using a slightly modified space of possible rules. In a first step, the task is to characterize the movies, which are falsely predicted as top ranking movies, allowing individuals (in this case persons) to appear in the rules. The result is shown in figure 3. It turns out that the 14 movies can be characterized by three individual persons acting in them as well as by a rule stating that a movie, which is additionally to being a drama also, a musical does not range among the top movies. Interestingly it turns out, that the individual actors, which are according to the rules responsible for the movies not to be

among the top movies, all acted in the movie series "Police Academy" before.

In a second step, the aim is to find the movies, which were not covered by the global pattern, but which are actually top ranking movies, again using a modified space of possible rules. Figure 4 shows the result of applying RDT to these 15 uncovered movies. Most interestingly, these rules cover mainly European movies. This is a very nice example of how interesting rules about local phenomena (as the European film) can be hidden behind a global rule covering only global phenomena (as the Hollywood films).

**$h_8$: actor(X,53) Ù botPerson(53)**
**® notop(X)**
**$h_9$: actor(X,17) Ù botPerson(17)**
**® notop(X)**
**$h_{10}$: actor(X,30) Ù botPerson(30)**
**® notop(X)**
**$h_{11}$ : usa(X) Ù drama(X) Ù musical(X)**
**® notop(X)**

**Figure 3**

**$h_6$ : italy(X) Ù drama(X)**
**® top(X)**
**$h_7$ : denmark (X) Ù drama(X)**
**® top(X)**
**$h_{12}$ : europe(X) Ù key_family(X)**
**® top(X)**
**$h_{13}$ : europe(X) Ù key_independent(X)**
**® top(X)**
**$h_{14}$ : europe(X) Ù key_bicycle(X)**
**® top(X)**
**$h_{15}$ : drama(X) Ù key_love(X)**
**® top(X)**

**Figure 4**

Beyond the problem of local and global patterns itself, the study of this problem reveals advantages and limitations of Knowledge Discovery and Visualization. To find global patterns, it is in many cases sufficient to provide the user with an adequate Visualization so she can find the patterns by herself. This has the advantage that no complicated and costly Knowledge Discovery is necessary. On the other hand Visualization shows its limitations when we go beyond global patterns (in our example beyond Hollywood films). Here even adequate Visualization makes it hard to find relevant information.

### 2.2 Visualization as adaptive interface to Knowledge Discovery results

In the last section we have shown, that methods from the field of Knowledge Discovery and Machine Learning can be very useful to enable Visualization, by connecting heterogeneous data sources, by aggregating massive amounts of data, by dealing with multi-relational data and by finding local patterns in data. In this section, we will emphasize on some aspects in which Visualization is essential in the process of making large amounts of data utilizable.

Certainly one of the most challenging problems is to make the results of a data analysis easily compressible to human users. Rule sets, for example, do often not contain just some few rules, as in the example above, but several dozens of rules. If they are just presented in a list, they are very hard to overlook. The same holds for many other types of aggregated data.

Though understandability is only one aspect of the presentation. Users may often not be interested in all of the discovered information but only in parts of it. In these cases interactive Visualization can enable advanced navigation in the result space, allowing to identify relevant information efficiently, by e.g. choosing the desired level of detail.

This second point is strongly connected with a third point concerning the adaptation to the users preferences and needs. This may include a customizable interface as well as possibilities to adapt the presentations to different media types used (e.g. cell phone displays).

These points exemplify that Visualization is essential, as the best and most interesting

information discovered in data do not fulfill the condition of making this data utilizable, if they are not easy to access and comprehensible for human users.

## 3. Extending the model

In the model presented in section 2, the data flow has a clear direction: from raw data through a Knowledge Discovery step to the interactive Visualization clients. Though not only the data sources themselves may contain valuable information. Many connections and relationship might not be included in the data itself, but are established only through the specific knowledge of individual users. As this knowledge is to some extend reflected in the user interactions, analyzing these interactions can reveal additional useful information. Therefore the model can be extended in a way, which allows interaction data to flow back from the clients to the servers, where it can be analyzed together with the content of the actual data sources. This principle is for example the basis for collaborative filtering systems, making recommendations for items, based on user ratings. (see e.g. [10]).

In the remainder of this section we will exemplify this approach by presenting some research results from the project AWAKE ([6]).

One of the central tasks for Knowledge Discovery in this project is to find similar and related items, which in this case are mainly documents. The information about similar items is the basis for various Visualization and interaction techniques.

There are two major possibilities to identify similar items.

Firstly, it is possible to use information about the items (such as author, year of publication, etc.). From this information, similarity concerning different aspects can be established, as documents can e.g. share authors or have the same year of publication. These aspects can be combined to obtain a single similarity measure among the items, to which we refer as content similarity.

The second possibility to identify similar items is by analyzing the user interactions on these items. In AWAKE, users have the possibility to group items together, denoting that these items belong in some way together. These implicit statements about connections of items, allow, together with some statistical processing, to obtain an alternative similarity measure among the items, which will be denoted as context similarity as it is derived from the use context of the items.

As for most methods a single similarity measure is needed, the challenging question is, how content based and context similarity should be combined into a single measure, preserving the advantages of both. The advantage of content based similarity is, that it is always applicable and does not rely on user generated data. Though content based similarity can lead to poor results, if the underlying items are too heterogeneous, e.g. are represented by different types of meta data. On the other hand, using context similarity, we avoid these problems completely. The disadvantage of context similarity is however, that if only few users look at a given item or if the contexts, in which it appears, diverge, we do not get any reliable evidence on the similarity of this item to other items.

A deeper discussion of this problem is beyond the scope of this paper. (A possible solution to the problem of combining content and context similarity based on a statistical test is described in [11])

## 4. Conclusion

The central thesis of this paper is, that Knowledge Discovery, as a means of finding useful patterns in data, and Visualization, as a means of making data understandable by presenting it adequately, are not just two alternative approaches to the common problem of making large amounts of data utilizable. Rather they both have their individual strengths and limitations. This suggests combining them for an optimal performance of the overall system. As we have shown, typical contributions of Knowledge Discovery are methods for the integration of heterogeneous sources, dealing with massive amounts of data, dealing with multi-relational data and detecting local patterns. Typical

contributions of Visualization include methods and models to make the results of Knowledge Discovery understandable, interactive methods for choosing relevant information, as well as adaptation to the user needs and preferences.

In the original model, Knowledge Discovery provides the input for Visualization. By pointing out, that the implicit knowledge of users can be revealed by analyzing their behavior in interactive Visualization, we have shown, that on the other hand Visualization can provide interesting input for the Knowledge Discovery methods.

## 5. Bibliography


[1] Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.: From data mining to knowledge discovery: An overview, In Advances in Knowledge Discovery and Data Mining, AAAI Press and the MIT Press, Chapter 1, 1-34.

[2] Grishman, R. : Information Extraction: Techniques and Challenges, SCIE-97.

[3] Rahm, E., Bernstein P.A. : A survey to automatic schema matching, VLDB Journal 10(4): 334-350, 2001.

[4] Kushmerick, N. : Wrapper Induction: Efficiency and Expressiveness, Artificial Intelligence Journal, 118, pp. 15-66, 2000.

[5] Morik, K.: Detecting Interesting Instances, in Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery.

[6] Novak, et al.: Augmenting the knowledge bandwidth and connecting heterogeneous communities through uncovering tacit knowledge, in Proceedings of the IEEE Workshop on Knowledge Media Networking (KMN 2002)

[7] Kohonen, T.; Kaski S.; et al.; Self Organization of a Massive Document Collection, IEEE Transactions on Neural Networks, Vol. 11, No. 3, May 2000.

[10] Resnick, P. et al; GroupLens: An Open Architecture for Collaborative Filtering of Netnews; Proceedings of ACM 1994 Conference on CSCW, Chapel Hill, North Carolina, 1994

[11] Wurst, M.; Novak, J.; Schneider, M.: Integrating Different Machine Learning Methods to Support Search in Cross-domain Information Sources - the Project Awake, in Proceedings of the FGML Workshop 2002.

[12] Agrawal, R.;Srikant. R..: Fast Algorithms for Mining Association Rules, in Proc. of the 20th International Conference on Very Large Databases, 1994.

[13] Morik,K.; Wrobel,S.;Kietz, J.U.;Emde,W: Knowledge Acquisition and Machine Leaming - Theory, Methods, and Applications, Academic Press, 1993.

[14] Dzeroski,S.; Lavrac, Nana (editors): Relational Data Mining, Springer, 2001

[15] www.imdb.com

[16] Brodlie, K.W. etal.: Scientific Visualization, Springer 1992