# A Roadmap for Web Mining: From Web to Semantic Web

Bettina Berendt[1], Andreas Hotho[2], Dunja Mladenic[3], Maarten van Someren[4], Myra Spiliopoulou[5], Gerd Stumme[2]

[1] Institute of Information Systems, Humboldt University Berlin, Germany. berendt@wiwi.hu-berlin.de
[2] Chair of Knowledge & Data Engineering, University of Kassel, Germany, {hotho, stumme}@cs.uni-kassel.de
[3] Jozef Stefan Institute, Ljubljana, Slovenia, Dunja.Mladenic@ijs.si
[4] Social Science Informatics, University of Amsterdam, The Netherlands, maarten@swi.psy.uva.nl
[5] Institute of Technical and Business Information Systems, Otto–von–Guericke–University Magdeburg, Germany, myra@iti.cs.uni-magdeburg.de

## 1 Introduction

The purpose of Web mining is to develop methods and systems for discovering models of objects and processes on the World Wide Web and for web-based systems that show adaptive performance. Web Mining integrates three parent areas: Data Mining (we use this term here also for the closely related areas of Machine Learning and Knowledge Discovery), Internet technology and World Wide Web, and for the more recent Semantic Web. The World Wide Web has made an enormous amount of information electronically accessible. The use of email, news and markup languages like HTML allow users to publish and read documents at a world-wide scale and to communicate via chat connections, including information in the form of images and voice records. The HTTP protocol that enables access to documents over the network via Web browsers created an immense improvement in communication and access to information. For some years these possibilities were used mostly in the scientific world but recent years have seen an immense growth in popularity, supported by the wide availability of computers and broadband communication. The use of the internet for other tasks than finding information and direct communication is increasing, as can be seen from the interest in "e-activities" such as e-commerce, e-learning, e-government, e-science.

Independently of the development of the Internet, Data Mining expanded out of the academic world into industry. Methods and their potential became known outside the academic world and commercial toolkits became available that allowed applications at an industrial scale. Numerous industrial applications have shown that models can be constructed from data for a wide variety of industrial problems (e.g. [1, 2]).

The World-Wide Web is an interesting area for Data Mining because huge amounts of information are available. Data Mining methods can be used to analyse the behaviour of individual users, access patterns of pages or sites, properties of collections of documents. Almost all standard data mining methods are designed for data that are organised

as multiple "cases" that are comparable and can be viewed as instances of a single pattern, for example patients described by a fixed set of symptoms and diseases, applicants for loans, customers of a shop. A "case" is typically described by a fixed set of features (or variables). Data on the Web have a different nature. They are not so easily comparable and have the form of free text, semi-structured text (lists, tables) often with images and hyperlinks, or server logs. The aim to learn models of documents has given rise to the interest in Text Mining [3]: methods for modelling documents in terms of properties of documents. Learning from the hyperlink structure has given rise to graph-based methods, and server logs are used to learn about user behavior.

The Semantic Web is a recent initiative, inspired by Tim Berners-Lee [4], to take the World-Wide Web much further and develop in into a distributed system for knowledge representation and computing. The aim of the Semantic Web is to not only support access to information "on the Web" by direct links or by search engines but also to support its *use*. Instead of searching for a document that matches keywords, it should be possible to combine information to answer questions. Instead of retrieving a plan for a trip to Hawaii, it should be possible to automatically construct a travel plan that satisfies certain goals and uses opportunities that arise dynamically. This gives rise to a wide range of challenges. Some of them concern the infrastructure, including the interoperability of systems and the languages for the exchange of information rather than data. Many challenges are in the are of knowledge representation, discovery and engineering. They include the extraction of knowledge from data and its representation in a form understandable by arbitrary parties, the intelligent questioning and the delivery of answers to problems as opposed to conventional queries and the exploitation of formerly extracted knowledge in this process. The ambition of representing content in a way that can be understood and consumed by an arbitrary reader leads to issues in which cognitive sciences and even philosophy are involved, such as the understanding of an asset's intended meaning.

The Semantic Web proposes several additional innovative ideas to achieve this:

**Standardised format.** The Semantic Web proposes standards for uniform metalevel description language for representation formats. Besides acting as a basis for exchange, this language supports representation of knowledge at multiple levels. For example, text can be *annotated* with a formal representation of it. The natural language sentence "Amsterdam is the capital of the Netherlands", for instance, can be annotated such that the annotation formalises knowledge that is implicit in the sentence, e.g. Amsterdam can be annotated as "city", Netherlands as "country" and the sentence with the structured "capital-of(Amsterdam, Netherlands)". Annotating textual documents (and also images and possibly audio and video) thus enables a combination of textual and formal representations of knowledge. A small step further is to store the annotated text items in a structured database or knowledge base.

**Standardised vocabulary and knowledge.** The Semantic Web encourages and facilitates the formulation of shared vocabularies and shared knowledge in the form of ontologies: if knowledge about university courses is to be represented and shared, it is

useful to define and use a common vocabulary and common basic knowledge. The Semantic Web aims to collect this in the form of ontologies and make them available for modelling new domains and activities. This means that a large amount of knowledge will be structured, formalised and represented to enable automated access and use.

**Shared services.** To realise the full Semantic Web, beside static structures also "Web services" are foreseen. Services mediate between requests and applications and make it possible to automatically invoke applications that run on different systems.

In this chapter, we concentrate on one thread of challenges associated with the Semantic Web, those that can be addressed with knowledge discovery techniques, putting the emphasis on the transition from Web Mining to mining the Semantic Web and on the role of ontologies and information extraction for this transition. Section 2 summarises the more technical aspects of the Semantic Web, in particular the main representation languages, section 3 summarises basic concepts from Data Mining, section 4 reviews the main developments in the application of Data Mining to the World Wide Web, section 5 extends this to the combination of Data Mining and the Semantic Web and section 6 reviews developments that are expected in the near future and issues for research and development. Each section has the character of a summary and includes references to more detailed discussions and explanations. This chapter summarises and extends [5], [6] and [7].

## 2 Languages for the Semantic Web

The Semantic Web requires a language in which information can be represented. This language should support (a) knowledge representation and reasoning (including information retrieval but ultimately a wide variety of tasks), (b) the description of document content, (c) the exchange of the documents and the incorporated knowledge and (d) standardisation. The first two aspects demand adequate expressiveness. The last two aspects emphasise that the Semantic Web, like the Web, should be a medium for the exchange of a wide variety of objects and thus allow for ease-of-use and for agreed-upon protocols. Naturally enough, the starting point for describing the Semantic Web has been XML. However, XML has not been designed with the intention to express or exchange knowledge. In this section, we review three W3C initiatives, XML, RDF(S) and OWL and their potential for the Semantic Web.

### 2.1 XML

XML (Extensible mark-up language) was designed as a language for mark-up or annotation of documents. An XML object is a labeled tree and consists of objects with attributes and values that can themselves be XML objects. Beside annotation for formatting, XML allows the definition of any kind of annotation, thus opening the way to annotation with ontologies and to use as data model for arbitrary information. This makes it extensible, unlike its ancestors like HTML.

XML Schema allows the definition of grammars for valid XML documents, and the reference to "name spaces", sets of labels that can be accessed via the internet. XML can also be used as a scheme for structured databases. The value of an attribute can be text but it can also be an element of a limited set or a number. XML is only an abstract data format.

Furthermore, XML does not include any procedural component. Tools have been developed for search and retrieval in XML trees. Tools can create formatted output from formatting annotations but in general any type of operation is possible. When tools are integrated in the Web and can be called from outside they are called "services". This creates a very flexible representation format that can be used to represent information that is partially structured.

Details about XML can be found in many books, reports and Web pages. In the context of the Semantic Web, the most important role for XML is that it provides a simple standard abstract data model that can be used to access both (annotated) documents and structured data (for example tables) and that it can be used as a representation for ontologies. However, XML and XML schema were designed to describe the structure of text documents, like HTML, Word, StarOffice, or LaTeXdocuments. It is possible to define tags in XML to carry meta data but these tags may not have a well-defined meaning. XML helps organizing documents by providing a formal syntax for annotation. Erdmann [8] provides a detailed analysis of the capabilities of XML, the shortcomings of XML concerning semantics and possible solutions. For Web Mining the standardisation created by XML simplifies the development of generic systems that learn from data on the web.

### 2.2 RDF(S)

The *Resource Description Framework (RDF)* is, according to the W3C recommendation [9], "a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web."

RDF documents consist of three types of entities: resources, properties, and statements. Resources may be Web pages, parts or collections of Web pages, or any (real-world) objects which are not directly part of the World-Wide Web. In RDF, resources are always addressed by URIs, Universal Resource Identifiers, a generalisation of URLs that includes services besides locations. Properties are specific attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource form an RDF statement. A value is either a literal, a resource, or another statement. Statements can thus be considered as object–attribute–value triples.

The data model underlying RDF is basically a directed labeled graph. RDF Schema defines a simple modeling language on top of RDF which includes classes, is-a relationships between classes and between properties, and domain/range restrictions for properties. XML provides the standard syntax for RDF and RDF Schema.

Summarising, RDF and RDF Schema provide base support for the specification of semantics *and* use the widespread XML as syntax. However, the expressiveness is limited, disallowing the specification of facts that one is bound to expect, given the long tradition of database schema theory. They include the notion of key, as in relational databases, as well as factual assertions, e.g. stating that each print of this book can be

either hardcover or softcover but not both. The demand for supporting more expressive semantics and reasoning is addressed in languages like DAML, OIL and the W3C recommendation OWL described below. More information on RDF(S) can be found on the W3C website (`www.w3.org`) and many books. As with XML, the standardisation provided by RDF(S) simplifies development and application of Web Mining.

### 2.3 OWL

Like RDF and RDF Schema, OWL is a W3C recommendation, intended to support more elaborate semantics. OWL includes elements from description logics and provides many constructs for the specification of semantics, including conjunction and disjunction, existentially and universally quantified variables and property inversion. Using these constructs, a reasoning module can make logical inferences and derive knowledge that was previously only implicit in the data. Using OWL for the Semantic Web implies that an application could invoke such a reasoning module and acquire inferred knowledge rather than simply retrieve data.

However, the expressiveness of OWL comes at a high cost. First, OWL contains constructs that make it undecidable. Second, reasoning is not efficient. Third, the expressiveness is achieved by increased complexity, so that ease-of-use and intuitiveness are no more given. These observations lead to two variations of OWL, *OWL DL* (stands for OWL Description Logic) and *OWL Lite*, which disallow the constructs that make the original *OWL Full* undecidable and at the same time aim for more efficient reasoning and for higher ease-of-use. To this end, OWL DL is more expressive than OWL Lite, while OWL Lite is even more restricted but easier to understand and to implement.

In terms of standardisation, it should be recalled that RDF and RDF Schema use XML as their syntax. OWL Full is upward compatible with RDF. This desirable aspect does not hold for OWL DL and OWL Lite. A legal OWL DL document is also a legal RDF document but not vice versa. This implies that reasoning and the targeted knowledge extraction are limited to the set of documents supporting OWL DL (resp. OWL Lite), while other documents, even if RDF Schema, cannot be taken into account in the reasoning process. For the transition of the Web to the Semantic Web, this is a more serious caveat than for other environments (e.g. institutional information sources) which need ontological support. More information on OWL can be found on the W3C website and many books.

The development of OWL and its application is still in an early stage. If it leads to the availability of large knowledge bases via the internet, this will increase the relevance of knowledge-intensive Data Mining methods, that combine data with prior (OWL) knowledge.

### 2.4 Ontologies

Beside the formal languages to be used for the Semantic Web there is the ambition to develop ontologies for general use. There are in practice two types of ontologies. The first type uses a small number of relations between concepts, usually the subclass relation and sometimes the part-of relation. Popular and commonly used are ontologies of Web documents, such as DMoz or Yahoo!, where the documents are hierarchically organized

based on the content. For each content topic, there is an ontology node, with more general topics placed higher in the hierarchy. For instance, one of the top level topics in DMoz is "Computers" that has as one of the subtopics "Data Formats. Under it, there is a subtopic "Markup Languages" that has "XML" as one of its subtopics. There are several hundred documents assigned to the node on "XML" or some of its subnodes. [6] Each Web document is very briefly described and this description together with the hyperlink to the document is placed into one or more ontology nodes. For instance, one item in the "XML" node is a hyperlink to W3C page on XML, http://www.w3.org/XML/, with the associated brief description: "Extensible Markup Language (XML) - Main page for World Wide Web Consortium (W3C) XML activity and information". We can say that here each concept (topic in this case) in the ontology is described by a set of Web documents and their corresponding short descriptions with hyperlinks. The only kind of relations that appear in such ontologies are implicit relations between more specific topic, that is a "subtopic of" a more general topic while the more general topic is a "supertopic of" a more specific topic.

The other kind of ontologies are rich with relations but have a rather limited description of concepts consisting usually of a few words. A well known example of a general, manually constructed ontology is the semantic network WordNet [10] with 26 different relations (e.g., hypernym, synonym). For instance, concepts such as "bird" and "animal" are connected with the relation "is a kind of", concepts "bird" and "wing" are connected with the relation "has part".

## 3 Data Mining

Before considering what the Semantic Web means with respect to Data Mining, we briefly review the main tasks that are studied in Data Mining. Data Mining methods construct models of data. These models can be used for prediction or explanation of observations or for adaptive behaviour. Reviews of the main methods can be found in textbooks such as [11–13]. The main tasks are classification, rule discovery, event prediction and clustering.

### 3.1 Classification

Classification methods construct models that assign a class to a new object on the basis of its description. A wide range of models can be constructed. In this context an important property of classification methods is the form in which objects are given to the data miner and the form of the models. Most learning methods take as input object descriptions in the form of attribute-value pairs where the scales of the variables are nominal or numerical. One class of methods, relational learning or Inductive Logic Programming, see for example [14], takes input in the form of relational structures that describe multiple objects with relations between them creating general models over structures.

Classification methods vary in the type of model that they construct. Decision tree learners construct models basically in the form of rules. A condition in a rule is a constraint on the value of a variable. Usually constraints have the form of identity (e.g.

[6] See http://dmoz.org/Computers/Data_Formats/Markup_Languages/XML/

colour = red) or an interval on a scale (age $> 50$). The consequent of a rule is a class. Decision trees have a variable at each node and a partitioning of the values of this variable. Each part of the values is associated with a subtree and the leaves of the tree are classes. Details on decision tree learning can be found in [11]

Bayesian methods construct models that estimate "a posteriori" probablity distributions for possible classes of an object using a form of Bayes Law. Popular models and methods are Naïve Bayes (assuming conditional independence of features within classes) and Bayesian Belief Networks. More details can be found in e.g. [11].

A third popular type of method is support vector machines. This is a method for minimising prediction errors within a particular class of models (the "kernel"). The method maximises the classification "margin": the distance between the data points of different classes that are closest to the boundary between the classes. The data points that are on the "right side" of the boundary and that are closest to the boundary are called "support vectors". SVM does not literally construct the boundary line (or in more dimensions, the hyperplane) that separates the classes but this line can be reconstructed from the support vectors. Implementations and more information are available from the Web, for example at http://www.kernel-machines.org/.

### 3.2 Rule Discovery

The paradigm of association rules discovery was first established through the work of Rakesh Agrawal and his research group, starting with [15, 16]. Association rules are based on the notion of "frequent itemset", i.e. a set of items occuring together in more data records than an externally specified frequency threshold. From a frequent itemset, association rules can be derived by positioning some of the items in the antecedent and the remaining ones in the consequent, thereby computing the confidence with which the former imply the latter. A popular algorithm is the Apriori algorithm [17].

One of the most popular applications for association rules discovery is market basket analysis, for which sets of products frequently purchased together are identified. In that context, the association rules indicate which products are likely to give rise to the purchase of other products, thus delivering the basis for cross-selling and up-selling activities. In the context of Web Mining, association rule discovery focusses on the identification of pages that are frequently accessed together but also on the discovery of frequent sets of application objects (such as products, tourist locations visited together, course materials etc).

### 3.3 Clustering

Clustering methods divide a set of objects into subsets, called *clusters*, such that the objects in any cluster are similar to those inside it and different from those outside it. Some methods are hierarchical and construct clusters of clusters. The objects are described with numerical or nominal variables. Clustering methods vary in the measure for similarity (within and between clusters), the use of thresholds in constructing clusters, whether they allow objects to belong to strictly to one cluster or can belong to

more clusters in different degrees and the structure of the algorithm. The resulting cluster structure is used as a result in itself, for inspection by a user, or to support retrieval of objects.

### 3.4 Sequence Discovery, Sequence Classification, and Event Prediction

In many applications, the data records do not describe sets of items but sequences of events. This is the case both for the navigation through a Web site and for carrying and filling a market basket inside a store. There are several Data Mining tasks involved here.

Sequence mining (or sequence discovery) extends the paradigm of association rules discovery towards the discovery of frequent sequences of events. Unlike conventional association rules, events are ordered. Hence, a rule emanating from a frequent sequence expresses the likelihood with which the last event will occur after the sequence of events in the antecedent. Sequence mining adds several new aspects to the original paradigm, such as the adjacency of events, the distance between the frequent events being observed and the sequentialisation of events recorded with different clocks. Methods for sequence mining are derived from rule discovery methods or from probabilistic models (Hidden Markov Models). A survey of sequence mining research is incorporated in the literature overview of [18].

Rules derived from frequent sequences can be used to predict events from a given series of observations. Data Mining can be used to find classifiers for frequent sequences. For example, certain types of sequences correspond to manufacturing errors or network intrusions. A classifier can be learned to predict which event will occur (immediately) after the sequence, enabling predictions. An example is page pre-fetching in file servers or Web servers e.g. [19, 20]. Just as discovered rules are not optimal for classification, rules derived from frequent sequences are not automatically appropriate for event prediction. For example, the frequent sequences “A-B-C-D” and “A-B-C-E” indicate that events D and E are likely to occur after A, B and C in that order and allow for a quantification of this likelihood. However, if the goal is to find events, whose appearance leads to E, the sequence “A-B-C” is not necessarily a good predictor, because this sequence may as well be followed by the event D. Nonetheless, solutions based on sequence mining have been devised to assist in prediction of given events as in the early works of [21, 22].

## 4 Data Mining and the World Wide Web

In this section we review how the Data Mining methods summarised in section 3 are used to construct models of objects and events on the World Wide Web and adaptive Web-based systems. Web Mining can involve the structured data that are used in standard Data Mining but it derives its own character from the use of data that are available on the Web. It should be kept in mind that data may be privacy-sensitive and that purpose limitations may preclude an analysis, in particular one that combines data from different sources gathered for different purposes.

### 4.1 Data on the Web

Learning methods construct models from samples of structured data. Most methods are defined for samples of which each element, each case, is defined as values of a fixed set of features. If we take documents as cases then a feature-value representation of a document must be constructed. A standard approache is to take words as features and occurrence of a word in a document as value. This requires a tokenising step (in which the series of basic symbols is divided into tokens - words, numbers, delimiters). Many documents are annotated with formatting information (often in HTML) that can be kept or removed. Refinements are the use of word stems, dropping frequent but uninformative words, merging synonyms, including combinations of words (pairs, triples, or more). In addition, other features of documents can be used (for example, length, occurrence of numbers, number of images).

The hyperlinks induce a graph structure on the set of Web pages. This structure has been used to identify central pages. The Page Rank algorithm [23] (implemented in the Google search engine) ranks for instance all Web pages based on the number of links from other important pages. When Google is answering a query, then it presents basically the answers to the query according to this order. The Hub & Authorities approach [24] follows a similar scheme, but differentiates between two types of pages. An authority is a page which is pointed to from many important hubs, and hubs are pages pointing to many important authorities.

Which data are recorded in user logs is an issue that has no general answer. . At the lowest level clicks on menu items and keystrokes can be recorded. At a higher level, commands, queries, entered text, drawings can be logged. The level of granularity and selection that is useful depends on the application and the nature of the interaction. The same is true of the context in which the user action is observed. The context can be the entire screen, a menu or other. The context can include textual documents. The content and usage of the Web can be viewed as single units but also as structures. The content consists of pieces that are connected to other pieces in several ways: by hyperlinks (possibly with labels), addresses, textual references, shared topics or shared users. Similarly, users are related by hyperlinks, electronic or postal addresses, shared documents, pages or sites. These relational data can be subject of Web Mining, modelling structural patterns, in combination with data about the components.

Web usage mining is characterised by the need of an extensive data preparation. Web server data are often incomplete, in the sense that important information is missing, including a unique association between a user and her activities and a complete record of her activities, in which also the order of retrieving locally cached objects is contained. Techniques to this end are either proactive, i.e. embedding to the Web server functionalities that ensure the recording of all essential data, or reactive, i.e. trying to reconstruct the missing data a posteriori. An overview of techniques for Web data preparation can be found in [25]. An evaluation of their performance is reported in [26].

### 4.2 Document classification

Classifying documents is one of the basic tasks in Web Mining. Given a collection of classified documents (or parts of documents), the task is to construct a classifier that

can classify new documents. Methods that are used for this task include the methods described in section 3.1. These methods are adapted to data on the Web, in particular textual documents. Features that are used to represent textual data mainly capture occurrences of words (or word stems) and in some cases also occurrence of word sequences. In the basic approach, all the words that occur in a whole set of documents are included in the feature set (usually several thousands of words). The representation of a particular document contains many zeros, as most of the words from the collection do not occur in particular document. To handle this, special methods are used (support vector machines) or relevant features are selected. The set of features is sometimes extended with, for example, text length, and features defined in terms of HTML tags (e.g. title, author). Document classification techniques have been developed and applied on different datasets including descriptions of US patents [27], Web documents [27, 28], and Reuters news articles [29]. An overview can be found in [30].

A related form is classification of structures of documents instead of single documents. Information on the Web is often distributed over several linked pages that need to be classified as a whole. Relational classification methods are appropriate for this.

Applications of document classification are adaptive spam filters where email messages are labelled as spam or not and the spam filter learns to recognise spam messages (e.g. [31]), adaptive automated email routing, where messages are labelled by the person or department that needs to deal with them to enable automated routing, identifying relevant Web pages or newspaper articles, and assigning documents to categories for indexing and retrieval.

### 4.3 Document clustering

Document clustering [32] means that large numbers of documents are divided into groups that are similar in content. This is usually an intermediate process for optimising search or retrieval of documents. The clusters of documents are characterised by documents features (single keywords or word combinations) and these are exploited to speed up retrieval or to perform keyword based search. Document clustering is based on any general data clustering algorithm adopted for text data by representing each document by a set of features in the same way as for document classification. The similarity of two documents is commonly measured by the cosine-similarity between the word-vector representation of the documents. Document clustering is used for collections that are at a single physical location (for example, the US national library of medicine) but also for search engines that give access to open collections over the internet. Document clustering is combined with document classification to enable maintenance: new documents are added to clusters by classifying them as cluster members. An example of an application of document clustering is [33].

### 4.4 Data Mining for Information Extraction

Information extraction means recognition of information in documents. Usually information extraction is combined with document classification: a document is recognised as a document that contains certain information and we need to find out where it is. A pattern is matched with the text in the document. If the pattern matches a fragment it

indicates where the target information is. Such patterns can be constructed manually or learned inductively from documents in which the information is labelled. The patterns can be strings of symbols but can also include features: linguistic features (e.g. part of speech, capital letters) or semantic features (e.g. person name, number greater than 1000).

Although this can be viewed as a form of classification (classify all word sequences of length up to N as being the sought information or not) the classification models above are not directly applicable. Documents lack the structure of objects for which learning methods were originally developed: they may vary in length and it is not obvious what should be the features of an instance. On the other hand, documents on the Web are often encoded in HTML which imposes some structure that flat texts do not have.

Standard Data Mining methods must be adapted to the less structured setting of information extraction and are combined with ideas from grammar induction. Wrapper induction is the inductive construction of a wrapper, a system that mediates between what we might call a client and a server. It translates requests from the client into calls to the server. If we are interested in information from a Web page, the wrapper translates an information request into the format of the Web page, extracts the information from the page and sends it to the client. The wrapper exploits the structure of the Web page. If the Web page is structured as HTML or XML the wrapper can exploit this, otherwise it has to use patterns in the language, or a combination of the two.

Wrappers and extraction patterns in general can be learned from examples in which the relevant information is marked. The learner compares patterns around relevant information with general structure in the text and inductively constructs the wrapper. Examples of systems that perform this task are RAPIER [34] and BWI [35].

An example of an application of information extraction is wrapper maintenance [36]. A wrapper types components of an object or procedure from an external perspective to interface it with other systems or users. Information extraction patterns can be viewed as wrappers for documents or Web pages. The layout of documents or certain pages changes regularly and then wrappers must be revised. This can be done by comparing pages with the old and the new layout, identifying the components in the new layout and using this to revise the wrappers.

The use of Data Mining methods for learning extraction patterns and wrappers is currently a very active area of research. For an overview see [37].

### 4.5 Usage Mining

Web usage mining is a Web Mining paradigm in which Data Mining techniques are applied to Web usage data. As for Data Mining in general, the goal can be to construct a model of users' behaviour or to directly construct an adaptive system. A potential advantage of an explicit user model is that it can be used for different purposes where an adaptive system has a specific function. For modelling user behaviour, usage mining is combined with other information about users. Many aspects of users can be modelled: their interaction with a system, their interests, their knowledge, their geographical behaviour and also of course combinations of these. Modelling preferences needs information about users preferences for individual objects. This is often problematic because

users are not always prepared to evaluate objects and enter the evaluations. Therefore other data are used like downloading, buying or time data.

Adaptive systems have the purpose to improve some aspect of the behaviour of the system. Improvements can be *system-oriented*, *content-oriented* (e.g. presenting information or products that relevant for the user), or *business-oriented* (e.g. presenting advertisements that the user is likely to buy, or that the vendor prefers to sell). Another dimension is wether the model or adaptation concerns individual users *personalisation* or generic system behaviour. An intermediate form is to use usage information obtained during a session to adapt system behaviour. This can also be combined with personalisation.

System-oriented adaptation based on usage mining is aimed at performance optimisation, e.g. for Web servers. This is of paramount importance in large sites that incur a lot of traffic. One of the factors leading to performance degradation is access to slow peripherals like disks, from which pages are pre-fetched upon user demand. Hence, it is of interest to devise intelligent pre-fetching mechanisms that allow for efficient caching. The problem specification reduces to a next-event prediction, where the next event is a page fetch request, combined with an appropriate mechanism for refreshing the cash, e.g. least frequently used or most frequently used. As described in the subsection 3.4 on event prediction above, the methods of choice here are Hidden Markov Models and, occasionally, sequence mining.

**Personalisation.** Although the terminology is not always used consistently, "personalisation" usually denotes adaptation to individual users that can be identified by the system (via a login step). Most systems have simple tools that a user can apply to adapt the interface to his preferences. For example, a user can store his favourite links to web pages. These are then later easily available when the user logs on to the system. Personalisation takes this further in two ways: (1) it includes aspects of systems that are less easy to specify with a few features and (2) the system automatically infers the preferences of the user and makes adjustments. Personalisation can be system-oriented, business-oriented or content-oriented and it can include a variety of data about an individual user.

Personalisation can be used for a variety of user tasks. A well-known example is shopping. The buying record in an electronic shop is used to infer user preferences of a user and direct advertising. Other applications center around information search. Examples are personalised newspapers (that include only material that is considered of interest to the user), personalised Web sites, active information gathering from the web, highlighting potentially interesting hyperlinks on a requested Web pages [28], query-expansion (by adding user-specific keywords to a query for a search engine), selection of TV programmes. Personalisation and recommending can be based on usage data, on documents that are associated with an individual user or a combination of these two.

**System adaptation.** Systems can also be adapted to a user community, rather than a single user, optimising average instead of user specific peformance.

For example, Etzioni et al. propose a clustering algorithm for correlated but not linked Web pages, allowing for overlapping clusters [38]; Alvarez et al. extend as-

sociation rules discovery to cope with the demands of online recommendations [39]; Mobasher et al propose two methods for modelling user sessions and corresponding distance functions, to cluster sessions and derive user and usage profiles [40]. Finally, dedicated Web usage mining algorithms are also proposed, focussing mainly on the discovery of Web usage patterns, as in [41–43].

An important subject in Web usage mining concerns the evaluation of adaptive systems based on Web Mining. Methods for the evaluation of Web sites with respect to user friendliness, interactivity and similar user-oriented aspects have been devised early, building upon the research on hypermedia and upon cognitive sciences [44]. The business-oriented perspective leads to other evaluation criteria derived from marketing.

### 4.6 Modelling networks of users

Users do not act in isolation - they are part of various social networks, often defined by common interests and by phenomena such as opinion leadership and, more generally, different degrees of influence on one another. An understanding of a user's surrounding social network(s) improves the understanding of that user and can therefore contribute to reaching various Web mining goals. For example, Domingos and Richardson [45] use a collaborative filtering database to understand the differential influence users have and propose to use this knowledge for "viral marketing": to preferentially target customers whose purchasing and recommendation behaviour are likely to have a strong influence on others. Other data sources include email logs [46] and publicly-available online information [47]. This research combines aspects of Web content mining and Web structure mining. The latter view closes a circle: link mining has its origins in social network analysis and is now being applied to analyze the social networks forming on the Web.

## 5 Data Mining for and with the Semantic Web

The standardized data format, the popularity of content-annotated documents and the ambition of large scale formalization of knowledge of the Semantic Web has two consequences for Web Mining. The first is that more structured information becomes available to which existing Data Mining methods can be applied with only minor modifications. The second is the possibility of using formalized knowledge (in the form of concept hierarchies in RDF but even more in the form of knowledge represented in OWL) in combination with Web data for Data Mining. The combination of these two gives a form of closed-loop learning in which knowledge is acquired by Web Mining and then used again for further learning. We briefly summarise the implications of this for the main Web Mining tasks.

### 5.1 Document classification for and with the Semantic Web

Document classification methods for the Semantic Web are like those for the World-Wide Web. Besides general features of documents, annotations can be used, as additional features or to structure features. Knowledge in the form of ontologies can be

used to infer additional information about documents, potentially providing a better basis for classification. This form of document classification uses classification learning with background knowledge and feature construction [48]. Document classes can be added to the annotation of documents. Classification be applied to predefined segments of documents.

Issues for current and future research are the use of non-textual data such as images. Images can be tagged more or less like textual documents (see [49]) giving rise to the same learning tasks and opening the opportunity for learning about combinations of text and images. Future issues are video, voice and sound. From the current state of the art it is likely that this will be possible in the next five years, enabling a wide range of new applications such as multimedia communication.

### 5.2 Document Clustering for and with the Semantic Web

Like document classification, clustering of annotated documents can exploit the annotations and it can infer extra information about documents from ontologies. An example is [50], where texts are preprocessed by adding semantic categories derived from Wordnet. Evaluation on Reuters newsfeeds shows an improvement of the results by using background knowledge.

Hierarchical document clusters and the descriptions of these clusters can be viewed as ontologies based on subconcept relations. In this sense hierachical clustering methods construct ontologies of documents and then maintain these ontologies [51, 7, 29] by classifying new documents in the hierachy. Characterising clusters supports the construction of ontologies because the description of a cluster reflects relations between concepts, see [52–54].

### 5.3 Data Mining for Information Extraction with the Semantic Web

Learning to extract information from documents can exploit annotations of document segments for learning extraction rules - assuming these have been assigned consistently - and it can benefit from knowledge in ontologies. The other way round, existing ontologies can support solving different problems including learning of other ontologies and assigning ontology concepts to text (text annotation). Ontology concepts are assigned either to whole documents, as in the case of already described ontologies of Web documents or to some smaller parts of text. In the letter case, researchers have been working on learning annotation rules from already annotated text. This can bee seen as a kind of information extraction, where the goal is not to fill in the database slots by extracted information (see Section 4.4) but to assign a label (slot name) to a part of text (see [55]). As it is non-trivial to obtain already annotated text, some researchers investigate other techniques, such as natural language processing or clustering to find text units (eg., groups of nouns, clusters of sentences) and map them upon concepts of the existing ontology [56].

### 5.4 Ontology mapping

Because ontologies are often developed for a specific purpose it is inevitable that similar ontologies are constructed and unifying these ontologies needs to be done to enable the

use of knowledge from one ontology in combination with knowledge in the other. This requires the construction of a mapping between the concepts, attributes, values and relations in the two ontologies, either as a solution or as a step towards a single unified ontology. Several approaches to this problems are explored by several researchers [57–62]. One line of attack is to first take information about concepts from the ontologies and then extract additional information for example from Web pages recognised as relevant for each concept. This information can then be used to learn a classifier for instances of a class. Applying this classifier to instances of concepts in the other ontology makes it possible to see which other concept (or combination of concepts) has most in common with the original concept.

### 5.5 User Modelling, Recommending, Personalisation and the Semantic Web

The Semantic Web opens up interesting opportunities for usage mining because ontologies and annotations can provide information about user actions and Web pages in a standardised form that enables discovery of richer and more informative patterns. Examples for recommending are the work by Mobasher and by Ghani (both this volume). The annotations of products that are visited (and bought) by users add information to customer segments and make it possible to discover the underlying general patterns. Such patterns can be used, for example, to predict reactions to new products from the description of the new product. This would not have been possible if only the name, image and price of the product had been available and mining can be done much more effectively using a uniform ontology than from documents that describe products.

Applications of usage mining such as usage-based recommending, personalisation and link analysis will benefit from the use of annotated documents and objects. Only a few technical problems need to be solved to extend existing methods this. Large scale applications need larger ontologies that can be maintained and applied semi-automatically. Designing or automatically generating ontologies for decribing user interests and user behaviour are more challenging problems that need to be addressed in this context.

### 5.6 Learning about services

The construction and design of ontologies for functions of Web services is an area that is currently topic of active research. As for descriptive concepts, a Web Mining approach can be applied to this problem. Requests to a service and the reaction of the server can be collected and learning methods can be applied to, at least for simple cases, reconstruct the function of the service. An illustration of this approach is shown in [63]. Advances to practical applications of this approach that are complex enough to make this approach competitive to manual construction of the service description are still beyond the state of the art and have to wait for suitable ontologies that can be used as background knowledge by the learner.

### 5.7 Infrastructure

In the sections above we reviewed research on Data Mining methods for the Semantic Web. Techniques and representations developed for the Semantic Web are not only ap-

plied as methods for which systems are developed. Notions from the Semantic Web are introduced in operating systems for single and distributed systems. These developments would facilitate the use and development of Web Mining systems and the unification imposed by system level standards will make it easier to exploit distributed ontologies and services. In this section we focus on the innovations in the infrastructure for systems that are based on such methods.

Current applications of Semantic Web ideas suffer partially from a lack of speed. The bigger problem is the lack of a large number of ontologies and annotations. Although access to ontologies and data via the internet is possible, existing applications strongly rely on local computing. Ontologies, instances, logfiles are imported and kept locally to achieve enough speed. This will clearly meet its limits when the Semantic Web will be used at a large scale. Bringing Semantic Web ideas into the lower level of the internet may allow distributed computing with distributed ontologies, instances and knowledge. This brings together the Semantic Web and Grid Computing and is pursued under the name of Semantic Grid.

Another development that will have a great influence in the Semantic Web is that the successor to the Windows operating system, the Longhorn operating system uses a version of XML to integrate the datamodel and applications. This is likely to make the Semantic Web languages known outside the current communities and also it will provide widely available support for Semantic Web tools. This in turn will create enormous opportunities for Web Mining methods both at the level of information and knowledge and at the level of systems.

## 6 Prospects

The future of Web Mining will to a large extent depend on developments of the Semantic Web. The role of Web technology still increases in industry, government, education, entertainment. This means that the range of data to which Web Mining can be applied also increases. Even without technical advances, the role of Web Mining technology will become larger and more central. The main technical advances will be in increasing the types of data to which Web Mining can be applied. In particular Web Mining for text, images and video/audio streams will increase the scope of current methods. These are all active research topics in Data Mining and Machine Learning and the results of this can be exploited for Web Mining.

The second type of technical advance comes from the integration of Web Mining with other technologies in application contexts. Examples are information retrieval, e-commerce, business process modelling, instruction, and health care. The widespread use of web-based systems in these areas makes them amenable to Web Mining.

In this section we outline current generic practical problems that will be addressed, technology required for these solutions, and research issues that need to be addressed for technical progress.

**Knowledge Management** Knowledge Management is generally viewed as a field of great industrial importance. Systematic management of the knowledge that is available in an organisation can increase the ability of the organisation to make optimal

use of the knowledge that is available in the organisation and to react effectively to new developments, threats and opportunities. Web Mining technology creates the opportunity to integrate knowledge management more tightly with business processes. Standardisation efforts that use Semantic Web technology and the availability of ever more data about business processes on the internet creates opportunities for Web Mining technology.

More widespread use of Web Mining for Knowledge Management requires the availability of low-threshold Web Mining tools that can be used by non-experts and that can flexibly be integrated in a wide variety of tools and systems.

**E-commerce** The increased use of XML/RDF to describe products, services and business processes increases the scope and power of Data Mining methods in e-commerce. Another direction is the use of text mining methods for modelling technical, social and commercial developments. This requires advances in text mining and information extraction.

**E-learning** The Semantic Web provides a way of organising teaching material, and usage mining can be applied to suggest teaching materials to a learner. This opens opportunities for Web Mining. For example, a recommending approach (as in [64]) can be followed to find courses or teaching material for a learner. The material can then be organized with clustering techniques, and ultimately be shared on the web again, e. g., within a peer to peer network [65]. Web mining methods can be used to construct a profile of user skills, competence or knowledge and of the effect of instruction. Another possibility is to use web mining to analyse student interactions for teaching purposes. The internet supports students who collaborate during learning. Web mining methods can be used to monitor this process, without requiring the teacher to follow the interactions in detail. Current web mining technology already provides a good basis for this. Research and development must be directed toward important characteristics of interactions and to integration in the instructional process.

**E-government** Many activities in governments involve large collections of documents. Think of regulations, letters, announcements, reports. Managing access and availability of this amount of textual information can be greatly facilitated by a combination of Semantic Web standardisation and text mining tools. Many internal processes in government involve documents, both textual and structured. Web mining creates the opportunity to analyse these governmental processes and to create models of the processes and the information involved. It seems likely that standard ontologies will be used in governmental organisations and the standardisation that this produces will make Web Mining more widely applicable and more powerful than it currently is. The issues involved are those of Knowledge Management. Also governmental activities that involve the general public include many opportunities for Web Mining. Like shops, governments that offer services via the internet can analyse their customers behaviour to improve their services. Information about social processes can be observed and monitored using Web Mining, in the style of marketing analyses. Examples of this are the analysis of research proposals for the European Commission and the development of tools for monitoring and structuring internet discussion fora on political issues (e.g. the E-presentation project at Fraun-

hofer Institute [66]). Enabling technologies for this are more advanced information extraction methods and tools.

**Health care** Medicine is one of the Web's fastest-growing areas. It profits from Semantic Web technology in a number of ways: First, as a means of organizing medical knowledge - for example, the widely-used taxonomy International Classification of Diseases and its variants serve to organize telemedicine portal content (e.g., http://www.dermis.net) and interfaces (e.g., http://healthcybermap.semanticweb.org). The Unified Medical Language System (http://www.nlm.nih.gov/research/umls) integrates this classification and many others. Second, health care institutions can profit from interoperability between the different clinical information systems and semantic representations of member institutions' organization and services (cf. the Health Level 7 standard developed by the International Healthcare XML Standards Consortium: http://www.hl7.org ). Usage analyses of medical sites can be employed for purposes such as Web site evaluation and the inference of design guidelines for international audiences [67, 68], or the detection of epidemics [69]. In general, similar issues arise, and the same methods can be used for analysis and design as in other content classes of Web sites. Some of the facets of Semantic Web Mining that we have mentioned in this article form specific challenges, in particular: the privacy and security of patient data, the semantics of visual material (cf. the Digital Imaging and Communications in Medicine standard: http://medical.nema.org), and the cost-induced pressure towards national and international integration of Web resources.

**E-science** In E-Science two main developments are visible. One is the use of text mining and Data Mining for information extraction to extract information from large collections of textual documents. Much information is "buried" in the huge scientific literature and can be extracted by combining knowledge about the domain and information extraction. Enabling technology for this is information extraction in combination with knowledge representation and ontologies. The other development is large scale data collection and data analysis. This also requires common concept and organisation of the information using ontologies. However, this form of collaboration also needs a common methodology and it needs to be extended with other means of communication, see [70] for examples and discussion.

**Webmining for images and video and audio streams** So far, efforts in Semantic Web research have addressed mostly written documents. Recently this is broadened to include sound/voice and images. Images and parts of images are annotated with terms from ontologies.

**Privacy and security** A factor that limits the application of Web Mining is the need to protect privacy of users. Web Mining uses data that are available on the web anyway but the use of Data Mining makes it possible to induce general patterns that can be applied to personal data to inductively infer data that should remain private. Recent research addresses this problem and searches for selective restrictions on access to data that do allow the induction of general patterns but at the same time preserves a preset uncertainty about individuals, thereby protecting privacy of individuals, e.g., [71, 72].

**Information extraction with formalised knowledge** In section 5.3 we briefly reviewed the use of concept hierarchies and thesauri for information extraction. If knowledge

is represented in more general formal Semantic Web languages like OWL, in principle there are stronger possibilities to use this knowledge for information extraction.

In summary, the main foreseen developments are:

- *The extensive use of annotated documents facilitates the application of Data Mining techniques to documents.*
- *The use of a standardised format and a standardised vocabulary for information on the web will increase the effect and use of Web Mining.*
- *The Semantic Web goal of large-scale construction of ontologies will require the use of Data Mining methods, in particular to extract knowledge from text.*

At the moment of writing the main issues to address are:

- *Methods for images and sound: an increasing part of the information on the web is not in textual form and methods for classification, clustering, rule and sequence learning and information extraction are needed, and thus require a combination with methods for text and structured data.*
- *Knowledge-intensive learning methods for information extraction from texts. Building powerful information extraction knowledge is likely to be a necessary condition to enable the Semantic Web.*

## References


1. Michalski, R., Bratko, I., (eds), M.K.: Machine Learning and Data Mining: methods and applications. John Wiley and Sons, Chichester (1998)
2. Paliouras, G., Karkaletsis, V., (eds), C.S.: Machine Learning and its Applications. Springer-Verlag, Heidelberg (2001)
3. Franke, J., Nakhaeizadeh, G., Renz, I., eds.: Text Mining, Theoretical Aspects and Applications. Physica-Verlag (2003)
4. Berners-Lee, T., Fischetti, M.: Weaving the Web. Harper, San Francisco (1999)
5. Berendt, B., Stumme, G., Hotho, A.: Usage mining for and on the semantic web. In: Data Mining: Next Generation Challenges and Future Directions. AAAI/MIT Press, Menlo Park, CA (2004) 467–486
6. Berendt, B., Hotho, A., Stumme, G.: Towards semantic web mining. In: [73]. (2002) 264–278
7. Mladenić, D., Grobelnik, M.: Feature selection on hierarchy of web documents. Journal of Decission support systems **35** (2003) 45–87
8. Erdmann, M.: Ontologien zur konzeptuellen Modellierung der Semantik von XML. Isbn: 3831126356, University of Karlsruhe (2001)
9. W3C: RDF/XML Syntax Specification (Revised). W3C recommendation, http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/ (2004)
10. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)
11. Mitchell, T.: Machine Learning. McGraw Hill (1997)
12. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. MIT Press (2001)
13. Weiss, M., Indurkhya, N.: Pedictive Data-Mining: A Practical Guide. organ Kaufmann, San Francisco (1997)
14. Lavrac, N., Dzeroski, S.: Inductive Logic Programming: Techniques and Applications. Ellis Horwood, New York (1994)

15. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD'93, Washington D.C., USA (1993) 207–216
16. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In Bocca, J.B., Jarke, M., Zaniolo, C., eds.: Proc. 20th Int. Conf. Very Large Data Bases, VLDB, Morgan Kaufmann (1994) 487–499
17. Adamo, J.M.: Data Mining and Association Rules for Sequential Patterns: Sequential and Parallel Algorithms. Springer, New York (2001)
18. Roddick, J., Spiliopoulou, M.: A survey of temporal knowledge discovery paradigms and methods. IEEE Trans. of Knowledge and Data Engineering (2002)
19. Lan, B., Bressan, S., Ooi, B.: Making web servers pushier. In: Proceedings WEBKDD-99. Springer Verlag, Berlin (2000) 108–122
20. Scheffer, T., Wrobel, S.: A sequential sampling algorithm for a general class of utility criteria. In: Knowledge Discovery and Data Mining. (2000) 330–334
21. Zaki, M., Lesh, N., Ogihara, M.: Mining features for sequence classification. In: Proc. of 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining KDD'99, ACM (1999) 342–346
22. Weiss, G.M., Hirsh, H.: Learning to predict rare events in event sequences. In Agrawal, R., Stolorz, P., Piatesky-Shapiro, G., eds.: Proc. of 4th Int. Conf. KDD, New York, NY (1998) 359–363
23. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine source. In: Proceedings of the seventh international conference on World Wide Web, Elsevier Science Publishers (1998)
24. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46** (1999) 604–632
25. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Journal of Knowledge and Information Systems **1** (1999) 5–32
26. Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M.: A framework for the evaluation of session reconstruction heuristics in web usage analysis. INFORMS Journal on Computing, Special Issue on "Mining Web-based Data for E-Business Applications" (eds. Rashid, Louiqa and Tuzhilin, Alex) (2003)
27. McCallum, A., Rosenfeld, R., Mitchell, T., Ng, A.: Improving text classification by shrinkage in a hierarchy of classes. In: Proceedings of the 15th International Conference on Machine Learning (ICML-98), Morgan Kaufmann, San Francisco, CA (1998)
28. Mladenic, D.: Web browsing using machine learning on text data. In Szczepaniak, P., ed.: Intelligent exploration of the web, 111,Physica-Verlag (2002) 288–303
29. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. In: Proceedings of the 14th International Conference on Machine Learning ICML97. (1997) 170–178
30. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34** (2002) 1–47
31. Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., Spyropoulos, C.: An evaluation of naive bayesian anti-spam filtering. In Potamias, G., Moustakis, V., van Someren, M., eds.: Proceedings of the workshop on Machine Learning in the New Information Age. (2000)
32. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In Grobelnik, M., Mladenic, D., Milic-Frayling, N., eds.: Proceedings of the KDD Workshop on Text Mining. (2000)
33. Zamir, O., Etzioni, O.: Web document clustering: A feasibility demonstration. In: Research and Development in Information Retrieval. (1998) 46–54
34. Califf, M.E., Mooney, R.J.: Bottom-up relational learning of pattern matching rules for information extraction. Journal of Machine Learning Research **4** (2003) 177–210

35. Freitag, D., Kushmerick, N.: Boosted wrapper induction. In: Proceedings AAAI-00. (2000) 577–583
36. X. Meng, D. Hu, C.L.: Schema-guided wrapper maintenance for web-data extraction. In: ACM Fifth International Workshop on Web Information and Data Management (WIDM 2003). (2003)
37. Kushmerick, N., Thomas, B.: Adaptive information extraction: Core technologies for information agents. In: Intelligent Information Agents R&D in Europe: An AgentLink perspective. Springer, Berlin (2004) 79–103
38. Perkowitz, M., Etzioni, O.: Adaptive web sites: Automatically synthesizing web page. In: Proc. of AAAI/IAAI'98. (1998) 727–732
39. Lin, W., Alvarez, S., Ruiz, C.: Efficient adaptive-support association rule mining for recommender systems. Data Mining and Knowledge Discovery **6** (2002) 83–105
40. Mobasher, B., Dai, H., Luo, T., Nakagawa, M.: Discovery and evaluation of aggregate usage profiles for web personalization. Data Mining and Knowledge Discovery **6** (2002) 61–82
41. Baumgarten, M., Büchner, A.G., Anand, S.S., Mulvenna, M.D., Hughes, J.G.: Navigation pattern discovery from internet data. In: Proceedings volume [74]. (2000) 70–87
42. Borges, J.L., Levene, M.: Data mining of user navigation patterns. In Spiliopoulou, M., Masand, B., eds.: Advances in Web Usage Analysis and User Profiling. Springer, Berlin (2000) 92–111
43. Spiliopoulou, M.: The laborious way from data mining to web mining. Int. Journal of Comp. Sys., Sci. & Eng., Special Issue on "Semantics of the Web" **14** (1999) 113–126
44. Cutler, M.: E-metrics: Tomorrow's business metrics today. In: KDD'2000, Boston, MA, ACM (2000)
45. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-01, New York, ACM (2001) 57–66
46. Schwartz, M., Wood, D.: Discovering shared interests using graph analysis. Communications of the ACM **36** (1993) 78–89
47. Kautz, H., Selman, B., Shah, M.: Referralweb: Combining social networks and collaborative filtering. Communications of the ACM **40** (1997) 63–66
48. Bloehdorn, S., Hotho, A.: Boosting for text classification with semantic features. In: Proc. of the Mining for and from the Semantic Web Workshop at KDD 2004. (2004)
49. Zaiane, O., Simoff, S.: Mdm/kdd: Multimedia data mining for the second time. SIGKDD Explorations **3** (2003)
50. Hotho, A., Staab, S., Stumme, G.: Wordnet improves text document clustering. In: Procs. of the SIGIR 2003 Semantic Web Workshop, Toronto, Canada (2003)
51. McCallum, A., Rosenfeld, R., Mitchell, T., Ng, A.: Improving text classification by shrinkage in a hierarchy of classes. In: Proceedings of the 15th International Conference on Machine Learning ICML98, Morgan Kaufmann (1998)
52. Hotho, A., Staab, S., Stumme, G.: Explaining text clustering results using semantic structures. In: Proc. of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD. (2003) 217–228
53. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems **16** (2001) 72–79
54. Cimiano, P., Hotho, A., Staab, S.: Comparing conceptual, partitional and agglomerative clustering for learning taxonomies from text. In: Proceedings of the European Conference on Artificial Intelligence (ECAI'04). (2004)
55. Handschuh, S., Staab, S.: Authoring and annotation of web page in CREAM. In: Proc. Of WWW Conference 2002. (2002)
56. Hotho, A., Staab, S., Stumme, G.: Explaining text clustering results using semantic structures. In: Proceedings of ECML/PKDD, Springer Verlag (2003) 217–228

57. Hovy, E.: Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In: Proc. 1st Intl. Conf. on Language Resources and Evaluation (LREC), Granada (1998)
58. Chalupsky, H.: Ontomorph: A translation system for symbolic knowledge. In: Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000). (2000) 471–482
59. McGuinness, D., Fikes, R., Rice, J., Wilder, S.: An environment for merging and testing large ontologies. In: In the Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000), Breckenridge, Colorado, USA (2000) 483–493
60. Noy, N., Musen, M.: Prompt: Algorithm and tool for automated ontology merging and alignment. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Austin, Texas (2000) 450–455
61. Stumme, G., Maedche, A.: Fca-merge: Bottom-up merging of ontologies. In: Proceedings 17th International Conference on Artificial Intelligence (IJCAI-01). (2001) 225–230
62. Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Ontology matching: A machine learning approach. In: Handbook on Ontologies. Springer, Berlin (2004) 385–404
63. Heß, A., Kushmerick, N.: Machine learning forannotating semantic web services. In: Proceedings of the First International Semantic Web Services Symposium. AAAI Spring Symposium Series 2. (2004)
64. Aguado, B., Merceron, A., Voisard, A.: Extracting information from structured exercises. In: Proceedings of the 4th International Conference on Information Technology Based Higher Education and Training ITHET03, Marrakech, Morocco. (2003)
65. Tane, J., Schmitz, C., Stumme, G.: Semantic resource management for the web: An elearning application. In: Proc. 13th International World Wide Web Conference (WWW 2004). (2004)
66. Althoff, K., Becker-Kornstaedt, U., Decker, B., Klotz, A., Leopold, E., Rech, J., Voss, A.: The indigo project: Enhancement of experience management and process learning with moderated discourses. In Perner, P., ed.: Data Mining in Marketing and Medicine. Springer, Berlin (2002)
67. Yihune, G.: Evaluation eines medizinischen Informationssystems im World Wide Web. Nutzungsanalyse am Beispiel www.dermis.net. PhD thesis, Ruprecht-Karls-Universitt Heidelberg (2003)
68. Kralisch, A., Berendt, B.: Cultural determinants of search behaviour on websites. In: Proceedings of the IWIPS 2004 Conference on Culture, Trust, and Design Innovation. (2004)
69. Heino, J., Toivonen, H.: Automated detection of epidemics from the usage logs of a physicians' reference database. In: Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'03), Berlin, Springer (2003) 180–191
70. Mladenic, D., Lavrac, N., Bohanec, M., Moyle, S., eds.: Data Mining and Decision Support: Integration and Collaboration. Kluwer Academic Publishers (2003)
71. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy preserving mining of association rules. In: [75]. (2002) 217–228
72. Iyengar, V.: Transforming data to satisfy privacy constraints. In: [75]. (2002) 279–288
73. Horrocks, I., Hendler, J.A., eds.: The Semantic Web. In Horrocks, I., Hendler, J.A., eds.: Proceedings of the First International Semantic Web Conference, Springer (2002)
74. Masand, B., Spiliopoulou, M., eds.: Advances in Web Usage Mining and User Profiling: Proceedings of the WEBKDD'99 Workshop. LNAI 1836, Springer Verlag (2000)
75. Hand, D., Keim, D., Ng, R., eds.: KDD - 2002 – Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, ACM (2002)