Pattern Based Browsing in Document Collections

Ronen Feldman, Willi Klösgen, Yaniv Ben-Yehuda, Gil Kedar and Vladimir Reznikov

Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel, 52900
German National Research Center for Information Technology(GMD), D-53757 St. Augustin

feldman@cs.biu.ac.il, kloesgen@gmd.de

Abstract

We present Document Explorer, a data mining system searching for patterns in document collections. These patterns provide knowledge on the application domain that is represented by the collection. A pattern can also be seen as a query that retrieves a set of documents. Thus the data mining tools can be used to identify interesting queries which can be used to browse the collection. The main pattern types, the system can search for, are frequent sets of concepts, association rules, concept distributions, and concept graphs. To enable the user to specify some explicit bias, the system provides several types of constraints for searching the vast implicit spaces of patterns that exist in the collection. The patterns which have been verified as interesting are structured and presented in a visual user interface allowing the user to operate on the results to refine and redirect search tasks or to access the associated documents. The system offers preprocessing tools to construct or refine a knowledge base of domain concepts and to create an internal representation of the document collection that will be used by all subsequent data mining operations. In this paper, we give an overview on the Document Explorer system. We summarize our methodical approaches and solutions for the special requirements of this document mining area.

1. Introduction

Most informal definitions (Fayyad et al. 1996) introduce *knowledge discovery in databases* (KDD) as the extraction of useful information from databases by large scale search for interesting patterns. The vast majority of existing KDD applications and methods deal with structured databases and thus exploits data organized in records structured by categorical, ordinal, and continuous variables. However, a tremendous amount of information is stored in documents that are nearly unstructured. The availability of document collections and especially of online information is rapidly growing, so that an information overload and analysis bottleneck often arises also in this area.

A document collection represents a specific domain and each document is related to some of the concepts that play a role in this domain. Therefore, our knowledge discovery approach for document collections is targeted at these concepts. The

patterns that are discovered describe (co-occurrence) relations between concepts of an application domain. Data mining methods thus extract knowledge about a document domain by searching and structuring interesting concept relations and by monitoring changes over time. Additionally, they supply new browsing possibilities, because inter-document information is contained in the patterns.

The basic pattern types in the Document Explorer system are frequent set of concepts and relations between these sets. The user can specify syntactical, background, quality and redundancy constraints to guide the search for interesting patterns. The search results are structured in a GUI (graphical user interface) giving the user the possibility to access the documents via the query that is associated to a selected pattern or to refine the search task.

Following this approach, it is obvious, that as for the mainstream KDD, also knowledge discovery from document collections is a process that involves preprocessing, data mining, and refinement tasks. Methods of term extraction or text categorization belong to the main preprocessing tasks in this area. Data mining methods, because typically based on large-scale brute force search, produce a lot of patterns. A main discovery task relates to constraining the search by operationalizing interestingness, especially to prevent the user from getting overwhelmed with too many results.

One rapidly developing application area is the exploitation of information made available in Internet and Intranet. Consider the document collection that was returned by a search engine using the search query "KDD or data mining". As a result of this Web-search with these general terms, several thousands of documents will be identified. Text based data mining methods, after identifying co-occurrence of concepts, could answer questions about methods that are typically applied, for example, in document discovery, or identify a network of researchers involved in the area of fraud detection. Another example of text mining is related to the analysis of e-mail messages of one institution to identify cooperation patterns between organizational units of the institution and could be used for reorganization decisions. In the political area, an analyst will be interested in exploiting news agency text collections, such as the Reuters articles that appeared in the newswire, to detect e.g. relations between countries in the context of economical or other fields.

When abstracting from these examples, the main analysis tasks in the document area (as in the traditional structured database area) consist of summarization, classification, prediction, and browsing. In the above examples, the knowledge discovered in the various document collections is used to describe relations between concepts relevant for the application domain, to predict a criminal activity, or to optimize an organizational structure. These analysis tasks pursue the general goal to derive knowledge on the domain that is represented by the document collections.

An additional goal aspires to provide a complemental retrieval approach. Traditionally, retrieval is supported by a query approach selecting all the documents in a collection that include some boolean combinations of keywords. Moreover,

clustering approaches are applied in the classical retrieval area to construct extensional clusters of documents using a distance function for documents. Documents can then be accessed via these query results or clusters. Data mining approaches provide complemental retrieval possibilities that are given by accessing the documents that support the detected patterns. Thus the data mining system detects collections of potentially interesting documents, and enables easy browsing of the collections. Such a group can be seen as a query that the system has identified and which the user was not aware of.

In traditional retrieval, it is assumed that the user knows in advance the concepts of documents he could be interested in, or that he selects a constructed cluster of documents (e.g. Salton, 1989; Cutting et al., 1993). Applying KDD tools like Document Explorer means that the system takes an active role in suggesting concepts of interest to the user, as well as supply new browsing methods that rely on inter-document information. The discovery framework of Document Explorer may thus be viewed as an intermediate point between user-specified retrieval queries and unsupervised document clustering. The user typically provides some guidance to the system about the type of patterns of interest, but then the system identifies groups of pattern instances applying filtering, ordering, generalization, statistical validation and clustering techniques.

The Document Explorer system builds upon the experiences that were gained from the KDT system (Feldman and Dagan 1995, Feldman et al. 1996), FACT (Feldman and Hirsh 1996, Feldman and Hirsh 1997), and Explora (Kloesgen 1992, Kloesgen 1996).

The rest of this paper is organized as follows. After summarizing the special requirements of data mining in document collections, a general overview on the preprocessing, mining and refinement tasks of Document Explorer is presented. Some examples of the GUI are then discussed in more detail to demonstrate the exploration approach of the system.

2. Special requirements for discovery in document collections

The data mining approach we propose in this paper for the analysis of document collections is based on a preprocessing step that extracts relevant keywords, terms or concepts from the documents. All the terms and keywords extracted from the documents are stored in a special Trie. The Trie proved to be very useful representation for databases with large repetitions of transactions. In this case the trie representation is significantly smaller than the original database, yet captures all its information in a more accessible manner. In addition, the algorithm can take advantage of pre-specified constraints to reduce the time it takes to generate all frequent sets and associations. Details regarding the construction of the Trie and generation of frequent sets and associations rules out of it, can be found in [Amir et al, 1997]. All data mining methods operate on this compact representation of the collection. The discovered patterns are presented in a GUI allowing the user to operate on these patterns to redefine a mining task and to browse the associated documents.

Besides providing the special preprocessing tools, the Document Explorer system must treat some special requirements that distinguish this mining task. These requirements are summarized in this section and solutions are then presented in the following sections.

A first property of data mining in document collections refers to the very large number of features. Typically some thousands of keywords or concepts may be relevant for an application domain. However, the features are sparse, i.e. only a small percentage of all possible features appears in a single document. Sparseness is also given for feature dimension. Some features appear only in a few documents, so that the support of many document subgroups built by combination of features is low.

Features can be arranged in a taxonomy which is implemented as a directed acyclical graph. This taxonomy is very important, since it allows grouping the patterns in a hierarchical way. Due to the very large feature set, the overabundance problem of identified patterns is still more relevant for these document applications than for usual structured data applications.

Relations between concept categories can easily introduce background knowledge. Thus an additional structure is given for the feature set. They are not just elements in a flat set as in most structured data applications.

These special characteristics of the data mining task for document collections stress the importance of efficient filtering (during and after search) or refinement techniques including suppressing, ordering, pruning, generalization and clustering approaches.

3. Preprocessing in the Document Explorer system

In the preprocessing stage, the collection of documents is transformed into a *target database,* which is implemented as a Trie. Besides this target database, the data mining tasks exploit the search specifications of the user and a knowledge base representing the concepts of the domain and their relations.

Special preprocessing tools have to be provided for the different types of document collections. Currently, we offer such tools for text collections and Web documents. For other document types, for example audio and image collections, quite different preprocessing tools will be necessary. For text collections, a source preprocessing and categorization module is available. This module includes the set of source converters and the text categorization software. It is responsible for converting the information fetched from each of the available sources into a canonical format and for tagging each document with the predefined categories and extracting all multi-word terms from the documents.

The first step in extracting information from the documents is to represent each of the documents using a set of keywords or phrases (multi-word terms). Each document is labeled with a set of keywords. The input text collections can be already labeled with such keywords, as is the case for the Reuters collection used in this paper. Alternatively, the collection is fed through a text categorization system (e.g., Iwayama and Tokunaga 1994; Apte et al. 1994) or term extraction algorithm that

augments each document with such keywords or phrases. In addition to keyword and phrases extracted directly from the document, the Document Explorer system can take advantage of a concept taxonomy that is supplied by the user when managing the knowledge base.

This *knowledge base* includes a directed acyclical graph of concepts of the domain and a set of relations for these concepts. For the Reuters collection, a concept graph is given by categories like countries, economic topics, persons which are arranged on several hierarchical levels. A concept graph for the application area *KDD* would include categories such as methods, tasks, applications, researchers, companies, and projects. A utility for term extraction from texts suggests a list of candidate terms to support the user in defining such a concept graph.

Relations defined on the concept categories represent additional domain knowledge and are used in the mining phase to build constraints for search. The knowledge base for the Reuters collection includes relations between pairs of countries (e.g. countries with land boundaries), between countries and persons (nationality), countries and commodities (exports) and so on. These relations can be defined by the user or transformed by special utilities from general available sources (such as the CIA World Fact Book, or companies home pages).

The concepts from the concept graph are used for constructing the patterns. The concepts are the main objects for which knowledge is to be discovered and expressed in patterns. Frequent concept sets, binary relations between frequent concept sets, and distributions of concepts are used as patterns and statistically validated in the document collection, resp. the target database generated from the document collection.

The target database is constructed for a document collection and a given set of all concepts that appear as leaves in the concept tree. For each application (or application type), a method is needed that generates the target database from the collection of documents. For a Web application, such a method can be arranged on top of a Web search engine (or even within the search engine). This method must determine for each document and each concept, if the concept is relevant for the document. In a simple realization, this method just checks, if the document contains the concept term. An elaborated method will rely on a more substantial text analysis. Similar methods could be available for audio and image applications.

The target database is represented as a compressed data structure, namely a Trie. The fact that we represent the documents as sets of phrases and keywords (concepts) has several implications. We have only binary attributes, and these attributes are sparse (i.e. only a fraction of the attributes appears in any given record; typically around 2-3 percent), and finally the number of records is of medium size (between 20,000-500,000 documents). The Trie is an efficient data structure that encapsulates all the information of the document collection. In this Trie, all aggregates existing in the target database are managed in a compressed format. In addition, the Trie provides an efficient approach to incrementally calculate all the aggregates, and to store and

access these aggregates. Several forms of tries can be used that treat in a different way the tradeoff between space of storing the Trie and time of calculating derived results from the Trie. For more details on our Trie methods we refer the reader to (Amir et al. 1997).

4. Data mining methods in Document Explorer

The data-mining layer of the system contains all the search and pruning strategies, which can be applied for mining patterns like frequent concept sets, associations, and distributions. The embedded search algorithms control the search for specific pattern instances within the target database. This level includes also the refinement methods that filter redundant information and cluster together closely related information. The data management layer is responsible for all access to the actual data stored in the tries. This layer encapsulates the target database from the rest of the system.

The data mining approach provided in Document Explorer is characterized by providing a set of pattern types to the analyst. These pattern types (frequent concept sets, association rules, near concept sets, concept distributions, and comparisons of concept distributions) are discussed now in more detail. The analyst can specify a search task to let the system discover all interesting instances of a pattern type that can be validated in the document collection. Four types of constraints (syntactic, background, quality, and redundancy constraints) are applied to operationalize interestingness.

The search algorithms of the mining layer have to process search spaces of instances for a selected pattern type. Search is organized in dependence of the specified search constraints, and appropriate search strategies and pruning techniques are chosen. All patterns can be studied in the context of a conditioning concept set or context free, i.e. for the general domain of the whole collection. Conditioning a search task therefore means, selecting a set of concepts that is used to restrict an analysis task, e.g. a restriction to documents dealing with USA and economic concepts.

4.1 Frequent sets of concepts

A first basic pattern that can be derived from the target database is a *frequent concept set.* This is defined as a set of concepts that is represented in the document collection with a minimal support (given as a threshold parameter *s*), i.e. all the concepts of the frequent concept set appear together in at least *s* documents. A frequent set can directly be seen as a query, given by the conjunction of concepts of the frequent set. Frequent sets can be partially ordered by their generality and hold the simple, but useful pruning property, that each subset of a frequent set is a frequent set.

4.2 Relations between frequent concept sets

Two types of binary relations can be studied: directed and undirected relations. Graph theoretical methods detecting pathes, cliques and other groups of concept sets can be placed on these relations (Feldman et al. 1997b).

Association rules belong to the directed relations. Associations rule A $\Rightarrow$ B, relating two frequent concept sets A and B, can be quantified by two basic measures: support and confidence. The *support* is the number of those documents that include all the concepts in A and B. The *confidence* is the percentage of documents that include all the concepts in B within the subset of those documents that include all the concepts in A. Association rules are used in many data mining applications. Methods for association rule generation and search strategies are, for example, discussed by Agrawal et al. (1993, 1996).

Near frequent concept sets establishes an undirected relation between two frequent sets of concepts. It can be quantified by measuring the degree of overlapping, e.g. based on the number of documents that include all the concepts of the two concept sets. This measure can be regarded as a distance function between the concept sets. Several distance functions can be introduced (e.g. based on the cosine of document vectors, Tanimoto distance, etc.).

4.3 Concept distributions

A group of concepts is associated with each internal node of the concept tree. This group includes all the leaf concepts under this node, e.g. all countries, all EU countries, all economic topics. A *concept distribution* can be defined for such a group of concepts as a vector of the proportions of the concepts in the document collection, i.e. the fractions of the documents that include a concept. Note that the proportions in a vector do not sum up to 1, since a document usually will include several concepts of the studied group of concepts. Nevertheless we use the term *distribution* to avoid introducing another term and because many of the connotations associated with this term still hold.

In the usual way, we can regard *conditional concept distributions* (simply referring to a subset of the document collection defined by a concept set as the conditional part). *Average concept distributions* refer to the averages of frequencies over all immediate children of a (conditional) node, e.g. the economic topic distribution of averages over all countries of the European Union.

As for frequent concept sets, we evaluate relations between concept distributions. We mainly deal with an undirected binary relation given by a distance function for distributions. Especially we use the Kullback-Leibler distance as an information theoretic measure (Feldman and Dagan 1995).

5. Browsing and Filtering in the Document Explorer System

Document Explorer offers a set of GUI based mining tools and graph based visualization techniques that enable the user a much easier access to the system. Now we present some examples of GUI based mining tools that implement the ideas presented above.

5.1 Browsing through Distributions

Traditional document retrieval systems allow a user to ask for all documents containing certain keywords — UK and USA, for example — but then present the entire set of matching documents with little information about the collection's internal structure other than perhaps sorting them by relevance score (which is typically a shallow measure computed from the frequency and position of keywords in the document) or chronological order. In contrast, the Distribution Browser Tool enables the user to investigate the contents of a document set by sorting it according to the child distribution of any node in the hierarchy, such as topics, countries, companies, etc. Once the documents are analyzed in this fashion and the distribution is displayed, the user can access the specific documents of each subgroup. In order to generate a distribution the user need to provide two boolean expressions. The first expression defines the selection condition for the documents. The second expression defines the distribution to be computed on the set of chosen documents. For instance the user can specify as the selection criteria the expression "USA and UK", only documents that contain both keywords will be selected for further processing. The distribution expression can be "topics" in which case, we will get a set of rules that correlated between USA and UK and any of the concepts defined under the node "topics" in the taxonomy. The results of this query is shown in Figure-1. The results are shown in a hierarchical way based on the structure of the taxonomy underneath "topics". We can see for instance a rule USA,UK $\Rightarrow$ acq 42/19.09% which means that in 19.09% of the documents in which both USA and UK are mentioned, the topic acquisition is mentioned too, which amount to 42 documents. The user can then click on that rule to get the list of 42 documents that support this rule.

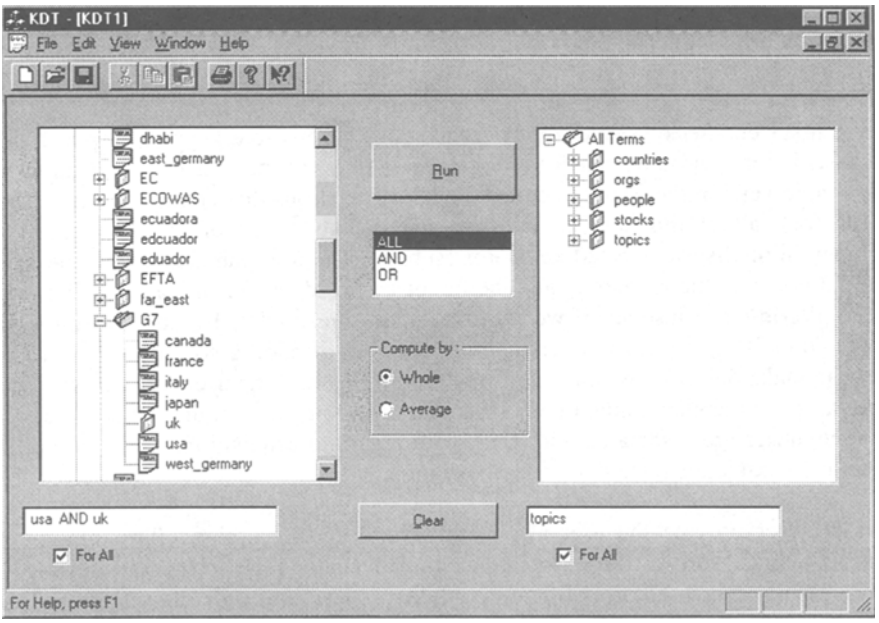

Fig. 1. Defining the Distribution Query

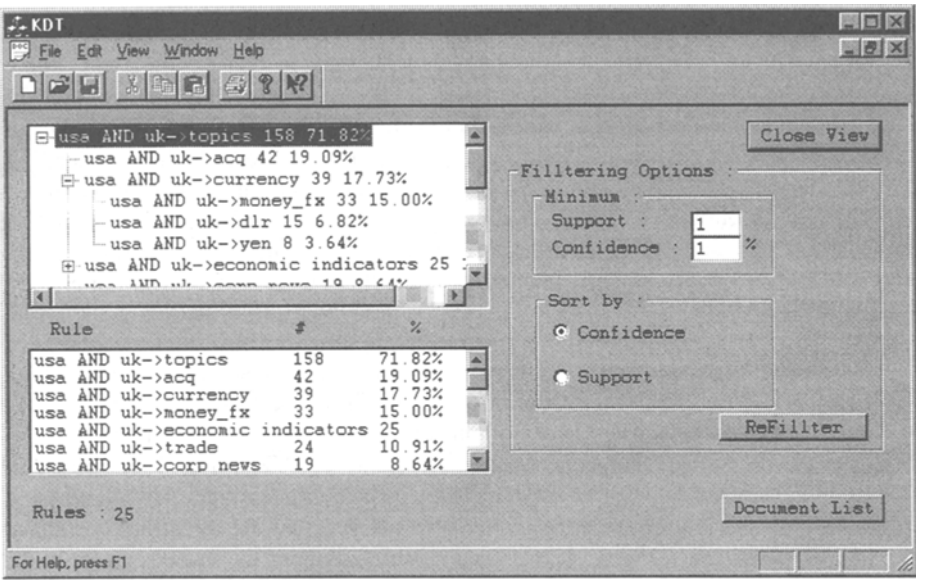

Fig. 2. Hierarchical Topic Distribution of USA and UK

5.2 Management of Association Rules

One of the major problems users have to face when using associations is the fact that usually the algorithms generate a huge number of associations that satisfy the support and confidence thresholds. In order to solve that difficulty we have developed the association browser. This tool is geared towards providing users with an easy way for finding associations, and then filtering and sorting them in different orders. This tool supports the specification of simple constraints on the presented associations. The user can select a set of keywords from the set of all possible keywords appearing in the associations and then select the logical test to performed on the associations. In this simple version the user can either see all associations that contain either of these words (or), all of this words (and), or that the keywords of the association are included in the list of selected keywords (subset). The user can also select one of the internal nodes in the taxonomy and the list of keywords under this node will be used in the filtering. For instance if we set the support threshold at 10, and the confidence threshold at 10%, we get an overwhelming number of 6560 associations. Clearly, no user can make out this amount of information. In Figure-3 the user chose to view only those associations that contain both USA and acq (company acquisition). We can see what countries are associated with USA with regard to acquisition, along with all the statistical parameters related to each association.

This browser provide the user also with several sorting options. Two options are rather obvious, sorting the associations in alphabetical ordering, and sorting the association in decreased order of their confidence. The third ordering scheme is based on the chi-square value of the association. In a way this measures how different is the

probability of seeing the RHS of the association given that we saw its LHS from the probability of seeing the RHS in the whole population.

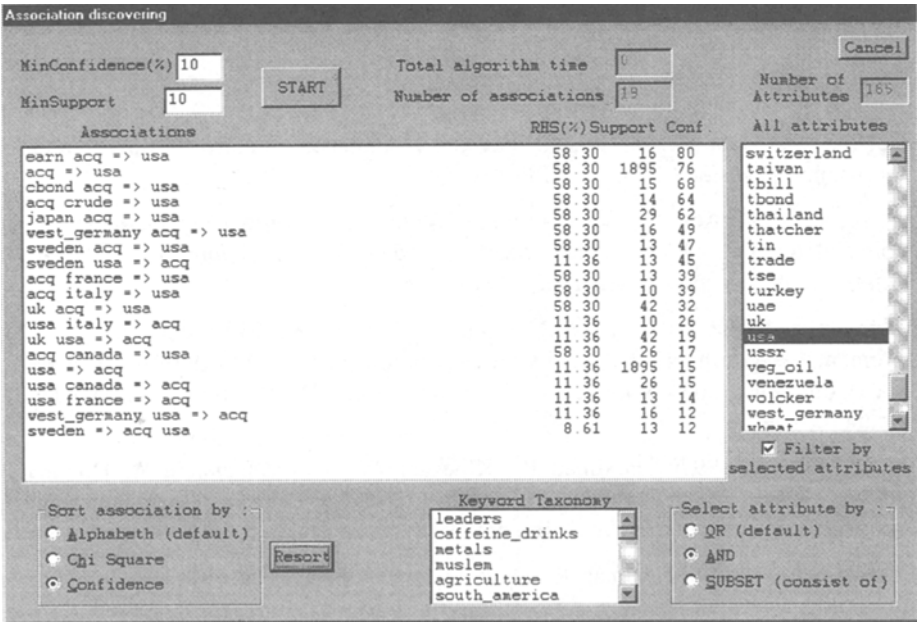

Fig. 3. A simple Browsing and Filtering Utility for Association Generation

6. Conclusion

We have presented in this paper an overview of the browsing mechanism of the Document Explorer system. The main pattern types, the system can search for, are frequent sets of concepts, association rules, concept distributions, and concept graphs. To enable the user to specify some explicit bias, the system provides several types of constraints for searching the vast implicit spaces of patterns that exist in the collection. The patterns which have been verified as interesting are structured and presented in a visual user interface allowing the user to operate on the results to refine and redirect search tasks or to access the associated documents. The system offers preprocessing tools to construct or refine a knowledge base of domain concepts and to create an internal representation of the document collection that will be used by all subsequent data mining operations. We have focused on the browsing and filtering capabilities of the Document Explorer System. These include the possibility to browse hierarchical relations between concepts in the taxonomy, and to browse and filter association rules.

We are currently working on extending the visual capabilities of the system to include various types of graphs that can depict the relationship between concept sets. In addition we are extending the set of statistical tests that can be applied to the discovered patterns.

7. References

[Amir et al., 1997] Amir A., Aumann Y., Feldman R., and Katz O. Efficient Algorithm for Association Generation. Technical Report, Department of Computer Science, Bar-Ilan University, Israel.

[Agrawal et al., 1995] Agrawal R., Mannila H., Srikant R., Toivonen H., and Verkamo I. Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, Eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, pages 307-328, AAAI Press.

[Apte et al., 1994] Apte C., Damerau F., and Weiss S. Towards language independent automated learning of text categorization models. In *Proceedings of ACM-SIGIR Conference on Information Retrieval*, 1994.

[Feldman et al., 1997a] Feldman R., Amir A., Aumann Y., Zilberstein A., Hirsh H. Incremental Algorithms for Association Generation. In Proceedings of the 1st Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD97), Singapore, 1997.

[Feldman et al., 1997b] Feldman R., Kloesgen W., and Zilberstein A. Document Explorer: Discovering Knowledge in Document Collections. Technical Report, Department of Computer Science, Bar-Ilan University, Israel.

[Feldman et al., 1996] Feldman R., Dagan I., and Kloesgen W. Efficient Algorithms for Mining and Manipulating Associations in Texts. In *Proceedings of EMCSR96*, Vienna, Austria, April 1996.

[Feldman and Dagan, 1995] Feldman R. and Dagan I. KDT - knowledge discovery in texts. In *Proceedings of the First International Conference on Knowledge Discovery (KDD-95)*, August 1995.

[Iwayama and Tokunaga, 1994] Iwayama M. and Tokunaga T. A probabilistic model for text categorization based on a single random variable with multiple values. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 1994.

[Klemettinen et al., 1994] Klemettinen M., Mannila H., Ronkainen P., Toivonen H., and Verkamo A. Finding Interesting Rules from Large Sets of Discovered Association Rules. *In Proceedings of the 3rd International conference on Information and Knowledge Management*, 1994.

[Kloesgen 1995] Klösgen W. Efficient Discovery of Interesting Statements. *The Journal of Intelligent Information Systems*, Vol. 4, No 1.

[Kloesgen 1996] Klösgen W. Explora: A Multipattern and Multistrategy Discovery Assistant. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Cambridge, MA: MIT Press.