# Visualization Techniques to Explore Data Mining Results for Document Collections

Ronen Feldman
Math and Computer Science Department
Bar-Ilan University
Ramat-Gan, ISRAEL 52900
feldman@macs.biu.ac.il

Willi Klösgen
GMD
German National Research Center
for Information Technology
D-53754 Sankt Augustin
Germany
kloesgen@gmd.de

Amir Zilberstein
Math and Computer Science Department
Bar-Ilan University
Ramat-Gan, ISRAEL 52900
zilbers@cs.biu.ac.il

## Abstract

Data mining has been informally introduced as large scale search for interesting patterns in data. It is often an explorative task iteratively performed within the process of knowledge discovery in databases. In this process, interactive visualization techniques are also successfully applied for data exploration. We deal in this paper with the synergy of these two complemental approaches. Whereas data mining typically relies on brute force, large scale and systematic search in hypotheses spaces, interactive visualization activates the visual capacities of an analyst to identify patterns in data. We demonstrate some possibilities to combine these approaches for the area of data mining in document collections. Document Explorer is a system that offers various preprocessing tools to prepare collections of text or multimedia documents which are available in distributed environments (e.g. Internet and Intranet) for data mining applications, and includes data mining methods based on searching for patterns like frequent sets or association rules. *Keyword graphs* are used in this system as an highly interactive technique to present the mining results. The user can operate on the visualized results, either to redirect the data mining process, to filter and structure the results, to link several graphs, or to browse into the document collection. Thus in the keyword graphs, the relations between interesting sets of keywords are presented (the sets may also be regarded as retrieval queries to be posed to the collection) and made operable to the analyst.

**Keywords:** visualization, frequent sets, association rules, redundancy elimination, keyword graphs

## 1. Introduction

Document Explorer (Feldman, Klösgen and Zilberstein 1997) is a data mining system searching for patterns in document collections. These patterns provide knowledge on the application domain that is represented by the document collection. Such a pattern can also be seen as a query or implying a query that, when addressed to the collection, retrieves a set of documents. Thus the data mining tools also identify interesting queries which can be used to browse the collection. The system searches for interesting concept sets and relations between concept sets, using explicit bias for capturing interestingness. The bias is provided by the user by specifying syntactical, background, quality and redundancy constraints to direct the search in the vast implicit spaces of pattern instances which 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 which will be used by all subsequent data mining operations. The source documents can be of text or multimedia type and be distributed, e.g. in Internet or Intranet.

A set of concepts (terms, phrases or keywords) directly corresponds to a query that can be placed to the document collection for retrieving those documents that contain all the concepts of the set. Various quality measures can be defined for a concept set which are considered as necessary conditions for the interestingness of the concept set. Document Explorer provides a broad spectrum of quality measures, ranging from simple support conditions to statistical indices for the concept set, based on distributions of (target) variables or change patterns for dynamic collections. Frequent sets (Agrawal et al. 1993) are the basic, simply support constrained concept sets in Document Explorer; they can be specialized by additional quality conditions.

Association rules (Agrawal et al. 1993) provide an useful construct for discovering sets of attributes that appear frequently together within the same row in the data base. In (Feldman and Hirsh 1995), we have shown that association rules can be applied also for analyzing textual collections. In that case, the FACT system identifies keywords (terms, concepts) that appear frequently within the same document in the collection. Association rules are special binary relations between concept sets. Document Explorer can search for various other binary relations, implied by similarity functions (section 4), or defined as maximal associations (Feldman et al. 1997b).

It is also useful in some occasions to analyze higher order connections between terms, e.g. by finding terms that are connected through another term or a chain of terms. So even though two companies do not appear together in the same document, still we may be interested in finding if they are connected through a third company, meaning they both occur in different documents with the third company. An analyst might use this information to infer about the indirect connection between the companies. This notion is useful also for structured databases. Consider a database reporting on transactions of money transfer performed between people. An IRS investigator is trying to find a connection between people, but they do not appear together in any single transaction. Then the investigator will try to show an indirect connection established through a third person or a chain of people. These higher order connections can be identified by the user in a graphical presentation of the basic associations or derived via a similarity function that captures these higher order connections.

A group of concept sets can be represented as a graph referring to the natural partial ordering of concept sets (ordering by generality). Relations between concept sets imply a graph structure as well. Various operations on these graphs allow the analyst in Document Explorer to redirect a mining task, to filter or group mining results, and to browse into the document collection. Thus this graph provides an interaction medium for the analyst based on interactive visualization techniques.

Interactive visualization techniques are already successfully applied within KDD processes. In the area of statistics (Wills 1997), interactive exploration graphs combine, for example, scatter diagrams with bar charts and link selected subgroups of items in one graph with the corresponding items in the other graph. New interaction methods such as parallel coordinates or netmap techniques let the user identify patterns by studying the graphical presentation of the data. Because of the well-established visual capabilities, it is of course much easier for the analyst to detect these patterns in the presented data visualizations than in the numerical raw data.

In this paper, we propose interactive visualization techniques to be applied for the exploration of data mining results. By this integrated data mining and visualization approach, we combine the complemental strengths of both methods. At first, we introduce the notion of keyword graphs. Then we summarize different types of keyword graphs (graphs for sets versus association graphs, singleton vertex graphs versus multi vertices graphs, single category versus multi categories graphs). In chapter 4, we describe several similarity functions for defining relations between sets of concepts to be presented in the association graphs. To control the complexity of the keyword graphs, we then introduce equivalence classes for associations and redundancy filters, overcoming the combinatorial explosion that is given by the large number of possible subsets of a set of concepts. Finally we present some more examples of keyword graphs and describe some classes of interactive operations the user can perform on the keyword graphs.

## 2. Definitions

In this section, we shortly define those fundamental constructs of Document Explorer to which we refer in the remaining sections. A knowledge base includes domain knowledge about the document area. It includes a *concept DAG* (directed acyclical graph) of the relevant concepts for the domain. Several *categories* of concepts are hierarchically arranged in this DAG. For the Reuters newswire collection, used in this paper as an application example, categories correspond to countries, persons, topics, etc. with subcategories like European Union, NATO, politicians, economic indicators, currencies. Additionally, the knowledge base contains *background relations*. These are binary relations between categories such as nationality (relation between persons and countries) or export partners (between countries). In preprocessing, the knowledge base and a *target database* are constructed. The target database contains binary tuples. A tuple represents a document and the concepts being relevant for the document. All data mining operations in Document Explorer are operated on a derived *trie* structure, that is an efficient data structure to manage all aggregates existing in the target database (Amir et al. 1997).

A *concept set* is simply a set of concepts. A set of concepts can be seen as an intermediate concept that is given by the conjunction of the concepts of the set. For example, the concepts ”data mining” and ”text analysis” define a joint concept which can be interpreted as ”data mining in text data”. *Frequent concept sets* are sets of concepts with a minimal *support*, i.e. all the concepts of the set must appear together in at least *s* documents. A *context* is given by a concept set and is used as a subselection of the document collection. Then only the documents in this subcollection are analyzed in a search task. The system derives, for example, patterns “in the context of crude oil” for the documents that contain crude oil as a phrase or are annotated by crude oil using text categorization algorithms.

A *binary relation* between concept sets is a subset of the crossproduct of the set of all concept sets. An *association* is a binary relation given by a *similarity function*. To measure the degree of connection (similarity) between two sets of concepts, we rely on the support of those documents in the collection, that include all the concepts of the two sets. If there is no document that contains all the concepts, then the two concept sets will have no connection (similarity = 0). If all the concepts of the two sets always appear together, the strongest connection measurable by the document collection (similarity = 1) is given. An *association rule* is a special association, defined as usual by a minimal support and *confidence*.

A *keyword graph* is a pair consisting of a set of nodes and a set of edges. Each edge connects two nodes. Quality measures are calculated for each node and each edge. A node corresponds to a concept set and an edge to an element of a binary relation. Special subsets of nodes and connections can be defined, e.g. a *clique* is a subset of nodes of a keyword graph, for which all pairs of its elements are connected by an edge. A *path* connects two nodes of a keyword graph by a chain of connected nodes.

A *search task* is specified in Document Explorer by defining sytactical, background, quality and redundancy constraints for searching spaces of concept sets or of associations (Feldman, Klösgen and Zilberstein 1997). The result of a search task is a group of concept sets or associations satisfying the specified constraints. These groups of results are arranged in keyword graphs offering to the user interactive operations on the elements of the graph.

## 3. Types of keyword graphs

Keyword graphs are used to present several types of connections between keywords (concepts) or between sets of keywords in an interactive way so that the user can operate on the graphs. At first we distinguish *set graphs* from *association graphs*. A set graph visualizes a subset of concept sets with respect to their partial ordering. An association graph presents associations.

Figure 1 shows a set graph for frequent sets arranged in a tree structure. The user can operate on this graph, e.g. by selecting nodes, opening and closing nodes, or defining new search tasks with respect to these nodes, for instance to expand the tree. The first level in figure 1 relates to country keywords, sorted by a simple quality measure (support of the frequent set). The node ”USA” (support: 12814 documents) is expanded by person keywords, further expansions relate to economical topic keywords (e.g. expansion of the node ”James Baker”: 124 documents, 0 %) and country keywords.

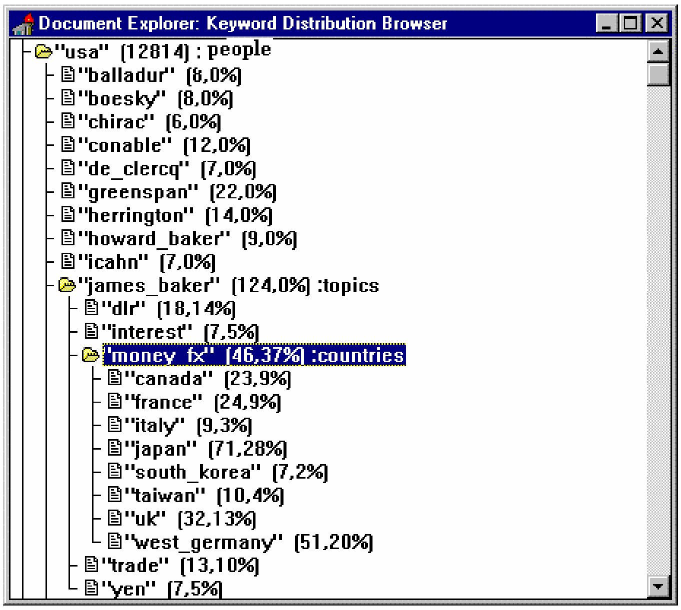


**Figure 1: Graph representing concept sets**

Next, we distinguish *simple association graphs* consisting of singleton vertices and *multi vertices graphs*, where the edges can consist of a set of several concepts. A simple association graph connects concepts of a selected category (e.g. countries category). At each vertex of a simple association graph, there is only one concept, and two concepts are connected by an edge, if their similarity with respect to a similarity function is larger than a given threshold. In Table 1, some similarity functions are summarized. Figure 2 shows this kind of keyword graph for the country category in the context of crude oil and a simple similarity function based on the number of documents in which the countries co-occur. The following observations can be easily studied: USA, Iran, and Iraq are the most connected countries, Greece and Turkey are the only members of a separate component of the graph, Canada is only connected to USA, etc.

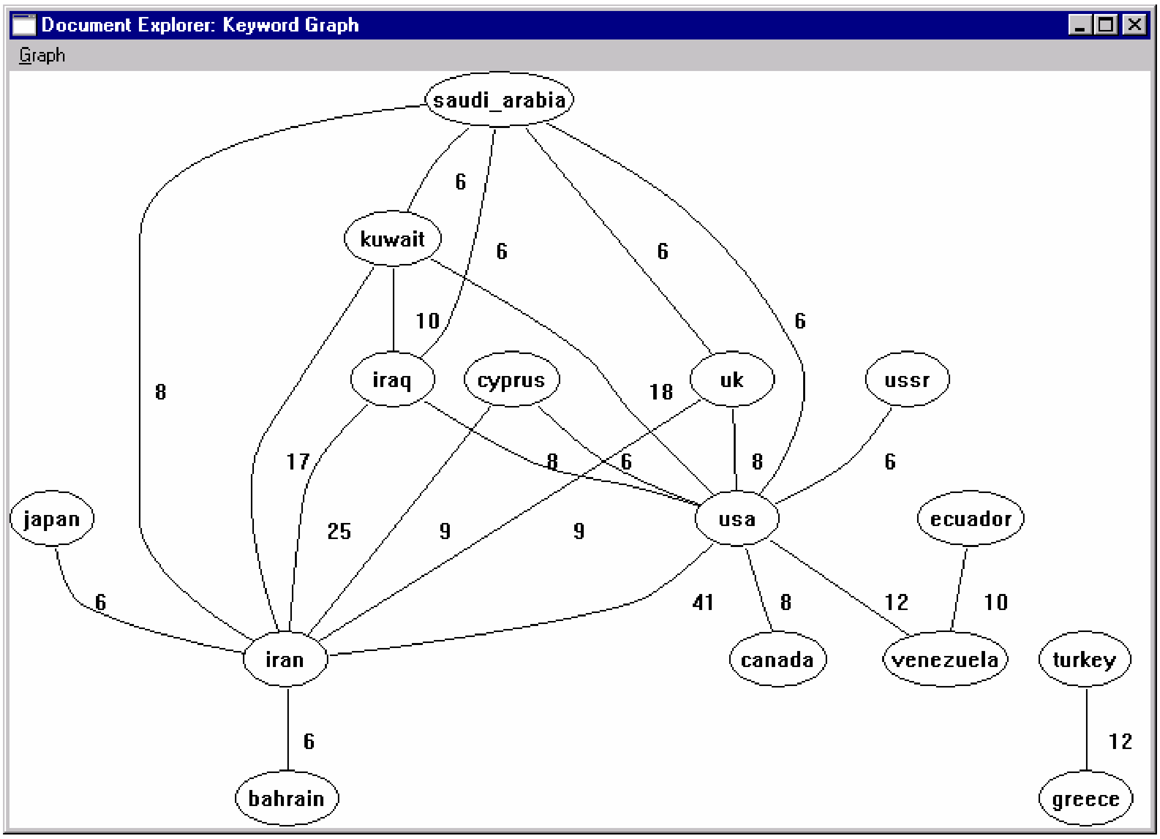


**Figure 2: Association graph (single vertex, single category) - countries in the context of crude oil**

This type of graph can be induced from the connection matrix calculated for a studied category and similarity function. A more complex type of graph allows that vertices will represent subsets of concepts from a category. Three subtypes of this graph can be defined by studying connections only between sets where one set is a subset of the other, between disjoint sets, or overlapping sets (but not subsets). In this case, vertices represent frequent sets of concepts and vertices are connected by an edge, if the union of their concepts has a significant support. Several criteria are selectable that define this significance (chapter 4).

Another type of graph presents the associations between different categories, e.g. contries and economical topics. The singleton vertix version of this graph is arranged in a netmap like technique, where different arcs of a circle are used to include the concepts of categories, and edges (between countries and topics) present the associations. Figure 3 shows an example of this kind of graph, where we use a bi-partite graph to represent the connections between G7 countries and economical indicators.

## 4. Similarity functions for association graphs

In this section, we summarize some quality constraints that are used in the search for interesting concept sets or associations. We concentrate here on associations; statistical patterns for set evaluation are described in more detail by Kloesgen (1995), and Feldman and Dagan (1995). Association rules are mostly introduced in an undirected way and specified by a support and a confidence threshold. A fixed confidence threshold is often not very reasonable, because it is independent from the support of the RHS of the rule. Therefore, an association should have a significantly higher confidence than the share of the RHS in the whole context to be considered as interesting. Significance is measured by a statistical test, e.g. t-test or chi-square.

With this addition, the relation given by an association rule is undirected. An association between two sets $A$ and $B$ in the direction $A \Rightarrow B$ implies also the association $B \Rightarrow A$. This equivalence can be explained by the fact that the construct of a statistically significant association is different from implication (which might be suggested by the notation $A \Rightarrow B$). It can easily be derived that, if $B$ is over-proportionally represented in $A$, then $A$ is also over-proportionally represented in $B$.

| Function | Similarity | Characteristic |
|---|---|---|
| support threshold | $d > d_0$ (step function) | evaluates only d, independent from $a - d$, $b - d$ |
| cosine | $s = d / \sqrt{a \cdot b}$ | low weight of $a - d$, $b - d$ |
| arithmetical mean | $s = 2d / (a + b)$ | middle point between cosine and Tanimoto |
| Tanimoto | $s = d / (a + b - d)$ | high weight of $a - d$, $b - d$ |
| information measure | weighted documents | only applicable, if weights are reasonable |
| statistical test | threshold statist. quality | typically for larger samples and covers |

Table 1: Similarity Functions for two Concept Sets $A$, $B$
$a$ = support($A$), $b$ = support ($B$), $d$ = support($A$,$B$)

As an example of differences of similarity functions, we compare the undirected connection graphs given by statistically significant association rules with the graphs based on the cosine function. The latter relies on the cosine of two vectors and is efficiently applied for continuous, ordinal and also binary attributes. In case of documents and concept sets, a binary vector is associated to a concept set with the vector elements corresponding to documents. An element holds the value 1, if all the concepts of the set appear in the document. As included in Table 1, the cosine similarity function in this binary case reduces to the fraction built by the support of the union of the two concept sets and the geometrical mean of the support of the two sets.

A connection between two sets of concepts is related to a threshold for the cosine similarity (e.g. 10%). This means, that the two concept sets are connected, if the support of the document subset that hold all the concepts of both sets is larger than 10% of the geometrical mean of the support values of the two concept sets. The threshold holds a property of monotony: if it is increased, some connections existing for a lower threshold disappear, but no new connections are established. This property is used as one technique to tune the complexity of a keyword graph. Another approach based on redundancy filters is described in the next section 5.

Let $f$ be the following factor: $f = N\, s(A,B) / s(A)\, s(B)$, where $s$ is the support for the two concept sets $A$ resp. $B$ and $N$ the number of documents in the collection (or a subcollection given by a selected context). In case of independence of the two concept sets, $f$ would be expected around the value 1. Thus $f$ is larger than 1 for a statistically significant association rule.

The cosine similarity of concept sets $A$ and $B$ can now be calculated as

$$S(A,B) = f \cdot \sqrt{q(A) \cdot q(B)}$$

i.e. as the geometrical mean of the relative supports of $A$ and $B$ ($q(A) = s(A) / N$) multiplied by the factor $f$, thus combining a measure for the relative support of the two sets (geometrical mean) with a significance measure (factor $f$).

By analysing this relation, it can be derived that the cosine similarity favors connections between concept sets with a large support (which need not necessarily hold a significant overlapping) and only includes connections for concept sets with a small support if the rule significance is high. This means, that the user should select the cosine similarity option, if there is a preference for connections between concept sets with a larger support. On the other side, the statistically based association rule connection should be preferred, if the degree of coincidence of the concepts has a higher weight for the analysis. Similar criteria for selecting an appropriate similarity function can be derived for the other options (see table 1).

## 5. Equivalence classes, partial orderings, redundancy filters

Very many pairs of subsets can be built from a given category of concepts, e.g. all pairs of country subsets for the set of all countries (given by the country category). Each of these pairs is a possible association between subsets of concepts, and even, if the threshold of the similarity function is increased, the resulting graph can have a too complex structure. We define now several equivalence relations to build equivalence classes of associations. Only a representative association from each class will then be included in the keyword graph in the default case.

A first equivalence is called *cover equivalence*. Two associations are cover-equivalent, iff they have the same cover. E.g., (Iran, Iraq) => (Kuwait, USA) is equivalent to (Iran, Iraq, Kuwait) => USA, because they both have the same cover (Iran, Iraq, Kuwait, USA). The association with the highest similarity is selected as the representative from a cover equivalence class.

*Context equivalence* is a next equivalence relation. Two associations are context-equivalent, iff they are identical up to a different context. That means that the two associations are identical, when from each association those concepts are eliminated which appear on both sides. E.g., (Iran, Iraq) => (Iran, USA) is equivalent to (Kuwait, Iraq) => (Kuwait, USA). The first association establishes a connection between Iraq and USA in the context of Iran, whereas the second association is related to the context of Kuwait. The context-free association (or the most general elements with a sufficient quality) are selected as the representatives from this equivalence class (e.g. Iraq => USA).

The next definition relates to a partial ordering of associations (not an equivalence relation). An association A1 is stronger than an association A2, iff the cover of A1 is a subset of the cover of A2. As special cases of this ordering, the right and left hand sides are treated separately.

Redundancy filters are defined for these equivalences and partial orderings. Selecting the representative of an equivalence class or the strongest associations is applied as a basic redundancy filter. Additionally criteria can refine these filters. E.g., for the context equivalence, a context conditioned association is selected additionally to the context-free association, iff the similarity of the context conditioned association is much higher (e.g. defined by a significance criterium).

There is a duality between frequent sets of concepts and associations. For a given set of frequent concepts, the implied set of all associations between freqent concepts of the set can be introduced. On the other side, for a given set of associations, the set of all frequent concepts appearing as left or right hand sides in the associations can be implied. In the application area of document collections, we are mainly interested in freqent concept sets, when we concentrate on retrieval or browsing aspects. These freqent concepts are considered as retrieval queries which are discovered as interesting by the system. When aspiring to gain some knowledge on the domain represented by the document collection, association rules are more interesting. In the keyword graphs, the concept sets are therefore included as active nodes (activating a query to the collection when selected by the user). Additionally, also complementary and intersection sets (e.g. related to the cover of an association) are provided as active nodes.

## 6. Some more examples of keyword graphs

A next example shows another association graph. Associations between two or more categories are presented in a netmap-like way. There, categories are arranged on arcs of a circle, and lines connect the (single-vertex) elements of the categories. Figure 3 demonstrates a simple bi-partite version for connections between contries and economic topics.

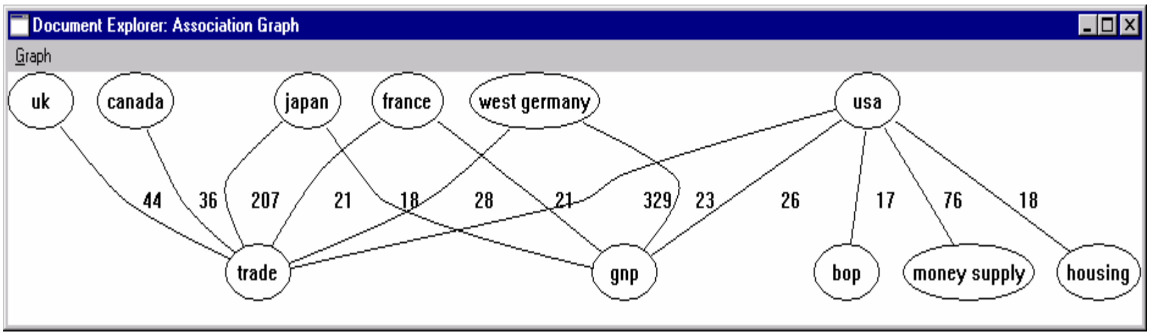


**Figure 3: Association graph (single vertex, several categories)**

The next example relates to connections between sets of concepts, e.g. between countries in the context of crude oil. The graph shown in figure 2 offers a first overview on the connections between individual countries. The user can handle the complexity of this graph by selecting subregions, limiting the number of connections either explicitely (e.g. top 30 connections) or implicitely by modifying the threshold of the similarity function. When a set connection graph is requested for the whole figure 2 graph with a relatively high similarity threshold (set to 0.30 in the example), the complexity of the resulting multiple vertices graph (see figure 4) is still moderate. The user can easily infer from this graph, that the strongest simple association (for pairs of countries) is given for Iran and Iraq which has a significant specialization in the context of Kuwait. There is another association between Kuwait and USA, valid only in the context of Iran, its generalization is valid at a level 0.19 which is below the threshold. Specialization and generalization relations can be underlined by special graphical elements (e.g. same colour).

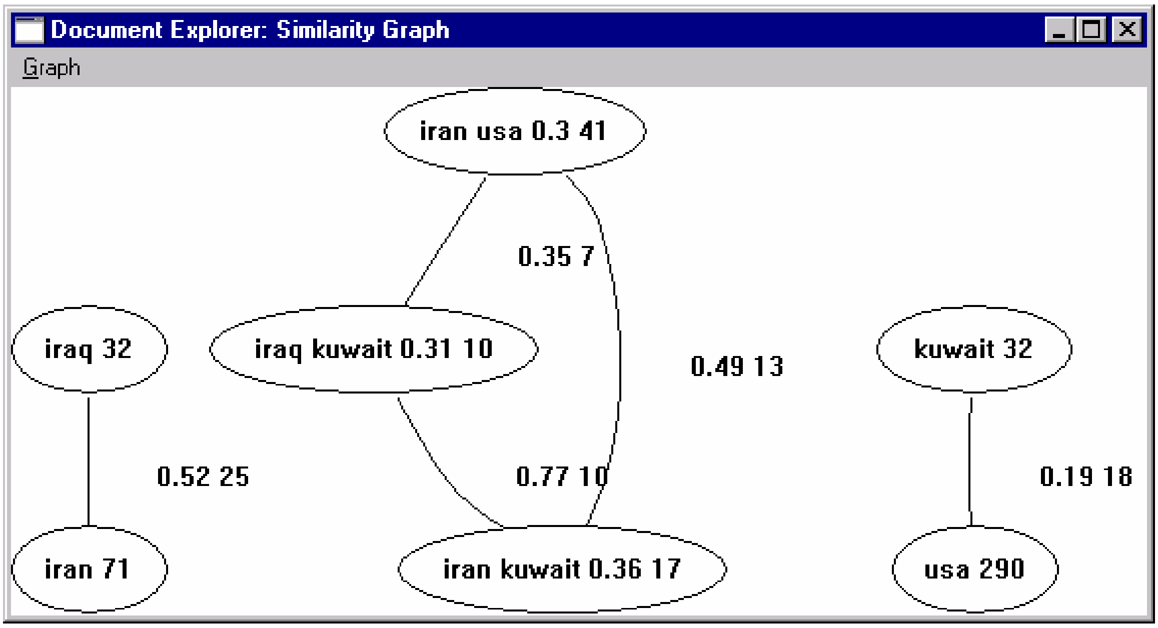


**Figure 4: Interesting concept sets and their associations**
**context: crude oil; categories: countries**

Figure 4 relies on the cosine similarity function with a threshold similarity of 0.3 (i.e. the union of two concept sets must have a support of at least 30% of the (geometrical) mean support of the two sets). This graph presents interesting frequent concept sets and a subset of non-redundant associations between frequent sets. In this case, only connections between (at most) pairs of countries satisfy the similarity threshold. The strongest bilateral connection holds for Iran and Iraq (similarity = 0.52, 25 documents in common). As a simple browse operation initiated on this graph, the user could now inspect the common documents or any complemental parts (e.g. the 7 documents on Iraq not related to Iran). Note that for this similarity measure, the connection between Iran and Iraq is higher than the connection between Iran and USA, although the support of Iran and USA is much higher (41 documents in common). Because USA has a high support (290 of the 620 documents on crude oil) and the support of Iraq is much lower (32 documents), the relative interdependency between Iran and Iraq is higher.

The strongest triplet cover consists of (Iran, Iraq, Kuwait), connecting the pairs (Iran, Kuwait) and (Iraq, Kuwait). Only the strongest connections belonging to a common cover have been included in this graph, due to the selection of the redundancy elimination option related to the cover equivalence. For example, the connection between (Iran, Iraq) and (Iraq, Kuwait) (with the same cover (Iran, Iraq, Kuwait) is not shown, because the similarity between the pairs (Iran, Iraq) and (Iraq, Kuwait) is lower. Another redundancy related to the context equivalence would not have included the relation between (Iran, Kuwait) and (Iraq, Kuwait), because this is a specialization of the already included relation between Iran and Iraq. This redundancy option has been chosen with an additional condition requiring a much higher similarity for the specialized connection (0.77 compared to 0.52 in this example).

However, the next triplet consisting of (Iran, Kuwait, USA) connecting (Iran, Kuwait) and (Iran, USA) is not a specialization of a more general rule related to context equivalence, because Kuwait and USA are not connected for the given threshold (similarity 0.19 is below the threshold of 0.30). Therefore Kuwait and USA are only connected in the context of Iran.

The quadrupel consisting of Iran, Iraq, Kuwait, USA connects (Iran, USA) and (Iraq, Kuwait) (highest similarity between all partitioning subsets of Iran, Iraq, Kuwait, USA). Again, the complementary sets could be interesting too, e.g. the three documents about (Iraq, Kuwait) which are not related to (Iran, USA). This connection would have been eliminated by the redundancy filter belonging to the partial ordering, because it is redundant compared to the connection between Iraq and Iran. In figure 4, this filter was not selected by the user, so the connection is included in the graph, but indicated as redundant (e.g. indicated by line type).

As a summary of many existing associations between countries in the context of crude oil, the system presents a subset of associations selected by filtering and redundancy elimination rules. This sets are arranged in a graph allowing the user to operate on the graph, e.g. by selecting nodes, edges or associated complemental nodes for browsing, by applying further filtering or redundancy elimination operations on the graph, or by defining new search tasks. As an example of a new search task, all specializations of a selected node (e.g. the association between Iran and Iran), also with respect to (other) categories, could be found.

Other graph operations on the connection graph shown in figure 2 are possible, e.g. identification of cliques or other distinguished subsets like stars. A clique of countries requires stronger constraints than those discussed above for figure 4. All pairwise relations must hold for a clique. The group consisting of Iran, Iraq, USA is not a clique, since Kuwait and USA are not connected (based on the cosine similarity measure and the selected threshold).

## 7. Operation types for keyword graphs

Keyword graphs are defined by their graphical presentation approaches and the types of interactive operations that can be performed on the graphs. In this section, we deal with these interactions. Some interactive operations on the keyword graphs have already been discussed in the previous examples. Now we will give a more systematical overview on several types of useful interactions. A first interaction class relates to diverse *presentation options* for the graphs. It includes such operations as sorting (e.g. different aspects of quality measures), expanding or collapsing, filtering or finding, zooming or unzooming nodes or edges.

*Browsing operations* enable access to the underlying document collections. A concept set corresponds to a query which can be forwarded to the collection retrieving those documents (or their titles as a first summary information) which include all the concepts of the set. Therefore, each concept set appearing in a graph can be activated for browsing purposes. Moreover, derived sets based on set operations (e.g. difference and intersection) can be activated for retrieval.
*Search operations* define new search tasks related to nodes or associations selected in the graph. A graph presents the results of a (former) seach task and thus puts together sets of concepts or sets of associations. In a graphical user interface (GUI), the user can specify the search constraints: syntactical, background, quality and redundancy constraints; see (Feldman, Kloesgen and Zilberstein 1997b) for details on the constraint language. The former search is now to be refined by a selection of reference sets or associations in the result graph. Some of the search constraints may be modified in the GUI for the scheduled refined search. In refinement operations, e.g., the user can increase the number of elements that are allowed in a concept set. For example, selected concept sets in figure 1 or selected associations in figure 4 can be expanded by modifying restrictions on the maximum number of elements in concept sets.

*Link operations* combine several keyword graphs. Elements in one graph are selected and corresponding elements are highlighted in the second graph. Three types of linked graphs can be distinguished: links between set graphs, between association graphs, and between set and association graphs. When linking two set graphs, one or several sets are selected in one graph and corresponding sets are highlighted in the second graph. A correspondence for sets can rely, e.g., on the intersections of a selected set with the sets in the other graph. Then all those sets are highlighted in the second graph, that have a high overlap with a selected set in the first graph. When selected elements in a set graph are linked with an association graph, associations in the second graph are highlighted which have a high overlap with a selected set. For instance, in a country graph (figure 2), all country nodes are highlighted that have a high intersection with a selected topic in a an economical topic graph.

Thus, linkage of graphs relies on the construct of a *correspondence* between two patterns (set or association). A correspondence between two sets is e.g. defined by a criterion referring to their intersection, a correspondence between a set and an association by a specialization condition for the more special association constructed by adding the set to the original association, and a correspondence between two associations by a specialization condition for an association constructed by combining the two associations.

## Conclusions

Keyword graphs have been introduced in this paper to summarize in one picture, for large collections of documents, the patterns which have been identified in a data mining search task as interesting. Interestingness is captured in a data mining task of Document Explorer by evaluating sytactical, background, quality and redundancy constraints. Further aspects of interestingness are then incorporated by the user who studies the keyword graphs and detects patterns in these graphs using her/his capacities for visual perceptions. Thus the systematical search approach of data mining is combined with a user centred approach stimulated by visualizations. Keyword graphs provide an efficient exploration tool for getting familiar with a document collection. The main benefit of these visualizations is their interactivity, i.e. the user can click on each node or edge and get the documents supporting them, or can initiate various other operations on the graphs. Exploration is further supported by linking several graphs. Thus the relevance of selected aspects of one graph can be studied in the context of another graph. Since many similar patterns will typically be found in search tasks of data mining, e.g. associations that contain similar elements, it is often very hard to make sense out of them. Therefore we introduced redundancy filters based on equivalence classes and partial orderings to extract nodes and relations for the visualizations helping us to get some order in the vast amount of mined patterns.

Document Explorer is a system available for PC with a first still qualitatively limited set of keywords graphs shown in the preceding figures. Currently we are extending the statistical quality evaluation functions, augmenting the graphical quality of the offered keyword graphs, and implementing further types and interaction operations for interactive result presentation.

## 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. 1993] Agrawal A., Imielinski T., and Swami A. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 207-216.

[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, Kloesgen and Zilberstein 1997] 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.

[Holsheimer et al. 1996] M. Holsheimer, M. Kersten, A. Siebes 1996. Data Surveyor: Searching the nuggets in parallel. In *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Cambridge, MA: MIT Press.

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

[Kloesgen 1996] Klˆsgen 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.

[Mannila et al. 1994] Mannila H., Toivonen H., and Verkamo A. Efficient Algorithms for Discovering association rules. In *KDD-94: AAAI workshop on Knowledge Discovery in Databases*, pages 181-192.

[Wills 1997] Wills G. Visual Exploration of Large Structured Data Sets. In: *New Technologies and Techniques for Statistics*, ed. by Eurostat, Amsterdam: IOS Press.