In Proceedings of the International Symposium on Document Analysis and Information Retrieval, 1996

Keyword-Based Browsing and Analysis of Large Document Sets

Ido Dagan and Ronen Feldman
Math and Computer Science Dept.
Bar-Ilan University
Ramat-Gan, ISRAEL
{feldman,dagan}@bimacs.cs.biu.ac.il

Haym Hirsh
Dept. of Computer Science
Rutgers University
Piscataway, NJ USA 08855
hirsh@cs.rutgers.edu

Abstract

Knowledge Discovery in Databases (KDD) focuses on the computerized exploration of large amounts of data and on the discovery of interesting patterns within them. While most work on KDD has been concerned with structured databases, there has been little work on handling the huge amount of information that is available only in unstructured textual form. This paper describes the KDT system for Knowledge Discovery in Texts. It is built on top of a text-categorization paradigm where text articles are annotated with keywords organized in a hierarchical structure. Knowledge discovery is performed by analyzing the co-occurrence frequencies of keywords from this hierarchy in the various documents. We show how this term-frequency approach supports a range of KDD operations, providing a general framework for knowledge discovery and exploration in collections of unstructured text.

Introduction

Traditional databases store large collections of information in the form of structured records, and provide methods for querying the database to obtain all records whose content satisfies the user's query. More recently, however, researchers in Knowledge Discovery in Databases (KDD) have provided a new family of tools for accessing information in databases (e.g. Brachman et al, 1993; Frawley et al, 1991; Kloesgen, 1992; Kloesgen, 1995b; Ezawa and Norton, 1995). The goal of KDD has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from given data" (Piatetsky-Shapiro and Frawley 1991). Work in this area includes applying machine-learning and statistical-analysis techniques towards the automatic discovery of patterns in databases, as well as providing user-guided environments for exploration of data.

Traditional database query tools allow a user to retrieve records based on the content of each record in isolation. In a hospital database, for example, a user might request all records for hospital stays that are less than one day with a cost greater than $10,000. Each retrieved record is selected because the information in that record, independent of any other record, satisfies the user's query. In contrast, KDD work provides tools for accessing information based on patterns appearing *across* records. For example, KDD tools might provide a user the ability to ask for records of patients whose medical care for some illness is much higher than typical (where "typical" is implicitly defined by the values of other records in the database), or to investigate if there exist some statistical patterns relating the length of patients' hospital stay and their family circumstances (whether the patient is married, how many children the patient has, etc.).

Although the goal of KDD work is to provide access to patterns and information in online information collections, most efforts have focused on knowledge discovery in structured databases. However, a tremendous amount of online information appears only in collections of unstructured text. Most research in Information Retrieval (IR) has developed methods for providing access to documents based on the information contained in a document in isolation (analogous to what traditional database query tools provide for databases). In this case, it is assumed that the user knows in advance the topic of documents of interest. Clustering methods were used to impose structure over a collection of documents, enabling the user to browse through the collection and select clusters of documents of interest (e.g. Salton 1989; Cutting et al, 1993). Visualization methods were also used for presenting some additional structures hidden in a document or a set of documents (Williamson and Shneiderman, 1992; Hearst 1995). However, there has been little work on providing KDD-style tools for browsing and analyzing text collections based on information appearing *across* documents. Applying such tools to texts means that the system would take an active role in suggesting topics of interest to the user, as well as supply new browsing methods that rely on inter-document information. A KDD framework for texts 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 makes unsupervised decisions in finding specific statistically motivated patterns.

This paper describes the Knowledge Discovery in Texts (KDT) system, which applies a novel knowledge discovery framework to textual databases. Our goal is to provide similar types of KDD operations previously provided for structured databases. To do so, we rely on a text-categorization paradigm where each document is labeled with sets of keywords, where each keyword comes from a hierarchy of terms. Unlike in traditional IR work, where keywords (category labels) are used in specification of retrieval (or routing) queries, KDT allows a user to access documents and recognize patterns across them based on the observed co-occurrence distributions of keywords in documents of the collection. A key insight in this work is that keyword co-occurence frequencies (or distributions) can provide the foundation for a wide range of KDD operations on collections of textual documents, including:

1. Summarization and Browsing: KDT allows the user to view the frequency of occurrence of keywords from some category in a collection of documents that contain particular keywords from some other category, and to browse the collection of documents based on these frequencies.

2. Comparing Document Distributions: KDT can compare the distributions of keywords in two collections of documents containing similar keywords and display the results using tables and graphs.
3. Trend Analysis: KDT can compare the distributions of keywords in documents from different points in time and display the results using tables and graphs.
4. Association Discovery: KDT can search for several types of associations (e.g. Toivonen et al, 1995) between classes of documents.
5. Further, KDT includes a browsing facility in which the user can click on any discovered pattern and get the list of documents that contributed to the pattern.

These operations can assist users that have to analyze and assimilate information spanning over a large number of documents, such as in business intelligence and economical analysis. For example, using the system an analyst can find out quickly the most active economical areas for certain countries, or major products of companies. Furthermore, the analyst can compare such a company “profile” to profiles of other companies in the same business area, and discover distinguishing aspects in the activity of the company. In business intelligence applications, the user may be interested in comparing profiles of different companies to identify relatively weak and strong areas in their activity, while in marketing applications an analyst may want to compare country profiles when looking for appropriate international markets for a product. Other types of KDT queries can answer questions like “find economical areas which are dominant in the economies of some (unspecified) countries”, or “find economical areas in which activity has increased or decreased in a specified period of time”. Investigation of issues such as mentioned above is not supported directly in conventional information retrieval systems, and typically requires a lot of manual effort in retrieving and analyzing a large number of documents.

It should be kept in mind that the answers to all KDT queries rely on document frequencies in some information source (such as a newswire or a professional magazine), which may introduce quantitative biases with respect to the real situation described in the texts. For example, an interesting story, from the media’s point of view, may be covered in a large number of articles, inflating the statistics of some items. To support verification of KDT’s finding, and for gaining further insights into them, the system provides a direct link from the results of each query to the documents which support that result.

Our general KDT framework was initially presented in (Feldman and Dagan 1995). The current paper both extends the KDT framework and its set of operations, and presents our interactive prototype system, which was implemented in Visual Prolog under Microsoft MS-Windows (all subsequent figures are screen dumps of this system). The system’s mode of operation involves three major steps:

1. Load input documents, annotated with keywords selected from a pre-existing hierarchy of meaningful category labels.
2. Compute the various co-occurrence frequencies of these keywords within the documents in the collection (typically performed as a pre-processing step).
3. Provide interactive tools that allow access to documents, discover patterns across documents, and perform other similar KDD operations, based on the co-occurrence frequencies computed in the previous step.

Keyword Tagging and the Keyword Hierarchy

Applying KDD operations to texts requires that documents will be represented in some structured way. We chose to base the current version of the system on the very simple representation scheme of annotating (or tagging) each document with a set of category label keywords. Category labels are commonly used in commercial and scientific text collections and information feeds, and provide a high level summary for the content of the document. For example, articles in hi-tech domains may be annotated with sets of keywords such as {IBM, product announcement, Power PC} and {Motorola, patent, cellular phone}. The annotation of documents with category labels may be either manual or automatic. Automatic text categorization has recently been the focus of substantial research in the IR and text processing communities (e.g. Apte et al 1994; Finch 1994; Iwayama and Tokunaga 1994). Altogether, we assume that having the documents of the collection annotated with category labels is a reasonable pre-requisite for the KDT system, which would hold for many text collections in the market.

KDT also requires that the category keywords would be organized in a hierarchical structure. This keyword hierarchy is a directed acyclic graph (DAG) of terms, where each of the terms is identified by a unique name. Figure 1 shows a portion of an example keyword hierarchy, the one used in our work with the Reuters data (see below), which will serve as a running example throughout this paper. In such a hierarchy an arc from A to B denotes that A is a more general term than B (i.e., *countries* → *G7* → *Japan*). We use a general DAG rather then a tree structure so that a keyword may belong to several parent nodes (e.g. *Germany* is both a *European-Community* and a *G7* country).

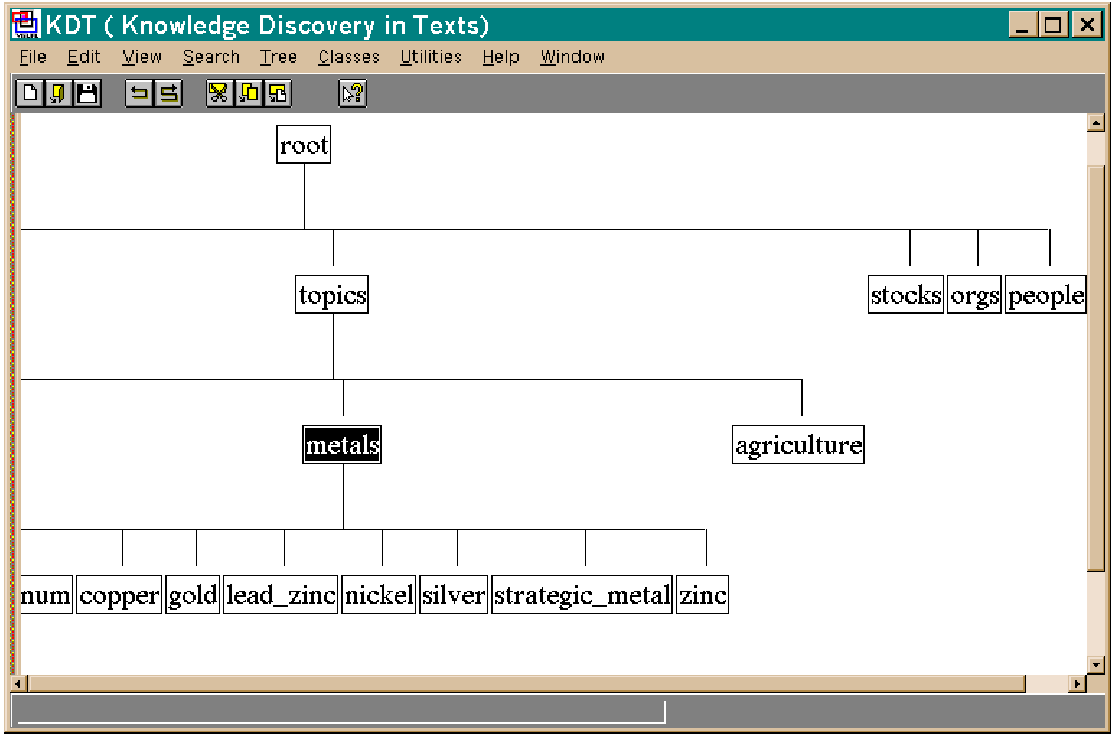


Figure A - Keyword Hierarchy for Reuters Data

It should be emphasized that the sole purpose of the keyword hierarchy is to enable generalizations and partitioning of KDD findings over sibling nodes. The structure of the hierarchy is typically simple, and reflects the basic generalizations common for the domain of interest. Such keyword hierarchies are commonly used by information providers (e.g. the Dialog service of Knight Ridder Information Inc. or the First service of Individual Inc.), and resemble in their form to a "subject index" in a yellow pages book. Rich hierarchies have been developed for several professional domains, such as the Medical Subject Heading (MeSH) hierarchy, and have been used to assist and augment free-text searching. The task of constructing, obtaining and modifying such hierarchies is thus relatively easy, and should not be confused with the task of constructing a semantically rich structure, such as a semantic network or a taxonomy in the "knowledge representation" sense. The KDT system provides a simple GUI for constructing and editing the hierarchy, supporting additions, deletions and modifications of nodes and links (Figure A is a screen dump of the hierarchy maintenance editor).

The Text Collection

As mentioned above, the KDT system expects as input documents which are annotated with category labels, where annotation might be achieved either manually or automatically. In the experiments described here we used the Reuters-22173 text categorization test collection, containing about 22,000 articles, totaling 25 megabytes. The documents in this collection appeared on the Reuters news wire in the late 1980's, and were assembled and indexed with categories by personnel from Reuters Ltd. and Carnegie Group, Inc. Further formatting and data file production was done in 1991 and 1992 by David D. Lewis and Peter Shoemaker.

The categories in this collection are classified only to five types of tags: countries, topics, people, organizations and stock exchanges. These five types provided us the skeleton of the keyword hierarchy, where each of the 5 types serves as an intermediate node in a two level hierarchy. We then enriched the hierarchy with some additional sub-types of categories, such as *agriculture* and *metals* as daughters of the *topics* node, and various international organizations (taken from the CIA Factbook on the Internet) as daughters of the *countries* node.

Keyword Co-Occurrence Distributions

All KDD operations supported by the KDT system are based on an analysis of the keywords that annotate the articles in the collection. More specifically, KDT computes the distribution of daughter terms relative to their siblings for all keywords in the hierarchy. For example, the annotations of documents with daughters of the keyword node *computers* may be distributed as follows: *mainframes*: 0.1; *work-stations*: 0.4; *PCs*: 0.5. In formal terms, we set a node C in the hierarchy to specify a discrete random variable whose values are denoted by its daughters, where each occurrence of a daughter provides a data point. We denote the distribution of the random variable by $P(C=c)$, where c ranges over all daughters of C. The event $C=c$ corresponds to the annotation of a document with the daughter category c. $P(C=c_i)$ is the proportion of annotations of documents with c_i among all annotations of documents with any daughter of C. In the example above we would say that $P(C=mainframes)=0.1$, where C denotes the random variable which corresponds to the node *computers*.

In KDT we are most interested in *conditional* keyword distributions of the form *P(C=c|x)*, where *x* is a conditioning event which denotes some other category keyword. Such distributions describe the co-occurrence of the category *x* with all daughters of *C*.

Figure B shows an example for such a distribution, where *C* stands for the node *topics* and *x* stands for *Argentina*. In other words, the figure presents the distribution of topic keywords (i.e., keywords that are daughters of the *topics* node) in articles that are annotated also with the keyword *Argentina*. In Figure B the distribution is presented as a pie-chart, along with the absolute frequency of each slice in the pie: 12 articles among all articles of *Argentina* are annotated with *sorghum*, 20 with *corn*, 32 with *grain*, etc. The KDT system presents distributions in several forms, graphical (e.g. bar-chart) or alphanumeric (see Figure C), listing absolute frequencies or probabilities (percentage).

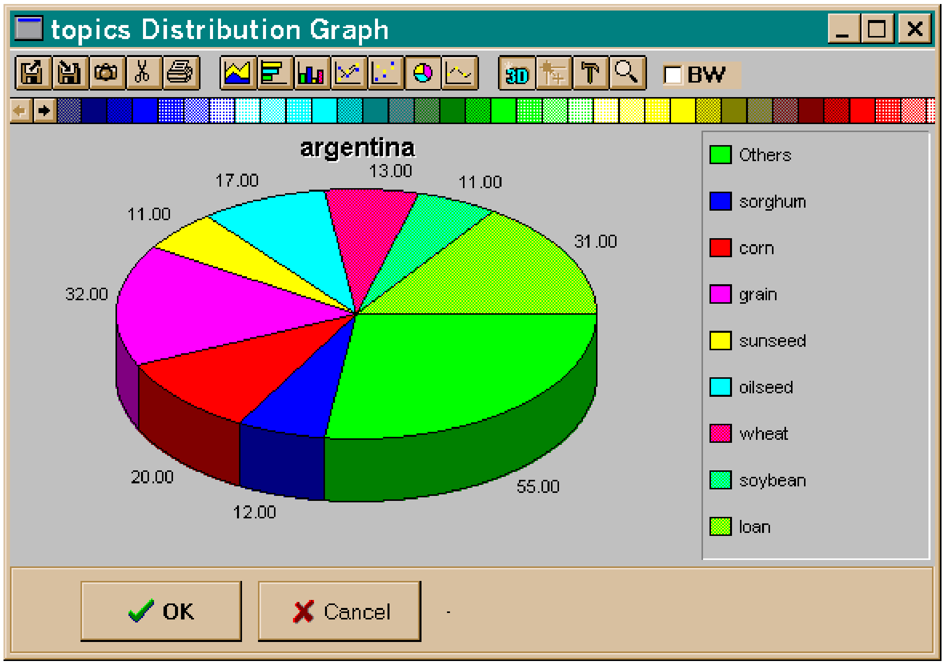
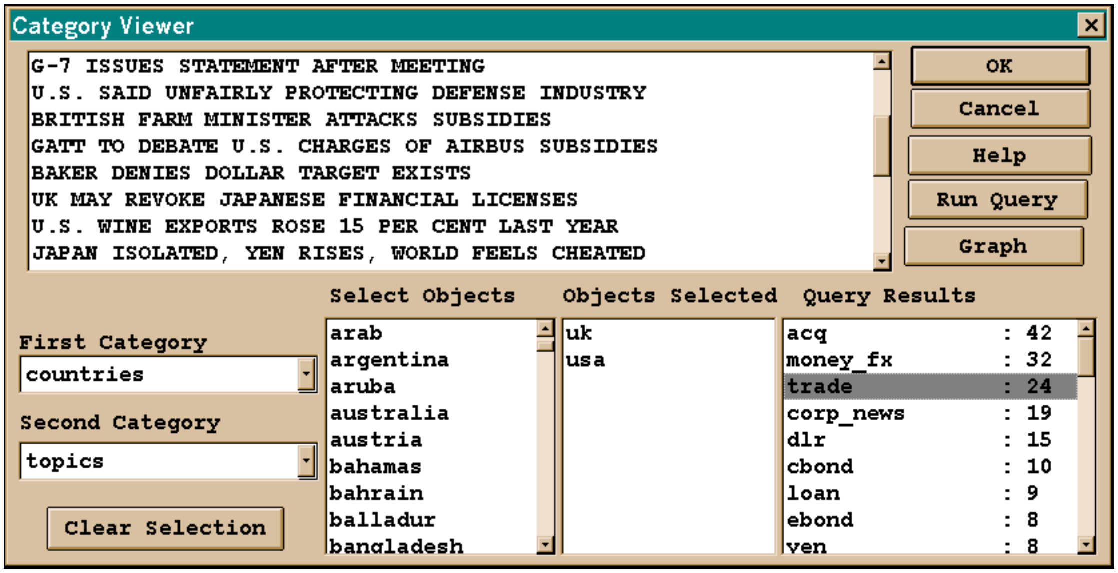


Figure B - Graphical representation of the Topic Distribution of Argentina

More generally, a keyword co-occurrence distribution may be conditioned by the joint occurrence of several category keywords, and not just one. For example, Figure C displays the distribution *P(C=c|x,y)*, where *C* stands for *topics*, *x* for *UK*, and *y* for *USA*. In other words, this is the distribution of topics in articles that deal with both *UK* and *USA*. The distribution is presented in the lower right window of the screen.

By letting the user specify and display conditional keyword co-occurrence distributions, as in Figure 2 and Figure C, the KDT system provides a powerful browsing mechanism for large subsets of documents. A traditional document retrieval system enables the user to ask for all documents containing the keywords *UK* and *USA*, but then presents the entire set of matching documents without describing its internal structure. Typically, the documents will be sorted by either relevance score, which would be determined in this case by the frequency and position of the given keywords in the document, or by chronological order. The KDT system, on the other hand, enables the user to investigate the contents of this document set by sorting it according to the daughter distribution of any node in the hierarchy, such as topics, countries, companies etc. Once the documents are sorted, and the distribution is displayed, the user can access the specific documents of each subgroup. In Figure C, for example, the user chose to click on the 24 documents annotated with *trade*, which led to the display of all titles of these documents (those annotated by *UK*, *USA*, and *trade*) in the upper window of the screen.



```
G-7 ISSUES STATEMENT AFTER MEETING
U.S. SAID UNFAIRLY PROTECTING DEFENSE INDUSTRY
BRITISH FARM MINISTER ATTACKS SUBSIDIES
GATT TO DEBATE U.S. CHARGES OF AIRBUS SUBSIDIES
BAKER DENIES DOLLAR TARGET EXISTS
UK MAY REVOKE JAPANESE FINANCIAL LICENSES
U.S. WINE EXPORTS ROSE 15 PER CENT LAST YEAR
JAPAN ISOLATED, YEN RISES, WORLD FEELS CHEATED
```

First Category: countries

Second Category: topics

Select Objects	Objects Selected	Query Results	
arab	uk	acq	: 42
argentina	usa	money_fx	: 32
aruba		trade	: 24
australia		corp_news	: 19
austria		dlr	: 15
bahamas		cbond	: 10
bahrain		loan	: 9
balladur		ebond	: 8
bangladesh		yen	: 8

Figure C - Viewing the Topic distribution of USA and the UK

Distribution Comparison

So far we have seen that the ability to specify keyword co-occurrence distributions provides the user with a useful mechanism for exploring subsets of documents. Taking a KDD perspective, we are interested not only in displaying an entire distribution to the user but also in identifying specific points in a distribution which are likely to be “interesting”.

We suggest to quantify the degree of “interest” of some data by comparing it to a given, or an “expected”, model. For example, we may want to compare the data regarding *IBM* to a model constructed by some averaging of the data regarding other computer manufacturers. Alternatively, we may want to compare the data regarding IBM in the last year to a model constructed from the data regarding IBM in previous years.

In our case, we use keyword distributions to describe the data. We therefore need a measure for comparing the distribution defined by the data to a model distribution. We chose to use the relative entropy measure (or Kullback-Leibler (KL) distance), defined in information theory, though we plan to investigate other measures as well. The KL-distance seems to be an appropriate measure for our purpose since it measures the amount of information that we lose if we model a given distribution p by another distribution q. Denoting the distribution of the data by p and the model distribution by q, the distance from p(x) to q(x) measures the amount of “surprise” in seeing p while expecting q. Formally, the relative entropy between two probability distributions p(x) and q(x) is defined as:

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(x)}{q(x)}$$

The relative entropy is always non-negative and is 0 if and only if *p=q*.
According to this view, interesting distributions will be those with a large distance to the model distribution. Interesting data points will be those that make a big contribution to the distance between the given distribution and the model (i.e., *x*'s whose contribution to the sum is large). The following sections show how various interesting patterns can be identified by measuring the relative entropy distance between a distribution and different baseline models.

Finding Strong Associations

A substantial part of the KDD literature deals with finding strong statistical associations (or correlations) between data elements in the collection (e.g. Toivonen et al 1995; Kloesgen 1995a; Feldman et al, 1996). Such associations were used for various applications, including:

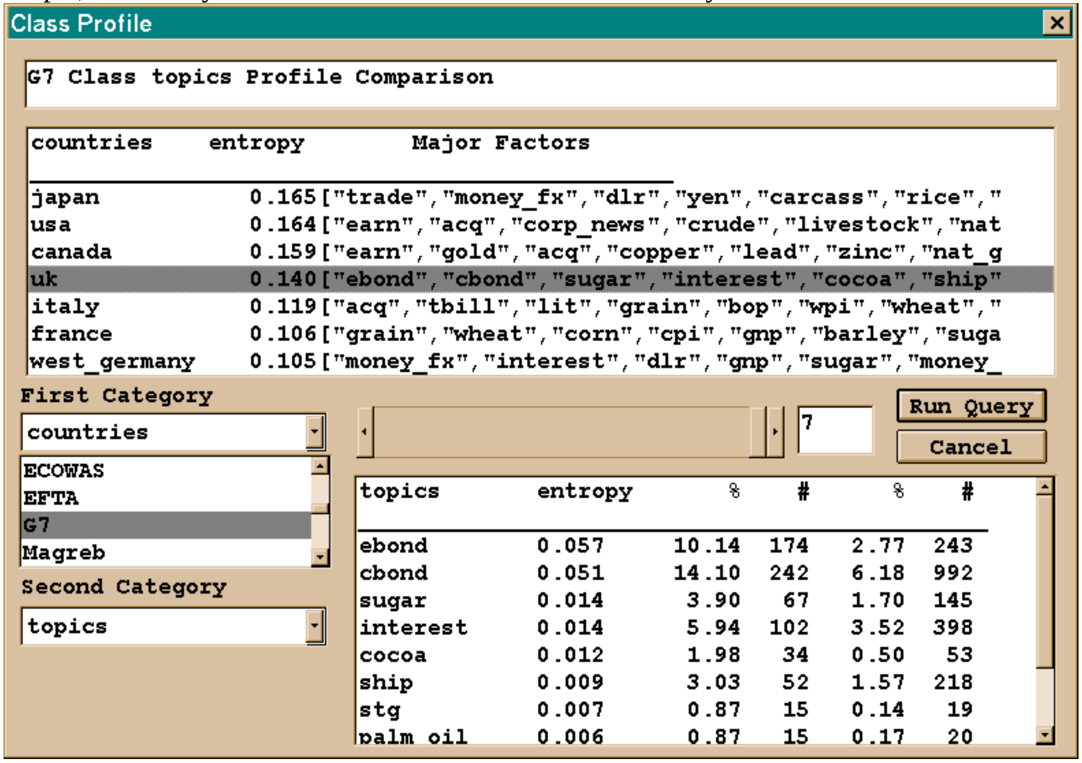
- Supermarket shopping list: finding correlations between user purchase preferences
- Identifying telecommunications alarm rules, as associations between system attributes and faults

In the KDT context, we are interested in finding statistical associations between various keywords. For example, we may identify the economical topics which are highly associated with a certain country. The comparative approach of the previous section enables us to focus on associations that are likely to be interesting, i.e. those associations that deviate from a baseline model. For example, we will give a higher rank to an association between a country and a topic only if this association is not typical for other countries as well.

Associations relative to a class

Consider a conditional distribution of the form *P(C=c | x_i)*, where x_i is a conditioning concept. In many cases, it is natural to expect that this distribution would be similar to other *distributions* of this form, in which the conditioning event is a sibling of x_i. For example, when *C* denotes the node *commercial-activity*, and *x_i=Ford* (the car manufacturer), we could expect a distribution that is quite similar to such distributions where the conditioning concept is another car manufacturer (a sibling of *Ford* in the hierarchy). To capture this reasoning, we use *Avg P(C=c | x)*, the *average sibling distribution*, as a model for *P(C=c | x_i)*, where *x* ranges over all siblings of x_i (including x_i itself). In the above example, we would measure the distance from the distribution *P(C=c | Ford)* to the average distribution *Avg P(C=c | x)*, where *x* ranges over all car manufacturers and *C* denotes the node *commercial-activity*. The distance between these two distributions would be large if the activity profile of Ford differs a lot from the average profile of other car manufacturers. Furthermore, specific points in the distribution (specific activities) that make a large contribution to the distance are activities which are associated with Ford much more than with other car manufacturers.

Figure D demonstrates this type of comparison, between the topic distribution of each G7 country and the average sibling distribution of topics for all G7 countries. The countries are sorted in decreasing order of their distance to the average distribution, revealing that Japan is the most "atypical" G7 country (with respect to its topic distribution) while Germany is the most typical one. The topics that made the largest contributions to the distance for each countries are also displayed. The user can then click on any class member and get an expanded view of the comparison between the topic distribution of this member and the average distribution. In Figure D we have expanded the topic list of the UK (at the bottom-right list box), providing the statistical detail for the strong associations between the UK and topics like bonds, sugar, cocoa etc. In addition to their value in finding associations, comparisons of this type provide a hierarchical browsing mechanism for keyword co-occurrence distributions. For example, an analyst that is interested in studying the topic distribution in articles dealing with G7 countries may first browse the average class distribution for G7, using a presentation as in Figures 2,3. This will reveal the major topics that are generally common for G7 countries. Then, the presentation of Figure D would reveal the major characteristics which are specific for each country.



G7 Class topics Profile Comparison

countries	entropy	Major Factors
japan	0.165	["trade","money_fx","dlr","yen","carcass","rice","
usa	0.164	["earn","acq","corp_news","crude","livestock","nat
canada	0.159	["earn","gold","acq","copper","lead","zinc","nat_g
uk	0.140	["ebond","cbond","sugar","interest","cocoa","ship"
italy	0.119	["acq","tbill","lit","grain","bop","wpi","wheat","
france	0.106	["grain","wheat","corn","cpi","gnp","barley","suga
west_germany	0.105	["money_fx","interest","dlr","gnp","sugar","money_

First Category

countries

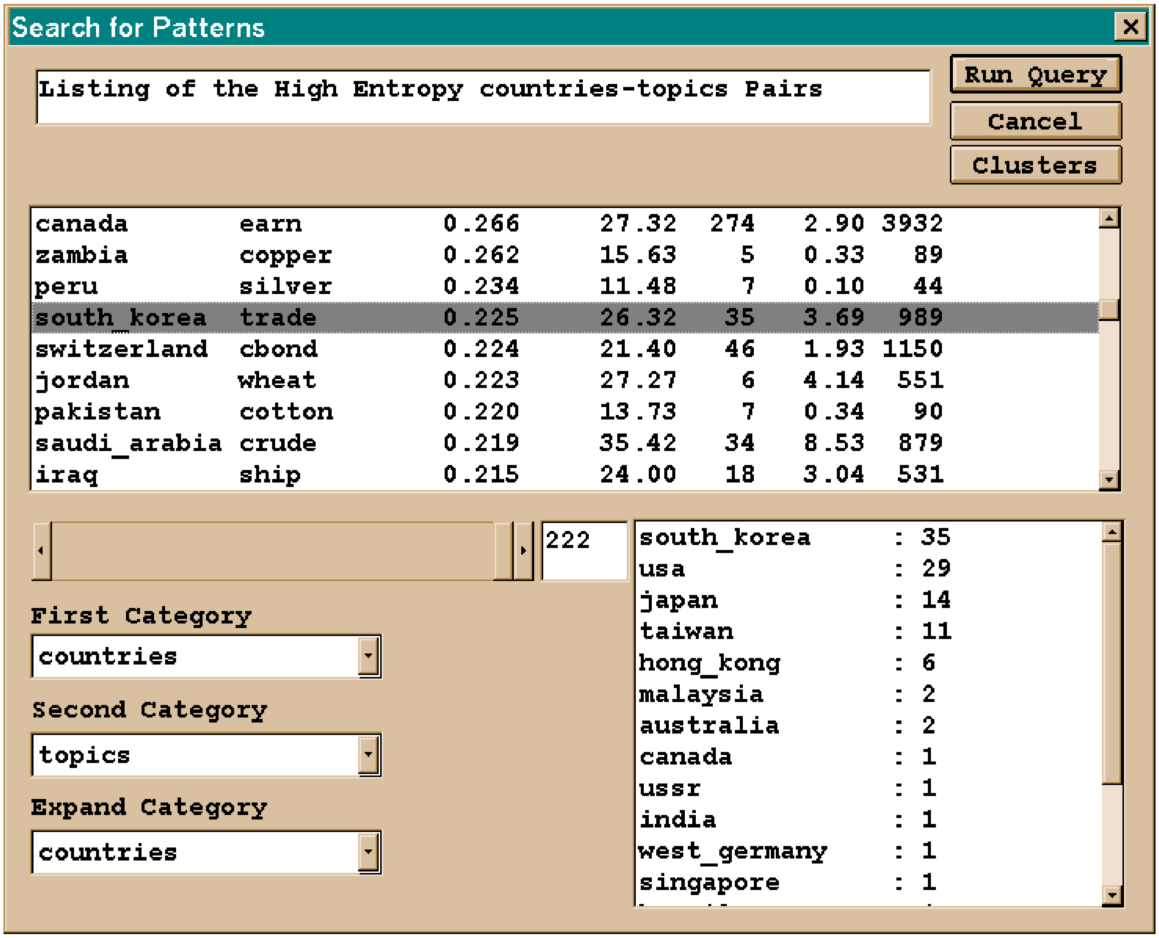
Second Category

topics

7

topics	entropy	%	#	%	#
ebond	0.057	10.14	174	2.77	243
cbond	0.051	14.10	242	6.18	992
sugar	0.014	3.90	67	1.70	145
interest	0.014	5.94	102	3.52	398
cocoa	0.012	1.98	34	0.50	53
ship	0.009	3.03	52	1.57	218
stg	0.007	0.87	15	0.14	19
palm oil	0.006	0.87	15	0.17	20

Figure D - Comparison of the topic distribution of members of the G7 organization vs. the average topic distributions of the G7. Entries in the top listbox are sorted in decreasing order of their relative entropy distance to the average topic distribution (2nd column). The 3rd column shows the major topics that contributed to that distance. In the lower-right listbox, we can see a detailed information about these topics, for a selected country (UK). The 2nd column shows the contribution of the topic to the relative entropy distance. The 3rd and 5th columns show respectively, the percentage that the topic takes from the topic distribution of the specific country (3rd) and from the average topic distribution of the G7 countries (5th). The 4th and 6th columns show, respectively, the total number of articles in which the topic appears with the specific country(4th), and with any G7 country(6th).



Listing of the High Entropy countries-topics Pairs

```
canada        earn      0.266   27.32  274   2.90 3932
zambia        copper    0.262   15.63    5   0.33   89
peru          silver    0.234   11.48    7   0.10   44
south_korea   trade     0.225   26.32   35   3.69  989
switzerland   cbond     0.224   21.40   46   1.93 1150
jordan        wheat     0.223   27.27    6   4.14  551
pakistan      cotton    0.220   13.73    7   0.34   90
saudi_arabia  crude     0.219   35.42   34   8.53  879
iraq          ship      0.215   24.00   18   3.04  531
```

222

First Category: countries

Second Category: topics

Expand Category: countries

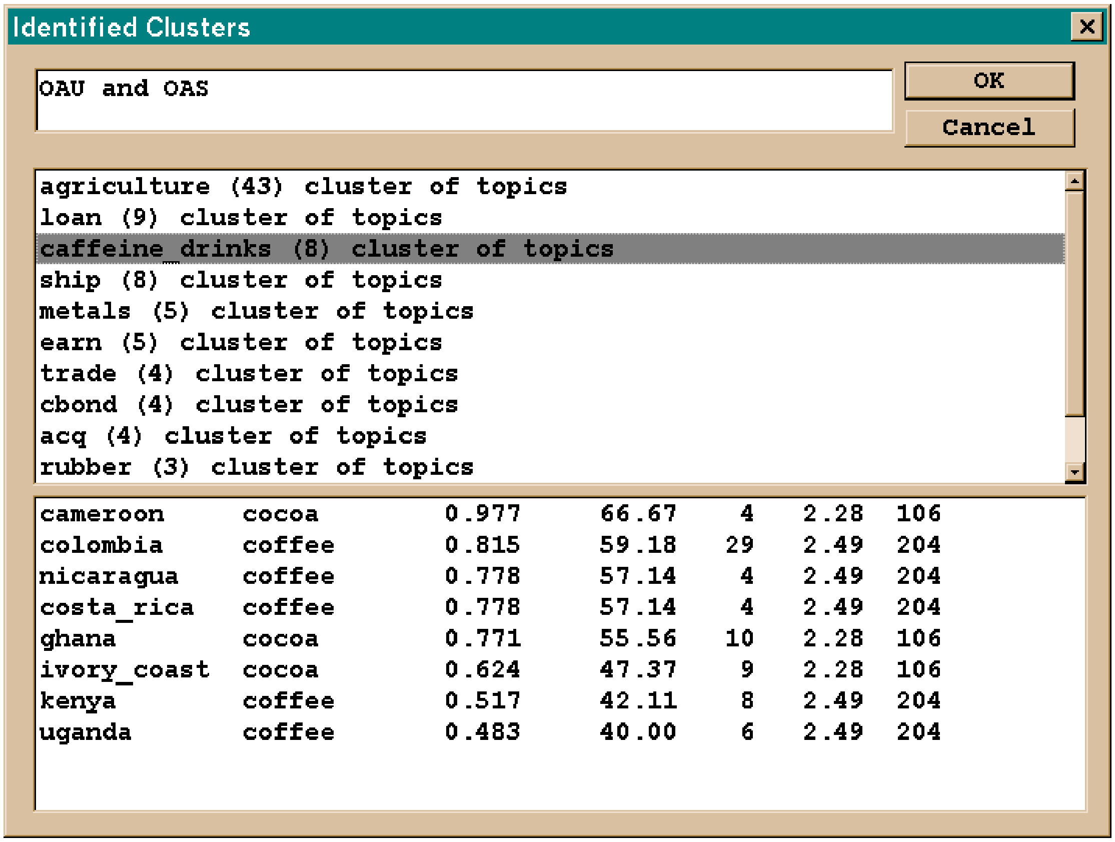
```
south_korea   : 35
usa           : 29
japan         : 14
taiwan        : 11
hong_kong     : 6
malaysia      : 2
australia     : 2
canada        : 1
ussr          : 1
india         : 1
west_germany  : 1
singapore     : 1
```

Figure E - Country-Topic associations with a high contribution to the relative entropy distance between the topic distribution of the country and the average topic distribution for all countries. Associations are sorted in decreasing order of the relative entropy distance to the global average (3rd column). The 4th and 6th columns show, respectively, the percentage that the topic takes from the topic distribution of the specific country (4th) and from the average topic distribution of all countries (6th). The 5th and 7th columns show, respectively, the total number of articles in which the topic appears with the specific country(5th) and with any country(7th).

General associations

Another form of association can be defined by taking as the baseline model the average distribution of the conditioned category over all possible instantiations of the conditioning category (in the formulation of the previous sub-section, *x* would range over all categories of the same type, rather than over all immediate siblings). This form is demonstrated in Figure E, which lists the strongest associations found between some country and some topic. The system also enables the user to investigate further the subset of documents which corresponds to a certain association. In Figure E we chose to explore the set of documents corresponding to the association between *South Korea* and *trade*, presenting the distribution of countries within this set (lower-right listbox, specified by the “Expand Category” pull-down menu). This reveals which countries are most prominent in articles dealing with both *South Korea* and *trade*, conveniently linking the browsing mechanism of Figure C to the association display screen.

In many cases, the system generates a very large number of associations, making it difficult to draw overall conclusions. To summarize the information, the system groups together correlations whose second component belongs to the same class in the hierarchy. Figure F shows the clusters that were formed by the system when grouping all the individual associations of Figure E. For example, in 43 associations of Figure E the right hand side of the association (the topic) was a daughter of the node *agriculture*. The user can examine any cluster and see the specific associations it contains (lower listbox, for the selected cluster *caffeine-drinks*). In addition, the system tries to provide a compact generalization for all the categories on the left hand side of the associations in the cluster. In our example, the system found that all countries that are highly correlated with caffeine drinks belong either to the OAU (African Union) or the OAS (South American countries) organizations.



Identified Clusters

OAU and OAS

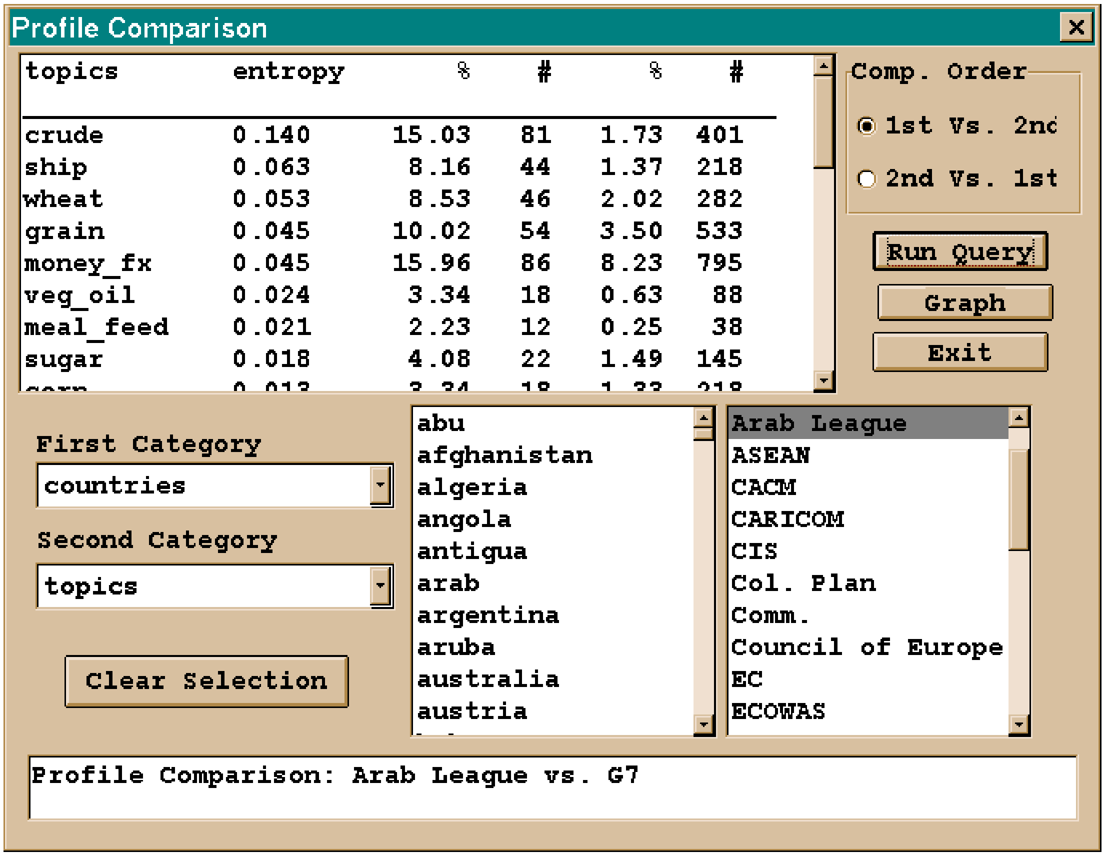
```
agriculture (43) cluster of topics
loan (9) cluster of topics
caffeine_drinks (8) cluster of topics
ship (8) cluster of topics
metals (5) cluster of topics
earn (5) cluster of topics
trade (4) cluster of topics
cbond (4) cluster of topics
acq (4) cluster of topics
rubber (3) cluster of topics
```

```
cameroon     cocoa      0.977    66.67    4    2.28   106
colombia     coffee     0.815    59.18   29    2.49   204
nicaragua    coffee     0.778    57.14    4    2.49   204
costa_rica   coffee     0.778    57.14    4    2.49   204
ghana        cocoa      0.771    55.56   10    2.28   106
ivory_coast  cocoa      0.624    47.37    9    2.28   106
kenya        coffee     0.517    42.11    8    2.49   204
uganda       coffee     0.483    40.00    6    2.49   204
```

Figure F - Clustering associations using the category hierarchy. In the upper listbox we can see all association clusters that were formed by the system along with their sizes (in parenthesis). In the lower listbox we see the members of the cluster that was selected in the upper listbox (caffeine drinks).

Specific comparisons

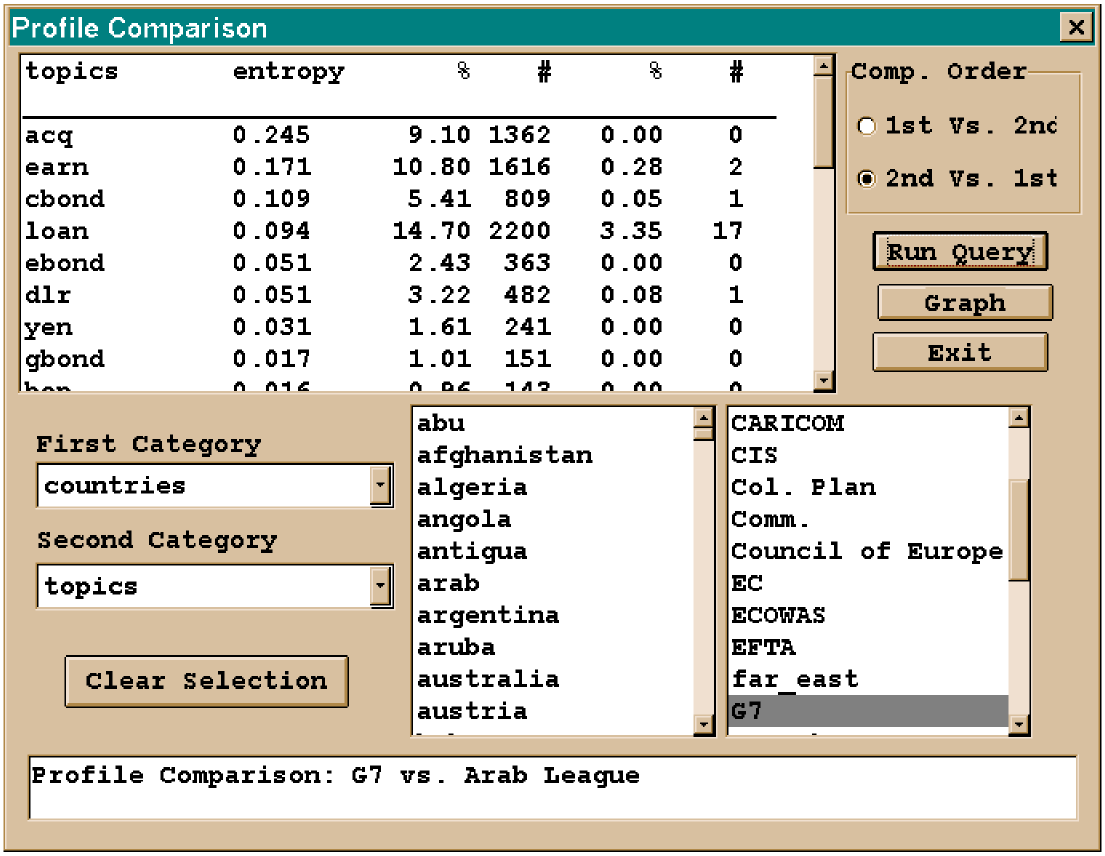
The mechanism for identifying strong associations relative to a model is also useful for comparing conditional distributions of two specific nodes in the hierarchy. In Figure G we measure the distance from the average topic distribution of Arab League countries to the average topic distribution of G7 countries. This reveals the topics with which Arab League countries are associated much more than G7 countries, like crude-oil and wheat. Figure H shows the comparison in the opposite direction, revealing the topics with which G7 countries are highly associated relative to the Arab League.



topics	entropy	%	#	%	#
crude	0.140	15.03	81	1.73	401
ship	0.063	8.16	44	1.37	218
wheat	0.053	8.53	46	2.02	282
grain	0.045	10.02	54	3.50	533
money_fx	0.045	15.96	86	8.23	795
veg_oil	0.024	3.34	18	0.63	88
meal_feed	0.021	2.23	12	0.25	38
sugar	0.018	4.08	22	1.49	145
corn	0.013	3.34	18	1.33	218

Profile Comparison: Arab League vs. G7

Figure G - Topics Profile Comparison of the Arab League countries vs. the G7 countries. Entries in the top listbox are sorted in decreasing order of their contribution to the relative entropy distance (2nd column). The 3rd and 5th columns show, respectively, the percentage of the topic in the average topic distribution of the Arab League countries and in the average topic distribution of the G7 countries. The 4th and 6th columns show, respectively, the total number of articles in which the topic appears with any Arab League country and any G7 country.



topics	entropy	%	#	%	#
acq	0.245	9.10	1362	0.00	0
earn	0.171	10.80	1616	0.28	2
cbond	0.109	5.41	809	0.05	1
loan	0.094	14.70	2200	3.35	17
ebond	0.051	2.43	363	0.00	0
dlr	0.051	3.22	482	0.08	1
yen	0.031	1.61	241	0.00	0
gbond	0.017	1.01	151	0.00	0
hon	0.016	0.96	143	0.00	0

Comp. Order

○ 1st Vs. 2nd

● 2nd Vs. 1st

First Category: countries

Second Category: topics

abu
afghanistan
algeria
angola
antigua
arab
argentina
aruba
australia
austria

CARICOM
CIS
Col. Plan
Comm.
Council of Europe
EC
ECOWAS
EFTA
far_east
G7

Profile Comparison: G7 vs. Arab League

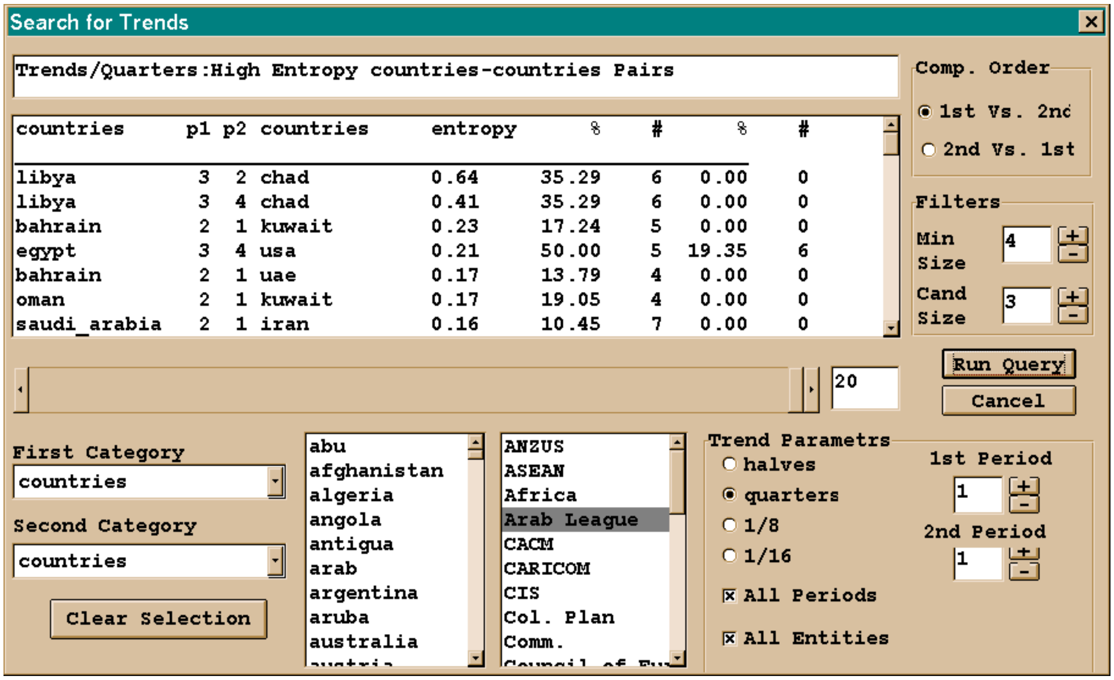
Figure H - Topics Profile Comparison of the G7 countries vs. the Arab League countries. The columns in the upper listbox are the same as in Figure G.

Finding Trends Over Time

One of the most important needs of an analyst is the ability to follow changes over time in the behavior of entities of interest. For example, a trend analysis tool should be able to compare the activities that a company performed in some domain in the past with its current activities in that domain. For example, a possible conclusion from such an analysis would be that a company is shifting its activities from one domain to another.

The KDT system identifies trends by comparing a distribution of data taken from one period of time to a corresponding model distribution which is constructed from data of another period. Trends are then discovered by searching for significant deviations from the expected model, as before. Figure 6 lists trends that were identified across the different quarters of the year. The program was directed to search for significant changes in the co-occurrence distributions of Arab League countries with any other country. For example, the first line of the top listbox shows that in the 3rd quarter there was a large increase in the proportion of articles that mention both Libya and Chad among all articles mentioning Libya (from 0% in the 2nd quarter to 35.29% in the 3rd quarter). The second line shows that the proportion of such articles in the 3rd quarter was also much higher than in the fourth quarter (a decrease over time, again to 0%). An analyst might then want to investigate what happened in the 3rd quarter regarding Libya and Chad. To facilitate such investigation, the system provides access to the specific articles that support the trend, by double clicking on the appropriate line.

Then, a listbox containing all titles of the relevant documents appears, as in Figure J, revealing that the cause for the trend was the fighting between Libya and Chad at that period.



Search for Trends

Trends/Quarters:High Entropy countries-countries Pairs

```
countries      p1 p2 countries      entropy    %     #      %     #

libya           3  2 chad           0.64     35.29   6    0.00    0
libya           3  4 chad           0.41     35.29   6    0.00    0
bahrain         2  1 kuwait         0.23     17.24   5    0.00    0
egypt           3  4 usa            0.21     50.00   5   19.35    6
bahrain         2  1 uae            0.17     13.79   4    0.00    0
oman            2  1 kuwait         0.17     19.05   4    0.00    0
saudi_arabia    2  1 iran           0.16     10.45   7    0.00    0
```

Comp. Order: ● 1st Vs. 2nd ○ 2nd Vs. 1st

Filters: Min Size 4; Cand Size 3

20

First Category: countries

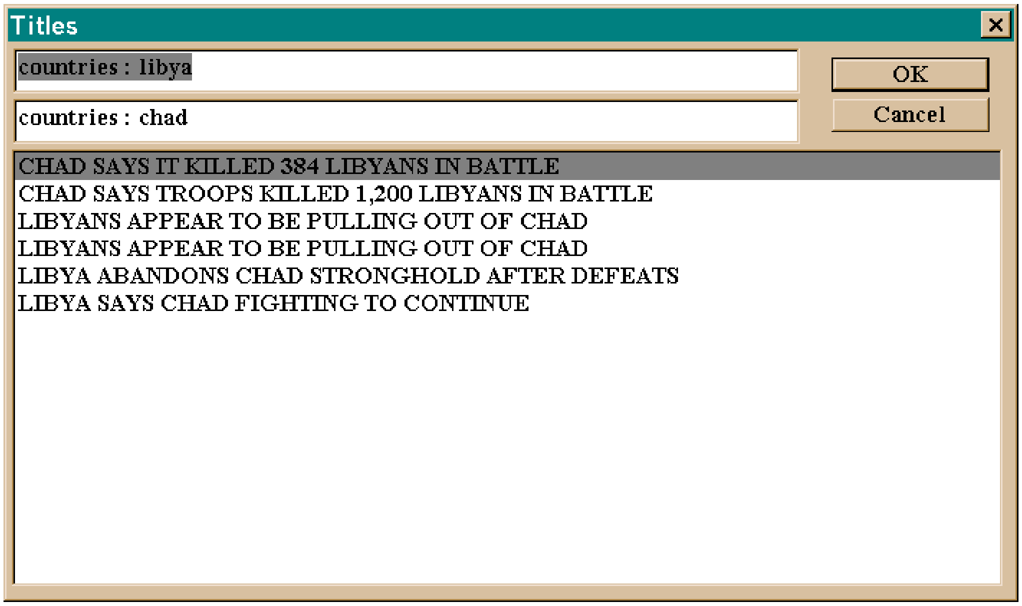
Second Category: countries

Trend Parametrs: ○ halves ● quarters ○ 1/8 ○ 1/16 ☒ All Periods ☒ All Entities

1st Period: 1

2nd Period: 1

Figure I - Trends in co-occurrence of Arab League countries with other countries. The distance is measured from the period (quarter) listed in the second column (P1) and the period in the third column (P2), where each line corresponds to a large contribution to this distance. The last five columns are as in previous figures.



Titles

countries : libya

countries : chad

CHAD SAYS IT KILLED 384 LIBYANS IN BATTLE
CHAD SAYS TROOPS KILLED 1,200 LIBYANS IN BATTLE
LIBYANS APPEAR TO BE PULLING OUT OF CHAD
LIBYANS APPEAR TO BE PULLING OUT OF CHAD
LIBYA ABANDONS CHAD STRONGHOLD AFTER DEFEATS
LIBYA SAYS CHAD FIGHTING TO CONTINUE

Figure J - Titles of all articles that include Libya and Chad

Finally, the user can request a graphical representation of co-occurrence frequencies of any 2 categories, in a desired level of granularity of time segments. Figure K displays the percentage of articles annotated with the category *crude* within the average topic distribution of OPEC countries, across different quarters.

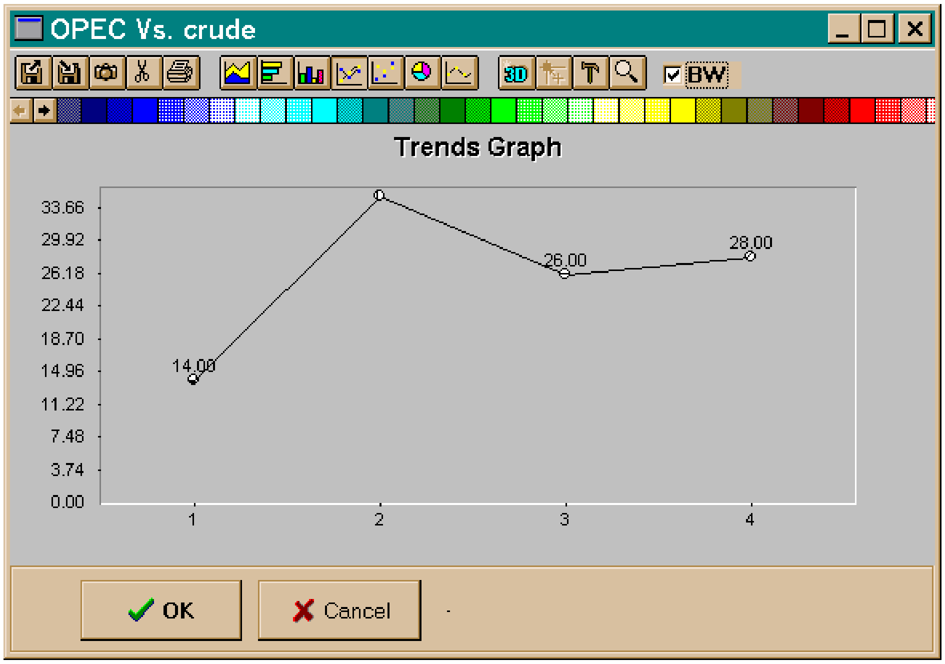


Figure K - Crude proportion of the topic distribution of OPEC across the year quarters

Conclusions

We have presented a framework and an implemented system for browsing and analyzing sets of documents which are annotated with category keyword labels. The system might be used as a support tool for domain experts that need to analyze and summarize large document sets. It may also be used in the regular query-and-browse cycle of a document retrieval session, to support the browsing phase. Currently, when users face the common response of the type “1000 documents match your query”, they need to guess in advance how they might restrict their query. In such cases the KDT system could provide much help in figuring out the content of these 1000 documents, and narrowing down the sets of target documents.

The KDT system is based on a compact model, which relies on rather modest assumptions. It requires annotation of documents with category keywords which are organized in a simple hierarchy. It also demonstrates the rich variety of KDD operations that can be based on keyword co-occurrence distributions and their comparison with the relative entropy distance measure. The simplicity of the model makes it rather easy to implement, and the pre-computation of keyword co-occurrence distributions makes online computations very efficient.

In future work we plan to extend the KDT framework to work also on co-occurrence distributions of terms and groups of terms that were extracted directly from the texts. This way we hope to combine these two levels of representation, namely category labels and document terms, in analogy to the way they are often combined in retrieval queries.

References

[1] C.Apte, F.Damerau, and S.Weiss. Towards language independent automated learning of text categorization models. In *Proceedings of ACM-SIGIR*, 1994.

[2] R.Brachman, P.Selfridge, L.Terveen, B.Altman, A.Borgida, F.Halper, T.Kirk, A.Lazar, D.McGuinness, and L.Resnick. Integrated support for data archeology. *International Journal of Intelligent and Cooperative Information Systems*, 1993.

[3] D.Cutting, D.Karger, and J.Pedersen. Constant interaction-time scatter/gather browsing of very large document collections. In *Proceedings of ACM-SIGIR*, 1993.

[4] Ido Dagan, Fernando Pereira, and Lillian Lee. Similarity-based estimation of word cooccurrence probabilities. In *Proc. of the Annual Meeting of the ACL*, pages 272-278, 1994.

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

[6] Ronen Feldman, Ido Dagan, and Willi Klosgen. Efficient algorithms for mining and manipulating associations in texts. To appear in *Proceedings of EMCSR-96 - Thirteenth European Meeting on Cybernetics and Systems Research,* Vienna, April 1996.

[7] S.Finch. Exploiting sophisticated representations for document retrieval. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 1994.

[8] W.J. Frawley, G.Piatetsky-Shapiro, and C.J. Matheus. Knowledge discovery in databases: an overview.

[9] M. Hearst. Tilebars: Visualization of term distribution information in fulltext information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Denver, CO, May 1995. ACM.

[10] In G.Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 1-27. MIT Press, 1991.

[11] M.Iwayama and T.Tokunaga. 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.

[12] W.Klosgen Problems for knowledge discovery in databases and their treatment in the statistics interpreter EXPLORA. *International Journal for Intelligent Systems*, 1992.

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

[14] W.Kloesgen 1995. Explora: A Multipattern and Multistrategy Discovery Assistant. In Advances in knowledge Discovery and Data Mining, MIT Press.

[15] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, H. Mannila:

Pruning and grouping discovered association rules. In Workshop Notes Statistics, Machine Learning and Knowledge Discovery in Databases, ECML-95.

[16] Gerard Salton. *Automatic Text Processing*. Addison-Wesley Publishing Company, 1989.

[17] C. Williamson and B. Shneiderman. The dynamic HomeFinder: Evaluating dynamic queries in a real-estate information exploration system. In *Proceedings of ACM-SIGIR*, 1992.