Journal of Intelligent Information Systems 10, 281–300 (1998)

Mining Text Using Keyword Distributions

RONEN FELDMAN feldman@cs.biu.ac.il
IDO DAGAN dagan@cs.biu.ac.il
Department of Mathematics and Computer Science Department, Bar-Ilan University, Ramat-Gan, ISRAEL

HAYM HIRSH hirsh@cs.rutgers.edu
Deptartment of Computer Science, Rutgers University, Piscataway, NJ USA 08855

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 Text, in which documents are labeled by keywords, and knowledge discovery is performed by analyzing the co-occurrence frequencies of the various keywords labeling the documents. We show how this keyword-frequency approach supports a range of KDD operations, providing a suitable foundation for knowledge discovery and exploration for collections of unstructured text.

Keywords: data mining, text mining, text categorization, distribution comparison, trend analysis

1. 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., Anand and Khan, 1993; Brachman et al., 1993; Frawley et al., 1991; Kloesgen, 1992; Kloesgen, 1995b; Ezawa and Norton, 1995). The goal of such work, often called *data mining*, 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.

However, 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, despite the tremendous amount of online information that appears only in collections of unstructured text. This paper addresses the problem of Knowledge Discovery from Text, and describes the KDT system, which provides for text the kinds of KDD operations previously provided for structured databases. Our approach is, first, to label documents with keywords taken from a controlled vocabulary that is organized into some meaningful hierarchical structure. Next, the keywords and higher-level entities in the hierarchy are used to support a range of KDD operations on the documents, to index into interesting subcollections, as well as to access and understand the various documents in a collection through keyword co-occurrence frequencies. A key insight in this work is that the frequency of occurrence of keywords can provide the foundation for a wide range of KDD operations on collections of textual documents, including analysis tools that allow a user to find patterns across sets of documents (such as tools for finding sets of documents whose keyword distributions differ significantly from the full collection, other related collections, or collections from other points in time) and presentation tools that allow a user to view the documents and information underlying them in convenient forms (such as tools for browsing a collection, viewing sets of underlying patterns in a structured way, or exploring the documents on which a pattern is based).

The focus of this paper is on analysis and presentation tools based on keyword co-occurrence frequencies. In particular, we do not concern ourselves in this paper with the initial step of labeling documents with keywords: in many commercial and scientific text collections and information feeds documents are already labeled with keywords taken from a hierarchy of controlled-vocabulary terms, to assist and augment free-text searching (e.g., the Dialog service of Knight Ridder Information Inc., the First service of Individual Inc., and the Medical Subject Heading (MeSH) hierarchy), and further, there is also a large body of work on automatically labeling documents with keywords (Lewis, 1992; Jacobs, 1992; Iwayama and Tokunaga, 1994; Apte et al., 1994; Lewis and Catlett, 1994). For example, the Reuters data used as a running example through this paper has been labeled with keywords from a controlled vocabulary through a combination of manual and automated methods. The work described in this paper begins with collections already labeled with keywords, showing how to use such keywords as the basis for knowledge discovery and exploration of collections of text.

The general architecture of the KDT system is shown in figure 1. The system takes two inputs: a collection of keyword-labeled documents, and a hierarchy with keywords as terminal nodes. The keyword hierarchy is a directed acyclic graph (DAG) of terms, where each of the terms is identified by a unique name. Figure 2 shows a portion of the keyword hierarchy used in our experiments with the Reuters data. 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 under both *European-Community* and *G7* in the hierarchy). Internal nodes in the hierarchy are used in two ways. First, each can be viewed as a keyword itself, labeling a document if any of the terms below it in the hierarchy label the document. Thus, for example, a document in the Reuters data may be thought of as being labeled by the *G7* term if it is labeled with one or more of keywords that appear below the *G7* node in the keyword hierarchy. In this context internal nodes can be viewed as keywords themselves. Second, internal nodes also serve as ways to specify sets of keywords. For example, we might be interested in computing the proportion of documents labeled by *gold* for each *G7* country. Rather than explicitly enumerating the *G7* countries, the token *G7* would be used to specify this set.[1] These two uses of internal nodes will usually be clear from context, although we try to identify which is being used when there is risk of confusion.

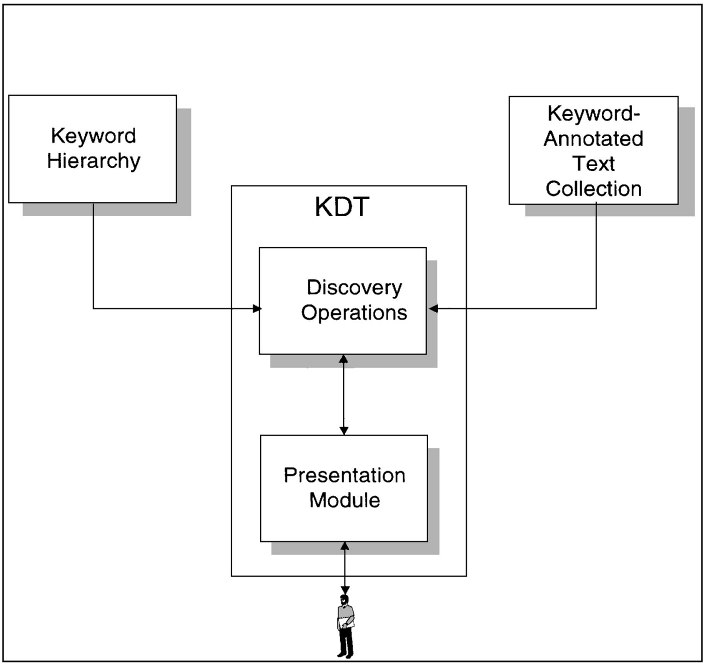


Figure 1. KDT system architecture.

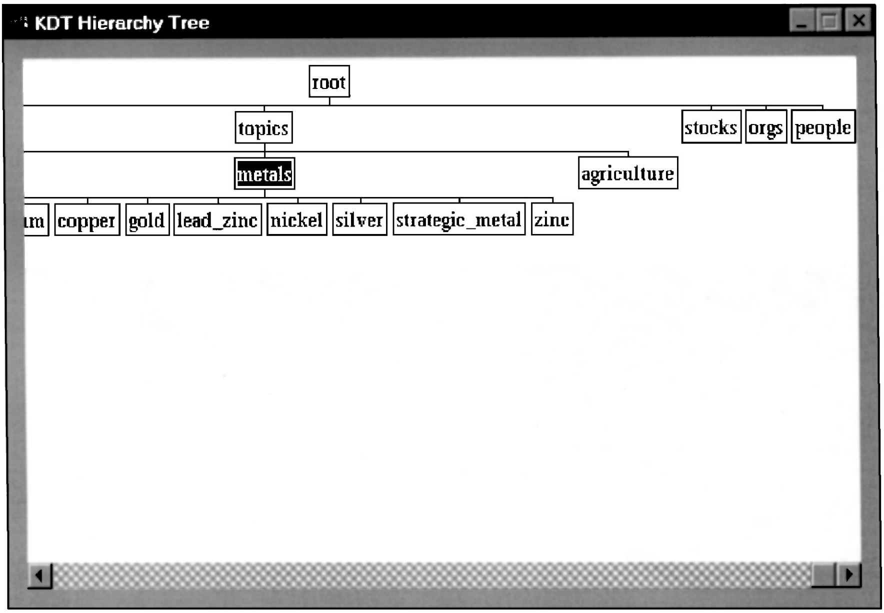


Figure 2. KDT display of part of the keyword hierarchy for Reuters data.

Most of the examples in this paper come from the use of the KDT system on the Reuters-22173 text collection, which contains over 20,000 articles that appeared on the Reuters newswire in the late 1980's, and were assembled and indexed with category keywords by personnel from Reuters Ltd. and Carnegie Group, Inc. (with further formatting and data file production performed in 1991 and 1992 by David D. Lewis and Peter Shoemaker (David Lewis, personal communication)). These keywords fall into five groups: countries, topics, people, organizations, and stock exchanges. We used these five keyword groupings as the skeleton for the keyword hierarchy given to KDT, with each of the five groupings serving as an intermediate node in an initial two-level hierarchy. This hierarchy was then enriched with some additional sub-groupings of keywords, such as *agriculture* and *metals* as daughters of the *topics* node,[2] and various international organizations (taken from the CIA World Factbook) as daughters of the *countries* node. This was the hierarchy that was then provided to KDT, together with the keyword-labeled collection of Reuters documents.

The remainder of this paper is structured as follows. We begin the paper in Section 2 with the basic terminology, notation, and concepts concerning keyword distributions that we will use through the rest of the paper. Section 3 then presents a range of KDD operations based on keyword distributions, with examples of how they are supported by the KDT system. Section 4 concludes the paper with some final remarks.

2. Keyword distributions

The basic idea in this work is to access and analyze collections of documents using frequencies of occurrence of various keywords labeling the documents. This section presents the basic concepts underlying our keyword-frequency approach to knowledge discovery from text. In all of our examples we will use R to represent the Reuters-22173 text collection.

2.1. Keyword selection

Given some collection of documents D, we will often want to refer to some subcollection of D that are labeled by one or more given keywords:

Definition 1.

> Keyword selection: If D is a collection of documents and K is a set of keywords, D/K is the subset of documents in D that are labeled with all of the keywords in K. When clear from context, given a single keyword, k, rather than writing $D/\{k\}$, we will use the notation D/k.

Thus, for example, the collection $R/\{\mathit{iran,nicaragua,reagan}\}$ contains a subset of the Reuters collection, namely those documents that are labeled with the keywords *iran*, *nicaragua*, and *reagan*, R/reagan contains the subset of documents that are labeled (at least) with *reagan*, and $R/G7$ contains those documents that are labeled with any terminal node under *G7* (i.e., labeled with any *G7* country)—*G7* is treated as a keyword here when doing keyword selection (rather than being viewed as the set of keywords under it, in which case it would have required *all* of its descendants to be present).[3]

2.2. *Keyword proportions*

We will also often want to know what proportion of a set of documents are labeled with a particular keyword.

Definition 2.

> Keyword proportion: If D is a collection of documents and K is a set of keywords, $f(D, K)$ is the fraction of documents in D that are labeled with all of the keywords in K i.e., $f(D, K) = (|D/K|)/(|D|)$. Given one keyword, k, rather than writing $f(D, \{k\})$, we will use the notation $f(D, k)$. When D is clear from context, we will drop it and write $f(k)$.

Thus, for example, $f(R, \{iran, nicaragua, reagan\})$ (which we can write as $f(\{iran, nicaragua, reagan\})$ since all of our examples concern the Reuters collection R) is the fraction of documents in the Reuters collection that are labeled with *iran*, *nicaragua*, and *reagan*, $f(reagan)$ is the proportion of the collection labeled with the keyword reagan, and $f(G7)$ is the proportion labeled with any *G7* country.

Given these definitions of selection and proportion we can already begin defining useful quantities for analyzing a set of documents. For example, the proportion of those documents labeled with K_2 that are also labeled by K_1 is designated by $f(D/K_2, K_1)$. This occurs often enough that we give it an explicit name and notation:

Definition 3.

> Conditional keyword proportion: If D is a collection of documents and K_1 and K_2 are sets of keywords, $f(D, K_1 \mid K_2)$ is the proportion of all those documents in D that are labeled with K_2 that are also labeled with K_1, i.e., $f(D, K_1 \mid K_2) = f(D/K_2, K_1)$. When D is clear from context, we will write this as $f(K_1 \mid K_2)$.

Thus, for example, $f(reagan \mid iran)$ is the proportion of all documents that are labeled by iran that are also labeled by *reagan*.

2.3. *Keyword-proportion distributions*

The operations supported by the KDT system are based on analyzing the distributions of keywords within sets of documents. For example, we may be interested in analyzing the distribution of keywords that denote economical topics—that is, descendants of the *topics* node in the keyword hierarchy. In particular, we will talk about various forms of distributions over sets of keywords. We will use $P_K(x)$ to refer to such distributions—it will assign to any keyword x in K a value between 0 and 1—and we will call these *keyword distributions*. (Note, however (as will be discussed shortly), we *do not* require the values to add up to 1.) In this subsection and the next we present a number of specific examples of such $P_K(x)$ distributions that will be used throughout this paper.

One particularly important keyword distribution that we will use is a keyword proportion distribution, which gives the proportion of documents in some collection that are labeled with each of a number of selected keywords:

Definition 4.

> Keyword-proportion distribution: If D is a collection of documents and K is a set of keywords, $F_K(D, x)$ is the proportion of documents in D that are labeled with x for any x in K. When D is clear from context, we will write this as $F_K(x)$.

Note the distinction between $P_K(x)$ and $F_K(x)$. We will use the former to refer generically to any function that is a keyword distribution. The latter is a specific keyword distribution defined by a particular keyword-labeled set of documents. Thus, for example $F_{\text{topics}}(R, x)$ would represent the proportions of documents in R that are labeled with keywords under the *topics* node in the keyword hierarchy. Observe that *topics* is used as shorthand for referring to a set of keywords, namely all those that occur under *topics*, rather than explicitly enumerating them all. Also, note that $F_{\{k\}}(D, k) = f(D, k)$, namely F_K subsumes the earlier-defined f when it is applied to a single keyword. However, unlike f, F_K is restricted to only refer to the proportion of occurrences of *individual* keywords (those occurring in the set K).[4] Thus f and F are incomparable.

As mentioned earlier, mathematically speaking, F is not a true frequency distribution, since each document may be labeled by multiple items in the set K. Thus, for example, a given document may be labeled by two (or more) *G7* countries, since occurrences of keywords are not disjoint events. Thus the sum of values in F_{G7} may be greater than one. In the worst case, if all keywords in K label all documents, the sum of the values in a distribution F can be as large as $|K|$. Furthermore, since some documents may contain none of the keywords in a given K, the sum of frequencies in F might also be smaller than one—in the worst case, 0 even. Nonetheless, we use the term “distribution” for F, since many of the connotations this term suggests still hold here.

Just as was the case for keyword proportions, we can consider conditional keyword-proportion distributions, which will be one of the central keyword distributions that we use:

Definition 5.

> Conditional keyword-proportion distribution: If D is a collection of documents and K and K' are sets of keywords, $F_K(D, x \mid K')$ is the proportion of those documents in D labeled with all the keywords in K' that are also labeled with keyword x (with x in K), i.e., $F_K(D, x \mid K') = F_K(D/K', x)$. We will often write this as $F_K(x \mid K')$, when D is clear from context.

Thus, for example, $F_{\text{topics}}(x \mid \mathit{Argentina})$ assigns any keyword x under *topics* in the hierarchy with the proportion of documents labeled by x within the set of all documents labeled by the keyword *Argentina*, and $F_{\text{topics}}(x \mid \{\mathit{UK}, \mathit{USA}\})$ is the similar distribution for those documents labeled with both the *UK* and *USA* keywords.

2.4. *Average keyword distributions*

Finally, when we compare distributions, one of the baseline distributions that we consider is the average distribution over a set of sibling nodes in the hierarchy. For example, when looking at the proportions of *loan* within South American countries such as $f(R,\ loan \mid Argentina)$, $f(R,\ loan \mid Brazil)$, and $f(R,\ loan \mid Columbia)$, the user may be interested in the average of all proportions of this form for all the South American countries, that is, the average of all proportions of the form $f(R,\ loan \mid k)$, where k ranges over all South American countries.

Definition 6.

> Average keyword proportion: Given a collection of documents D, a keyword k, and an internal node in the hierarchy n, an *average keyword proportion*, denoted by $a(D, k \mid n)$, is the average value of $f(D, k \mid k')$ where k' ranges over all immediate children of n, i.e., $a(D, k \mid n) = Avg_{\{k' \text{ is a child of } n\}}\{f(D, k \mid k')\}$. When D is clear from context, this will be written $a(k \mid n)$.

For example, $a(loan \mid South\_America)$ is the average keyword proportion of $f(loan \mid k')$ as k' varies over each child of the node *South_America* in the keyword hierarchy, i.e., it is the average conditional keyword proportion for *loan* within South American countries. Note that this quantity does *not* average the values weighted by the number of documents labeled by each child of n. Instead, it represents equally each descendant of n, and should be viewed as summary of what a typical keyword proportion is for a child of n.

And, as before, the user may be interested in the distribution of averages for each economic topic within South American countries. This is just another keyword distribution:

Definition 7.

> Average keyword distribution: Given a collection of documents D, and two internal nodes in the hierarchy n and n', an *average keyword distribution*, denoted by $A_n(D, x \mid n')$ is the distribution that, for any x that is a child of n, averages x's proportions over all children of n', i.e., $A_n(D, x \mid n') = Avg_{\{k' \text{ is a child of } n'\}}\{F_n(D, x \mid k')\}$. When clear from context, this will be written $A_n(x \mid n')$.

Consider for example $A_{\text{topics}}(x \mid South\_America)$ (which can be read as "The average distribution of topics within South American countries"). For any topic x this gives its average proportion within all South-American countries.

2.5. *Comparing keyword distributions*

In addition to allowing a user to request particular keyword distributions, we would also like to identify distributions that are likely to be "interesting" for the user in some context. We quantify the potential degree of "interest" in some piece of information by comparing it to a given "expected" model, which serves as a baseline for the investigated distribution. For example, we may want to compare the data regarding IBM to an averaged model constructed for a group of 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.

Since we use keyword proportions and distributions to describe the data, we therefore need measures for quantifying the distance between an investigated distribution to another distribution that serves as a baseline model. Since our distributions are discrete, we simply use sum-of-squares to measure the distance between two models:

$$D(p' \parallel p) = \sum_x (p'(x) - p(x))^2,$$

where the target distribution is designated by p and the approximating distribution by p', and the x in the summation is taken over all objects in the domain. This measure is always non-negative and is 0 if and only if $p' = p$.

Given this measure, we can now use it as a heuristic device for judging keyword-distribution similarities:

Definition 8.

Keyword distribution distance: Given two keyword distributions $P'_K(x)$ and $P_K(x)$, the distance $D(P'_K \parallel P_K)$ between them is defined by:

$$D(P'_K(x) \parallel P_K(x)) = \sum_{x \in K} (P'_K(x) - P_K(x))^2.$$

We will also sometimes be interested in the value of the difference between two distributions at a particular point:

Definition 9.

Keyword proportion distance: Given two keyword distributions $P'_K(x)$ and $P_K(x)$, and a keyword k in K, the distance $d(P'_K(k) \parallel P_K(k))$ between them is defined by:

$$d(P'_K(k) \parallel P_K(k)) = P'_K(k) - P_K(k).$$

Thus another way to state $D(P'_K \parallel P_K)$ is $\sum_{x \in K} [d(P_K(x) \parallel P_K(x))]^2$. As an example, the distance between the distribution of *topics* within *Argentina* and the distribution of *topics* within *Brazil* would be written as $D(F_{\text{topics}}(x \mid \textit{Argentina}) \parallel F_{\text{topics}}(x \mid \textit{Brazil}))$, and the distance between the distribution of *topics* within *Argentina* and the average distribution of *topics* within *South-America* is written as $D(F_{\text{topics}}(x \mid \textit{Argentina}) \parallel A_{\text{topics}}(x \mid \textit{South\_America}))$.

3. Mining text using keyword distributions

Given the various concepts and definitions of the previous section concerning keyword distributions, we can begin considering various knowledge-discovery tasks that they support.

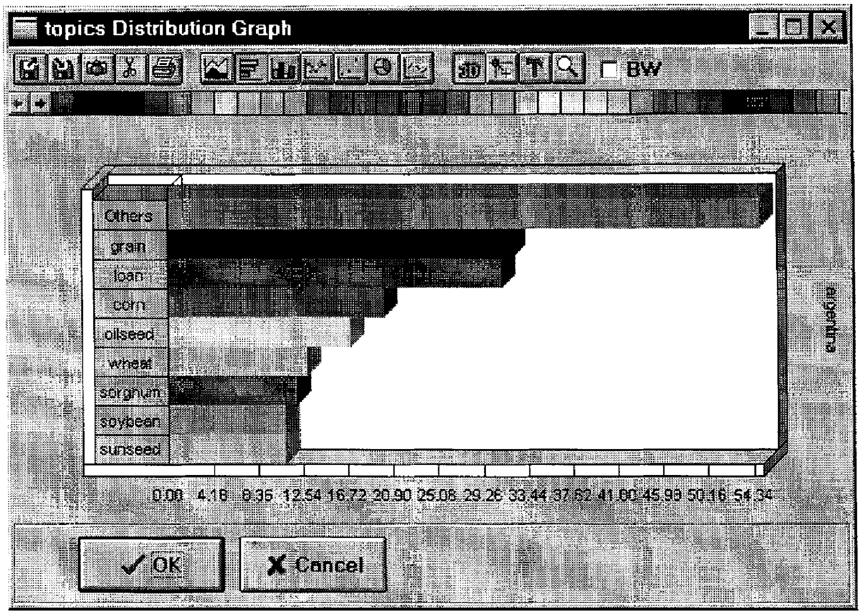


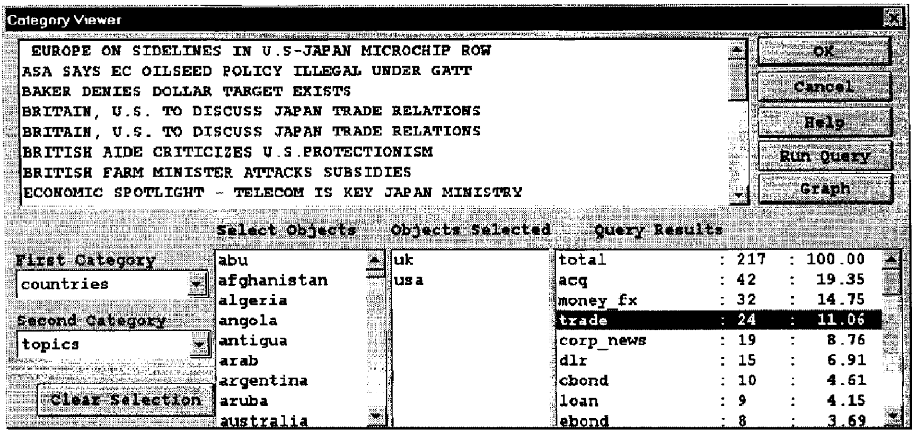
Figure 3. Graphical representation of the topic distribution of Argentina.

This section demonstrates a number of knowledge-discovery operations made possible by considering keyword distributions, and how they are supported by the KDT system.

3.1. Conditional keyword-proportion distributions

The most basic operation using keyword distributions that the KDT system supports is the display of conditional keyword-proportion distributions. For example, a user may be interested in seeing the proportion of documents labeled with each child of *topics* for all those documents labeled by the keyword Argentina, i.e., what proportion of *Argentina* documents are labeled with each topic keyword. This distribution would be designated by $F_{\text{topics}}(R, x \mid \mathit{Argentina})$, and the graphical display of this distribution that would be generated by KDT is given in figure 3. The distribution is presented as a bar-chart: 12 articles among all articles of *Argentina* are annotated with *sorghum*, 20 with *corn*, 32 with *grain*, etc., providing a summary of the areas of economical activity of Argentina, as reflected in the text collection. KDT presents distributions in several forms, graphical (e.g., pie-chart) or alphanumeric, listing absolute counts or proportions.

Conditional keyword-proportion distributions can also be conditioned on *sets* of keywords. Figure 4 shows the result KDT would give for the keyword distribution $F_{\text{topics}}(x \mid \{\mathit{UK}, \mathit{USA}\})$—the distribution of proportions for each *topics* amongst documents labeled with both the *UK* and *USA* keywords. Here the user has chosen to display the distribution in tabular form. The distribution itself is presented in the lower right window of the screen, with the distribution request specified to its left. This form of display also allows a user to



EUROPE ON SIDELINES IN U.S-JAPAN MICROCHIP ROW
ASA SAYS EC OILSEED POLICY ILLEGAL UNDER GATT
BAKER DENIES DOLLAR TARGET EXISTS
BRITAIN, U.S. TO DISCUSS JAPAN TRADE RELATIONS
BRITAIN, U.S. TO DISCUSS JAPAN TRADE RELATIONS
BRITISH AIDE CRITICIZES U.S.PROTECTIONISM
BRITISH FARM MINISTER ATTACKS SUBSIDIES
ECONOMIC SPOTLIGHT - TELECOM IS KEY JAPAN MINISTRY

First Category: countries

Second Category: topics

Select Objects: abu, afghanistan, algeria, angola, antigua, arab, argentina, aruba, australia

Objects Selected: uk, usa

Query Results

```
total      : 217  : 100.00
acq        : 42   :  19.35
money_fx   : 32   :  14.75
trade      : 24   :  11.06
corp_news  : 19   :   8.76
dlr        : 15   :   6.91
cbond      : 10   :   4.61
loan       : 9    :   4.15
ebond      : 8    :   3.69
```

Figure 4. Viewing the topic distribution of *USA* and the *UK.*

access documents based on the displayed distribution—for example, by clicking on any of the keywords in the distribution to see the articles that are so labeled. Here, 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.

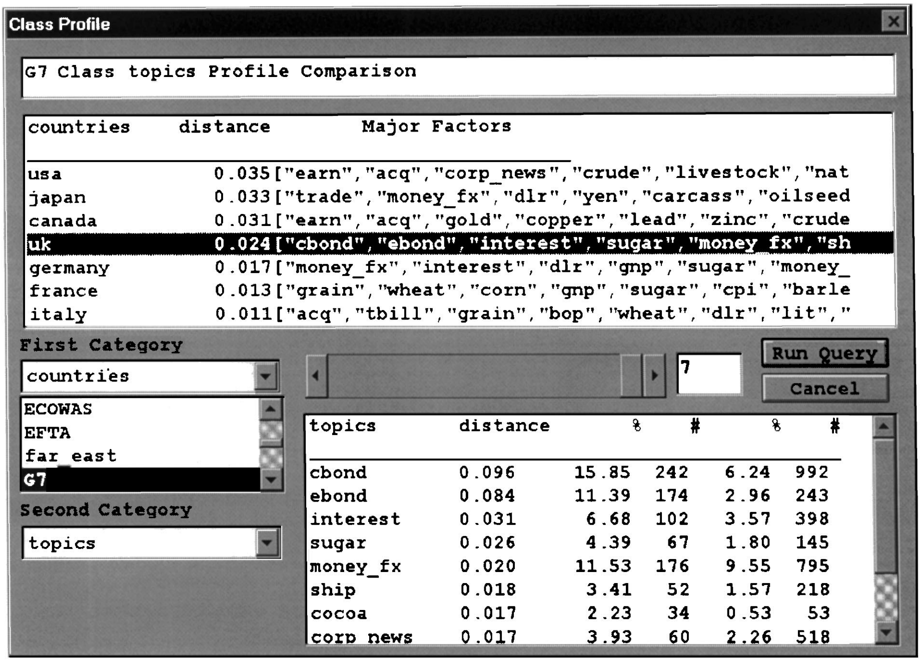
In some sense this type of operation can be viewed as a more refined form of traditional keyword-based retrieval. Rather than simply requesting all documents labeled by Argentina or by both *UK* and *USA*, the user can see the documents at a higher level, by requesting documents labeled by Argentina, for example, and first seeing what proportions are labeled by keywords from some secondary set of keywords that are of interest, with the user being able to access the documents through this more fine-grained grouping of Argentina-labeled documents.

3.2. *Comparing to average distributions*

Consider a conditional proportion of the form $F_K(D, x \mid k)$, the distribution over K of all documents labeled with some keyword k (not necessarily in K). It is natural to expect that this distribution would be similar to other distributions of this form, over conditioning events k' that are siblings of k. When they differ substantially it is a sign that the documents labeled with the conditioning keyword k may be of interest.

KDT supports this kind of comparison of keyword-labeled documents to the average of those labeled with the keyword and its siblings. A user can specify two internal nodes of the hierarchy, and compare individual distributions of keywords under one of the nodes conditioned on the keyword set under the other node, i.e., compute $D(F_n(x \mid k) \parallel A_n(x \mid n'))$ for each k that is a child of n'.

Figure 5 demonstrates this type of comparison, between the topic distribution of each *G7* country and the average distribution of topics for all *G7* countries, i.e., $D((F_{\text{topics}}(x \mid k) \parallel A_{\text{topics}}(x \mid G7))$ for each keyword k that is a child of the *G7* node in the hierarchy. In the



G7 Class topics Profile Comparison

countries	distance	Major Factors
usa	0.035	["earn","acq","corp_news","crude","livestock","nat
japan	0.033	["trade","money_fx","dlr","yen","carcass","oilseed
canada	0.031	["earn","acq","gold","copper","lead","zinc","crude
uk	0.024	["cbond","ebond","interest","sugar","money_fx","sh
germany	0.017	["money_fx","interest","dlr","gnp","sugar","money_
france	0.013	["grain","wheat","corn","gnp","sugar","cpi","barle
italy	0.011	["acq","tbill","grain","bop","wheat","dlr","lit","

First Category

countries

ECOWAS
EFTA
far_east
G7

Second Category

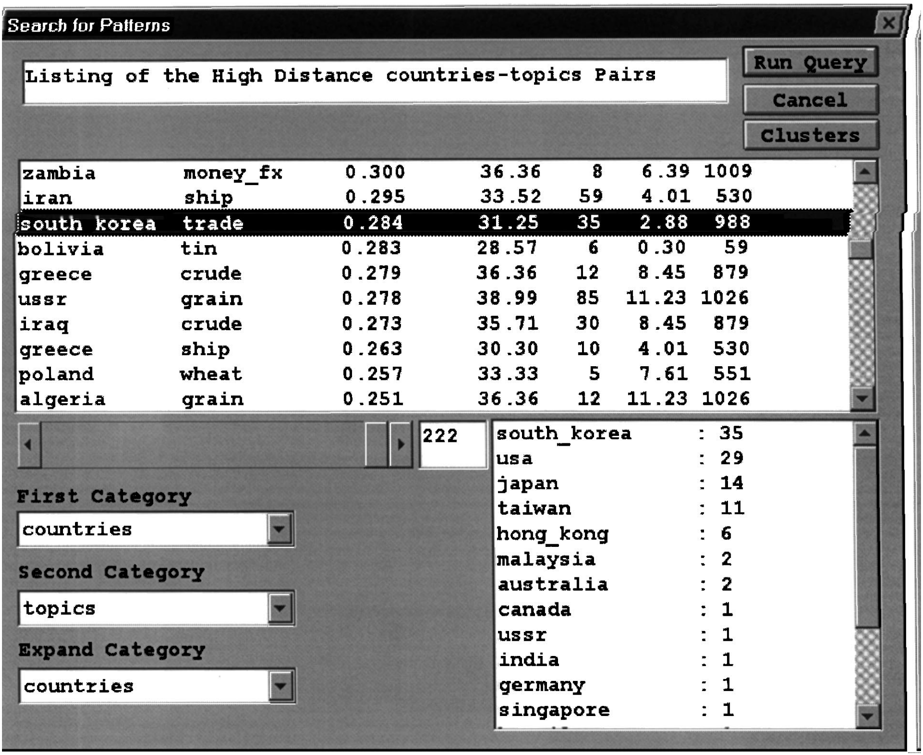
topics

7

topics	distance	%	#	%	#
cbond	0.096	15.85	242	6.24	992
ebond	0.084	11.39	174	2.96	243
interest	0.031	6.68	102	3.57	398
sugar	0.026	4.39	67	1.80	145
money_fx	0.020	11.53	176	9.55	795
ship	0.018	3.41	52	1.57	218
cocoa	0.017	2.23	34	0.53	53
corp news	0.017	3.93	60	2.26	518

Figure 5. Comparison of the topic distribution of members of the *G7* organization vs. the average topic distributions of the *G7*.

large box in the upper half of the figure countries are sorted in decreasing order of their distance to the average distribution (column 2), revealing that USA is the most "atypical" *G7* country (with respect to its topic distribution) while Italy is the most typical one. For each country k, the topics k' that made the largest contributions to the distance are also displayed (column 3), i.e., they are sorted by $d(F_{\text{topics}}(k' \mid k) \parallel A_{\text{topics}}(k' \mid G7))$. 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 5, we have expanded the topic list of the *UK* (at the bottom-right listbox), displaying $F_{\text{topics}}(x \mid UK)$. The first column there shows topic names. The second column shows the contribution of the topic to the distance. The third column shows, respectively, the proportion of *UK*-labeled documents also labeled with that topic keyword ($f(k' \mid UK)$ for each topic) with the corresponding absolutely number of documents in column four. The final two columns display the comparable figures for the average distribution ($a(k' \mid UK)$). In addition to their value in finding possible interesting keyword labelings, comparisons of this type also 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 3 and 4. This reveals the major topics that are generally common for *G7* countries. Then, the presentation of figure 5 could then be used to reveal the major characteristics that are specific for each country.



Listing of the High Distance countries-topics Pairs

zambia	money_fx	0.300	36.36	8	6.39	1009
iran	ship	0.295	33.52	59	4.01	530
south korea	trade	0.284	31.25	35	2.88	988
bolivia	tin	0.283	28.57	6	0.30	59
greece	crude	0.279	36.36	12	8.45	879
ussr	grain	0.278	38.99	85	11.23	1026
iraq	crude	0.273	35.71	30	8.45	879
greece	ship	0.263	30.30	10	4.01	530
poland	wheat	0.257	33.33	5	7.61	551
algeria	grain	0.251	36.36	12	11.23	1026

222

First Category
countries

Second Category
topics

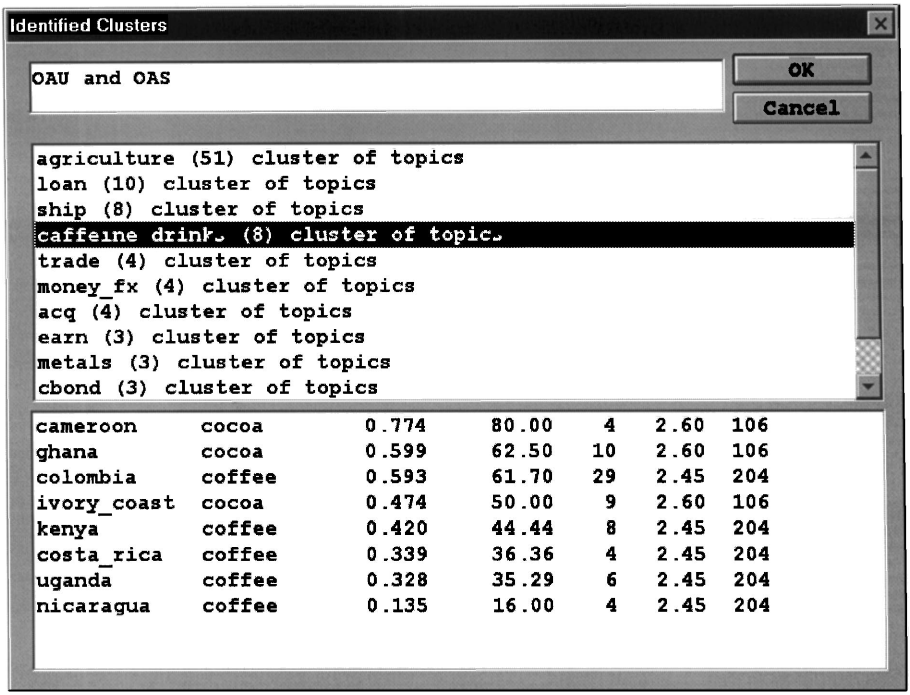
Expand Category
countries

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

Figure 6. Country-topic associations.

KDT also allows a user to see this information by listing pairs of conditioning and conditioned keywords that contribute significantly to the above distance measure. For example, figure 6 lists (in the box in the upper half of the figure) those country/topic pairs whose terms are largest in their distribution's distance (they are sorted in decreasing order of their contribution to the distance from the topic distribution of the given country to the average topic distribution of all countries (third column)—i.e., by $d(F_{\text{topics}}(k \mid k') \parallel A_{\text{topics}}(k \mid \mathit{countries}))$ for each topic k and country k'). The remaining columns display the same information as in the final four columns at the bottom of figure 5. When the line for any pair of keywords is selected, as is shown in the figure for *South_Korea* and *trade*, KDT gives the conditional keyword distribution from which it comes (in absolute-frequency form) in the lower-right part of the display.

Finally, in many cases KDT can generate a large number such results. To summarize the information, the system uses the keyword hierarchy to group together results whose second component falls under the same node in the hierarchy. Figure 7 shows the clusters that were formed by the system when grouping the results of figure 6, along with their sizes (in parentheses). For example, in 43 cases the second component was a daughter of the node agriculture. The user can examine any cluster and see the specific items that it contains (lower listbox, for the selected cluster caffeine-drinks). (The columns of



OAU and OAS

```
agriculture (51) cluster of topics
loan (10) cluster of topics
ship (8) cluster of topics
caffeine drinks (8) cluster of topics
trade (4) cluster of topics
money_fx (4) cluster of topics
acq (4) cluster of topics
earn (3) cluster of topics
metals (3) cluster of topics
cbond (3) cluster of topics
```

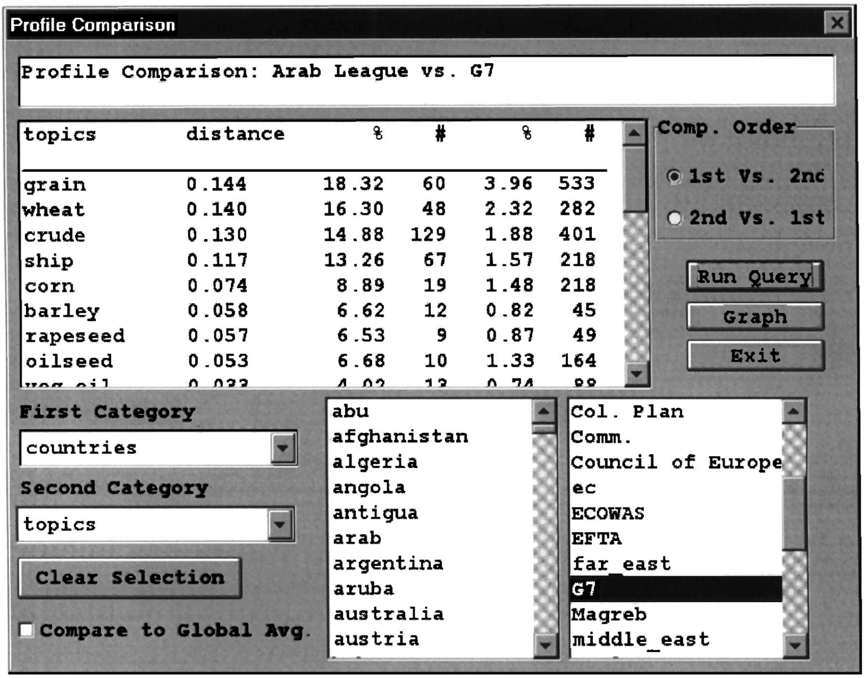
```
cameroon     cocoa    0.774   80.00   4   2.60  106
ghana        cocoa    0.599   62.50  10   2.60  106
colombia     coffee   0.593   61.70  29   2.45  204
ivory_coast  cocoa    0.474   50.00   9   2.60  106
kenya        coffee   0.420   44.44   8   2.45  204
costa_rica   coffee   0.339   36.36   4   2.45  204
uganda       coffee   0.328   35.29   6   2.45  204
nicaragua    coffee   0.135   16.00   4   2.45  204
```

Figure 7. Clustering associations using the category hierarchy.

the lower listbox are the same as in figure 5.) In addition, the system tries to provide a compact generalization for the first component of each result 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.

3.3. *Comparing specific distributions*

The preceding mechanism for comparing distributions to an average distribution is also useful for comparing conditional distributions of two specific nodes in the hierarchy. In figure 8, we measure the distance from the average topic distribution of *Arab League* countries to the average topic distribution of *G7* countries (in the upper half of the figure). Entries are sorted in decreasing order of their contribution to the distance (second column), namely $d(A_{\text{topics}}(k \mid Arab\_League) \parallel A_{\text{topics}}(k \mid G7))$. The third and fifth columns show, respectively, the percentage of the topic in the average topic distribution of the *Arab League* countries $(A_{\text{topics}}(x \mid Arab\_League))$ and in the average topic distribution of the *G7* countries $(A_{\text{topics}}(x \mid G7))$. The fourth and sixth columns show, respectively, the total number of articles in which the topic appears with any *Arab League* country and any *G7* country. This reveals the topics with which *Arab League* countries are associated much more than *G7* countries, such as crude-oil and wheat. Figure 9 shows the comparison in the opposite



Profile Comparison: Arab League vs. G7

topics	distance	%	#	%	#
grain	0.144	18.32	60	3.96	533
wheat	0.140	16.30	48	2.32	282
crude	0.130	14.88	129	1.88	401
ship	0.117	13.26	67	1.57	218
corn	0.074	8.89	19	1.48	218
barley	0.058	6.62	12	0.82	45
rapeseed	0.057	6.53	9	0.87	49
oilseed	0.053	6.68	10	1.33	164
veg_oil	0.033	4.02	13	0.74	88

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

First Category: countries

Second Category: topics

□ Compare to Global Avg.

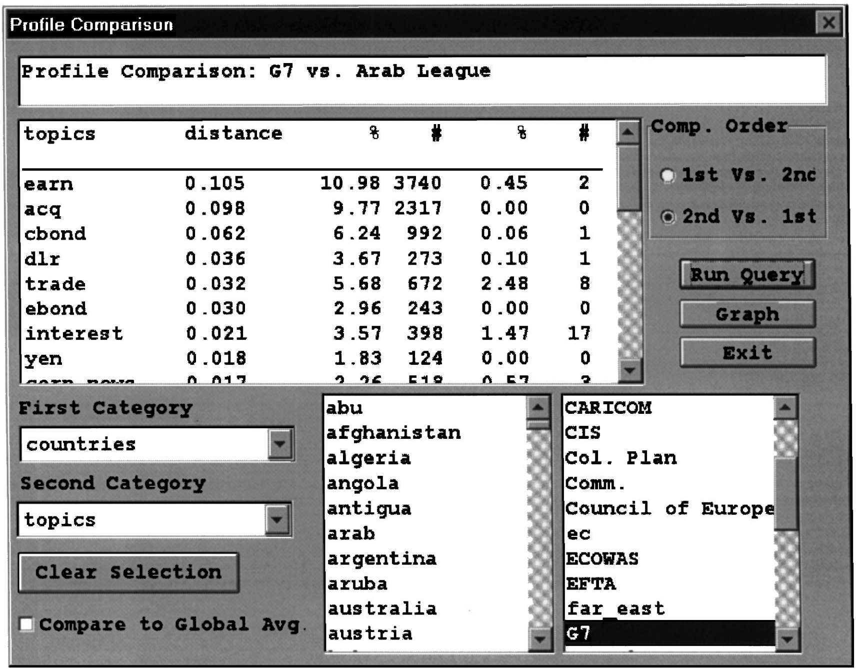
Figure 8. Topics profile comparison of the Arab league vs. the *G7*.

direction, revealing the topics with which *G7* countries are highly associated relative to the *Arab League*.

3.4. *Trend analysis*

Although we haven't focused on it so far, the various keyword distributions are functions of collections of documents. It is therefore possible to compare two distributions that are otherwise identical except that they are for different collections. One notable example of this is when the two collections are from the same source (such as from a news-feed), but from different points in time. For example, we can compare the distribution of *topics* within *Argentina*-labeled documents, as formed by documents published in the first quarter of 1987, to the same distribution formed by documents from the second quarter of 1987. This comparison will highlight those economical topics whose proportion changed between the years, directing the attention of the user to specific trends or events in the economical activity of Argentina. If R_1 is used to designate the portion of the Reuters newswire data from the first quarter of 1987, and R_2 designates the portion from the second quarter of 1987, this would correspond to comparing $F_{\text{topics}}(R_1, x \mid \textit{Argentina})$ and $F_{\text{topics}}(R_2, x \mid \textit{Argentina})$.

Figure 10 shows how KDT supports this knowledge-discovery operation, listing trends that were identified across different quarters in the time period represented by the Reuters collection, computing $D(F_{\text{countries}}(D_1, x \mid \textit{countries}) \parallel F_{\text{countries}}(D_2, x \mid \textit{countries}))$ where D_1



Profile Comparison

Profile Comparison: G7 vs. Arab League

topics	distance	%	#	%	#
earn	0.105	10.98	3740	0.45	2
acq	0.098	9.77	2317	0.00	0
cbond	0.062	6.24	992	0.06	1
dlr	0.036	3.67	273	0.10	1
trade	0.032	5.68	672	2.48	8
ebond	0.030	2.96	243	0.00	0
interest	0.021	3.57	398	1.47	17
yen	0.018	1.83	124	0.00	0

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

First Category: countries

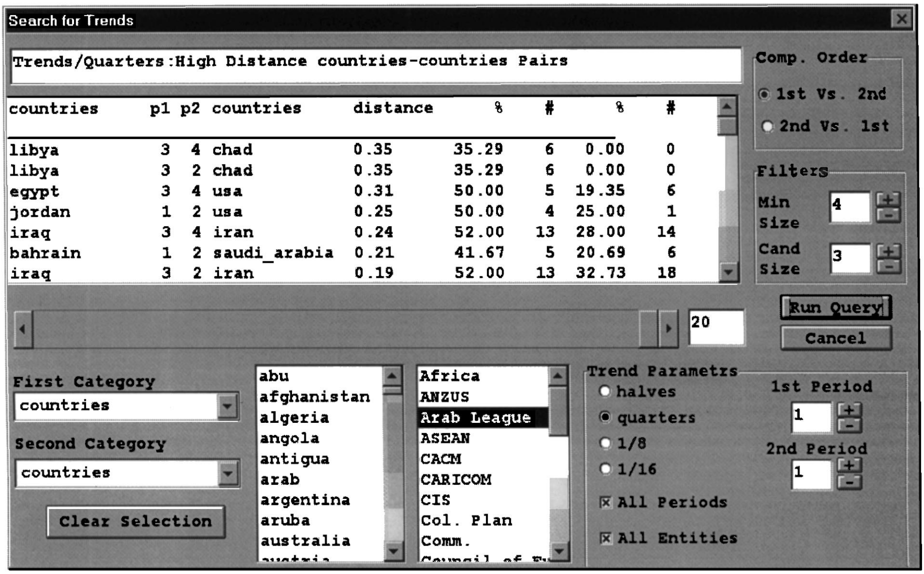
Second Category: topics

□ Compare to Global Avg.

abu
afghanistan
algeria
angola
antigua
arab
argentina
aruba
australia
austria

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

Figure 9. Topics profile comparison of the *G7* vs. the Arab league.



Search for Trends

Trends/Quarters:High Distance countries-countries Pairs

countries	p1	p2	countries	distance	%	#	%	#
libya	3	4	chad	0.35	35.29	6	0.00	0
libya	3	2	chad	0.35	35.29	6	0.00	0
egypt	3	4	usa	0.31	50.00	5	19.35	6
jordan	1	2	usa	0.25	50.00	4	25.00	1
iraq	3	4	iran	0.24	52.00	13	28.00	14
bahrain	1	2	saudi_arabia	0.21	41.67	5	20.69	6
iraq	3	2	iran	0.19	52.00	13	32.73	18

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

Filters
Min Size: 4
Cand Size: 3

20

First Category: countries

Second Category: countries

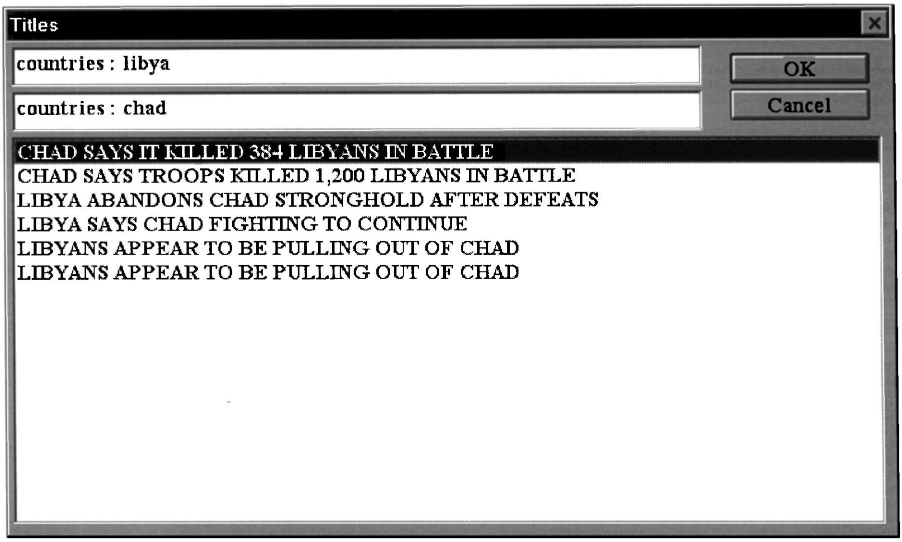
abu
afghanistan
algeria
angola
antigua
arab
argentina
aruba
australia

Africa
ANZUS
Arab League
ASEAN
CACM
CARICOM
CIS
Col. Plan
Comm.

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

1st Period: 1
2nd Period: 1

Figure 10. Trends in co-occurrence of Arab league countries with other countries.



countries : libya

countries : chad

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

Figure 11. Titles of all articles that include Libya and Chad in the 3rd quarter.

and D_2 correspond to different subcollections from different quarters (identified by the second and third columns, labeled $p1$ and $p2$).[5] The sixth and seventh columns show, respectively, the percentage and absolute frequency for $F_{\text{countries}}(x \mid \textit{countries})$ for each such pair of collections. The first line of the top listbox, for example, shows that in the third 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 second quarter to 35.29% in the third quarter). The second line shows that the proportion of such articles in the third quarter was also much higher than in the fourth quarter (a decrease over time, again to 0%).

Given such results, an analyst might then want to investigate what happened in the third quarter regarding *Libya* and *Chad*. To facilitate such an 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 11, which could help reveal that the cause for the trend was the fighting between *Libya* and *Chad* at that period.

Finally, the system can display a graphical representation of a sequence of values of the same proportion, which correspond to a sequence of time periods, in a desired level of granularity of time. Figure 12 displays the proportion of articles annotated with the category *crude* within the average topic distribution of *OPEC* countries, across different quarters.

4. Concluding remarks

Although much information can be found in online repositories of unstructured text, little work has addressed the problem of finding interesting patterns and information underlying large quantities of such textual data. This paper has described an approach to knowledge

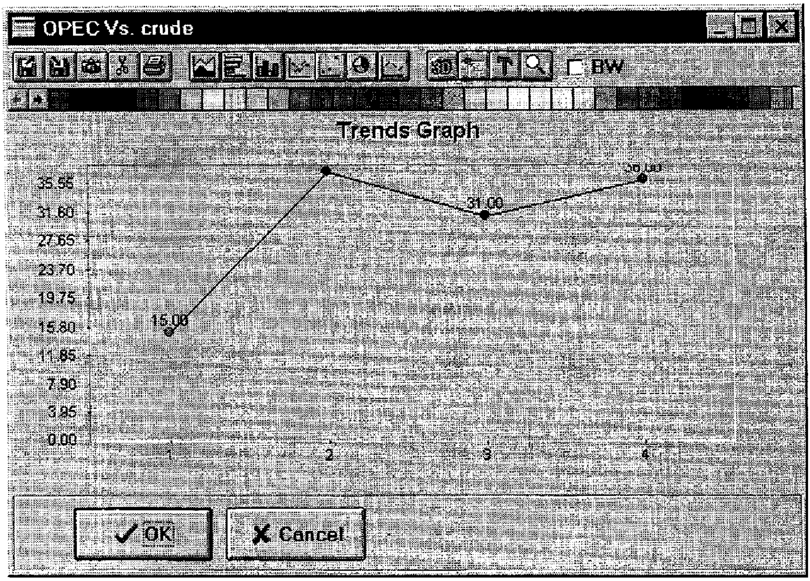


Figure 12. Crude proportion in the topic distribution of OPEC across quarters.

discovery for text that begins with documents labeled by keywords selected from a hierarchy of keywords. A user can then explore potentially interesting collections of documents by exploring the distribution of labels on the documents. We have described how this approach can support a range of mining operations, as well as how they are instantiated within KDT, an implemented system for knowledge discovery from text. This includes tools for comparing the distribution of keywords under some node in the keyword hierarchy for subcollections of the full set of documents (selected via the keywords as well) to average distributions, as well as comparing distributions for collections from different points in time. The KDT system also provides a range of display methods for presenting such distributions and accessing the documents that give rise to them.

Our work here focuses on comparisons of keyword distributions for different subsets of a document collection. In contrast, our related work on the FACT system (Feldman and Hirsh, 1996) focuses on finding associations (e.g., Agrawal et al., 1993; Mannila et al., 1994; Toivonen et al., 1995) between the keywords labeling a single collection of documents. Our work is also related to efforts in the information-retrieval community to structure and display collections of documents to help a user browse the collection and to display additional structures hidden in the documents (e.g., Salton, 1989; Cutting et al., 1993; Williamson and Shneiderman, 1992; Hearst, 1995). Here we use a different source of power to support such functionality—keyword co-occurrence frequency. Further, rather than simply presenting a tool for structuring and displaying documents, a higher-level point of this paper is that a keyword-frequency approach supports a range of useful knowledge discovery operations (in addition to those that have simply been implemented in our system). Our use of hierarchies to structure the values being explored by our discovery tools is similar to the work of Srikant and Agrawal (1995) and Han and Fu (1995), where a taxonomy is imposed on the items that occur in transactions and knowledge discovery attempts to find associations between items at any level of the taxonomy. Given a hierarchy over items in transactions, the KDT approach would also apply. However, KDT would additionally use the hierarchy as a "vocabulary" of useful sets of keywords for structuring a user's discovery operations. Finally, Kloesgen (1995a, 1995b) also uses distribution comparisons in the EXPLORA system, to discovery interesting statements in a database.

Our focus in this work has been on the development of tools particularly well-suited to collections of keyword-labeled textual documents. In future we plan to explore the development of similar tools for structured databases, exploring distributions of attribute values amongst various (sub)sets of records in a database. We also plan to investigate possible synergistic relationships between automatic keyword labeling and discovery methods that use such keyword labels, in the hope of developing keyword-labeling algorithms that are tailored to keyword-based knowledge discovery from text. Complementary to this, we also plan to use the KDT approach when the "keywords" labeling documents represent the presence or absence of selected words or phrases in a document, with the goal of performing knowledge discovery using both forms of keywords. Finally, we plan to continue our development of presentation tools for displaying the results of our distribution-based discovery tools, such as through more sophisticated use of clustering methods.

Acknowledgments

This research was supported by NSF grant IRI-9509819 and by grant 8615-1-96 from the Israeli Ministry of Science. The authors would like to thank the reviewers for helpful comments given on drafts of this paper.

Notes

1. Moreover, in many contexts in the KDT system sets of keywords may only be specified through the use of internal nodes in the hierarchy. The assumption is that the hierarchy maintains those subsets of the keywords that are interesting, by virtue of the fact that they have been placed under a single node in the hierarchy. To specify additional groups of keywords a user must add an internal node for them in the hierarchy, through a hierarchy editor included with the system—it is a simple graphical user interface for constructing and editing keyword hierarchies, supporting additions, deletions and modifications of nodes and links. Indeed, figure 2 is a screen dump of this hierarchy maintenance editor.
2. It is unfortunate that, although all keywords in some sense represent topics that might arise in documents in the collection, the token *topics* was used by Reuters to designate those keywords that are economical topics, and for consistency we maintain its use in that way here as well.
3. Throughout this paper we primarily consider subsets of a collection of documents that are selected by whether they are labeled with particular keywords. Although all our definitions generally apply to arbitrary sets of documents—indeed, we exploit this fact when comparing documents from different points of time in Section 3.4—we focus primarily on keyword-selected document sets.
4. Although it is quite simple to define a similar notion for *sets* of keywords (for example, by computing the proportions for each subset of a set K), we have not found it necessary for any of the operations supported by KDT.
5. Although this is our first example doing this, it is quite fair to ask for a distribution $F_K(x \mid K)$, which analyzes the co-occurrences of different keywords under the same node of the hierarchy. Thus, for example, $F_{\text{countries}}(x \mid \textit{countries})$ would analyze the co-occurrences of country labels on the various documents.

References

Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (pp. 207–216).

Anand, T. and Kahn, G. (1993). Opportunity explorer: Navigating large databases using knowledge discovery templates. In *Proceedings of the 1993 workshop on Knowledge Discovery in Databases*.

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

Brachman, R., Selfridge, P., Terveen, L., Altman, B., Borgida, A., Halper, F., Kirk, T., Lazar, A., McGuinness, D., and Resnick, L. (1993). Integrated Support for Data Archeology. *International Journal of Intelligent and Cooperative Information Systems*.

Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory*, John Wiley and Sons.

Cutting, C., Karger, D., and Pedersen, J. (1993). Constant interaction-time scatter/gather browsing of very large document collections. In *Proceedings of ACM-SIGIR Conference on Information Retrieval*.

Dagan, I., Pereira, F., and Lee, L. (1994). Similarity-based estimation of word co-occurrence probabilities. In *Proceedings of the Annual Meeting of the ACL* (pp. 272–278).

Dagan, I., Feldman, R., and Hirsh, H. (1996). Keyword-based browsing and analysis of large document sets. To appear In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR-96)*. Las Vegas.

Ezawa, K. and Norton, S. (1995). Knowledge discovery in telecommunication services data using Bayesian Network Models. In *Proceedings of the First International Conference on Knowledge Discovery (KDD-95)*.

Feldman, R. (1996). The KDT system—using prolog for KDD. To appear In *Proceedings of PAP'96 (Practical Applications of Prolog)*. London, UK.

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

Feldman, R., Dagan, I., and Klöesgen, W. KDD tools for mining associations in textual databases. To appear. In *Proceedings of the 9th International Symposium on Methodologies for Intelligent Systems*.

Feldman, R., Dagan, I., and Klöesgen, W. (1996). Efficient algorithms for mining and manipulating associations in texts. To appear, *Research and Cybernetics*.

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

Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C.J. (1991). Knowledge Discovery in Databases: An Overview. In G. Piatetsky-Shapiro and W.J. Frawley (Eds.), *Knowledge Discovery in Databases*. MIT Press, pp. 1–27.

Han, J. and Fu, Y. (1995). Discovery of multiple-level association rules from large databases. In *Proc. of 1995 Int. Conf. on Very Large Data Bases (VLDB'95)* (pp. 420–431). Zürich, Switzerland.

Hearst, M. (1995). Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. Denver, CO: ACM.

Iwayama, M. and Tokunaga, T. (1994). 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*.

Jacobs, P. (1992). Joining statistics with NLP for text categorization. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*.

Klösgen, W. (1992). Problems for Knowledge Discovery in Databases and Their Treatment in the Statistics Interpreter EXPLORA, *International Journal for Intelligent Systems*, 7(7), 649–673,

Klösgen, W. (1995a). EXPLORA: A Multipattern and Multistrategy Discovery Assistant. In U. Fayyad, G. Piatetsky-Shapiro, and R. Smyth (Eds.), *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press, pp. 2249–271.

Klösgen, W. (1995b). Efficient Discovery of Interesting Statements in Databases, *Journal of Intelligent Information Systems*, 4, 53–69.

Lewis, D. (1992). An evaluation of phrasal and clustered representations on a text categorization problem. In *Proceedings of ACM-SIGIR Conference on Information Retrieval*.

Lewis, D. and Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the 11th International Conference on Machine Learning*.

Mannila, H., Toivonen, H., and Verkamo, A. Efficient algorithms for discovering association rules. In *KDD-94: AAAI workshop on Knowledge Discovery in Databases* (pp. 181–192).

Salton, G. (1989). *Automatic Text Processing*, Addison-Wesley Publishing Company.

Srikant, R. and Agrawal, R. 1995. Mining generalized association rules. In *Proc. of the 21st Int'l Conference on Very Large Databases*. Zurich, Switzerland, Sept. 1995. Expanded version available as IBM Research Report RJ 9963.

Toivonen, H., Klemettinen, M., Ronkainen, P., Hatonen, K., and Mannila, H., Pruning and grouping discovered association rules. In *Worksop Notes Statistics, Machine Learning and Knowledge Discovery in Databases, ECML-95*.

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