Knowledge Discovery in Textual Databases (KDT)

Ronen Feldman and Ido Dagan

Math and Computer Science Dept.
Bar-Ilan University
Ramat-Gan, ISRAEL 52900
{feldman,dagan}@bimacs.cs.biu.ac.il

Abstract

The information age is characterized by a rapid growth in the amount of information available in electronic media. Traditional data handling methods are not adequate to cope with this information flood. Knowledge Discovery in Databases (KDD) is a new paradigm that focuses on computerized exploration of large amounts of data and on discovery of relevant and interesting patterns within them. While most work on KDD is concerned with structured databases, it is clear that this paradigm is required for handling the huge amount of information that is available only in unstructured textual form. To apply traditional KDD on texts it is necessary to impose some structure on the data that would be rich enough to allow for interesting KDD operations. On the other hand, we have to consider the severe limitations of current text processing technology and define rather simple structures that can be extracted from texts fairly automatically and in a reasonable cost. We propose using a text categorization paradigm to annotate text articles with meaningful concepts that are organized in hierarchical structure. We suggest that this relatively simple annotation is rich enough to provide the basis for a KDD framework, enabling data summarization, exploration of interesting patterns, and trend analysis. This research combines the KDD and text categorization paradigms and suggests advances to the state of the art in both areas.

Introduction

Knowledge discovery is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from given data [Piatetsky-Shapiro and Frawley 1991]. Algorithms for knowledge discovery ought to be efficient and discover only interesting knowledge. In order to be regarded as efficient, the complexity of the algorithm must be polynomial (with low degree) both in space and time. Algorithms that can not meet this criteria won't be able to cope with very large databases. Knowledge would be regarded as interesting if it provides some nontrivial and useful insight about objects in the database. There are two main major bodies of work in knowledge discovery. The first is concentrated around applying machine learning and statistical analysis techniques towards automatic discovery of patterns in knowledge bases, while the other body of work is concentrated around providing a user guided environment for exploration of data. Among the systems that belong to the first group we can mention EXPLORA (Klosgen, 1992), KDW (Piatetsky-Shapiro and Matheus, 1992), and Spotlight (Anand and Kahn, 1991). Among the systems the belong to the second group we can mention IMACS (Brachman et al, 1992) and Nielsen Opportunity Explorer (Anand and Kahn 1993).

Most previous work in knowledge discovery was concerned with structured databases. In reality a large portion of the available information does not appear in structured databases but rather in collections of text articles drawn from various sources. However, before we can perform any kind of knowledge discovery in texts we must extract some structured information from them. Here we show how the Knowledge Discovery in Texts (KDT) system is using the simplest form of information extraction, namely the categorization of the topics of a text by meaningful concepts. While more complex types of information have been extracted from texts, most notably in the work presented at the series of Message Understanding Conferences (MUC), text categorization methods were shown to be simple, robust and easy to reproduce. Therefore text categorization can be considered as an acceptable pre-requisite for initial KDT efforts, which can be later followed by the incorporation of more complex data types.

Data Structure: the Concept Hierarchy

In order to perform KDD tasks it is traditionally required that the data will be structured in some way. Furthermore, this structure should reflect the way in which the user conceptualize the domain that is described by the data.

Most work on KDD is concerned with structured data bases, and simply utilizes the given database structure for the KDD purposes. In the case of unstructured texts, we have to decide which structure to impose on the data. In doing so, we have to consider very carefully the following tradeoff. Given the severe limitations of current technology in robust processing of text we need to define rather simple structures that can be extracted from texts fairly automatically and in a reasonable cost. On

the other hand, the structure should be rich enough to allow for interesting KDD operations.
In this paper, we propose a rather simple, data structure, which is relatively easy to extract from texts. As described below, this data structure enables interesting KDD operations. Our main goal is to study text collections by viewing and analyzing various concept distributions. Using concept distributions enables us to identify distributions that highly deviate from the average distribution (of some class of objects) or that are highly skewed (when expecting a uniform distribution). After identifying the limits of using this data structure it will be possible to extract further types of data from the text, enhance the KDD algorithms to exploit the new types of data and examine their overall contribution to the KDD goals.

The Concept Hierarchy

The concept hierarchy is the central data structure in our architecture. The concept hierarchy is a directed acyclic graph (DAG) of concepts where each of the concepts is identified by a unique name. An arc from concept A to B denotes that A is a more general concept than B (i.e., `communication` → `wireless communication` → `cellular phone`, `company` → `IBM`, `activity` → `product announcement`). A portion of the "technology" subtree in the concept hierarchy is shown in Figure 1 (the edges point downward).
The hierarchy contains only concepts that are of interest to the user. Its structure defines the generalizations and partitioning that the user wants to make when summarizing and analyzing the data. For example, the arc `wireless communication` → `cellular phone` denotes that at a certain level of generalization, the user wants to aggregate the data about cellular phones with the data about all other daughters of the concept "wireless communication". Also, when analyzing the distribution of data within the concept "wireless communication", one of the categories by which the data will be partitioned is "cellular phones". Currently, the concept hierarchy is constructed manually by the user. As future research, we plan to investigate the use of document clustering and term clustering methods (Cutting et al, 1993; Pereira et al. 1993) to support the user in constructing a concept hierarchy that is suitable for texts of a given domain.

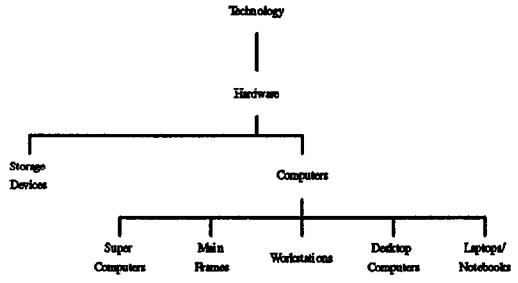

Figure 1 - Concept Hierarchy for technological concepts

Tagging the text with concepts

Each article is tagged by a set of concepts that correspond to its content (e.g. {IBM, product announcement, Power PC}, {Motorola, patent, cellular phone}). Tagging an article with a concept entails implicitly its tagging with all the ancestors of the concept in the hierarchy. It is therefore desired that an article will be tagged with the lowest concepts possible. In the current version of the system these concept sets provide the only information extracted from an article, each set denoting the joint occurrence of its members in the article.

For the KDD purposes, it does not matter which method is used for tagging. As was explained earlier, it is very realistic to assume automatic tagging by some text categorization method. On the other hand, tagging may be semi-automatic or manual, as common for many text collections for which keywords or category labels are assigned by hand (like Reuters, ClariNet and Individual).

KDD over concept distributions

Concept Distributions

The KDD mechanism summarizes and analyzes the content of the concept sets that annotate the articles of the database. The basic notion for describing this content is the distribution of daughter concepts relative to their siblings (or more generally, the distribution of descendants of a node relative to other descendants of that node). Formally, we set a concept node C in the hierarchy to specify a discrete random variable whose possible values are denoted by its daughters (from now on we relate to daughters for simplicity, but the definitions can be applied for any combination of levels of descendants). We denote the distribution of the random variable by $P(C=c)$, where c ranges over the daughters of C. The event $C=c$ corresponds to the annotation of a document with the concept c. $P(C=c_i)$ is the proportion of documents annotated with c_i among all documents annotated with any daughter of C.

For example, the occurrences of the daughters of the concept C="*computers*" in the text corpus may be distributed as follows: $P(C$="*mainframes*"$)=0.1$; $P(C$="*work-stations*"$)=0.4$; $P(C$="*PCs*"$)=0.5$.

We may also be interested in the joint distribution of several concept nodes. For example, the joint distribution of C_1=*company* and C_2="*computers*" may be as follows (figures are consistent with those of the previous example): $P(C_1=IBM, C_2=mainframe)=0.07$; $P(C_1=Digital, C_2=mainframe)=0.03$; $P(C_1=IBM, C_2=work\text{-}stations)=0.2$; $P(C_1=Digital, C_2=work\text{-}stations)=0.2$; $P(C_1=IBM, C_2=PCs)=0.4$; $P(C_1=Digital, C_2=PCs)=0.1$. A data point of this distribution is a joint occurrence of daughters of the two concepts *company* and "*computers*".

The daughter distribution of a concept may be conditioned on some other concept(s), which is regarded as a conditioning event. For example, we may be interested in the daughter distribution of C="*computers*" in articles which discuss announcements of new products. This distribution is denoted as $P(C=c \mid announcement)$, where *announcement* is the conditioning concept. $P(C=mainframes \mid announcement)$, for example, denotes the proportion of documents annotated with both *mainframes* and *announcement* among all documents annotated with both *announcement* and any daughter of "computers"[1].

Concept distributions provide the user with a powerful way for browsing the data and for summarizing it. One form of queries in the system simply presents distributions and data points in the hierarchy. As is common in data analysis and summarization, a distribution can be presented either as a table or as a graphical chart (bar, pie or radar). In addition, the concept distributions serve to identify interesting patterns in the data. Browsing and identification of interesting patterns would typically be combined in the same session, as the user specifies which portions of the concept hierarchy she wishes to explore.

Comparing Distributions

The purpose of KDD is to present "interesting" information to the user. We suggest to quantify the degree of "interest" of some data by comparing it to a given, or an "expected", model. Usually, interesting data would be data that deviates significantly from the expected model. In some cases, the user may be interested in data that highly agrees with the model.

In our case, we use concept 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 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 this distance, in one or several distributions. Below we identify three types of model distributions, with which it is interesting to compare a given distribution of the data.

Model Distributions

The Uniform Distribution

Comparing with the uniform distribution tells us how much a given distribution is "sharp", or heavily concentrated on only few of the values it can take. For example, regard a distribution of the form $P(C=c \mid x_i)$, where C=*company* and x_i is a specific product (a daughter of the concept *product*). Distributions of this form will have a large distance from the uniform distribution for products x_i that are mentioned in the texts only in connection with very few companies (e.g., products that are manufactured by only few companies).

Using the uniform distribution as a model means that we establish our expectation only on the structure of the concept hierarchy, without relying on any findings in the data. In this case, there is no reason to expect different probabilities for different siblings (a uniformative prior).

Notice that measuring the KL-distance to the uniform distribution is equivalent to measuring the entropy of the given distribution, since $D(p||u) = log(N) - H(p)$, where u is the uniform distribution, N is the number of possible values in the (discrete) distribution, and H is the entropy function. Looking at $D(p||u)$ makes it clear why using entropy to measure the "interestingness", or the "informativeness" of the given distribution is a special case of the general framework, where the expected model is the uniform distribution.

[1] A similar use of conditional distributions appears in the EXPLORA system (Klosgen 1993). Our conditioned variables and conditioning events are analogous to Klosgen's dependent and independent varibales.

Sibling Distribution

Consider a conditional distribution of the form $P(C=c \mid x_i)$, 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, for *C=activity*, and x_i=*Ford*, we could expect a distribution that is quite similar to such distributions where the conditioning concept is another car manufacturer.

To capture this reasoning, we use $Avg\ P(C=c \mid x)$, the *average sibling distribution*, as a model for $P(C=c \mid x_i)$, 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 $P(C=activity \mid Ford)$ to the average distribution $Avg\ P(C=activity \mid x)$, where x ranges over all car manufacturers. 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.

In some cases, the user may be interested in comparing two distributions which are conditioned by two specific siblings (e.g. Ford and General Motors). In this case, the distance between the distributions indicates how much these two siblings have similar profiles, with regard to the conditioned class C (e.g. companies that are similar in their activity profile). Such distances can also be used to cluster siblings, forming subsets of siblings that are similar to each other[2].

Past Distributions (trend analysis)

One of the most important tools for an analyst is the ability to follow trends in the activities of companies in the various domains. For example, such a trend analysis tool should be able to compare the activities that a company did in certain domain in the past with the activities it is doing in those domains currently. An example conclusion from such analysis can be that a company is shifting interests and rather than concentrating in one domain it is moving to another domain.

Finding trends is achieved by using a distribution which is constructed from old data as the expected model for the same distribution when constructed from new data. Then, trends can be discovered by searching for significant deviations from the expected model.

Finding Interesting Patterns

Interesting patterns can be identified at two levels. First, we can identify interesting patterns by finding distributions that have a high KL-distance to the expected model, as defined by one of the three methods above. Second, when focusing on a specific distribution, we can identify interesting patterns by focusing on those components that mostly affect the KL-distance to the expected model. For example, when focusing on the distribution $P(C=activity \mid Ford)$, we can discover which activities are mentioned most frequently with Ford (deviation from the uniform distribution), in which activities Ford is most different than an "average" car manufacturer (deviation from the average sibling distribution), and which activities has mostly changed their proportion over time within the overall activity profile of Ford (deviation from past distribution).

A major issue for future research is to develop efficient algorithms that would search the concept hierarchy for interesting patterns of the two types above. In our current implementation we use exhaustive search, which is made feasible by letting the user specify each time which nodes in the hierarchy are of interest (see examples below). It is our impression that this mode of operation is useful and feasible, since in many cases the user can, and would actually like to, provide guidance on areas of current interest. Naturally, better search capabilities would further improve the system.

Implementation and Results

In order to test our framework, we have implemented a prototype of KDT in LPA Prolog for Windows. The prototype provides the user a convenient way for finding interesting patterns in the Text Corpora. The Corpora we used for this paper is the Reuters-22173 text categorization test collection. The documents in the Reuters-22173 collection appeared on the Reuters newswire in 1987. The 22173 documents were assembled and indexed with categories by personnel from Reuters Ltd. and Carnegie Group, Inc. in 1987. Further formatting and data file production was done in 1991 and 1992 by David D. Lewis and Peter Shoemaker.

The documents were tagged by the Reuters personnel with 135 categories from the Economics domain. Our prototype system converted the document tag files into a set of prolog facts. Each document is represented as prolog fact which includes all the tags related to the document. There are 5 types of tags: countries, topics, people, organizations and stock exchanges. The user can investigate the prolog database using this framework. The examples in this paper are related to the country and topic tags of the articles (which are the largest tag groups), although we have found interesting patterns in the other tag groups as well.

Typically the user would start a session with the prototype by either loading a class hierarchy from a file or by

[2]Notice that the KL-distance is an asymmetric measure. If desired, a symmetric measure can be obtained by the summing the two distances in both directions, that is, $D(p\|q)+D(q\|p)$.

building a new hierarchy based on the collection of tags of all articles. The following classes are a sample of classes that were built out the collection of countries mentioned in the articles: South America, Western Europe and Eastern Europe. In the next phase we compared the average topic distribution of countries in South America to the average topic distribution of countries in Western Europe. In the terms of the previous Section, we compared for all topics t the expression *Avg P(Topic = t | c)* where *c* ranges over all countries in South America to the same expression where c ranges over all countries in Western Europe. In the next tables we see the topics for which we got the largest KL-Distance between the suitable averages over the 2 classes. In Table 1 we see topics which have a much larger share in South America than in Western Europe. In Table 2 we see the topics which have a much larger share in Western Europe than in South America.

Table 1 - Comparing South America to Western Europe

Topic	Rel Entropy	% / # in S.A	% / # in W.E
coffee	0.414	21.6 / 201	0.3 / 37
loan	0.160	18.2 / 169	2.4 / 102
crude	0.055	13.3 / 124	5.1 / 86
coppe	0.023	2. 8 / 26	0.4 / 12
silver	0.017	1.4 / 12	0.1 / 7

Table 2 - Comparing Western Europe to South America

Topic	Rel Entropy	% / # in W.E	% / # in S.A
acq	0.119	9.5 / 373	0.5 / 9
cbond	0.067	5.8 / 230	0.4 / 6
earn	0.052	5.2 / 204	0.5 / 22
corp_news	0.035	1.8 / 71	0.05 / 1
money_fx	0.031	4.9 / 191	1.1 / 13
interest	0.029	2.6 / 101	0.2 / 4

We can see that (according to this text collection) countries South America have much larger portion of agriculture and rare metals topics, while Western Europe countries have a much larger portion of financial topics. In the next phase, we went into a deeper analysis of comparing the individual topic distribution of the countries in South America to the average topic distribution of all countries in South America. In Table 3 we see the topics in which the country topic distribution deviated considerably from the average distribution (i.e., the topics that mostly affected the KL-distance to the average distribution). From the table we can infer the following information:

- Columbia puts much larger emphasis on coffee than any other country in South America. (it is interesting to note that Brazil that has 47 articles about coffee, more than any other country, is below the class average for Coffee).
- Both Brazil and Mexico (not shown) have a large proportion of articles that talk about loans.

Table 3 - Comparing Topic Distributions of Brazil, and Columbia to *Avg P(Topic = t | South America)*

Topic	Relative Entropy	% (#) in Brazil	Avg % (#) in S.A.
ship	0.065	7.4 (27)	1.0 (32)
loan	0.063	29.6 (108)	18.2 (223)
earn	0.057	5.5 (20)	0.5 (22)
coffee	-0.029	12.9 (47)	21.6 (91)
orange	0.025	2.2 (8)	0.2 (8)
Topic	**Relative Entropy**	**% (#) in Colum**	**Avg % (#) in S.A.**
coffee	0.259	59.2 (29)	21.6 (91)
loan	-0.029	6.1 (3)	18.2 (223)
crude	0.014	16.3 (7)	13.3 (66)
cpi	0.013	4.1 (1)	2.0 (14)
trade	0.006	4.1 (1)	2.9 (26)

In Table 4 we see the results of a similar analysis that was done from the opposite point of view. In this case we built a class of all agriculture related topics and computed the distribution of each individual topic and compared it to the average distribution of topics in the class. We picked 2 of the topics that got the highest relative entropy and listed the countries that that mostly affected the KL-distance to the average country distribution.

Table 4 - Comparing Country Distributions of cocoa, and coffee to *Avg P(Country = c | Agriculture)*

Country	Relative Entropy	% (#) in cocoa	Avg % (#) in Agr.
uk	0.207	32.1 (34)	7.2 (252)
ghana	0.114	9.4 (10)	0.6 (16)
ivory coast	0.098	8.5 (9)	0.6 (16)
usa	-0.049	8.5 (9)	32.5 (1301)
Country	**Relative Entropy**	**% (#) in coffee**	**Avg % (#) in Agr.**
brazil	0.178	23.0 (47)	3.9 (132)
colombia	0.171	14.2 (29)	0.9 (42)
usa	-0.051	10.3 (21)	32.5 (1301)
uganda	0.038	2.9 (6)	0.2 (6)

Finding Elements with Small Entropy

Another KDD tool is aimed at finding elements in the database that have relatively low entropy, i.e., elements that have "sharp" distributions (a "sharp" distribution is a distribution that is heavily concentrated on a small fraction of the values it can take).

When the system computed the entropy of the topic distribution of all countries in the database we found that Iran (according to text collection used) that appears in 141 articles has an entropy of 0.508, where 69 of the articles are about crude, 59 are about ship, the other 13 times in which Iran appears belong to 13 different topics. Another country which has relatively low topic is

Columbia. In this case 75.5% of the topics in which Columbia is mentioned are crude (59.2%) and coffee(16.3%).

When the system computed the entropy of the country distribution of all topics we notice that the topic "earn" has very high concentration in 6 countries. More than 95% of the articles that talk about earning involve the countries USA, Canada, UK, West Germany, Japan and Australia. The other 5% are distributed among another 31 countries.

Summary

We have presented a new framework for knowledge discovery in texts. This framework is based on three components: The definition of a concept hierarchy, the categorization of texts by concepts from the hierarchy, and the comparison of concept distributions to find "unexpected" patterns. We conjecture that our uniform and compact model can become useful for KDD in structured databases as well. Currently, we are performing research in text categorization which has some similarity to that of (Hebrail and Marsais, 1992), which is geared to make the KDT system more feasible and accurate. In addition, we are building another layer to the system that will provide the user with textual conclusions based on the distribution analysis it is performing. We plan to use the KDT system for filtering and summarizing new articles. We conjecture that the concept distributions of articles marked as interesting by the user can be used for updating the user's personal news profile and for suggesting subscribing to news groups of similar characteristics.

Acknowledgments

The authors would like to thank Haym Hirsh and the anonymous reviewers for helpful comments. Ronen Feldman is supported by an Eshkol Fellowship.

References

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 Archaeology. International Journal of Intelligent and Cooperative Information Systems.

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.

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

Feldman R., 1994. Knowledge Discovery in Textual Databases. Technical Report. Bar-Ilan University, Ramat-Gan, Israel.

Frawley W.J., Piatetsky-Shapiro G., and Matheus C.J., 1991. Knowledge Discovery in Databases: An Overview. In knowledge Discovery in Databases eds. G. Piatetsky-Shapiro and W. Frawley, 1-27. Cambridge, MA: MIT Press.

Hebrail G., and Marsais J. Experiments of Texual Data Analysis at Electricite de France. In Proceedings of IFCS-92 of the International Federation of Classification Societies.

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

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

Lewis D. and Gale W., 1994. Training text classifiers by uncertainty sampling. In Proceedings of ACM-SIGIR Conference on Information Retrieval.

Mertzbacher M. and Chu W., 1993. Pattern-Based Clustering for Databases Attribute Values. In Proceedings of the 1993 workshop on Knowledge Discovery in Databases.