CERIAS Tech Report 2001-90
Data mining on text
by Christopher Clifton
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

Data Mining on Text

Dr. Chris Clifton
The MITRE Corporation
Bedford, MA 01730-1420 USA
clifton@mitre.org

Dr. Rick Steinheiser*
rickster11@msn.com

Abstract

Data mining technology is giving us the ability to extract meaningful patterns from large quantities of structured data. Information retrieval systems have made large quantities of textual data available. Extracting meaningful patterns from this data is difficult. Current tools for mining structured data are inappropriate for free text. We outline problems involved in Knowledge Discovery in Text, and present an architecture for extracting patterns that hold across multiple documents. The capabilities that such a system could provide are illustrated.

Data mining technology has created a new opportunity for exploiting the information in databases. Much of the success has been in support of marketing. Patterns in the data, such as associations among similar items purchases, enables targeting marketing to focus on what customers are likely to purchase.

The Department of Defense has similar needs for knowledge, even if it does not have "customers" or "marketing" in the traditional sense. However, we can see obvious analogies. One view of commercial uses of data mining is to increase sales, however another view is improving service to the customer. Within the DOD, a "customer" may be another organization within the DOD; improving service to this organization is desirable and improves the efficiency of the DOD as a whole. As another analogy, we can view the "customer" as an enemy; in this case "targeted marketing" is literally choosing the targets most likely to end a war with minimal effort/loss of life.

The difficulty with this analogy is that the Department of Defense (as with many large organizations) stores much of its "corporate knowledge" as textual data. Corporate customer databases are generally engineered, and contain structured data that is easily used by data mining tools. Collections of text, however, are not as amenable to data mining. We are working on extending data mining technology to textual data.

The goal of this work is to find patterns that hold across a variety of documents. This is distinct from information retrieval, where the goal is to find a document (or set of documents) containing the desired information.[1] We assume that the knowledge we desire is not contained in a single document, but can only be found through the discovery of patterns holding across many documents.[2] For example, we may find that documents that mention the term **yyy** are also likely to mention the organization **xxx**. This pattern (with follow-on investigation of the documents) can lead to knowledge such as "Organization **xxx** is the strongest proponent of **yyy**".

1. System Overview

Current data mining solutions are highly optimized for a single pattern specification (or at best a limited set). This ill-suits Knowledge Discovery in Text, as the variability in information in a text bases means we will need flexibility in defining the type of pattern that interests us. We want to see industry develop a base that provides general services for data mining; this will ease the development of flexible KDT systems. We define an architecture (Figure 1) that supports processing of a range of declaratively-specified KDT problems, using such general services.

This two-part architecture connects an information retrieval systems to a "generalized data mining" engine. This approach limits the need to develop KDT-specific solutions. In addition, advances in both information retrieval and data mining technology will di-

*This work is based on discussions held with Dr. Steinheiser while visiting at the MITRE corporation.

[1] The term "Text Mining" has been used to refer to information retrieval, for example IBM's Textminer product is actually an information retrieval engine [3].

[2] This view was originally expressed in [2], and further agreed on in the ECML'98 Text Mining Workshop [4].

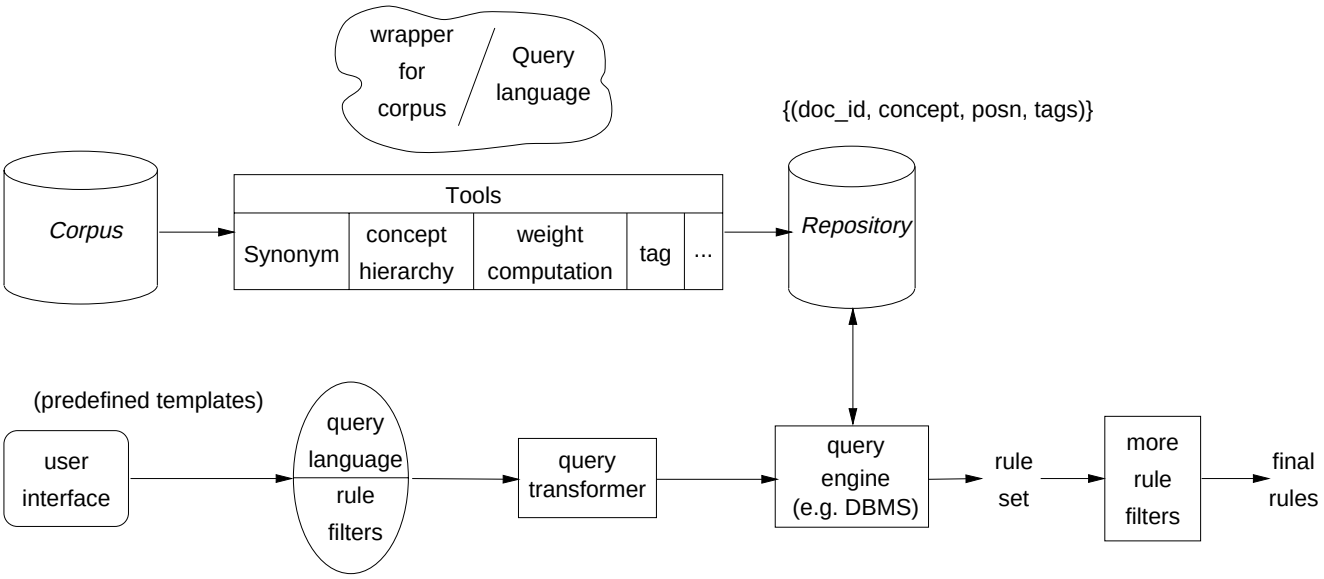

Figure 1. Text Mining System Architecture

rectly result in improvements in KDT capabilities.

We expect an end user will want the information retrieval and KDT systems to be integrated. The upper half of Figure 1 is based on information retrieval technology. Many of the tools, such as synonym matching and tagging, are already used as part of information retrieval products. As these tools improve in their support of information retrieval, they will also provide additional capabilities for KDT. The Repository may either be virtual (computed "on-the-fly" to support specific KDT requests), or may be precomputed. Some information retrieval systems already incorporate Information Extraction tools and precompute and store a repository that meets many of our requirements.

The lower half of the figure deals with the pattern detection process. Here there are significant differences from existing structured data mining tools. We view pattern generation as answering a specialized query. Existing data mining systems use a specific algorithm to find a specific type of pattern; we believe that text analysis demands greater flexibility. Instead we view the algorithmic issues as a query optimization problem (see [9] for further details). This allows us greater flexibility in the types of analysis performed. This "query optimization" approach is not limited to Knowledge Discovery in Text; such a system would support integration of data warehouses and data mining, thus we expect vendors will support a data mining approach based on query optimization.

The end user of such a system will not want to write complex specifications to define the type of patterns to be searched for. We expect that the user will see a selection of predefined templates. The templates can be used as components to create a pattern specification, giving the user considerable flexibility in defining the desired search. Handling pattern search using a common back-end rather than independent algorithms allows the templates to be combined and modified in ways not supported by current data mining tools.

We are currently developing a prototype / testbed based on this architecture. This makes use of named entity tagging (person, location, organization) to give a limited set of concepts. The prototype will use a fixed weighting rule (presence in title, number of appearances of a concept in a document). We use a relational database for the repository, with a hand-written wrapper to connect the Alembic output to the repository.

This prototype will serve as a testbed, primarily for experimenting with filtering techniques and interest measures. The use of a relational database allows us to use SQL as a filtering mechanism; for certain types of queries (e.g. association rules from the typed weighted market basket model) we can hand-optimize SQL queries to simulate the compiler. This will allow us to determine where future effort should be focused.

Another area needing development is the user interface. We plan to implement a forms-based interface supporting specific types of patterns and rule filtering, integrated with an information retrieval system if possible. This will give us the capability to demonstrate what a KDT system would look like, using live data and obtaining real results.

2. Sample Results

The data is from MITRE's Broadcast News Navigator [5], and consists of concepts extracted from closed-caption text of actual news broadcasts using the Alembic named-entity tagger [1]. The corpus consists of 6612 stories, with 13,737 distinct concepts mentioned. The associations between pairs of concepts (ranked by support: the number of documents containing the pair) are shown in Table 1. All correlations have at least 50% confidence: At least 50% of the stories mentioning one item in a pair also mention the other. Note that the highest support associations are between CNN co-anchors, this shows a problem with support as an interest measure for KDT patterns (we believe a choice of interest measures is the right approach).

Table 1 points out some issues addressed by the typed weighted market basket model. One is the co-reference problem: note the "association" found between Dole and Bob Dole. (There were 12 such associations with a support of at least 5; we have eliminated all but one for clarity.) Information Extraction tools do address this co-reference problem, however they are not yet as successful as with named entity tagging [8]. Even as information extraction tools improve, we still have a co-reference problem between documents; here the lack of synonym matching leads to false drops (an association would not be recognized because different documents use different terms for the concept). The Gramm/Alexander association shows this: The support should be higher, except that Alexander and Lamar Alexander were treated as different concepts.

A second difference between this model and "plain association rules" is the ability to use tags and varying relevance measures to restrict results to the patterns of interest. If we instead look for specific three-way associations between pairs of people and a location, and rank the results by the *correlation strength* (ratio of joint probability to independent probabilities), we end up with more focused results. Table 2 shows the results in this case. Note that all the associations show what we are looking for – conflict in a particular location. Those listed as "Hampshire" (actually the U.S. state of New Hampshire) reflect a election battle, rather than warfare.

This association is much closer to our original goal (significant warlords in a region), and is in fact probably all we could expect from such a broad "world news" corpus. Note that this association can be gleaned from Table 1, however the ability to clearly specify what we are looking for weeds out a lot of chaff.

Figure 2 gives an example user interface that would allow such flexibility in specifying the type of associations to search for. Note that this supports a broader notion of "pattern specification" than existing data mining tools (including, for example, capturing sequential associations as well as "normal" associations.)

We see another issue in Table 1: The problem posed by errors in tagging. We can characterize this error, and determine its effect on the patterns found (e.g. potential variability in the significance of results). A related problem, however, is more difficult: erroneous information in the corpus itself. To the extent we can characterize the reliability of the corpus, we can apply the same techniques to determine the effect of corpus reliability on the results.

3. Conclusions

A top priority is to increase the density of relevant results. The ability to easily and effectively focus the type of pattern to be returned (e.g., statistically significant associations between two people and a place, rather than simply "associations"), and to apply additional tests (such as excluding known patterns, or working in the context of external knowledge as in [7]), substantially increases the density of relevant results. Current data mining systems (products) use a specific algorithm to find a specific type of pattern. Knowledge Discovery in Text creates a need for greater flexibility.

Some of the specific technical challenges we see in extending data mining to Text are:

- Finding "low support" patterns. Many of the patterns of interest in text bases are relatively infrequent; this leads to questions of statistical significance of what is found. Once it is established what it means to be significant (which may not exactly match standard statistical meanings of significance), we have the question of how to *efficiently* find such patterns.

- Eliminating "common knowledge". Data mining within the context of a knowledge base (e.g. extended concept hierarchies [7]) can help to eliminate true, but trivial or obvious, patterns.

- Identifying *causality*. One problem with association rules is that associations imply a relationship, but not a causal relationship. Although it is extremely difficult to identify causal relationships, we can identify associations that are likely to be effects, rather than causes, thus weeding out some of the "chaff" [6].

One result of this project has been "query flocks" technology [9] that addresses the flexibility problem by

Table 1. Pairwise Associations in News Stories

Type	*Value*	*Type*	*Value*	*Support*
Person	Natalie Allen	Person	Linden Soles	117
Person	Leon Harris	Person	Joie Chen	53
Person	Ron Goldman	Person	Nicole Brown Simpson	19
Person	Dole	Person	Bob Dole	18
Person	Forbes	Person	Dole	16
Person	Forbes	Organization	New Hampshire	15
Person	Bill Cosby	Location	Cosby	14
Person	Pat Buchanan	Person	Forbes	12
Person	Steve Forbes	Organization	New Hampshire	12
Person	Mobutu Sese Seko	Location	Kinshasa	10
Person	Kelly Flinn	Organization	Adultery	9
Person	Mobutu Sese Seko	Person	Laurent Kabila	9
Person	Laurent Kabila	Location	Kinshasa	9
Person	Lamar Alexander	Person	Buchanan	9
Person	Gramm	Person	Buchanan	8
Person	Martin Luther King	Location	Rifle	8
Person	Buchanan	Location	Hampshire	7
Person	Buchanan	Person	Alexander	7
Person	Lamar Alexander	Person	Gramm	7
Person	Gramm	Person	Alexander	6
Person	Libby Robert	Person	Bupropion	6
Person	Liggett	Location	Liggett	6
Person	Phil Gramm	Organization	God	6
Person	Prince William	Person	Prince Harry	6
Person	Moines	Person	Des	6
Location	Moisture	Location	And East	5
Person	Delmar Simpson	Location	Army	5

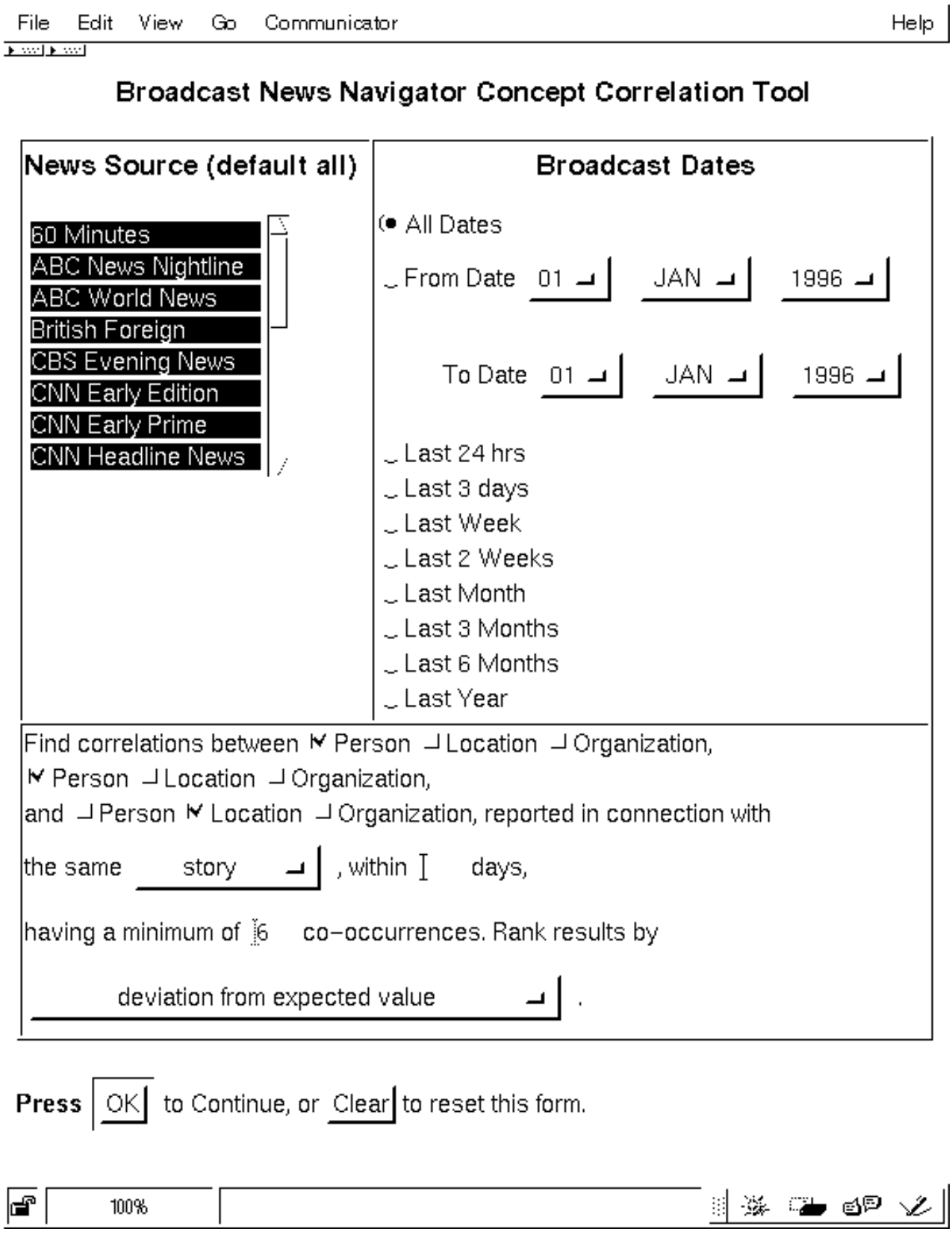

Figure 2. Interface to request correlations between PERSON, PERSON, and LOCATION

Table 2. Correlations between PERSON, PERSON and LOCATION, occurring in the same story having a minimum of 6 co-occurrences. Results ranked by deviation from expected value (top 20 shown).

Person 1	*Person 2*	*Location*	*co-occurrences*
Lamar Alexander	Buchanan	Hampshire	6
Mobutu Sese Seko	Laurent Kabila	Kinshasa	7
Pat Buchanan	Buchanan	Hampshire	7
Forbes	Buchanan	Hampshire	6
Dole	Buchanan	Hampshire	7
Buchanan	Bob Dole	Hampshire	7
Ray	James Earl Ray	Rifle	6
Mobutu Sese Seko	Laurent Kabila	Zaire	9
Pat	Buchanan	Hampshire	6
Pat Buchanan	Lamar Alexander	Hampshire	6
Lamar Alexander	Dole	Hampshire	6
Pat Buchanan	Forbes	Hampshire	6
Pat	Lamar Alexander	Hampshire	6
Lamar Alexander	Bob Dole	Hampshire	6
Forbes	Dole	Hampshire	7
Pat Buchanan	Dole	Hampshire	7
Steve Forbes	Dole	Hampshire	6
Forbes	Bob Dole	Hampshire	7
Pat Buchanan	Bob Dole	Hampshire	7
Steve Forbes	Bob Dole	Hampshire	6

treating data mining algorithmic issues as query optimization problems. In addition to providing the flexibility needed for text mining, this supports integration of data warehouses and data mining. This is a first step, but we feel more is necessary before Knowledge Discovery in Text will become a reality.

References

[1] D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain. Mixed initiative development of language processing systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., Mar. 1997.

[2] R. Feldman and I. Dagan. Knowledge discovery in textual databases (KDT). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 112–117, Montreal, Canada, Aug. 1995. AAAI.

[3] IBM. Advanced Search Engine: TextMiner. http://www.software.ibm.com/data/iminer/fortext/advancedSE.html, 1998.

[4] Y. Kodratoff, editor. *Proceedings of the European Conference on Machine Learning Workshop on Text Mining*, Chemnitz, Germany, Apr. 1998.

[5] M. Maybury, A. Merlino, and D. Morey. Broadcast news navigation using story segments. In *ACM International Multimedia Conference*, Seattle, WA, Nov. 1997.

[6] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structure. In *Proceedings of the 24th International Conference on Very Large Databases*, New York City, USA, Aug. 1998.

[7] L. Singh, P. Scheuermann, and B. Chen. Generating association rules from semi-structured documents using an extended concept hierarchy. In *Proceedings of the Sixth International Conference on Information and Knowledge Management*, Las Vegas, Nevada, Nov. 1997.

[8] B. M. Sundheim. Oerview of results of the MUC-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 13–31, Columbia, MD, Nov. 1995.

[9] D. Tsur, J. D. Ullman, S. Abiteboul, C. Clifton, R. Motwani, S. Nestorov, and A. Rosenthal. Query flocks: A generalization of association rule mining. In *Proceedings of the 1998 ACM SIGMOD Conference on Management of Data*, pages 1–12, Seattle, WA, June 1998.