1

A Survey of Emerging Trend Detection in Textual Data Mining

April Kontostathis
Leon M. Galitsky
William M. Pottenger
Soma Roy
Daniel J. Phelps

Abstract

In this chapter we describe several systems that detect emerging trends in textual data. Some of the systems are semi-automatic, requiring user input to begin processing, others are fully-automatic, producing output from the input corpus without guidance. For each Emerging Trend Detection (ETD) system we describe components including linguistic and statistical features, learning algorithms, training and test set generation, visualization and evaluation. We also provide a brief overview of several commercial products with capabilities for detecting trends in textual data, followed by an industrial viewpoint describing the importance of trend detection tools, and an overview of how such tools are used.

This review of the literature indicates that much progress has been made toward automating the process of detecting emerging trends, but there is room for improvement. All of the projects we reviewed rely on a human domain expert to separate the emerging trends from noise in the system. Furthermore, we discovered that few projects have used formal evaluation methodologies to determine the effectiveness of the systems being created. Development and use of effective metrics for evaluation of ETD systems is critical.

Work continues on the semi-automatic and fully-automatic systems we are developing at Lehigh University [HDD]. In addition to adding formal evaluation components to our systems, we are also researching methods for automatically developing training sets and for

merging machine learning and visualization to develop more effective ETD applications.

1.1 INTRODUCTION

What is an emerging trend? An emerging trend is a topic area that is growing in interest and utility over time. For example, Extensible Markup Language (XML) emerged as a trend in the mid 1990s. Table 1.1 shows the results of an INSPEC®[INS] database search on the keyword 'XML' from 1994 to 1999 (no records appeared before 1994). As can be seen from this table, XML emerged from 1994 to 1997; by 1998 it was well represented as a topic area.

Year	**Number of documents**
1994	3
1995	1
1996	8
1997	10
1998	170
1999	371

Table 1.1. Emergence of XML in the mid-1990s.

Knowledge of emerging trends is particularly important to individuals and companies who are charged with monitoring a particular field or business. For example, a market analyst specializing in biotech companies might want to review technical and news-related literature for recent trends that will impact the companies he is tracking. Manual review of all the available data is simply not feasible. Human experts who are tasked with identifying emerging trends need to rely on automated systems as the amount of information available in digital form increases.

An Emerging Trend Detection (ETD) application takes as input a collection of textual data and identifies topic areas that are either novel or are growing in importance within the corpus. Current applications in ETD fall generally into two categories: fully-automatic and semi-automatic. The fully-automatic systems take in a corpus and develop a list of emerging topics. A human reviewer then peruses these topics and the supporting evidence found by the system to determine which are truly emerging trends. These systems often include a visual component that allows the user to track the topic in an intuitive manner [DHJ+98], [SA00]. Semi-automatic systems rely on user input as a first step in detecting an emerging trend [PD95], [RGP02]. These systems then provide the user with evidence that indicates whether the input topic is truly emerging, usually in the form of

user-friendly reports and screens that summarize the evidence available on the topic.

We begin with a detailed description of several semi-automatic and fully-automatic ETD systems in Section 1.2. We discuss the components of an ETD system including linguistic and statistical features, learning algorithms, training and test set generation, visualization and evaluation. In Section 1.3 we review the ETD capabilities in commercial products. Our conclusions are presented in Section 1.4. In Section 1.5, Dr. Daniel J. Phelps, Leader of Eastman Kodak's Information Mining Group, describes the role of ETD systems in modern corporate decision-making environments.

1.2 ETD SYSTEMS

As mentioned above, ETD systems can be classified as either fully-automatic or semi-automatic. Semi-automatic systems require user input as a first step in detecting the emerging trends in a topic area. As part of our ongoing research at Lehigh University, we have developed both fully- and semi-automatic systems that have successfully identified emerging trends. In this section we provide an overview of the components that are included in most ETD systems (input data sets, attributes used for processing, learning algorithms, visualization, evaluation), followed by a detailed description of several ETD systems.

We begin with a discussion on the data that is used in ETD systems. The most commonly used data repository for ETD emerged from the Topic Detection and Tracking (TDT) project [TDT] that began in 1997. TDT research develops algorithms for discovering and threading together topically related material in streams of data, such as newswire and broadcast news, in both English and Mandarin Chinese. The TDT project, while not directly focused on emerging trend detection, has nonetheless encouraged the development of various fully-automated systems that track topic changes through time. Several of those algorithms will be described in this section.

As part of the TDT initiative several data sets have been created. The TDT data sets are sets of news stories and event descriptors. Each story/event pair is assigned a relevance judgment. A relevance judgment is an indicator of the relevance of the given story to an event. Table 1.2 portrays several examples of the relevance judgment assignment to a story/event pair. Thus, the TDT data sets can be used as both training and test sets for ETD algorithms. The Linguistic Data Consortium (LDC) [Lin] currently has three TDT corpora available for system development, the TDT Pilot study (TDT-Pilot), the TDT Phase 2 (TDT2), the TDT Phase 3 (TDT3), as well as the TDT3 Arabic supplement.

Story Description	**Event**	**Relevance Judgment**
Story describes survivor's reaction after Oklahoma City Bombing	Oklahoma City Bombing	Yes
Story describes survivor's reaction after Oklahoma City Bombing	US Terrorism Response	No
Story describes FBI's increased use of surveillance in government buildings as a result of the Oklahoma City Bombing	Oklahoma City Bombing	Yes
Story describes FBI's increased use of surveillance in government buildings as a result of the Oklahoma City Bombing	US Terrorism Response	Yes

Table 1.2. Story/event pairs.

Not all of the systems we describe rely on the TDT data sets. Other approaches to creation of test data have been used, such as manually assigning relevance judgments to the input data and comparing the system results to the results produced by a human reviewer. This approach is tedious and necessarily limits the size of the data set. Some of the systems we present use databases such as INSPEC®, which contains engineering abstracts, or the United States patent database [US], which allows searching of all published US patents. The input data set, along with the selection of appropriate attributes that describe the input, is a critical component of an ETD system. Attribute selection is at the core of the tracking process, since it is the attributes that describe each input item and ultimately determine the trends.

The attributes obtained from the corpus data are input to the methods/techniques employed by each ETD system we describe below. As will become apparent, some research groups use traditional Information Retrieval (IR) methodologies to detect emerging trends, while others have focused more on traditional machine learning approaches such as those used in data mining applications.

Work in the areas of visualization-supported trend detection has explored multiple techniques for identifying topics. When a user is trying to understand a large amount of data, a system that allows an overview, at multiple

levels of detail and from multiple perspectives, is particularly helpful. One of the simplest approaches is a histogram, where bars indicate discrete values of actual data at some discrete point in time. Information visualization is meant to complement machine learning approaches for trend detection. Plotting the patterns along a timeline allows one to see the rate of change of a pattern over time. For each algorithm described below, we will discuss the visualization component, showing how the component enhances the trend detection capabilities of the system.

The evaluation of an emerging trend detection system can be based on formal metrics, such as precision (the percentage of selected items that the system got right) and recall (the proportion of the target items that the system found), or by less formal, subjective means (e.g., answers to usability questions such as: Is the visualization understandable?). The particulars of an evaluation are related to the goals of the method and thus can vary greatly, but some justification and interpretation of the results should always exist to validate a given system.

1.2.1 Technology Opportunities Analysis (TOA)

Alan L. Porter and Michael J. Detampel describe a semi-automatic trend detection system for technology opportunities analysis in [PD95]. The first step of the process is the extraction of documents (such as INSPEC®abstracts) from the knowledge area to be studied. The extraction process requires the development of a list of potential keywords by a domain expert. These keywords are then combined into queries using appropriate Boolean operators to generate comprehensive and accurate searches. The target databases are also identified in this phase (e.g., INSPEC®, COMPENDEX®[COM], US Patents, etc.).

The queries are then input to the Technology Opportunities Analysis Knowbot (TOAK), a custom software package also referred to as TOAS (Technology Opportunities Analysis System). TOAK extracts the relevant documents (abstracts) and provides bibliometric analysis of the data. Bibliometrics uses information such as word counts, date information, word co-occurrence information, citation information and publication information to track activity in a subject area. TOAK facilitates the analysis of the data available within the documents. For example, lists of frequently occurring keywords can be quickly generated, as can lists of author affiliations, countries, or states.

In [PD95], the authors present an example of how the TOAK system can be used to track trends in the multichip module sub field of electronic manufacturing and assembly. Figure 1.1 ([PD95]) shows a list of keywords that appear frequently with 'multichip module' in the INSPEC®database. The authors observed that multichip modules and integrated circuits (particularly hybrid integrated circuits) co-occurred very frequently. An additional search using the US Patent database showed that many patents had been

issued in the area of multichip modules. Furthermore, the integrated circuits activity was more likely to be US based, while large scale integration activity was more likely to be based in Japan.

TOAK is meant to be used by a human expert in an interactive and iterative fashion. The user generates initial queries, reviews the results and is able to revise the searches based on his/her domain knowledge. TOA represents an alternative approach to the time-consuming literature search and review tasks necessary for market analysis, technology planning, strategic planning or research [PD95].

Multichip Module Keywords and Frequencies [INSPEC Database]

Keyword	Number of articles	Keyword	Number of articles
Multichip modules	842	Circuit layout CAD	69
Packaging	480	Tape automated bonding	68
Hybrid integrated circuits	317	Printed circuit manufacture	66
Module	271	Printed circuit design	65
Integrated circuit technology	248	Thin film circuit	62
Integrated circuit testing	127	CMOS integrated circuits	56
Substrates	101	Soldering	50
VLSI	98	Optical interconnections	48
Surface mount technology	93	Lead bonding	44
Flip-chip devices	93	Integrated optoelectronics	43
Integrated circuit manufacture	88	Printed circuits	42
Ceramics	85	Production testing	41
Circuit reliability	80	Reliability	41
Polymer films	79	Microassembling	38
Cooling	70	Circuit CAD	35
Metallisation	69	Microprocessor chips	35

Figure 1.1. Co-occurrences with 'multichip modules' ([PD95]).

Input Data and Attributes

The INSPEC®database serves as the primary corpus for TOA and its related software, TOAK. Two opportunities exist for attribute selection. First (Table 1.3), a list of keywords (a single word or multiple words, termed *n*-grams[1]) and their possible combinations (using Boolean operators) are supplied to TOAK, which retrieves all relevant items. The number of keyword occurrences and keyword co-occurrences - the appearance of two keywords in the same item - are calculated per year and over all years. A second pass (Table 1.4) involves selecting all phrases (single- and multiword) from a specific field and calculating the number of items that contain each phrase. For example, every phrase in the `keyword` field of each item may be counted, or each phrase in the `affiliation` field [PD95].

[1] An *n*-gram is a sequence of n words. For example, the phrase 'stock market' is a bigram (or 2-gram).

Attribute	Detail	Generation
n-grams	E.g., multichip modules, ball grid array	Manual
Frequency	Count of n-gram occurrence	Automatic
Frequency	Count of n-gram co-occurrence	Automatic
Date	Given by year	Automatic

Table 1.3. TOA First Pass Attributes.

Attribute	Detail	Generation
Field	A section of an item (e.g., an indexing term or city name)	Manual
Frequency	Count of n-gram occurrence in a field	Automatic

Table 1.4. TOA Second Pass Attributes.

Learning Algorithms

Like most of the systems that facilitate trend detection in textual collections, TOA relies on the expertise of the user who is researching a given area. TOAK provides access to many different data sources, including INSPEC®, COMPENDEX®, US Patents and others, but is necessarily limited as not all R&D work is patented or published. The power of TOAK resides in the visual interface and easy access to different views of the data. There are no inherent learning algorithms present in the system; the user is solely responsible for trend detection.

Visualization

Visualizations in TOA include frequency tables, histograms, weighted ratios, log-log graphs, Fisher-Pry curves, and technology maps [PD95]. These tools present information graphically using various linking and clustering approaches such as multi-dimensional scaling. In multi-dimensional scaling the goal is to reduce an n dimensional space to two or three dimensions. For example, Figure 1.2 ([PJ01]) shows a mapping of the organizations that are affiliated with research in nanotechnology. In this case there are 40 affiliations, and the complete map (not shown) has 40 nodes. TOA can also present maps based on other attributes that are available in the data. Attributes such as source, country of origin or author are commonly used. Similar techniques are used to generate maps of keywords that represent relationships among frequently occurring index terms, and principal components maps that represent relationships among conceptual clusters.

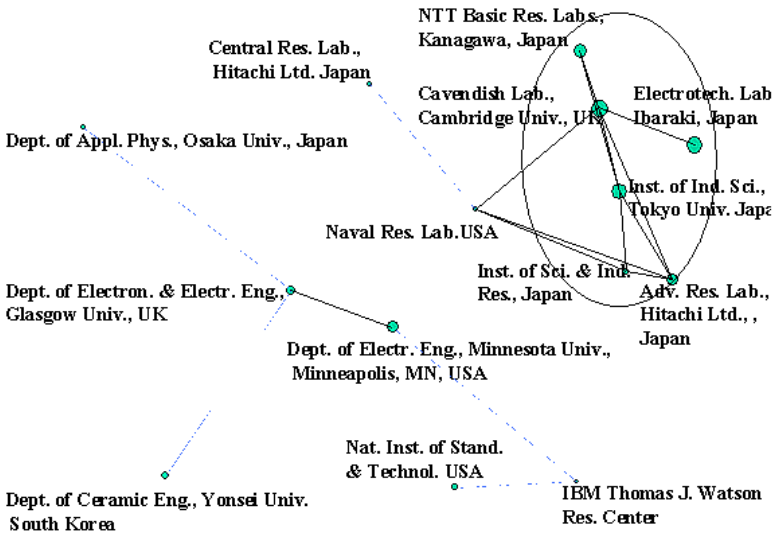

Figure 1.2. Organizations affiliated with nanotechnology research ([PJ01]).

These maps represent co-occurrence and correlative information gathered from within the dataset.

Evaluation

Identification of trends is left to the user in this semi-automatic method. TOA could, however, be evaluated on how well it presents information to the user. Visualizations are meant to significantly increase understanding of the data, and intuitively do. TOA developers, however, provide no evidence for the efficacy of these tools, apart from various author's claims. Solutions do however exist for evaluating this type of method. For example, results of usability studies and focus groups can strengthen arguments that visualization is indeed helpful. ThemeRiver^{TM}(Section 1.2.5) employs this usability approach for evaluation. Formal metrics, even with a semi-automatic method, can also be utilized as in CIMEL (Section 1.2.2).

1.2.2 CIMEL: Constructive, Collaborative Inquiry-based Multimedia E-Learning

CIMEL is a multi-media framework for constructive and collaborative inquiry-based learning that we, the authors of this survey, have developed [BPK$^+$01, BPK$^+$02, CIM]. Our semi-automatic trend detection methodology described in [RGP02] has been integrated into the CIMEL system in order to enhance computer science education. A multimedia tutorial has been developed to guide students through the process of emerging trend detection. Through the detection of incipient emerging trends, students see

the role that current topics play in course-related research areas. Early studies of this methodology, using students in an upper-level computer science course, show that use of the methodology improves the number of incipient emerging trends identified.

Our semi-automatic algorithm employs a more robust methodology than TOA because the user base is assumed to be individuals who are learning a particular area, as opposed to domain experts. The methodology relies on web resources to identify candidate emerging trends. Classroom knowledge, along with automated 'assistants', help students to evaluate the identified candidate trends. This methodology is particularly focused on incipient trends.

1. Identify a main topic area for research (e.g., Object databases)
2. Identify recent conferences and workshops in this area, e.g., OOPSLA for Object Oriented Programming
3. Review content and create a list of candidate emerging trends
4. Evaluate each emerging trend identified in step 3, using general web research tools (e.g., Google^{TM}search)
5. For each candidate emerging trend remaining after step 4, verify the trend using an INSPEC®database search

Table 1.5. Methodology for detecting emerging trends.

The methodology is outlined in Table 1.5. In step two of this methodology (after a main topic area has been identified) the user is directed to recent conferences and workshops online and instructed to review the content and develop a list of candidate emerging trends. Next, the user is directed to a general-purpose web search engine to find other references to candidate emerging trends identified in step three. Searches using the candidate trend phrase, along with terms such as 'recent research,' 'approach,' etc., are employed to improve the precision of the search results. The user is provided with a detailed algorithm that includes parameters for evaluation of the pages returned from the search engine. The candidate emerging trend may be rejected as a result of this search. In addition, other candidate emerging trends may be identified in this step.

Finally, the user is asked to verify candidate emerging trends using document count and author and publication venue spread based on an INSPEC®database search. To make the trend detection process easier, this step has been automated [Gev02]. The user need only enter a candidate emerging trend (Figure 1.3) identified in steps three and/or four, and the database search tool automatically generates document count, unique author sets, unique co-author sets and a list of unique venues (all across time) that pertain to the chosen candidate emerging trend. The tool also

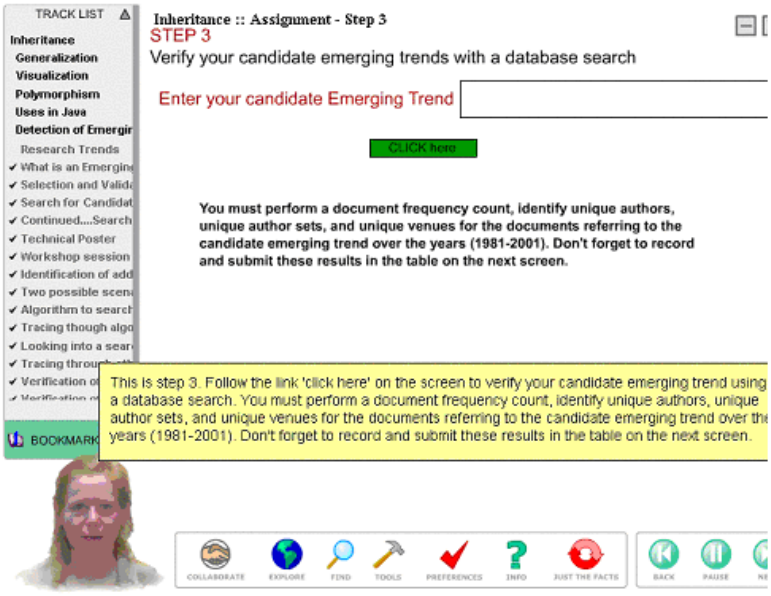

Figure 1.3. Emerging Trend Detection Tutorial ([CIM]).

provides a link to the corresponding abstracts, which can be accessed by clicking on individual document titles. This feature of the tool is important, as the user still has to decide whether a given candidate trend is truly emerging based on heuristics provided in the tutorial.

For example, students in an upper-level Object Oriented Software Engineering course might be asked to find an emerging trend in the field of Object Databases. Several conference web sites would be provided, including the Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA) website. A manual review of the content of papers presented at OOPSLA '01 leads the student to the candidate emerging trend 'XML Databases.' A search of the web using Google™results in additional papers related to XML databases, providing further evidence that 'XML Databases' is an emerging trend. Finally, the student is directed to the INSPEC®database. A search using XML <and> Databases <and> Object-oriented reveals the information depicted in Table 1.6. Further inspection reveals multiple author sets and publication venues, confirming that 'XML Databases' is an incipient emerging trend in the field of Object Databases.

Input Data and Attributes

The corpus for this semi-automatic methodology can be any web resource. A description of the main topic is chosen, which can consist of any text. An initial search of recent conferences and workshops is performed to identify

Year	Number of documents
Prior to 1999	0
1999	5
2000	11
2001	5

Table 1.6. 'XML Databases' is an Emerging Trend ([RGP02]).

candidate emerging trends. Using a web search engine, phrases associated with emerging trends[2] are used in conjunction with either the main topic or the candidate emerging trends to uncover additional evidence for the given candidate and/or to identify other candidate trends. Several attributes guide this initial decision-making process (Table 1.7), including the current year, the number of times either the main topic or candidate emerging trend appears on the page, the number of supporting terms on the page, and the line or paragraph containing the main topic/candidate emerging trend and supporting term [RGP02]. The validation step (Table 1.8) involves automatically calculating four frequencies across time: the number of unique documents, unique authors, unique author sets, and unique venues [Gev02]. These frequencies help the user make a final emerging trend determination. For example, an increase in the number of documents that reference the main topic and candidate emerging trend over time is indicative of a true emerging trend. On the other hand, if one or two documents appear in different years by the same author, the candidate trend may not actually be emerging [RGP02].

Learning Algorithms

Like TOA, the CIMEL system relies on the user to detect emerging trends. No machine-learning component is employed. Instead CIMEL relies on a precisely defined manual process. Like TOA, this system is restricted by the electronic availability of documentation in a given subject area. Furthermore, the INSPEC®query tool is currently based on abstracts that are downloaded to a local database, which must be periodically refreshed. Unlike TOA, CIMEL provides specific parameters for identifying an emerging trend, rather than relying solely on the domain expertise of the user.

Visualization

At the current time the visualization component for trend detection in CIMEL is under development.

[2]'Supporting' terms - the list of current associated 'supporting' terms: most recent contribution, recent research, a new paradigm, hot topics, emergent, newest entry, cutting edge strategies, first public review, future, recent trend, next generation, novel, new approach, proposed, current issues.

Attribute	Detail	Generation
n-grams	Main topic, e.g., 'object databases'	Manual
n-grams	Candidate trend, e.g., 'XML Databases'	Manual
n-grams	Supporting terms	Automatic
n-grams	Search item - any Boolean <and> combination of the previous attributes, e.g., 'XML <and> novel'	Automatic
Date	Given by year	Automatic
Frequency	Count of main topic/candidate trend in page	Automatic
Frequency	Count of 'supporting' terms	Automatic
n-grams	Line or paragraph containing the main topic/candidate trend and 'supporting' term in a given document	Manual

Table 1.7. CIMEL Initial Step Attributes.

Attribute	Detail	Generation
Frequency	Count of unique authors, per year	Automatic
Frequency	Count of unique documents, per year	Automatic
Frequency	Count of unique author sets, per year	Automatic
Frequency	Count of unique venues, per year	Automatic

Table 1.8. CIMEL Validation Step Attributes.

Evaluation

Several experiments have been conducted to evaluate the utility of the ETD component of the CIMEL system. In one such experiment, two groups of students in a Programming Languages class were asked to identify emerging trends in the area of Inheritance in Object Oriented Programming. Group B (experimental) viewed a multimedia tutorial on the methodology that included a case study; Group A (control) did not. Hypothesis testing was performed using the standard metric of precision. Precision for a student was calculated by dividing the number of actual emerging trends found (zero, one or two for this experiment) by the number of total

trends found (two, if the student completed the assignment successfully). Recall was not determined since a complete list of emerging trends was not available. A lower tail t-test concluded with 95% confidence that the mean precision of students that used the methodology (Group B) was significantly greater than the mean precision of students that did not use the methodology (Group A). These results provide convincing evidence that the ETD methodology employed in the CIMEL system is effective at detecting emerging trends [Roy02].

1.2.3 TimeMines

The TimeMines system [SJ00] takes free text data, with explicit date tags, and develops an overview timeline of statistically significant topics covered by the corpus. Figure 1.4 presents sample output from TimeMines. TimeMines relies on Information Extraction (IE) and Natural Language Processing (NLP) techniques to gather the data. The system employs hypothesis-testing techniques to determine the most relevant topics in a given timeframe. Only the 'most significant and important' information (as determined by the program) is presented to the user.

TimeMines begins processing with a default model that assumes the distribution of a feature depends only on a base rate of occurrence that does not vary with time. Each feature in a document is compared to the default model. A statistical test is used to determine if the feature being tested is significantly different than what the model would expect. If so, the feature is kept for future processing, otherwise it is ignored.

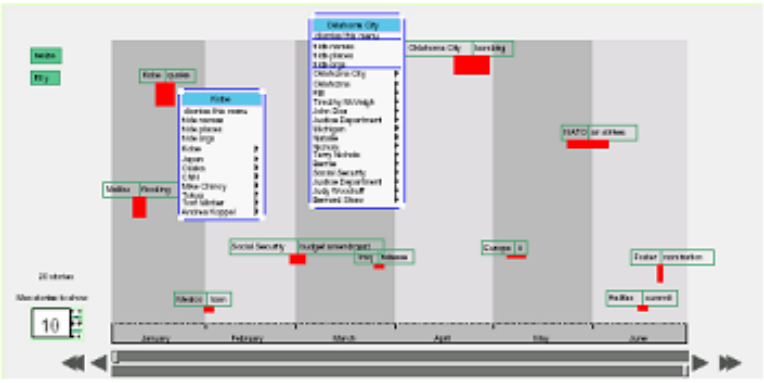

Figure 1.4. TimeMines Sample Output ([SJ00]).

The reduced set of features that is developed using the first round of hypothesis testing is then input into a second processing phase which groups related features together. The grouping again relies on probabilistic techniques that combine terms that tend to appear in the same timeframes

into a single topic. Finally, a threshold is used to determine which topics are most important and these are displayed via the timeline interface (Figure 1.4). The threshold is set manually, and is determined empirically.

Like TOA, TimeMines presents a model of the data without drawing any specific conclusions about whether or not a topic is emergent. It simply presents the most statistically significant topics to the user, and relies on the user's domain knowledge for evaluation of the topics.

Input Data and Attributes

In [SJ00], the TDT and TDT-2 corpora were date tagged and part-of-speech tagged with JTAG [XBC94]. (TDT-2 was preliminarily tagged with Nymble [BMSW97].) In the TimeMines system, an initial attribute list of all 'named entities' and certain noun phrases is generated. A named entity is defined as a specified person, location, or organization (extracted using the Badger IE system [FSM+95]). Noun phrases match the regular expression (N|J)* N for up to five words, where N is a noun, J is an adjective, | indicates union, and * indicates zero or more occurrences. The documents are thus represented as a 'bag of attributes', where each attribute is true or false (i.e., whether the named entity or noun phrase is contained in the document or not). The attributes are shown in Table 1.9.

Attributes	**Detail**	**Generation**
Named Entity	Person, location, or organization	Automatic
n-grams	Follows (N|J)*N pattern for up to five words, e.g., 'textual data mining'	Automatic
Presence	'True' if the named entity or *n*-gram occurs in the document, else 'False'. Each document has a presence attribute for each named entity and *n*-gram.	Automatic
Date	Given by day	Automatic

Table 1.9. TimeMines Attributes.

Learning Algorithms

There are two separate machine learning aspects present in the TimeMines application. First, TimeMines must select the 'most significant and important information' to display. To do this, TimeMines must extract the 'most significant' features from the input documents.

TimeMines uses a statistical model based on hypothesis testing to choose the most relevant features. As noted, the system assumes a stationary ran-

dom model for all features (n-grams and named entities) extracted from the corpus. The stationary random model assumes that all features are stationary (meaning their distributions do not vary over time) and the random processes generating any pair of features are independent. Features whose actual distribution matches this model are considered to contain no new information and are discarded. Features that vary greatly from the model are kept for further processing. The hypothesis testing is time dependent. In other words, for a specific block of time, a feature either matches the model (at a given threshold) or violates the model. Thus the phrase 'Oklahoma City Bombing' may be significant for one time slice, but not significant for another.

After the feature set has been pruned in this manner, TimeMines uses another learning algorithm, again based on hypothesis testing. Using the reduced feature set, TimeMines checks for features within a given time period that have similar distributions. These features are grouped into a single 'topic'. Thus each time period may be assigned a small number of topic areas, represented by a larger number of features.

One potential drawback of ranking the general topics derived from the significant attributes is discussed in [SJ00]. The occurrence of an attribute is measured against all other occurrences of it in the corpus. As a result a consistently heavily used attribute may not distinguish itself properly. The Kenneth Starr-President Clinton investigation is unquestionably the most covered story in the TDT-2 corpus, yet ranked 12^{th} because it is so prevalent throughout. Against a longer time period, including time after coverage had died down, the story probably would have ranked 1^{st}.

Like all of the algorithms we present here, the final determination of whether or not a topic is emerging is left to the user, but unlike CIMEL and TOA, the user does not direct the TimeMines system. This system is completely automated; given a time-tagged corpus it responds with a graphical representation of the topics that dominate the corpus during specific time periods.

Visualization

TimeMines generates timelines automatically for both visualization of temporal locality of topics and the identification of new information within a topic. The x-axis represents time, while the y-axis represents the relative importance of a topic. The most statistically significant topic appears near the top of the visualization (Figure 1.4). Each block in the visualization interface includes all the terms used to describe a topic and thus indicates the coverage within the corpus. Clicking on a term (named entity or n-gram) pops up a menu of all the associated features of that type within the topic, and a sub-menu option allows the user to choose this feature as the label, or to obtain more information about the feature. However no effort

is made to infer any hierarchical structure in the appearance of the feature in the timeline.

Evaluation

Two hypotheses are evaluated in [SJ00]: do term occurrence and co-occurrence measures properly group documents into logical time-dependent stories, and, are the stories themselves meaningful to people? A randomization test [Edg95] was conducted to support the first hypothesis. The documents were shuffled and assigned an alternate date, but were otherwise left intact. From an IR standpoint the corpus looked the same, since term frequency and inverse document frequency were preserved. The authors concluded that the results of this test overwhelming suggest the groupings are logical and not random.

The second hypothesis was explored with two methods of evaluation but results were inconclusive. The first evaluation method used precision and recall metrics from IR. The January 1996 *Facts on File* [Fac] listed 25 major stories, which were used as the 'truth' set to compare with the TimeMines-generated major stories. Recall was defined as the number of *Facts on File* major stories identified by TimeMines divided by the total number of *Facts on File* major stories. Precision was defined as the number of *Facts on File* major stories identified by TimeMines divided by the total number of TimeMines-identified major stories. A relatively low precision of 0.25 and a similarly low recall of 0.29 resulted.

The second evaluation method attempted to tune the threshold. Four students manually determined whether the automatic groupings related to zero, one, or multiple topics. Based on a pair wise Kappa statistic, however, the manual results could not be distinguished from random results [SA00].

1.2.4 New Event Detection

New event detection, also referred to as first story detection, is specifically included as a subtask in the TDT initiative. New event detection requires identifying those news stories that discuss an event that has not already been reported in earlier stories. New event detection operates without a predefined query. Typically algorithms look for keywords in a news story and compare the story with earlier stories. The approach taken in [APL98] implies that the input be processed sequentially in date order: i.e., only past stories can be used for evaluation, not the entire corpus.

A new event detection algorithm based on a single pass clustering algorithm is presented in [APL98]. The content of each story is represented as a query. When a new story is processed, all the existing queries (previous stories) are run against it. If the 'match' exceeds a predefined threshold (discussed below) the new story is assumed to be a continuation of the query story. Otherwise it is marked as a new story.

An interesting characteristic of news stories is that events often occur in bursts. Figure 1.5 ([YPC98]) portrays a temporal histogram of an event where the x-axis represents time in terms of days (1 through 365) and the y-axis is the story count per day.

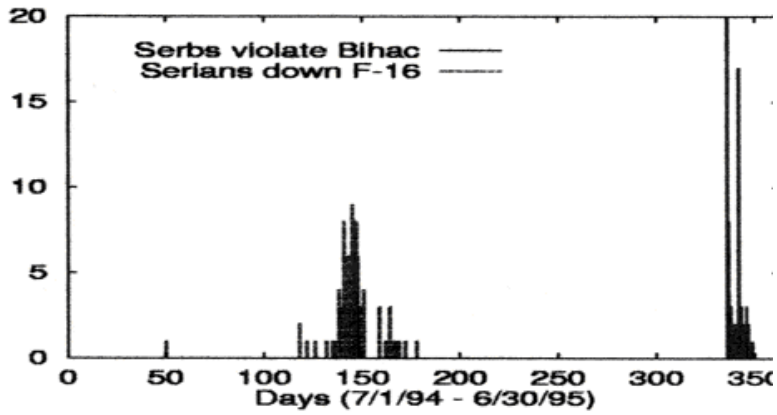

Figure 1.5. Temporal histogram of news data ([YPC98]).

News stories discussing the same event tend to be in temporal proximity and hence lexical similarity and temporal proximity are considered to be two criteria for document clustering. Also, a time gap between the bursts as exemplified in Figure 1.5 discriminates between distinct events, and the system is more likely to match stories that appear in the same timeframe.

As reported in [APL98], with proper tuning the algorithm was able to separate news stories related to the Oklahoma City Bombing from those about the World Trade Center bombing. However, some stories could not be detected. For example, the crash of Flight 427 could not be distinguished from other airplane accidents, and the OJ Simpson trial could not be separated from other court cases.

Input Data and Attributes

All stories in the TDT corpus deemed relevant to 25 selected 'events' were processed. For new event detection, each story is represented by a query and threshold. Table 1.10 lists all the attributes required for computing the query and threshold. The n most frequent single words comprise the query, and are weighted and assigned a 'belief' value by the Inquery system [ABC$^+$95], indicating the relevance of each word in the story to the query. Belief is calculated using term frequency and inverse document frequency. Term frequency is derived from the count of times the word occurs in the story, the length of the story, and the average length of a story in the collection. Inverse document frequency is derived from the count of stories in the collection and the count of stories that contain the word.

Attribute	Detail	Generation
Unigram	A single word	Automatic
Frequency	Number of times unigram occurs, per story	Automatic
Count	Total number of unigrams, per story	Automatic
Mean	Average number of unigrams per story	Automatic
Frequency	Number of stories in which unigram occurs	Automatic
Count	Number of stories	Automatic
Date	Given by available granularities	Automatic

Table 1.10. New Event Detection Attributes

Learning Algorithms

As noted, the approach presented in [APL98] is based on a single-pass clustering algorithm that detects new stories by comparing each story processed to all of the previous stories/queries detected. As each incoming story is processed, all previous 'queries' are run against it. If a story does not match any of the existing queries, the story is considered a new event.

The system relies on a threshold to match the queries to the incoming stories. The initial threshold for a query is set by evaluating the query with the story from which it originated. If a subsequent story meets or exceeds this initial threshold for the query, the story is considered a match. The threshold is used as input to a thresholding function based on the Inquery system described above [ABC+95]. Since new event detection implies that documents are processed in order, however, traditional IR metrics that are usually applied to an entire corpus (such as the number of documents containing the term and average document length) are not readily available. To overcome this problem, an auxiliary collection is used to provide this information to the Inquery system. The thresholding function takes advantage of the time dependent nature of the news story collection by using a time penalty that increases the value required to 'match' a story as stories grow further apart in time.

Like the TimeMines system, the new event detection system described here is completely automated. Given a corpus, it provides a list of 'new events' in the form of news stories that first describe an occurrence of an event. New event detection differs somewhat from ETD in that it is focused on the sudden appearance of an unforeseen event rather than the (more gradual) emergence of a trend.

Visualization

The new event detection system is based on Lighthouse [LA00b, LA00a], an interactive information retrieval system that provides a ranked list of search results together with two and three-dimensional visualizations of inter-document similarities. After events are extracted, a visual timeline is constructed to show how these events occur in time and relate to one another.

Evaluation

[APL98] evaluated their system using miss (false negative) and false alarm (false positive) rates as well as the metrics of precision and recall. Arriving at meaningful thresholds for these rates was difficult, and as a complement, Detection Error Tradeoff (DET) curves [MDOP97] were studied. DET curves highlight how miss and false alarm rates vary with respect to each other (each is plotted on an axis in a plane). A perfect system with zero misses and false alarms would be positioned at the origin, thus, DET curves 'closer' to the origin are generally better. 'Close' was defined as the Euclidean distance from the DET curve to the origin in [APL98]. Using nearly all (400) single-word attributes in the queries resulted in averages of 46% for the miss rate, 1.46% for the false alarm rate, 54% for recall, and 45% for precision.

1.2.5 ThemeRiver™

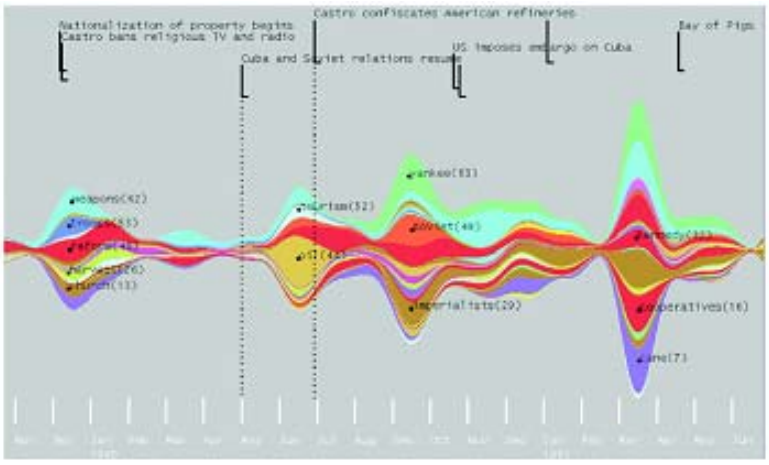

Figure 1.6. ThemeRiver™ sample output ([HHWN02]).

Similar to TimeMines, ThemeRiver™ [HHWN02] summaries the main topics in a corpus and presents a summary of the importance of each topic

via a graphical user interface. The topical changes over time are shown as a 'river' of information. The river is made up of multiple streams. Each stream represents a topic and each topic is represented by a color and maintains its place in the river relative to other topics. Figure 1.6 portrays an example visualization.

The river metaphor allows the user to track the importance of a topic over time (represented on the horizontal axis). The data represented in Figure 1.6 is from Fidel Castro's speeches. You can see that Castro frequently mentioned oil just before American oil refineries were confiscated in 1960 (shown as the second vertical line from the left in Figure 1.6). Oil is the large bubble immediately preceding this dotted line in the middle of the river. At no other time did Castro dwell on that topic in the 18-month period represented by this corpus.

Such patterns in the data may confirm or refute the user's knowledge of hypotheses about the collection. Like TOA and TimeMines, ThemeRiver^{TM}does not presume to indicate which topics are emergent. The visualization is intended to provide the user with information about the corpus. ThemeRiver^{TM}presents a topic- or feature-centered view of the data. This topic-centered view is a distinguishing characteristic of the ETD approaches surveyed in this chapter. Related areas in information retrieval, such as text filtering and text categorization are usually document-centered.

Input Data and Attributes

The corpus in the example presented in [HHWN02] consisted of speeches, interviews, articles, and other text about Fidel Castro over a 40-year period. ThemeRiver^{TM}automatically generates a list of possible topics, called theme words, of which a subset is manually chosen as attributes (the example in [HHWN02] narrowed the list to 64). Counts of the number of documents containing a particular theme word for each time interval provide the input for the method. An alternate count, using the number of occurrences of the theme word for each time interval is suggested but not implemented in [HHWN02].

An automatic method for generating the initial list of theme words was not specified, nor was the procedure for deciding which or how many of the theme words should be included in the subset. Theme word frequencies are computed after these attributes are chosen, effectively making attribute selection a manual process (i.e., not automatic based strictly on the counts, see Table 1.11).

Learning Algorithms

The ThemeRiver^{TM}application does not use a learning algorithm per se. Like TOA, it provides a view of the data that an experienced domain expert can use to confirm or refute a hypothesis about the data. ThemeRiver^{TM}begins by binning time-tagged data into time intervals. A

Attribute	Detail	Generation
Unigram	A single word	Manual
Frequency	Number of documents in which unigram occurs, per time interval	Automatic
Date	Given by month	Automatic

Table 1.11. ThemeRiver^{TM} Attributes.

set of terms, or themes, that represent the data is chosen and the river is developed based on the strength of each theme in the collection. As noted, the themes are chosen by automatically developing a list of words that are present in the data and then manually selecting a subset that represent various topics. The number of documents containing the word determines the strength of each theme in each time interval. Other methods of developing the themes and strengths are possible. The visual component of ThemeRiver^{TM} is the most important aspect of this work, particularly as it applies to trend detection.

Visualization

The ThemeRiver^{TM} system uses the river metaphor to show the flow of data over time (Figure 1.6). While the river flows horizontally, vertical sections of the river contain colored currents that identify topics or themes. The width of the river changes with the emergence or disappearance of topics, thereby making the system effective in cases where there is no major variation in topic.

The curves in Figure 1.6 show how interpolation is done to obtain a river metaphor. The idea is to produce a smooth curve with positive stream width for better visual tracking of the stream across time. Even though this technique aids human pattern recognition, a simple histogram can be more accurate. The algorithm interpolates between points to generate smooth curves (continuity in the flow of the river).

ThemeRiver^{TM} makes judicious use of color, leveraging human perceptual and cognitive abilities. Themes are sorted into related groups, represented by a color family. This allows viewing of a large number of (related) themes that can easily be separated due to color variation. For example in Figure 1.7, "germany", "unification", "gdr" and "kohl" are represented by different shades of green and hence can easily be identified as being related.

Evaluation

Evaluation, or usability in such visual applications, was conducted with two users in [HHWN02]. After being given some background information about the data, the users were asked about specifics related to the following five general questions:

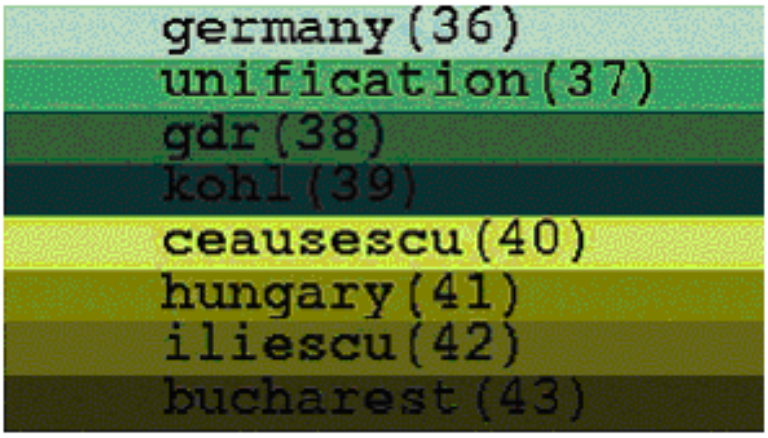

Figure 1.7. Color family representation in ThemeRiver^{TM}([HHWN02]).

- Did the users understand the visualization?
- Could they determine differences in theme discussion?
- Did the visualization prompt new observations about the data?
- Did the users interpret the visualization in any unexpected ways?
- How did the interpretation of the visualization differ from that of a histogram?

Observation, verbal protocol, and a questionnaire were used to gather feedback. This evaluation method is formalized well, but it lacks significance due to the small sample consisting of just two users.

1.2.6 PatentMiner

The PatentMiner system was developed to discover trends in patent data using a dynamically generated SQL query based upon selection criteria input by the user [LAS97]. The system is connected to an IBM DB2 database containing all granted United States (US) patents. There are two major components to the system, phrase identification using sequential pattern mining [AS95, SA96] and trend detection using shape queries.

Input Data and Attributes

As noted, in [LAS97] an IBM DB2 database containing all US Patents served as the basis for the corpus. Several procedures prepare the data for attribute extraction. Stop-words are removed. Identifiers are assigned to the remaining words, indicating position in the document and occurrences of sentence, paragraph, and section boundaries. After a subset of patents is

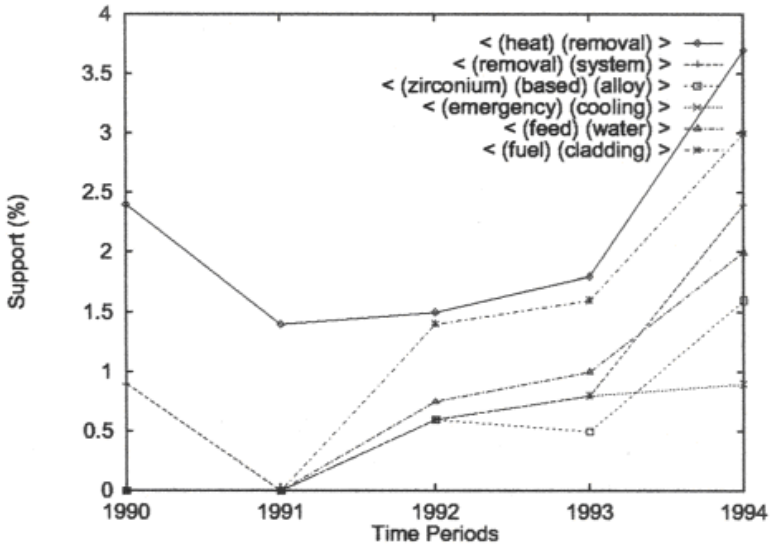

Figure 1.8. PatentMiner Sample Output ([LAS97]).

specified by category and date range, the Generalized Sequential Patterns (GSP) algorithm [SA96] selects user-defined attributes, called 'phrases'. Only phrases with support greater than a user-defined minimum are considered. A phrase can be any sequence of words, with a minimum and maximum 'gap' between any of the words. Gaps can be described in terms of words, sentences, paragraphs, or sections. For example, if the minimum sentence gap is one for the phrase 'emerging trends', than 'emerging' and 'trends' must occur in separate sentences. Or if the maximum paragraph gap is one, than 'emerging' and 'trends' must occur in the same paragraph. A time window indicates the number of words to group together before counting gap length. Finally, a shape definition language (SDL) [APWZ95] specifies which types of trends (e.g., upwards, spiked, etc.) are displayed. Table 1.12 summarizes these attributes.

The number of phrases selected can be substantial, given their very open-ended nature. Two pruning methods are discussed in [LAS97]. A sub-phrase of a phrase may be ignored if the support of the two phrases is similar. Or, a sub-phrase (general, higher-level) might be preferred over a longer phrase (specific, lower-level) initially, after which specific lower-level phrases could be easier to identify. This has the flavor of the technique used in CIMEL in which a 'main topic' is combined with a 'candidate trend' in order to improve the precision of the results.

Attribute	Detail	Generation
n-grams	Search phrase, e.g., emerging trends	Manual
Size	Minimum gap, with distinct gaps for words, sentences, paragraphs, and sections.	Manual
Size	Maximum gap, with distinct gaps for words, sentences, paragraphs, and sections.	Manual
Size	Time window, groups words in a phrase before determining gaps	Manual
Ratio	Support, number of search phrases returned divided by total number of phrases	Manual
Date	Given by available granularities	Manual
Shape	Graphical trend appearance over time, e.g., spiked or downwards	Manual

Table 1.12. PatentMiner Attributes.

Learning Algorithms

Most of the systems presented in this survey use traditional IR techniques to extract features from the text corpus that serves as input; the PatentMiner system takes a different approach. PatentMiner adapts a sequential pattern matching technique that is frequently used in data mining systems. This technique treats each word in the corpus as a transaction. The pattern matching system looks for frequently occurring patterns of words. The words may be adjacent, or separated by a variable number of other words (up to some maximum that is set by the user). This technique allows the system to identify frequently co-occurring terms and treat them as a single topic. [LAS97] refers to the resulting set of words (that make up a topic) as a 'phrase'.

As with TimeMines, documents in the input data set are binned into various collections based on their date information. The above technique is used to extract phrases from each bin and the frequency of occurrence of each phrase in all bins is calculated. A shape query is used to determine which phrases to extract, based on the user's inquiry.

The shape query processing is another learning tool borrowed from data mining [APWZ95]. In the PatentMiner system, the phrase frequency counts represent a data store that can be mined using the shape query tool. The shape query has the ability to match upward and downward slopes based on frequency counts. For example, a rapidly emerging phrase may occur frequently in two contiguous time slices, then level off, before continuing on an upward trend. The shape query allows the user to graphically define

various shapes for trend detection (or other applications) and retrieves the phrases with frequency distributions that match the query.

Like ThemeRiver^{TM}, TimeMines and others, the PatentMiner system presents a list of phrases to the user. The domain expert must then identify those that represent emerging trends.

Visualization

The system is interactive; a histogram is displayed showing the occurrences of patents by year based on the user's selection criteria. The user has the ability to focus on a specific time period and to select various shape queries to explore the trends as described above.

The phrases that match an increasing usage query on US patents in the category "Induced Nuclear Reactions: Processes, Systems and Elements" are shown in Figure 1.8.

Evaluation

Like TOA, the presentation of PatentMiner in [LAS97] lacks an evaluation component. While it automatically generates and displays potential trends, no claim is made as to the validity of these trends. The visualization is intuitive, but no user study on its effectiveness is reported in [LAS97]. In addition, no metrics are employed in [LAS97] to verify that the trends discovered are correctly identified.

1.2.7 HDDI^{TM}

Our research has led to the development of the Hierarchical Distributed Dynamic Indexing (HDDI^{TM}) system [PKM01, BCG^{+}01, BP00]. The HDDI^{TM}system supports core text processing including information/feature extraction, feature subset selection, unsupervised and supervised text mining & machine learning as well as evaluation for many applications, including ETD [HDD].

In [PY01] we describe our approach to the detection of emerging trends in text collections based on semantically determined clusters of terms. The HDDI^{TM}system is used to extract linguistic features from a repository of textual data and to generate clusters based on the semantic similarity of these features. The algorithm takes a snapshot of the statistical state of a collection at multiple points in time. The rate of change in the size of the clusters and in the frequency and association of features is used as input to a neural network that classifies topics as emerging or non-emerging.

Initially we modeled the complex non-linear classification process using neural networks. The datasets, which included three years of abstracts related to processor and pipelining patent applications, were separated by year and a set of concepts and clusters was developed for each year. In order to develop a training set, 14,530 concepts were extracted and manually

labeled. The system was, for example, able to correctly identify 'Low power CMOS with DRAM' as an emerging trend in the proper timeframe.

In follow-on experiments we were able to duplicate the precision achieved by the neural network with the C4.5 decision tree learning algorithm [Zho00]. The run time performance for training was significantly better with the decision tree approach. These experiments show that it is possible to detect emerging concepts in an on-line environment.

Like most other algorithms that we have reviewed, our approach relies on a domain expert for the final determination; thus the goal of the system is to identify emerging topics whenever possible (i.e., maximize recall) while not sacrificing precision. Unlike the first story detection algorithms, our research focuses on integrative or non-disruptive emergence of topics, as opposed to the sudden appearance of completely new topics.

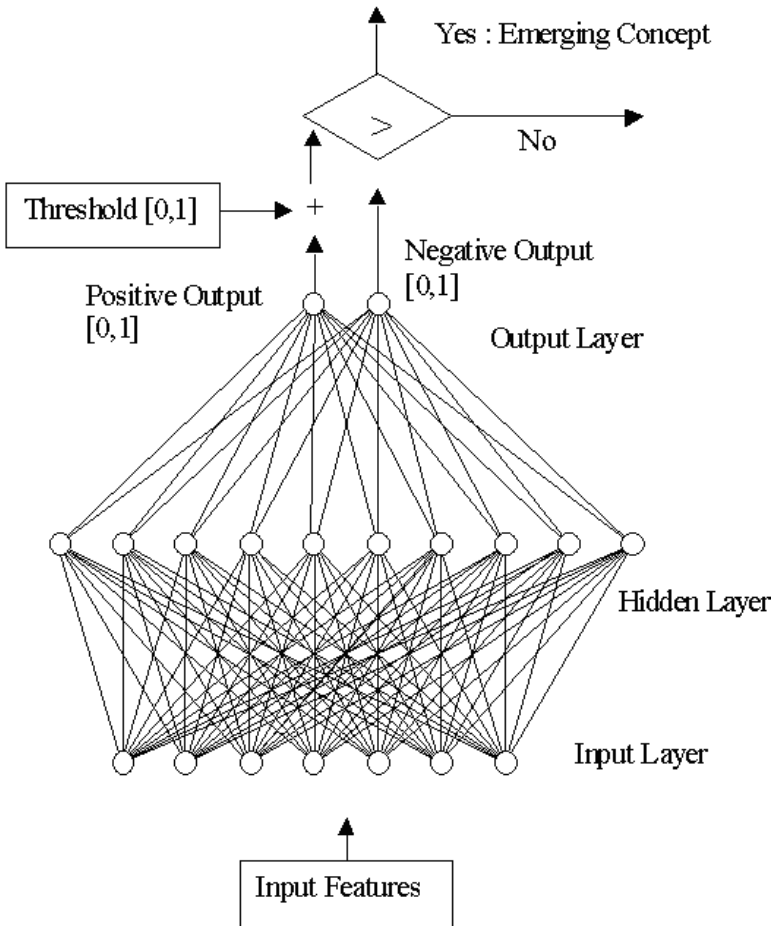

Figure 1.9. Using a neural net to detect emerging trends ([PY01]).

Input Data and Attributes

Four databases were used to formulate a corpus in [PY01]; the US patent database, the Delphion patent database [Del], the INSPEC®database, and the COMPENDEX®database. Initial attribute selection (Table 1.10) requires parsing and tagging before extraction. The parser retains only relevant sections of the original documents. The tagger maps a part-of-speech label to each word using lexical and contextual rules [Bri92]. A finite-state machine extracts complex noun phrases (concepts) according to the regular expression

$$C?(G|P|J) * N + (I * D?C?(G|P|J) * N+)* \tag{1.1}$$

where C is a cardinal number, G is verb (gerund or present participle), P is a verb (past participle), J is an adjective, N is a noun, I is a preposition, D is a determiner, ? indicates zero or one occurrence, | indicates union, * indicates zero or more occurrences, and + indicates one or more occurrence [BCG+01]. Counts of each concept and counts of co-occurrence of concept pairs are recorded at this point [PY01].

An asymmetric similarity between concept pairs is calculated based on a cluster weight function described in [CL92]. The concepts are then grouped into regions of semantic locality using sLoc, an algorithm we describe in [BP00]. The maximum, mean and standard deviation of the similarity, along with a parameter α that is a multiplication factor of the number of standard deviations, determine the threshold τ used in the first step of the sLoc algorithm. Cluster size is used in the last step (both are pruning mechanisms). As α decreases, τ increases and the number of connections between pairs of concepts decreases, resulting in smaller but more focused semantic regions. Too small a value of α could produce too many regions, while too large a value may result in only a single, large region. Thus, statistically finding the optimum value for α (and the threshold τ) is worthwhile, and work continues in this area. Empirical research supports an optimum value of $\alpha = 1.65$ [Yan00]. The identification of regions of semantic locality is an unsupervised learning step that produces values used as attributes in the ETD supervised learning process discussed following (see Table 1.13 for detail).

An emerging concept satisfies two principles: it should grow semantically richer over time (i.e., occur with more concepts in its region), and it should occur more often as more items reference it [PY01]. Using a cluster-based rather than an item-based approach, the artificial neural network model takes seven inputs (and one tuning threshold parameter) to classify a concept as emerging or not [PY01]. The seven inputs are described in Table 1.14.

Attribute	Detail	Generation
Regular expression	A concept (see text for definition), e.g., 'emerging trend detection'	Automatic
Frequency	Number of times each concept occurs over all documents	Automatic
Frequency	Number of co-occurrences of concept pairs over all documents	Automatic
Similarity	Arc weight between concepts	Automatic
Mean	Average arc weight	Automatic
Standard Deviation	Arc weight standard deviation	Automatic

Table 1.13. HDDI^{TM} Attributes for Regions of Semantic Locality.

Attribute	Detail	Generation
Frequency	Number of times concept occurs in trial year	Automatic
Frequency	Number of times concept occurs in the year before trial year	Automatic
Frequency	Number of times concept occurs in the year two years before trial year	Automatic
Frequency	Total number of times concept occurs in all years before trial year	Automatic
Count	Number of concepts in region containing the concept in trial year	Automatic
Count	Number of concepts in region containing the concept in the year before trial year	Automatic
Count	Number of words in the concept with length at least four	Automatic

Table 1.14. HDDI^{TM} Attributes for Emerging Trend Detection.

Learning Algorithms

As mentioned above, our fundamental premise is that computer algorithms can detect emerging trends by tracing changes in concept frequency and association over time. Our approach involves separating the data into time-determined bins (like PatentMiner and TimeMines) and taking a snapshot of the statistical relationships between terms. Two particular features were important in our model. Similar to other algorithms, the frequency of oc-

currence of a term should increase if the term is related to an emerging trend. Also, the term should co-occur with an increasing number of other terms if it is an emerging trend. To our knowledge, only our system has exploited term co-occurrence for automatic ETD.

As noted above, the first learning model we employed is a feed-forward, back-propagation artificial neural network (Figure 1.9). We used a standard three-layer network (one input layer, one hidden layer, one output layer). The number of hidden neurons was varied to optimize our results.

The attributes were extracted as described in the previous section and used as input to both the neural network model [PY01], and to various other data mining algorithms such as a decision tree, support vector machine, etc. [Zho00]. In all cases, we determined that the algorithms could be trained to detect emerging trends. As with other systems, precision was fairly low (although much better than the baseline) and final determination as to whether or not a term displayed by the system represents an emerging trend must be left to a domain expert.

Visualization

Visualization is ongoing for trend detection within the HDDI^{TM}system.

Evaluation

Both concept extraction and trend detection evaluations were performed. For concept extraction [BCG$^+$01], mean precision (number of system identified correct concepts / total number of system identified concepts) and mean recall (number of system identified correct concepts / total number of human expert-identified concepts) were calculated for several collections. Two of the test collections (drawn from the Grainger DLI database [UIU], the US Patent Office and the aforementioned commercial patent database Delphion) had precision ranges of [95.1, 98.7] and [95.2, 99.2] respectively, and recall ranges of [77.4, 91.3] and [75.6, 90.6] respectively, with 95% confidence.

Automatic trend detection performance in the HDDI^{TM}system was measured by precision, recall, and F_β [PY01]. An average precision of 0.317 constituted a 4.62 factor of improvement over random baseline precision; recall averaged 0.359. Either metric could be improved by altering the neural network threshold parameter. Since good recall was the primary focus, F_β, a weighted average of precision and recall with parameter β, was also examined. β is the precision weight and

$$F_\beta = \frac{(1+\beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}. \tag{1.2}$$

1.2.8 Other Related Work

[FD95] proposes a technique for development of a hierarchical data structure from text databases. This data structure then facilitates the study of concept distributions in text. The authors propose comparing the concept distributions from adjacent time periods. This approach to trend analysis seems promising; however, we were not able to obtain a more detailed description of the approach, or the experimental results, so we are unable to present a more comprehensive summary. Feldman has also been active in the development of commercial products for emerging trend detection (Section 1.3).

We have focused on research efforts that identify trends based primarily on the use of words and phrases; however, several research groups are using a different approach. [CC99], [PFL+00] and [Ley02] present algorithms that primarily employ citation information for trend detection.

Several systems focus more on the visualization of textual data and can be adapted to trend detection at the discretion of the user. One such system, Envision [NFH+96], allows users to explore trends graphically in digital library metadata (including publication dates) to identify emerging concepts. It is basically a multimedia digital library of computer science literature, with full-text searching and full-content retrieval capabilities. The system employs the use of colors and shapes to convey important characteristics of documents. For example, the interface uses color to show the degree of relevance of a document.

Plaisant et al. describe a visual environment called LifeLines for reviewing personal medical histories in [PMS+98]. The visualization environment presented in their work exploits the timeline concept to present a summary view of patient data.

The EAnalyst [LSL+00] analyzes two types of data, textual and numeric, both with timestamps. The system predicts trends in numeric data based on the content of textual data preceding the trend. For example, the system predicts the trend in stock prices based on articles published prior to the appearance of the (numeric) trend.

1.3 COMMERCIAL SOFTWARE OVERVIEW

Commercial software products are available to aid a company interested in ETD. Some companies specialize in providing content [Lex, Mor, Nor, IDC, Gar, Fac], some provide general purpose information retrieval capabilities such as feature extraction [Appa, Cap, Loc], document management [HyB, Ban], search and retrieval [Ban, HyB, Lex, Tex], categorization [INS, Int, Sem, Ser, SPSb, Str, Ver], and clustering [Appb, Cle, INS, Tex, Tho]. Although all of these products can be used to facilitate an ETD effort,

only a few have capabilities specifically geared toward trend analysis and detection. These products are briefly introduced in this section.

1.3.1 Autonomy

The Clusterizer™tool provided by Autonomy [Appb] provides support for ETD. Clusterizer™[Clu] takes a corpus as input and produces a set of clusters of information from the data. These clusters can be used to identify new topics in the data by taking a set of clusters from a previous time period and comparing it to a set of clusters in the current period. In essence, Clusterizer™displays a view of clusters over time. The tool is thus designed to show trends in clusters, including the appearance and disappearance of clusters, as well as changes in cluster size.

The pattern-matching algorithms of non-linear adaptive digital signal processing, Claude Shannon's principles of information theory, Bayesian inference, and neural networks form the core of Autonomy's technology [Cla99]. Since natural language contains much duplication, concepts that are repeated less frequently in a document are assumed to correspond to the essence of that document. Autonomy thus describes a document with patterns based on usage and frequency of terms. Adaptive probabilistic concept modeling is used to determine relevance between documents and further train the system [Aut].

The Autonomy Clusterizer™module assists a human domain expert in detecting trends. The Breaking News pane automatically finds new clusters of information by comparing clusters from different time periods. The Spectrograph pane is a visualization tool that plots clusters as lines over time. The color and width of the lines indicate the size and quality of the cluster; changes signal an increase or decrease in significance.

Like the research tools described in the previous section, Spectrograph is designed to provide the user with a view into the data. The domain expert must then use the data to form conclusions as to the validity of a given trend. In terms of evaluation, to the best of our knowledge no formal assessment has been conducted of the performance of these tools when used for ETD.

1.3.2 SPSS LexiQuest

LexiQuest products use advanced natural language processing technology to access, manage and retrieve textual information [SPSb]. LexiQuest LexiMine is a text mining tool designed to help the user obtain new insights by identifying key concepts and the relationships between them. It employs a combination of dictionary-based linguistic analysis and statistical proximity matching to identify key concepts (i.e., terms) as well as the degree of relationship between concepts [Bry02].

Concept identification is achieved through the use of unabridged dictionaries and thesauri that contain (multi-word) terms used to match terms in LexiMine's input text. Term occurrences and term co-occurrences are counted either by paragraph or document and are used to build relational concept maps. Although no machine learning algorithms are employed per se, LexiQuest's term similarity formulas are akin to those used in association mining [Bry02].

The concept relationships are portrayed in a graphical map that displays the cumulative occurrence of concepts. The map can be utilized to investigate trends. Further analysis can be achieved by importing LexiMine data into the related Clementine [SPSa] tool. As with many of the research and commercial tools discussed in this survey, the validity of trends is ultimately left to a human domain expert and tool performance is neither quantified nor evaluated in any formal way [Bry02].

1.3.3 ClearForest

ClearForest provides a platform and products to extract relevant information from large amounts of text and to present summary information to the user [Cle]. Two products, ClearResearch and ClearSight, are useful for ETD applications. ClearResearch is designed to present a single-screen view of complex inter-relationships, enabling users to view news and research content in context. ClearSight provides simple, graphic visualizations of relationships between companies, people and events in the business world. It also provides real-time updates of new product launches, management changes, emerging technologies, etc. in any specified context. Users can drill down further into each topic to view more information or read related articles [Gra02].

ClearForest uses a rule-based approach to identify 'entities' and 'facts'. An entity is a sequence of one or more words corresponding to a single concept. A fact is a relationship between entities. Rules can contain part-of-speech or stem identifiers for words, references to dictionaries and lexicons, and structural characteristics. The extraction engine applies rules to input documents and outputs tagged information. The precise location of all information extracted is recorded. Occurrence and co-occurrence of concepts are used by the analytic engine to summarize information. Visualization tools display this information in various levels of detail [Gra02].

ClearForest divides the detection process into four stages. The first, gathering information, is performed by search engines. The second and third, extracting and consolidating information, is managed by ClearForest. The last, identifying a valid trend, is handled by a human domain expert using ClearForest's Trends Graph display. In terms of evaluation, like LexiQuest, ClearForest has not performed formal evaluation of the ETD component [Gra02].

1.4 CONCLUSIONS AND FUTURE WORK

We have described several semi-automatic and fully-automatic ETD systems, providing detailed information related to linguistic and statistical features, training and test set generation, learning algorithms, visualization and evaluation. This review of the literature indicates that much progress has been made toward automating the process of detecting emerging trends, but there remains room for improvement. All of the systems surveyed rely on human domain expertise to separate emerging trends from noise in the system. As a result, research projects that focus on creating effective processes to both semi- and fully-automatically detect emerging trends, develop effective visualizations, and applying various learning algorithms to assist with ETD can and should continue.

In addition, we discovered that few systems, whether research or commercial in nature, have employed formal evaluation metrics and methodologies to determine effectiveness. The development and use of metrics for evaluation of ETD systems is critical. The results published to date simply do not allow us to compare systems to one another. In a step designed to address this issue we are in the process of building the HDDI^{TM} textual data mining software infrastructure that includes algorithms for formal evaluation of ETD systems (hddi.cse.lehigh.edu).

Wider use of the TDT [TDT] data sets will also be helpful in the process of standardizing evaluation of ETD systems. In addition, usability studies need to be conducted for visualization systems. Additional training sets geared specifically toward trend detection also need to be developed. Toward this end, we have developed a back end to our CIMEL system (Section 1.2.2) that gathers data generated by students who use the ETD component of CIMEL. This will aid us in developing techniques to automatically generate training sets for use in machine learning approaches to ETD.

We also note that projects tend to focus either on applying machine learning techniques to trend detection, or on the use of visualization techniques. Both techniques, when used alone, have proved inadequate thus far. Techniques that blend the use of visualization with machine learning may hold more promise. As a result, we are extending our HDDI^{TM} system to include a visualization component for trend detection. Early prototypes hold promise, but, as noted above, usability studies must be conducted to prove the effectiveness of our approach.

A final point: to the best of our knowledge, no formal studies have been conducted of the (manual) processes employed by domain experts in ETD. Such a study would employ standard tools such as surveys and focus groups

to develop a (manual) methodology for ETD[3]. We plan to pursue the execution of such a study in the near term in order to define a robust methodology that can be automated using extensions of the various techniques employed in our previous work [BCG+01, BPK+01, BPK+02, BP00, PY01, PCP01, PKM01, RGP02].

1.5 INDUSTRIAL COUNTERPOINT: IS ETD USEFUL? Dr. Daniel J. Phelps, Leader, Information Mining Group, Eastman Kodak Company

Background: The Information Mining Group at Eastman Kodak Company has been following developments in the text mining field since 1998. Initially, our interest was in using text mining tools to help us do a better job of understanding the content of patents. More recently, we have expanded our interest to include mining science and technology literature. We have had practical experience identifying suitable data sources, working with both custom and commercial tools, and presenting information in a form that our clients find useful. This background gives me a good perspective for commenting on the potential usefulness of Emerging Trend Detection (ETD) tools and some of the challenges that will arise in trying to use them in the corporate environment.

The objective of ETD is to provide an automated alert when new developments are happening in a specific area of interest. It is assumed that a detected trend is an indication that some event has occurred. The person using the ETD software will look at the data to determine the underlying development. Whether the development is important or not is a call of judgment that depends on the situation and the particular information needs of the person evaluating the data.

The need to become aware of new developments in science, technology, or business is critical to decision makers at all levels of a corporation. These people need to make better data-driven decisions as part of their daily work. They need data that is complete and available in a timely manner. Traditionally, people have learned about a majority of the new developments by reading various types of text documents or by getting the information from others who have read the documents. As the pace of new developments accelerates and the number of documents increases exponentially, it will no longer be possible for an individual to keep up with what is happening by using manual processes. There is a clear need for new

[3]Note that in [RGP02] we developed such a methodology, which has been partially automated in the CIMEL system (Section 1.2.2).

tools and methodologies to bring some level of automation to detect trends and new developments. ETD tools have the potential to play an important role in identifying new developments for corporate decision makers. These tools should help make it possible to look through more data sources for new developments and do it in less time than with current manual methods.

To better understand what capabilities an ETD tool must have to be useful, one has to look at who will be using the tool. There are three broad groups of potential users in a corporation. The first group is the analysts or information professionals who work to fulfill the information needs of others. The second is the individual contributors looking for information relevant to their own projects. The third is the managers who need to make strategic and/or tactical decisions.

Analysts work with information as the main component of their jobs. These people work on projects specified by clients. The output of a given project will be a report delivered to the client for use in the decision-making process. Analysts are trained in information retrieval techniques, text mining techniques, etc., and are familiar with the various information sources needed to complete a given project. They have to be able to communicate the results of the work in a form clients can easily use and understand. Taking time to learn new tools and methodologies is an expected part of the job.

An ETD tool that is targeted for use by analysts can be complex. The analysts will have to be given sufficient training to become proficient in its use. Because the analysts will use the tool for multiple projects, they will learn the capabilities and limitations of the tool and be able to recognize those areas where its application is appropriate. One would expect a sophisticated user interface that would allow analysts to access the relevant data sources, process the underlying text, and display the results in a meaningful way using computer graphics visualization techniques. The visualization scheme used must draw the analysts' attention to trends and allow them to drill down into the data to find out what developments lead to what trends. The determination of whether or not a detected trend is important is complicated by the fact that the analysts do not always know what clients will judge to be important. Interaction between the analysts and the clients is critical to ensure the needs of the clients are met. This is typically an iterative process as the analysts learn more about what information the clients need, and the clients find out what information is actually available. Once the analysis is done, the ETD tool should have the ability to export information to facilitate report generation.

The scientists, engineers, or business people who want to use ETD tools to obtain project-specific information need a tool that is easy to learn and intuitive to use. Connecting to the appropriate data sources and processing the data must be transparent to the users. This user group will typically have limited training in the tool and will use it only occasionally. They will not have the time to learn all the nuances of using the software. The infor-

mation that is used will be that which is delivered automatically. A simple, graphical user interface with easily interpreted graphical visualizations is required. These people have the advantage that they are performing the work for themselves, therefore, they can make the determination whether newly detected trends are actually important.

An ETD tool meant to be used by management personnel must automatically be connected to the appropriate data sources, have an intuitive user interface, be very easy to learn, and provide output in a consistent format with which the managers are comfortable. Extremely sophisticated visualizations that are difficult to interpret and require high levels of interaction will not be useful in this environment. In the current corporate culture, managers do not have the time to engage in anything but the most cursory training in new tools. This means they will probably not be able to operate the ETD tool effectively enough to complete the analysis themselves. They are generally more comfortable with having an analyst assemble the information and provide an executive summary. The summarized information and the underlying data could be presented to them using a browser format that would allow them to look at the high-level results and then drill down into the detail when they find something in which they are particularly interested.

No matter how capable an ETD tool becomes, the quality of the output will depend upon the quality of the input data. Because ETD tools are supposed to identify new developments, the data processed through the tool must be current. There are several news services, like Factiva, that supply all types of news on a continuous basis. Currently, there are not equivalent services for fundamental science and technology information. One has to search a variety of sources to find the required information. The sampling frequency used to extract the data from the news feed needs to reflect the rapidity in which things change in the area. Business and financial events happen much more frequently than changes in technology. One might set up the ETD tool to collect news data each day and process it to look for new trends. Since developments in science and technology occur at a slower pace, it might be appropriate to work in blocks of one week or one month. Processing information in one-year blocks is adequate for a retrospective look at what has happened, but it is not acceptable for most decision-making situations.

ETD systems will have to be customizable to meet the needs of specific clients. The ETD algorithms will eventually become proficient at determining when something has happened. However, whether or not the development is important depends on the needs of the person who is looking at the data. Broadly speaking, there are two types of client profile information that must be obtained. The first is the definition of the area of interest, and the second defines what is characterized as an important development in that area. Traditional alerting systems handle the first problem by setting up a user profile containing a search query that is run on a periodic

basis against specified data sources. Typically, the query is built on a trial-and-error basis by the information professional and the client. This is an iterative process. Some of the newer knowledge-management systems use training sets to determine the characteristics of the documents of interest and build the 'query'. Each new document is checked against the target characteristics and a decision is automatically made whether the document belongs to the area of interest or not. What remains to be seen is which approach will work best with a given ETD tool. Either process can take a significant amount of time for the client who wants the information. There is also the problem that the client's areas of interest will expand and change over time. Each time this happens, an existing profile will have to be modified or a new profile will have to be generated.

The problem of determining what is a significant event in a given area is handled in interactive systems by having the decision makers operate the tools themselves. If the decision makers are unable or unwilling to work directly with the tool, analysts will have to interview the decision makers and obtain the basic guidelines with which to work. The analysts will perform the analysis and compile a list of potentially significant developments for the decision maker's review. It would be best if this processed information were presented in a browser format that would allow the decision makers to drill down into the detail underlying any development they find to be of interest.

It is too early to predict the cost of a commercial ETD software system. If it is comparable to the knowledge management and text database mining software of today, it will cost tens of thousands to hundreds of thousands of dollars. It probably will carry the same sort of fee structure as the high-end software packages available today. Vendors charge an initial license purchase price and require an annual maintenance fee to provide technical support and updates of the software. Sometimes it is possible to buy the software individually by the 'seat', but often the vendors push to sell a corporate license. If only a few people will be using the software, then purchasing seats makes sense. If the software is actually going to be used across the enterprise, then a corporate license is probably the better choice. Another cost that is often overlooked is the impact on the corporate IT infrastructure. There can be a capital cost to purchase the high-performance hardware needed to run calculation-intensive ETD applications. Even when the application is run on existing in-house servers, there is usually the need to have a system administrator, and possibly a database administrator, available to keep the application up and running.

To get a picture of what an ETD tool might look like in the future, it is helpful to examine a perfect-world scenario for an executive information system that would include an ETD capability. The characteristics of such a scenario are depicted in Table 1.15.

Executive decision makers are extremely busy and want to make good, data-driven decisions as fast as possible. This means they cannot take the

1. Raw data processed into useful information
2. Sufficient information presented to meet current need
3. No irrelevant information presented
4. All information available immediately when needed
5. Information prioritized for the current need
6. Information presented in a format that is intuitively easy to understand
7. Information can be viewed at different levels of detail
8. Information can be viewed from multiple perspectives

Table 1.15. Perfect-world Scenario.

time to assemble and process the raw data themselves. They want complete, timely, processed information, sorted in an appropriate prioritized order for the decisions at hand. They do not want to waste time looking at redundant or irrelevant information. The information needs to be presented in a format that is intuitively easy to understand and can be looked at in different levels of detail and from multiple perspectives. An excellent ETD tool will have these same characteristics and will meet the needs of all three groups of potential users in the corporation.

There continues to be good progress made in knowledge management and text mining tools. Because ETD systems make use of these types of tools, I think there is a good possibility that practical ETD systems will eventually become available for fixed-information needs. Building a system that will keep up with a given decision maker's changing information needs will be difficult, unless a good method is found to automatically translate the subject areas of interest and the important developments criteria from the words of the user to the ETD system. It will always be a challenge to assure that data sources available for processing are adequate to support the needs of the decision maker.

In this section, I have reviewed some of the practical aspects of working with an ETD tool in a corporate environment. The real test for an ETD system is whether or not it provides useful information about new developments to the decision maker in a timely manner. The current systems do not seem to provide the required level of performance to be used extensively in the corporate environment. There is hope that new generations of ETD tools will be useful to corporate decision makers when they become available.

1.6 ACKNOWLEDGMENTS

It is with genuine gratitude that we thank the following individuals: Michael Berry, Editor of this volume and Chair of the 2001 SIAM Textmine

Workshop, and Doug Bryan at SPSS (LexiQuest) and Barry Graubart at ClearForest for providing timely information included in our commercial products section. We also wish to acknowledge the National Science Foundation for their role in funding the CIMEL project (Section 1.2.2) and our HDDI^{TM} textual data mining software infrastructure [HDD] under Grant Number 0087977. Finally, co-author William M. Pottenger would like to express his sincere gratitude to his Lord and Savior, Jesus Christ, for His continuing help in his life.

References

[ABC+95] J. Allan, L. Ballesteros, J. Callan, W. Croft, and Z. Lu.Recent experiments with inquery.In *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, pages 49–63, 1995.

[APL98] J. Allan, R. Papka, and V. Lavrenko.On-line new event detection and tracking.In *Proceedings of ACM SIGIR*, pages 37–45, 1998.

[Appa] Applied Semantics.Applied Semantics [online, cited July 2002].Available from World Wide Web: `www.appliedsemantics.com`.

[Appb] Applied Semantics.Autonomy [online, cited July 2002].Available from World Wide Web: `www.autonomy.com`.

[APWZ95] R. Agrawal, G. Psaila, E.L. Wimmers, and M. Zait.Querying shapes of histories.In *Proceedings of the 21st International Conference on Very Large Databases*, Zurich, Switzerland, Sep 1995.

[AS95] R. Agrawal and R. Srikant.Mining sequential patterns.In *Proceedings of the International Conference on Data Engineering (ICDE)*, Taipei, Taiwan, Mar 1995.

[Aut] Autonomy.Autonomy [online, cited July 2002].Available from World Wide Web: `www.autonomy.com/Content/Technology/Background/IntellectualFoundations`.

[Ban] Banter.Banter [online, cited July 2002].Available from World Wide Web: `www.banter.com`.

[BCG+01] R. Bader, M. Callahan, D. Grim, J. Krause, N. Miller, and W.M. Pottenger.The Role of the HDDI^{TM} Collection Builder in Hierarchical Distributed Dynamic Indexing.In *Proceedings of the Textmine '01 Workshop, First SIAM International Conference on Data Mining*, Apr 2001.

[BMSW97] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel.Nymble: a high-performance learning name-finder.In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201, 1997.

[BP00] F. Bouskila and W.M. Pottenger.The Role of Semantic Locality in Hierarchical Distributed Dynamic Indexing.In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI 2000)*, Las Vegas, Nevada, Jun 2000.

[BPK+01] G.D. Blank, W.M. Pottenger, G.D. Kessler, M. Herr, H. Jaffe, S. Roy, D. Gevry, and Q. Wang.Cimel: Constructive, collaborative inquiry-based multimedia e-learning.In *Proceedings of the 6th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE)*, Jun 2001.

[BPK+02] G.D. Blank, W.M. Pottenger, G.D. Kessler, S. Roy, D.R. Gevry, J.J. Heigl, S.A. Sahasrabudhe, and Q. Wang.Design and evaluation of multimedia to teach java and object-oriented software engineering.*ASEE*, Jun 2002.

[Bri92] E. Brill.A simple rule-based part of speech tagger.In *Proceedings of the Third Conference on Applied Natural Language Processing*. ACL, 1992.

[Bry02] D. Bryan, Jul 2002.Email correspondence.

[Cap] Captiva.Captiva [online, cited July 2002].Available from World Wide Web: `www.captivacorp.com`.

[CC99] C. Chen and L. Carr.A semantic-centric approach to information visualization.In *Proceedings of the 1999 International Conference on Information Visualization*, pages 18–23, 1999.

[CIM] CIMEL.CIMEL [online, cited July 2002].Available from World Wide Web: `www.cse.lehigh.edu/~cimel`.

[CL92] H. Chen and K.J. Lynch.Automatic construction of networks of concepts characterizing document databases.*IEEE Transaction on Systems, Man and Cybernetics*, 22:885–902, 1992.

[Cla99] S. Clarke.Knowlege Suite (Review) [online].1999 [cited July 2002].Available from World Wide Web: `www.autonomy.com/Extranet/Marketing/AnalystWhitePapers/ButlerReportonAutonomySuite200299.pdf`.

[Cle] ClearForest.ClearForest [online, cited July 2002].Available from World Wide Web: `www.clearforest.com`.

[Clu] Clusterizer^{TM}.Clusterizer^{TM} [online, cited July 2002].Available from World Wide Web: `www.autonomy.com/Extranet/Technical/Modules/TBAutonomyClusterizer.pdf`.

[COM] COMPENDEX®.COMPENDEX® [online, cited July 2002].Available from World Wide Web: `edina.ac.uk/compendex`.

[Del] Delphion.Delphion [online, cited July 2002].Available from World Wide Web: `www.delphion.com`.

[DHJ+98] G.S. Davidson, B. Hendrickson, D.K. Johnson, C.E. Meyers, and B.N. Wylie.Knowledge mining with VxInsight^{TM}: Discovery through interaction.*Journal of Intelligent Information Systems*, 11(3):259–285, 1998.

[Edg95] E. Edgington.*Randomization Tests*.Marcel Dekker, New York, NY, 1995.

[Fac] Factiva.Factiva [online, cited July 2002].Available from World Wide Web: `www.factiva.com`.

[FD95] R. Feldman and I. Dagan.Knowledge discovery in textual databases.In *Proceedings of the First International Conference on Knowledge Discovery (KDD-95)*. ACM, Aug 1995.

[FSM+95] D. Fisher, S. Soderland, J. McCarthy, F. Feng, and W. Lehnert.Description of the UMASS Systems as Used for MUC-6.In *Proceedings of the 6th Message Understanding Conference*, pages 127–140, Nov 1995.

[Gar] GartnerG2.GartnerG2 [online, cited July 2002].Available from World Wide Web: `www.gartnerg2.com/site/default.asp`.

[Gev02] D. Gevry.Detection of emerging trends: Automation of domain expert practices.Master's thesis, Department of Computer Science and Engineering at Lehigh University, 2002.

[Gra02] B. Graubart.White paper, turning unstructured data overload into a competitive advantage, Jul 2002.Email attachment.

[HDD] HDDI™.HDDI™ [online, cited July 2002].Available from World Wide Web: `hddi.cse.lehigh.edu`.

[HHWN02] S. Havre, E. Hetzler, P. Whitney, and L. Nowell.Themeriver: Visualizing thematic changes in large document collections.*IEEE Transactions on Visualization and Computer Graphics*, 8(1), Jan - Mar 2002.

[HyB] HyBrix.HyBrix [online, cited July 2002].Available from World Wide Web: `www.siemens.com/index.jsp`.

[IDC] IDC.IDC [online, cited July 2002].Available from World Wide Web: `www.idc.com`.

[INS] INSPEC®.INSPEC® [online, cited July 2002].Available from World Wide Web: `www.iee.org.uk/Publish/INSPEC`.

[Int] Interwoven.Interwoven [online, cited July 2002].Available from World Wide Web: `www.interwoven.com/products`.

[LA00a] A. Leuski and J. Allan.Lighthouse: Showing the way to relevant information.In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, pages 125–130, 2000.

[LA00b] A. Leuski and J. Allan.Strategy-based interactive cluster visualization for information retrieval.*International Journal on Digital Libraries*, 3(2):170–184, 2000.

[LAS97] B. Lent, R. Agrawal, and R. Srikant.Discovering trends in text databases.In *Proceedings of the Third International Conference On Knowledge Discovery and Data Mining*, California, 1997.

[Lex] LexisNexis.LexisNexis [online, cited July 2002].Available from World Wide Web: `www.lexisnexis.com`.

[Ley02] L. Leydesdorff.Indicators of structural change in the dynamics of science: Entropy statistics of the sci journal citation reports.*Scientometrics*, 53(1):131–159, 2002.

[Lin] Linguistic Data Consortium.Linguistic Data Consortium [online, cited July 2002].Available from World Wide Web: `www.ldc.upenn.edu`.

[Loc] Lockheed-Martin.Lockheed-Martin [online, cited July 2002].Available from World Wide Web: `www.lockheedmartin.com`.

[LSL+00] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan.Mining of concurrent text and time-series.In *ACM KDD-2000 Text Mining Workshop*, 2000.

[MDOP97] A. Martin, T.K.G. Doddington, M. Ordowski, and M. Przybocki.The det curve in assessment of detection task performance.In *Proceedings of EuroSpeech97*, volume 4, pages 1895–1898, 1997.

[Mor] Moreover.Moreover [online, cited July 2002].Available from World Wide Web: `www.moreover.com`.

[NFH+96] L.T. Nowell, R.K. France, D. Hix, L. S Heath, and E.A. Fox.Visualizing search results: Some alternatives to query-document similarity.In *Proceedings of SIGIR'96*, Zurich, 1996.

[Nor] Northern Light.Northern Light [online, cited July 2002].Available from World Wide Web: `www.northernlight.com`.

[PCP01] W.M. Pottenger, M.R. Callahan, and M.A. Padgett.Distributed information management.*Annual Review of Information Science and Technology (ARIST)*, 35, 2001.

[PD95] A.L. Porter and M.J. Detampel.Technology opportunities analysis.*Technological Forecasting and Social Change*, 49:237–255, 1995.

[PFL+00] A. Popescul, G.W. Flake, S. Lawrence, L. Ungar, and C.L. Giles.Clustering and identifying temporal trends in document databases.In *Proceedings of IEEE Advances in Digital Libraries*, 2000.

[PJ01] A.L. Porter and D. Jhu.Technological mapping for management of technology.In *International Symposium on Technology*, 2001.

[PKM01] W.M. Pottenger, Y. Kim, and D.D. Meling.HDDI™: Hierarchical Distributed Dynamic Indexing.*Data Mining for Scientific and Engineering Applications, Robert Grossman, Chandrika Kamath, Vipin Kumar and Raju Namburu, Eds.*, Jul 2001.

[PMS+98] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman.Lifelines: Using visualization to enhance navigation and analysis of patient records.In *Proceedings of the 1998 American Medical Informatic Association Annual Fall Symposium*, pages 76–80, 1998.

[PY01] W.M. Pottenger and T. Yang.Detecting emerging concepts in textual data mining.*Computational Information Retrieval, Michael Berry, Ed.*, Aug 2001.

[RGP02] S. Roy, D. Gevry, and W.M. Pottenger.Methodologies for trend detection in textual data mining.In *Proceedings of the Textmine '02 Workshop, Second SIAM International Conference on Data Mining*, Apr 2002.

[Roy02] S. Roy.A multimedia interface for emerging trend detection in inquiry-based learning.Master's thesis, Department of Computer Science and Engineering at Lehigh University, May 2002.

[SA96] R. Srikant and R. Agrawal.Mining sequential patterns: Generalizations and performance improvements.In *Proceedings of the*

Fifth International Conference on Extending Database Technology (EDBT), Avignon, France, 1996.

[SA00] R. Swan and J. Allan.Automatic generation of overview timelines.In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, 2000. Association for Computing Machinery.

[Sem] Semio.Semio [online, cited July 2002].Available from World Wide Web: `www.semio.com`.

[Ser] Ser Solutions.Ser Solutions [online, cited July 2002].Available from World Wide Web: `www.sersolutions.com`.

[SJ00] R. Swan and D. Jensen.Timemines: Constructing timelines with statistical models of word usage.In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.

[SPSa] SPSS Clementine.SPSS Clementine [online, cited July 2002].Available from World Wide Web: `www.spss.com/spssbi/clementine`.

[SPSb] SPSS LexiQuest.SPSS LexiQuest [online, cited July 2002].Available from World Wide Web: `www.spss.com/spssbi/lexiquest`.

[Str] Stratify.Stratify [online, cited July 2002].Available from World Wide Web: `www.stratify.com`.

[TDT] TDT.TDT [online, cited July 2002].Available from World Wide Web: `www.nist.gov/speech/tests/tdt/index.htm`.

[Tex] TextAnalyst.TextAnalyst [online, cited July 2002].Available from World Wide Web: `www.megaputer.com/products/ta/index.php3`.

[Tho] ThoughtShare.ThoughtShare [online, cited July 2002].Available from World Wide Web: `www.thoughtshare.com`.

[UIU] UIUC Digital Library Initiative.UIUC Digital Library Initiative [online, cited July 2002].Available from World Wide Web: `dli.grainger.uiuc.edu`.

[US] US Patent Site.US Patent Site [online, cited July 2002].Available from World Wide Web: `www.uspto.gov/main/patents.htm`.

[Ver] Verity.Verity [online, cited July 2002].Available from World Wide Web: `www.verity.com`.

[XBC94] J. Xu, J. Broglio, and W.B. Croft.The design and implementation of a part of speech tagger for english.Technical report, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, 1994.Technical Report IR-52.

[Yan00] T. Yang.Detecting emerging conceptual contexts in textual collections.Master's thesis, Department of Computer Science at the University of Illinois at Urbana-Champaign, 2000.

[YPC98] Y. Yang, T. Pierce, and J. Carbonell.A study on retrospective and on-line event detection.In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, 1998.

[Zho00] L. Zhou.Machine learning classification for detecting trends in textual collections.Master's thesis, Department of Computer Science at the University of Illinois at Urbana-Champaign, December 2000.