# Methodologies for Trend Detection in Textual Data Mining

Soma Roy, David Gevry, William M. Pottenger
Computer Science and Engineering, Lehigh University
{sor4, drg2, billp @lehigh.edu}

## *Abstract*

We present two methodologies for the detection of emerging trends in the area of textual data mining. These manual methods are intended to help us improve the performance of our existing fully automatic trend detection system [3].

The first methodology uses citations traces with pruning metrics to generate a document set for an emerging trend. Following this, threshold values are tested to determine the year that the trend emerges.

The second methodology uses web resources to identify incipient emerging trends. We demonstrate with a confidence level of 99% that our second approach results in a significant improvement in the precision of trend detection.

Lastly we propose the integration of these methods for both the improvement of our existing fully automatic approach as well as in the deployment of our semi-automated CIMEL [20] prototype that employs emerging trends detection to enhance multimedia-based Computer Science education.

## *1.0 Introduction*

Emerging trend detection is an exciting area of research in text mining. An emerging trend is a topic area for which one can trace the growth of interest and utility over time. An example of such a trend is XML, a technology that emerged in the mid 1990's.

The necessity for automated methods for detecting emerging trends has grown with the increasing availability of digital information. What makes it difficult is that it is not only based on data collected / explored but also on the experience or domain expertise of the person involved in the detection process. Currently too much data is available for a human expert to examine manually and not risk missing some vital piece of information. Trending of this nature is thus primarily based on human-expert analysis of sources (e.g., patent, trade, and technical literature) combined with bibliometric and text mining techniques that employ both semi [1] and fully automatic methods [3, 6, 11].

With the continued increases in the performance of computational technologies, more aggressive implementations of trend detection methodologies are becoming possible. This has spurred our research into the development of more sophisticated methodologies for the detection of emerging trends. Such emerging trends are defined as topic areas that have grown in size and variety at an increasing rate over time. We are specifically interested in incipient trends (trends that occur for the first time). This article discusses our research in the area of trend detection methodologies. We also present initial case studies of these methodologies in the field of data mining. These methodologies will form the groundwork for our ultimate goal of enhancing the performance of our existing fully automatic trend detection system [3].

The development of these methodologies will result in precise and efficient metrics and methods to identify and characterize trends as emerging or non-emerging. We are currently in the process of integrating semi-automatic trend detection (based on a combination of the two methodologies) in the CIMEL [20] prototype that employs trend detection to enhance Computer Science education. CIMEL is a multimedia framework for constructive and collaborative, inquiry-based learning.

## *2.0 Related Work and Motivation*

In our previous work, [3, 9, 10, 13] we examined the usage of various linguistic and statistical features to track trends across time. The HDDI™ system [4,14] 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 rate of change in the size of clusters and in the frequency and association of features is used as input to machine learning techniques to classify topics as emerging or non-emerging.

However, a domain expert does not use linguistic features exclusively to detect an emerging trend. The research for this paper is motivated by the desire to better characterize a domain expert approach to the detection of emerging trends. Through this research we aim to identify features and methods to enhance the automatic detection of emerging trends.

Several research projects are exploring solutions to the detection of emerging trends. ThemeRiver [5] enables users to visualize trends and detect emerging trends. It is a prototype (mock-up) that visualizes thematic variations over time across a collection of documents. As it flows through time, the river changes width to depict changes in the thematic strength of temporally collocated documents. The river is within the context of a timeline and a corresponding textual representation of external events.

Another project [6] presents a method of tracking sequential patterns across time. This method extracts content bearing words from the corpus it is using and generates sequential patterns within a selected time-interval based on a minimal support threshold between content bearing words. The authors also present a system for visualizing these patterns.

The Envision system [7] allows users to explore trends in digital library metadata (including publication dates) graphically 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 TDT project [8] is an 'Event Tracking' mechanism, which tracks topical information in a stream consisting of news stories using speech processing. The goal of [8] is essentially to detect changes in topics – disruptive events exhibiting discontinuities in semantics in localized data sources such as newscasts. Our research [3, 4, 9, 10] focuses on integrative or non-disruptive emergence of topics that build on previously existing topics. There is a significant difference in the goal of these research projects: unlike the TDT research, our goal is to detect novel trends that are globally incipient in a given domain.

TimeMines [11] is an automated system that generates overview timelines for topics in free text news corpora. These timelines are used to indicate the key topics involved in the corpora and their coverage with a ranking function of how important a particular topic is within that area. In contrast, our research goal is not to identify all topics that are important but rather identify selected emerging trends that are incipient.

TOAS [1] extracts information about particular emerging technologies through a process of search and retrieval from abstract databases (e.g., INSPEC, Medline, etc.) with manually generated queries. Following this initial data collection, various data processing techniques are used to generate reports on the topic of the search. TOAS incorporates the ideas of 'Monitoring' and 'Bibliometrics' in a complementary fashion for the detection of emerging trends. Monitoring involves tracking of data for a specific purpose, the implication of which will subsequently be interpreted by a domain expert. On the other hand, bibliometrics uses counts of citations in publications, patents or citations to measure and interpret scientific and technological advances [2].

In [19] a discussion of studies of patterns in citations concludes that active research fronts develop in citations between recent years. This is an important characteristic that can be leveraged to enhance our fully automatic approach to emerging trend detection.

## *3.0 Approach*

The methodologies we present are based on our own intuition supported by our domain knowledge in the text mining trend detection area [3, 4, 9, 10, 13, 14]. We expect that the results of our study of manual methodologies can be used to increase the precision of our existing automated approach to emerging trend detection. As part of our future work we plan to conduct more extensive research that involves focus group studies and surveys of domain experts to gain a better understanding of how domain experts perform trend detection.

We approach the problem of trend detection from two different perspectives. The first methodology uses citations to generate a document/trend set for a selected topic. This set is then analyzed to detect the initial emergence point of a trend. The second methodology uses web resources to identify candidate emerging trends. Domain knowledge is used to validate potential incipient trends as emerging. The results for this second methodology are discussed in the evaluation section.

## *3.1 Approach 1: Tracing a Trend via Citation Linkages*

In what follows we outline the seven steps in this first methodology.

**1. Determination of a potential trend and/or selection of a topic of interest.**

A topic of interest is selected and is characterized in one or two descriptive sentences[1]. This description is used for comparison to the documents retrieved in later steps. Various sources are used to retrieve recent documents on the topic (e.g., csindex.com). This allows initial validation of a topic's worth based on document and citation counts. Additionally, key authors on a topic can be identified based on counts of citations to their work.

The documents retrieved from this step are examined to verify that they discuss the topic of

[1] This is the manual approach to describing a topic. In an automated system such topic description would be facilitated by use of a set of keywords and linguistic features entered by the user or extracted from a chosen set of related documents.

interest and are then used to determine keywords related to the topic.

### 2. Initial Citation Traversal Backward in Time

The references of the retrieved papers from various sources (e.g., csindex.com) are examined and from these references a subset is selected based on the titles and the author names that appeared more frequently in the papers. For this case any repeated reference or repeated author is selected due to the limited amount of citation information overlap. Once the papers are selected, citation link information is used to retrieve the abstract of this set of papers from various online resources (e.g., Science Citation Index). The abstracts are then examined for relation to the topic, and those papers that do not discuss the topic are pruned. This comparison is accomplished by examining the title of the document and the abstract for relation to the description of the topic formed in step one. One method that may enhance this process is finding the subject sentence in the abstract. This sentence usually starts with particular catch phrases (e.g. This paper, We present, The author discusses, etc). This helps determine whether the citation should be used for examination of the trend. If an abstract is not available from online sources then the reference is used conditionally to trace citations forward in time.

### 3. Tracing Citations Forward in Time

Citations to the papers retrieved from the initial traversal are looked up (e.g., with Science Citation Index). This search returns a large number of documents, which requires initial pruning steps. First we assume that pruning based on venue will yield a viable set of documents that is representative of the topic. Thus venue pruning rejects all documents published in venues outside of a selected domain, which is the overall goal of this pruning. This pruning action helps in reducing the total number of documents retrieved and also to restrict the documents retrieved to a particular area of interest. We are currently investigating how venue spread affects emerging trend detection. This is a first step into the examination of how domains can be represented by venues.

The next pruning step examines the title and keywords of the documents for similarity to the topic description sentence formed in step 1. Finally the abstracts are examined as in step 2 to determine whether to include or exclude a document from the trend set (the set of documents related to the topic description sentence which represent the trend). These last two steps can be combined if there is a small enough collection of documents that cited a particular source. If the documents from step 2 that were used conditionally to trace citations forward in time do not yield useful documents they are pruned from the trend set.

### 4. Tracing Backwards in Time

The references for the papers obtained in steps 2 and 3 are examined and a subset of these references is formed. Each set of papers is handled separately, with the set from step 3 being examined first. First the author names are examined for the set. Author names that are referenced by three or more documents are selected. Next we obtain the repeated references for articles written by one of the selected authors. New papers that were not previously found are selected. If the abstracts or titles are obtainable this set is pruned based on similarity as in step 2. If there is not an abstract available for a paper it is conditionally added. This process is repeated for the set of papers from step 2.

### 5. Set Improvement

Online repositories with citation linkage information (e.g., Science Citation Index) are queried with terms from the topic description to determine if there are additional documents missed by the citation tracing. The results are pruned on similarity to the topic and duplicates are removed. If there are remaining papers these are added to the trend set and their references are examined to identify potential matches with the sets obtained so far. The citation information of the retrieved documents is combined with that of the trend set. The process ends with a final query to additional online sources (e.g., Inspec, Compendex and csindex.com) with terms from the topic description sentence. This final search retrieves documents that are not covered by the sources that contained the necessary citation information for the previous steps.

### 6. Identification of Emergence Time

Upon completion of the previous step duplicate documents are identified and removed from the trend set. The document frequency, number of repeated authors, and number of new venues is then graphed by year. We then select the years with an overall higher document frequency. It is our premise that these 'candidate years' have a higher likelihood of being points where the trend is emerging.

### 7. Thresholds for Emerging Trend Detection

Using candidate years as an emergence point for the trend we then apply a series of thresholds. These thresholds are based on our own intuition and domain expertise [3, 4, 9, 10, 13, 14], but in future work we plan to conduct more extensive research that involves focus group studies and surveys of domain experts to gain a better understanding of how domain experts perform trend detection. The heuristic thresholds that we have identified in our research are as follows:

1. A document frequency of five or greater is required for the candidate year. This is used to prune out candidate years where a trend has not developed to the point of emergence.
2. The candidate year is required to be the largest document year in all years prior to the candidate. The candidate year should represent the largest amount of work to date on a trend. Therefore we prune out candidate years that do not exceed the years prior to them in document frequency.
3. The candidate year is required to contain 20% of all documents in the trend set, prior to and including the candidate. The candidate year should have a high level of representation for the work to date on a topic.
4. The candidate year is required to contain 10% of all documents in the trend set for all the years studied. If the candidate year being examined is not the current year then this threshold is used to assert the overall importance of the candidate year.
5. 25% of all documents in prior years must occur in the three years prior to the candidate year. The trend should have an increase over a short period of time to be considered emerging. Additionally the majority of documents in a trend that is emerging should occur close to the emergence point.
6. Venue variety increases in the candidate year. This increase in venue variety indicates an increase in the activity of a trend.
7. There should be at least one repeated author present in the trend. The trend needs to have the beginnings of a community of authors.
8. There should be at least 10 venues present in the trend. A core set of venues is required for the trend to be considered emerging.

### *3.1.1 Analyses of Case Studies Selection of Decision Trees*

The main topic of our case studies was decision trees. From this topic two sub-topics were selected; Inductive Decision Trees (IDT) and Fuzzy Decision Trees (FDT). The topic of decision trees was selected due to the attention it has received in data mining and machine learning literature and research (e.g., [12, 15, 16, 17, 18]).

**Inductive Decision Tree Case Study**

Our first case study considered the trend of IDT in the domain of Data Mining. Since the trend of IDT has emerged already this gave us a good starting point to examine the patterns surrounding its emergence. The methodology was followed to yield a trend set of documents. Then thresholds were examined to determine when the initial point of emergence occurred. Figure 1 shows the document frequency for the IDT trend set across time.

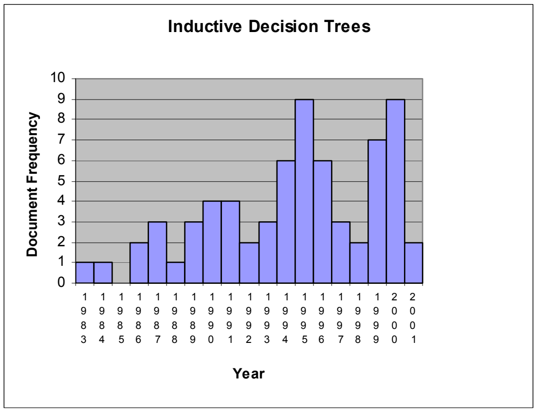

Figure 1: IDT Document Frequency

From this graph the years 1991, 1995, and 2000 were selected as potential candidate years of emergence due to their document frequency in respect to previous years. However, the year 1991 was later excluded due to not meeting the document frequency threshold of at least 5 documents. Next, the year 2000 was also removed from the candidate year set because it did not have the required 20% of previous documents. These thresholds are used to maintain a level of representation in the candidate year. The rational behind this is to prevent a candidate year from being identified as emerging when the bulk of the documents occur in prior years. Similarly the threshold for the past three years (threshold 5 of part 6 in the Methodology Section) is used to maintain a larger percentage of the documents close to the candidate year.

The next threshold we applied required that all candidate years contain at least one repeated author. In order for a trend to be considered emerging it had to have the beginnings of an author base or community. Figure 2 shows the growth of the number of repeated authors over time.

Finally the restriction of an increase in the number of new venues is applied to determine how active the trend is in the domain. There should be a core venues set that the topic occurs in to guarantee that the topic has a foothold in a domain before it is considered emerging. Additionally, there should be a relative growth within the domain for the topic. The number of new venues a topic acquires in the trend set represents this growth. Figure 3 shows the inclusion of new venues into the trend set of induction decision trees in relation to time.

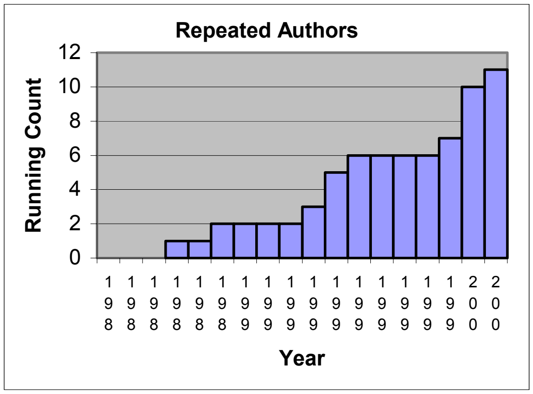

Figure 2: IDT Repeated Authors

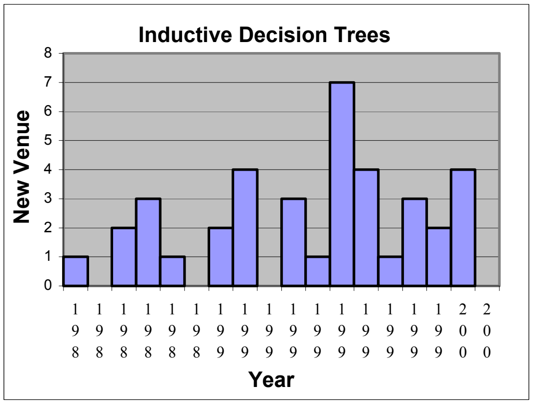

Figure 3: IDT New Venues

The candidate year 1995 was selected as emerging for the inductive decision tree trend by these threshold metrics.

## Fuzzy Decision Tree Case Study

Following the same methodology the document set for the topic of FDT was generated. Figure 4 shows the document frequency for the trend.

The candidate years for this topic were 1992, 1996, 1998, and 2000. The years 1998 and 2000 were pruned by threshold (2) because they were not the largest years for document frequency.

All candidate years contained at least one repeated author showing that the trend was beginning to receive attention from a group of authors. Figure 5 shows the increase of repeated authors with relation to time.

The year 1992 was removed from the set of candidate years because it did not reach the threshold for venues included in the trend. There were only 7 different venues present by 1992. The year 1996 did however contain the required number of venues and had an increase in the number of new venues over the year prior to it. Therefore the year 1996 was selected as the year of emergence for FDT.

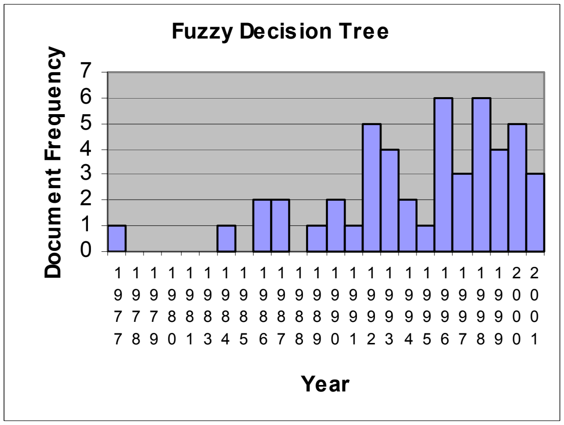

Figure 4: FDT Document Frequency

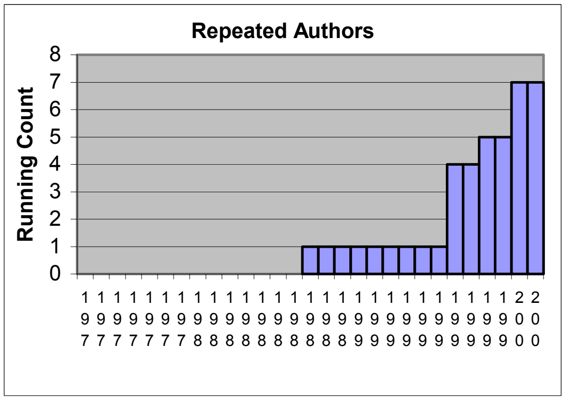

Figure 5: FDT Repeated Authors

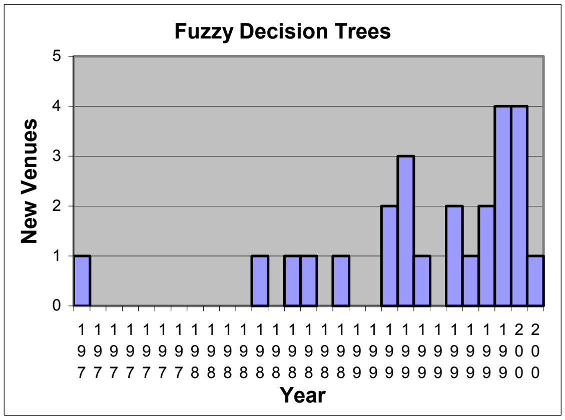

Figure 6: FDT New Venues

### *3.2 Approach 2: Identification of emerging trends using web resources*

Following is the second blueprint for a methodological approach to the manual detection of emerging trends. Like the first methodology, we plan to employ what we learn from this methodology to both improve the performance of our existing fully automatic trend detection algorithms as well as in our semi-automatic CIMEL [20] prototype that employs emerging trends detection to enhance multimedia-based Computer Science education. We present a case study of this approach and its evaluation in sections 3.2.1 and 4.0.

**1. Selection and validation of main topic area.**

Detection of emerging trends starts with the selection of a main topic area. Knowledge in this area is required as the use of domain knowledge at various stages of identification of emerging trends is necessary. The objective is to discover emerging trends in the area of interest. An INSPEC Database search on the main topic is performed to confirm it as a possible area of research.

**2. Search for candidate emerging trends.**

Recent conferences and workshops are searched for discussion on the main topic area giving special attention to workshop websites and technical papers for possible emerging trends (i.e. topics within the domain of the main topic area). Links that deal with the recently discovered potential / candidate emerging trends can also be traced (using step 3).

**3. Candidate emerging trend verification.**

A web search engine (e.g. Google, Yahoo, etc.) is used to find additional trends and find further evidence of references to the candidate trends.

Following is a list of words associated with emerging trends (also called "helper" terms):

| | |
|---|---|
| most recent contribution | Future |
| recent research | recent trend |
| a new paradigm | next generation |
| hot topics | novel |
| Emergent | new approach |
| newest entry | proposed |
| cutting edge strategies | current issues |
| first public review | |

Two possible scenarios are:

a) If step 2 identified any candidate emerging trends, a search is made using any of the popular web search engines like Yahoo or Google using a candidate trend <and> any of the helper terms from the above list (where <and> indicates a logical AND of search terms).

Otherwise,

b) If step 2 did not identify any candidate emerging trends, a search is made using any of the popular web search engines like Yahoo or Google using main topic area <and> any of the helper terms from above list.

The algorithm in Figure 7, which uses web-based resources to validate candidate emerging trends, is followed at this stage. In the algorithm, "main topic" should be read as "candidate emerging trend" if case 3(a) applies. The algorithm is meant to assist the user in accepting or rejecting web links identified by a search engine while simultaneously extracting candidate emerging trends from pages (or links) of interest. These candidate emerging trends are maintained in a list and each trend is later verified in step 4 (Verification of Algorithmic Results). In completing this step several candidate emerging trends are identified. Also, further references to candidate emerging trends identified in step 2 can be found.

**4. Verification of Algorithmic Results**

An INSPEC Database search is performed using main topic area <and> newly found candidate emerging trend from the year of origin of the main topic area to the current year. If the frequency of documents referencing the search terms increases over the years, the candidate emerging trend is confirmed as a bona fide trend with respect to the main topic.[2]

If few documents appear in different years (say one or two), the authors of the articles are also investigated. If the same author is writing (may be as a follow up report to recent research, etc.), it's NOT an emerging trend.

**5. Additional Trends**

Steps 3 and 4 are repeated with combinations of other helper terms and/or other candidate emerging trends until all the desired emerging trends are found.

### *3.2.1 Case Study on the topic of Object Databases*

In this section a common example of how the above methodology would be used is presented. The

[2] Note: The objective here is to find the year of origin of a candidate emerging trend within the main topic.

following case study is a manual trace through the methodology we have developed. For this example the main topic area is chosen to be "Object Databases".

**1. Selection and validation of main topic area.**

First following step 1 of the methodology, a main topic area is chosen which in this case is Object Databases. Following this selection an INSPEC database search is performed to verify the selected topic area for its potential to contain emerging trends.

| **1988** | **1989** | **1990** | **1991** | **1992** | **1993** | **1994** |
|---|---|---|---|---|---|---|
| 10 | 13 | 28 | 38 | 24 | 41 | 73 |
| **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** |
| 36 | 46 | 47 | 54 | 39 | 36 | 16 |

**INSPEC Search: Object Databases.**

The above table shows the document counts by year from an INSPEC query on Object Databases. The coverage of this topic over recent years suggests that new innovative enhancements to Object Databases are being researched. Therefore Object Databases is a valid main topic area for the identification of emerging trends.

**2. Search for candidate emerging trends.**

The next step is to explore workshops and conferences that are related to the discussion of this main topic area. This was initiated by examining OOPSLA (oopsla.acm.org). Through the examination of OOPSLA's 2001 conference as well as additional sources, "XML Databases" was found to be a candidate emerging trend.

Following are a few excerpts from key conference and workshop papers, which discuss the area of Object Databases. Phrases in the relevant portions of the excerpts are highlighted (bold) to provide a visual cue to the detection of XML Databases as a candidate emerging trend.

i) (oopsla.acm.org/oopsla2001/fp/workshops/17.html)
"During the past few years, there has been a **considerable interest and growth in a number of new and emerging technologies, such as XML**. For many organizations already using object-orientation with database management systems, XML data adds a new dimension that brings considerable flexibility and promise, … The **recent trend towards XML servers, native XML databases** and support for XML in existing relational databases is a testimony to the importance of this issue for the vendor community as well."

ii) EDBT 2002 Workshop (XMLDM)
"… As database systems increasingly start talking to each other over the Web, there is **a fast growing interest in using the eXtensible Markup Language (XML) as the standard exchange format**. As a result, many relational database systems can export data as XML documents and import data from XML documents. XML is on its way to becoming the communication standard of the Web. Moreover, there is an **increasing trend to store XML-data in database systems** and, by this, make it easier to access and maintain."

iii) 1st ECOOP Workshop (XOT)
"XML has many similarities with object-oriented data models and languages. However, whereas the object-oriented technology has reached a great level of maturity, **XML is still in its infancy**."

Additional Sources that were examined include:
iv) Web Databases 2001
v) WebDB Workshops
vi) ACM SIGIR 2000 Workshop on XML and Information Retrieval

The way in which XML Databases is referred to in the above excerpts identifies it as a candidate emerging trend in the area of Object Databases.

**3. Candidate emerging trend verification.**

A web search engine (e.g. Google, Yahoo, etc.) was used to find additional trends and to find further evidence of the candidate trends that were found in the previous step (in this case only XML Databases). A query was formed which combined "Object Databases" <and> various helper terms (where <and> indicates a logical AND of search terms), which were listed in the methodology. The methodology did not require step 3(b) to be followed due to the detection of a candidate emerging trend in step 2 of the methodology, however it can be followed to gain additional candidate trends and to assist in the validation of candidates found in the previous steps.

The algorithm in Figure 7 was followed and the candidate emerging trend (XML Databases) was discovered as an emerging trend that was widely referred to in recent research work.

**4. Verification of Algorithmic Results**

An INSPEC Database search using the "Object-oriented" <and> "XML" <and> "Database" was performed to verify the algorithmic results.

| **1988** | **1989** | **1990** | **1991** | **1992** | **1993** | **1994** |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1995** | **1996** | **1997** | **1998** | **1999** | **2000** | **2001** |
| 0 | 0 | 0 | 0 | 5 | 11 | 5 |

**INSPEC Verification Search**

The frequency of documents referencing the search term increases over the years; hence XML Databases is an emerging trend with reference to Object Databases in Object Oriented Software Engineering.

## *4.0 Evaluation*

An experiment was conducted to test this methodology of emerging trends detection. 21 students participated in this experimental evaluation. The subjects were all students of a graduate course in Object Oriented Software Engineering. They were asked to identify three emerging trends in the area of Design Patterns. The main topic area was given to them to simplify the experimental evaluation and also to make it more relevant to their coursework. A survey was conducted before starting the experiment to find out how many students knew the definition of the term ‘literature search’ and how many of them had used web-based resources before to search for information. 19 students responded to the survey. While all of them said that they do use the web to search for information, 57.8% said that they didn’t know what a literature search was.

## *4.1 Methodology for evaluation*

The class was divided into two groups, groups A and B, each group having an approximately equal number of students.

Students from both groups A and B were expected to have attended the lectures of the class. They were also expected to have introductory knowledge in the main topic area of Design Patterns before participating in this experiment. This was necessary as at different steps of the experiment they needed to apply their domain knowledge to justify their choices of emerging versus non-emerging trends. Also, all the students had access to their textbooks, reference books and handouts given in the class. Both groups A and B attempted an exercise that involved identification of three emerging trends in the area of Design Patterns of Object Oriented Software Engineering. In addition, group B was provided with the methodology. Group B was also provided with a practical case study that demonstrated the process of detecting emerging trends as outlined in the methodology. After completing the task, students in group A were given the methodology and case study and required to resubmit their solutions using the methodology.

The standard metrics of evaluation in text mining are precision and recall and the exercise was evaluated using precision. In the following, the variable RET (retrieved) is the set of all trends the student has identified (note that this was a maximum of three trends for this experiment – i.e., assuming that the student has completed their assignment, RET == 3). RETREL (retrieved relevant) is the set of the retrieved trends that are truly emerging (true positive). Then, precision is defined as:

**precision = RETREL / RET**

Possible values of precision are 0% (RETREL = 0), 33.33% (RETREL = 1), 66.67% (RETREL = 2) and 100% (RETREL = 3).

Note that measuring recall would be quite difficult considering the broad range of research conducted in the area of Design Patterns. We did not have the resources to obtain a complete list of emerging trends at the time of this experimental evaluation, nor was it our purpose to have students retrieve all trends so recall was not measured.

**Hypothesis:**
Precision will be significantly higher for Group B.

## *4.2 Results*

| | Precision | Resubmits |
|---|---|---|
| **Group B (methodology)** | 66.67 | |
| | 0 | |
| | 100 | |
| | 0 | |
| | 100 | |
| | 100 | |
| | 100 | |
| | 33.33 | |
| | 33.33 | |
| | 66.67 | |
| **Group A** | 66.67 | |
| | 0 | 100 |
| | 33.33 | |
| | 100 | |
| | 66.67 | 100 |
| | 100 | |
| | 33.33 | |
| | 33.33 | 100 |
| | 0 | |

**Table 1:** Precision results of samples from Group A, Group B and resubmissions.

| | Group A (R1) | Group B (R2) | Group B (R3) | Group B (R4) | Group B (R5) |
|---|---|---|---|---|---|
| Mean | 54.54 | 60 | 69.23 | 100 | 100 |
| Standard Error | 11.26 | 12.96 | 10.98 | 0 | 0 |
| Median | 66.67 | 66.67 | 100 | 100 | 100 |
| Standard Deviation | 37.34 | 40.98 | 39.58 | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| Sample Variance | 1394.01 | 1679.06 | 1566.99 | 0 | 0 |
| Range | 100 | 100 | 100 | 0 | 0 |
| Minimum | 0 | 0 | 0 | 100 | 100 |
| Maximum | 100 | 100 | 100 | 100 | 100 |
| Sum | 600 | 600 | 900 | 400 | 600 |
| Count | 11 | 10 | 13 | 4 | 6 |
| Confidence Level(95.0%) | 25.08 | 29.31 | 23.92 | 0 | 0 |

**Table 2:** Analysis of Group A and Group B precision results (R2) entire group B, (R3) entire group B including resubmissions, (R4) actually followed methodology, and (R5) actually followed methodology including resubmissions

**Hypotheses:**

1. Group B (R5: actually followed methodology including resubmissions) will perform significantly better than group A (R1) in terms of precision of their results.

2. Group B (R4: actually followed methodology excluding resubmissions) will perform significantly better than group A (R1) in terms of precision of their results.

### 4.2.1 Result 1: (R1 vs. R5)

*Lower Tail test*

Population 1 sample corresponds to Group A (R1: without methodology); Population 2 Sample corresponds to Group B (R5: actually followed methodology including resubmissions from Group A).

*Hypothesis 1*: (Mean precision of sample 1) – (Mean precision of sample 2) ≥ 0

| | |
|---|---|
| **Hypothesized Difference** | **0** |
| **Level of Significance** | **0.01** |
| **Population 1 Sample** | |
| **Sample Mean** | **54.54** |
| **Sample Size** | **11** |
| **Sample Standard Deviation** | **37.34** |
| **Population 2 Sample** | |
| **Sample Mean** | **100** |
| **Sample Size** | **6** |
| **Sample Standard Deviation** | **0** |
| Population 1 Sample Degrees of Freedom | 10 |
| Population 2 Sample Degrees of Freedom | 5 |
| Total Degrees of Freedom | 15 |
| Pooled Variance | 929.52 |
| Difference in Sample Means | -45.46 |
| *t*-Test Statistic | -2.94 |
| | |
| **Lower-Tail Test** | |
| **Lower Critical Value** | **-2.60** |
| ***p*-Value** | **0.00509** |
| **Reject the null hypothesis** | |

Table 3

*Conclusion*: Mean difference between sample 1 (without methodology) and sample 2 (actually followed methodology including resubmissions from Group A) is less than 0 with a confidence level of 99%. Thus sample 2 precision results are significantly greater than sample 1.

### 4.2.2 Result 2: (R1 vs. R4)

*Lower Tail test*

Population 1 sample corresponds to Group A (R1: without methodology); Population 2 Sample corresponds to Group B (R4: actually followed methodology excluding resubmissions from Group A).

*Hypothesis 2:* (Mean precision of sample 1) – (Mean precision of sample 2) ≥ 0

| | |
|---|---|
| **Hypothesized Difference** | **0** |
| **Level of Significance** | **0.05** |
| **Population 1 Sample** | |
| **Sample Mean** | **54.54** |
| **Sample Size** | **11** |
| **Sample Standard Deviation** | **37.34** |
| **Population 2 Sample** | |
| **Sample Mean** | **100** |
| **Sample Size** | **4** |
| **Sample Standard Deviation** | **0** |
| Population 1 Sample Degrees of Freedom | 10 |
| Population 2 Sample Degrees of Freedom | 3 |
| Total Degrees of Freedom | 13 |
| Pooled Variance | 1072.52 |
| Difference in Sample Means | -45.46 |
| *t*-Test Statistic | -2.37743 |
| | |
| **Lower-Tail Test** | |
| **Lower Critical Value** | **-1.77093** |
| ***p*-Value** | **0.016734** |
| **Reject the null hypothesis** | |

Table 4

*Conclusion:* Mean difference between sample 1 (without methodology) and sample 2 (who actually followed methodology excluding resubmissions from Group A) is less than 0 with a confidence level of 95%. Thus sample 2 precision results are significantly greater than sample 1.

### 4.3 Analysis of Results

The sample data collected is based on precision of the results of student performance on the assignment. Precision of the results is calculated based on the number of correct trends identified by each student versus number of total trends identified.

We found with a confidence level of 99% that the mean precision of sample 2 (actually followed methodology including resubmissions from Group A) is significantly greater than the mean precision of sample 1 (without methodology). Also, with a confidence level of 95%, mean precision of sample 2 (actually followed methodology excluding resubmissions from Group A) is significantly greater than sample 1 (without methodology). Initial results are indeed promising.

In our experiment initially we did not find a significant difference in precision results between Group A (R1: without methodology) and Group B (R2: entire group B with methodology). However, after a critical study of the results and the experimental methodology, we found that there are some variables that we were unable to account for in this first set of experiments. For example, some students, even with the methodology, did not actually finish the assignment and reported only one or two out of three required emerging trends.

To address this concern we held a focus group discussion with the students and determined that at least some of them decided to stop after putting in seven hours of time on the assignment. This gave us some confidence to use these partial results because we had explicitly directed the students to stop after seven hours.

A second variable that we were unable to control was whether the students in group B actually followed the methodology. Based on the focus group discussion with the students we learned that despite the fact that they were required to follow the methodology, several of them had difficulty understanding it and *did not follow the methodology at all*. As a result, we performed a critical study of group B results, and were able to determine which students had actually followed the methodology.[3] In future experiments we will modify our methods to take into account the usability of the methodology.

## *5.0 Conclusion*

We have developed two methodologies for the manual detection of emerging trends. These two methodologies are based on our own intuition and domain knowledge in the manual identification and characterization of emerging trends. The first methodology exploits citation linkages and pruning methods to generate a document set for a trend of interest. This trend set can then be expanded with the use of web-based repositories (INSPEC, etc.). Trend emergence is validated through a series of thresholds. This methodology is still under development.

The second methodology uses web-based resources and an algorithmic approach to identify incipient emerging trends. We demonstrated at a confidence level of 99% that the use of this methodology improves the detection of incipient emerging trends.

The development of these two manual methodologies for emerging trend detection has given us insight into different ways to characterize trends. We believe that we have taken important new steps towards understanding the emergence of trends. Our next step is to combine these two methodologies and perform a controlled usability study of the approach. We expect these results to aid us in our efforts to improve the performance of our existing fully automatic trend detection algorithms [3, 4, 9, 10, 13, 14].

## *6.0 Future Work*

The long-term goal of our research is to extend our previous work [3, 4, 9, 10, 13, 14] to develop a more sophisticated approach to the automatic detection of emerging trends using the methodologies presented in this paper. We are also nearing completion of a multimedia-based interactive system for semi-automatic emerging trend detection as part of the CIMEL [20] prototype that employs emerging trends detection to enhance Computer Science education. Finally, we plan to add a visualization module to our existing fully automatic trend detection system as an extension to our previous work [3, 13].

## *7.0 Acknowledgements*

This research was supported in part by the National Science Foundation, Grant CRCD-0087977 in the Engineering Education Program. Many thanks to Program Directors Anita LaSalle, Mary Poats and Harry Hedges for their assistance.

We also wish to acknowledge the faculty, students and staff in the Lehigh University CSE Department for their assistance in completing this work as well as independent evaluation experts Dr. Deb Dirkson and Dr. Morgan Jennings. We especially thank Professors Glenn Blank and Drew Kessler, and the entire CIMEL Team! Co-author William M. Pottenger would like to express his gratitude to his Lord and Savior, Jesus Christ, for His continual guidance and support throughout his life.

[3] The assignment required students to log the steps they took to detect the emerging trends.

## *References*

[1] Alan L. Porter and Michael J. Detampel. Technology Opportunities Analysis. *Technological Forecasting and Social Change*, Vol 49, 237-255, 1995.

[2] H. D. White and K. W McCain. *Bibliometrics*. Annual Review of Information Science and Technology, Elseiver, Amsterdam, Vol 24, 119-186, 1989.

[3] William M. Pottenger and Ting-hao Yang. *Detecting Emerging Concepts in Textual Data Mining*. Computational Information Retrieval, Michael Berry, Ed., SIAM, Philadelphia, PA, August 2001.

[4] Fabien Bouskila, William 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, June 2000.

[5] Susan Havre, Beth Hetzler and Lucy Nowell. ThemeRiver$^{TM}$: In Search of Trends, Patterns, and Relationships. Battelle Pacific Northwest Division. Presented at *IEEE Symposium on Information Visualization,* InfoVis '99, San Francisco CA, October 25-26 1999.

[6] Pak Chung Wong, Wendy Cowley, Harlan Foote, Elizabeth Jurus, Jim Thomas. Visualizing Sequential Patterns for Text Mining. Pacific Northwest National Laboratory. *In Proceedings of IEEE Information Visualization* 2000, October 2000.

[7] 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

[8] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B.T Archibald, X. Liu. Learning Approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval*, Vol 14(4), 32-43, 1999.

[9] T. Yang, *Detecting Emerging Contextual Concepts in Textual Collections.* M.S. Thesis, Department of Computer Science at the University of Illinois at Urbana-Champaign, 2000.

[10] L. Zhou, *Machine Learning Classification For Detecting Trends In Textual Collections.* M.S. Thesis, Department of Computer Science at the University of Illinois at Urbana-Champaign, 2001.

[11] Russel Swan and David Jensen. TimeMines: Constructing Timelines with Statistical Models of Word Usage. In *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.

[12] S.K. Murthy. *On growing better decision trees from data*. Doctoral dissertation, University of Maryland, 1997.

[13] Yumi Jin. *Graphical User Interface and Information Visualization Techniques for Detection of Emerging Concepts*. M.S. Thesis, Department of Computer Science at the University of Illinois at Urbana-Champaign, December 2000

[14] W. Pottenger, Y. Kim, and D. Meling. *Data Mining for Scientific and Engineering Applications,* chapter HDDI™. Hierarchical Distributed Dynamic Indexing. R. Grossman and C. Kamath and V. Kumar and R. Namburu, Eds., 2001.

[15] Quinlan, J. R. Induction of decision trees. *Machine Learning*, 1:81--106, 1986.

[16] Quinlan, J. R. *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann 1993.

[17] Cezary Z. Janikow. Exemplar Learning in Fuzzy Decision Trees. In *Proceedings of FUZZ-IEEE 1996*, pp. 1500-1505.

[18] Yuan, Y.F. and Shaw, M.J. Induction of Fuzzy Decision Trees. *Fuzzy Sets and Systems*. Vol 69, Iss 2, pp 125-139, 1995.

[19] Price, D.D. Networks of Scientific Papers. *Science.* Vol. 149, pp 510-515, 1965.

[20] Glenn D. Blank, William M. Pottenger, G. Drew Kessler, Soma Roy, David R. Gevry, Jeff Heigel, Shreeram Sahasrabudhe and Qiang Wang. Design and Evaluation of Multimedia to Teach Java and Object-Oriented Software Engineering. In *2002 ASEE Annual Conference and Exposition,* Session 1526. June 2002.

**Figure 7**

```
While (there are links to follow from the search engine retrieved pages <or> the desired number of trends has not been found)
{       Make an empty list L2. // use this to store all candidate emerging trends.
      Click on link = 1 // first link of interest in the search results
       If (year page last modified == current year or (current year -1) or (current year –2))
      {   DO WHILE  a page of interest is found
               {
                   Make a list L1:
                      number of occurrences of  the term "main topic area" in the page : "main topic area" : m
                      number of occurrences of the helper words used to do the search in the   page : <helper term> : n
                   If (m, n > 1)
                   {
                        The page is of interest.

                        Add "main topic area" to L2 if it is a candidate emerging trend.
                        List the frequency of occurrences of all the phrases in the page.
                        Look for the phrases (except general phrases) with the highest frequency of occurrence.  Add them
                        to L2 if they qualify as candidate emerging trends (use domain knowledge).
                        Give special attention to the line (or paragraph) containing the word set : main topic area <and>
                        helper  term. Add to the list L2, phrases appearing in that paragraph (or sentence) that are judged to
                        be candidate emerging trends (use domain knowledge).
                    }
                   Else
                   {
                         If (m > 1 && (find any other helper term in the page))
                         {
                                  The page is of interest.
                                  Add "main topic area" to L2 if it is a candidate emerging trend.
                                  List the frequency of occurrences of all the phrases in the page.
                                  Look for the phrases (except general phrases) with the highest frequency of occurrence.
                                  Add them to L2 if they qualify as candidate emerging trends (use domain knowledge).
                                  Give special attention to the line (or paragraph) containing the word set : main topic area
                                  <and> helper  term. Add to the list L2, phrases appearing in that paragraph (or sentence)
                                  that are judged to be candidate emerging trends (use domain knowledge).
                         }
                         Else
                         {
                                  Add "main topic area" to L2 if it is a candidate emerging trend.
                                  List the frequency of occurrences of all the phrases in the page.
                                  Look for the phrases (except general phrases) with the highest frequency of occurrence.
                                  Add them to L2 if they qualify as candidate emerging trends (use domain knowledge).
                                  Give special attention to the line (or paragraph) containing the word set : main topic area
                                  <and> helper  term. Add to the list L2, phrases appearing in that paragraph (or sentence)
                                  that are judged to be as candidate emerging trends (use domain knowledge).

                                  If (found candidate emerging trend)
                                            The page is of interest
                                  Else
                                            Reject page.
                         }
                   }
               }  // Close Do While loop
      } // Close check year if statement
      Else
      {
                Perform an INSPEC database search to confirm that term is not emerging. Reject page.
                // An INSPEC database search should show an increasing number of documents referencing the term over
                the years if the candidate is truly an emerging trend. If not, it is not emerging.
       }
     Click on link++ or exit and proceed to Step 4
         // click on next link of interest or exit and do step 4
}
```