Detecting Emerging Concepts in Textual Data Mining*

***William M. Pottenger*[†‡] *and Ting-hao Yang*[§]**

1 Introduction

Recent advances in computer technology are fueling radical changes in the nature of information management. Increasing computational capacities coupled with the ubiquity of networking have resulted in widespread digitization of information, thereby creating fundamentally new possibilities for managing information. One such opportunity lies in the budding area of textual data mining. With roots in the fields of statistics, machine learning and information theory, data mining is emerging as a field of study in its own right. The marriage of data mining techniques to applications in textual information management has created unprecedented opportunity for the development of automatic approaches to tasks heretofore considered intractable.

This article summarizes our research to date in the automatic identification of emerging trends in textual data. Applications are numerous: the detection of trends in warranty repair claims, for example, is of genuine interest to NCSA industrial partners Caterpillar and Boeing. Technology forecasting is another example with numerous applications of both academic and practical interest. In general, trending analysis of textual data can be performed in any domain that involves written records of human endeavors whether scientific or artistic in nature.

Trending of this nature is primarily based on human-expert analysis of sources (e.g., patent, trade, and technical literature) combined with bibliometric techniques

*This work was supported by Lehigh University and the Eastman Kodak Company in partnership with the National Center for Supercomputing Applications at the University of Illinois.

†Co-author William M. Pottenger expresses his sincere gratitude to his Lord and Savior Jesus Christ for His help in this work.

‡Department of Electrical Engineering and Computer Science, Lehigh University, Bethlehem, PA 18015, E-mail: billp@eecs.lehigh.edu.

§National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

that employ both semi and fully automatic methods [1]. Automatic approaches have not focused on the actual content of the literature primarily due to the complexity of dealing with large numbers of words and word relationships. With advances in computer communications, computational capabilities, and storage infrastructure, however, the stage is set to explore complex interrelationships in content as well as links (e.g., citations) in the detection of time-sensitive patterns in distributed textual repositories.

Semantics are, however, difficult to identify unambiguously. Computer algorithms deal with a digital representation of language and we do not have a precise interpretation of the semantics. The challenge thus lies in mapping from this digital domain to the semantic domain in a temporally sensitive environment. In fact, our approach to solving this problem imbues semantics to a statistical abstraction of relationships that change with time in literature databases.

2 An Overview of Our Approach to Detecting Trends in Textual Information

Our research objective is to design, implement, and validate a prototype for the detection of emerging content through the automatic analysis of large repositories of textual data. Trends in warranty repair claims, technology forecasting, and automatic detection of emerging interpretations in case law are just a few examples of the variety of applications in which the techniques can be applied. Technology forecasting, as a specific example, employs collections of trade, technical, and patent literature. Such collections are partitioned into topical domains of knowledge that we refer to as regions of semantic locality [2]. These topical domains of knowledge are traced over time to detect emerging trends in conceptual content.

The process of detecting emerging conceptual content that we envision is analogous to the operation of a radar system. A radar system assists in the differentiation of mobile vs. stationary objects, effectively screening out uninteresting reflections from stationary objects and preserving interesting reflections from moving objects. In the same way, our proposed techniques will identify regions of semantic locality in a set of collections and screen out topic areas that are stationary in a semantic sense with respect to time. As with a radar screen, the user of our proposed prototype must then query the identified hot topic regions of semantic locality and determine their characteristics by studying the underlying literature automatically associated with each such hot topic region.

The following steps are involved in the process: concept identification/extraction; concept co-occurrence matrix formation; knowledge base creation; identification of regions of semantic locality; the detection of emerging conceptual content; and a visualization depicting the flow of topics through time. Each of these steps is outlined in more detail below. Several aspects of this approach reflect our initial intuition on how the problem should be addressed. Each of these six steps will be addressed in the course of our research in order to refine our approach.

3 Technical Details

This section of the article deals with our approach in detail.

3.1 Concept identification/extraction

Our approach to concept identification/extraction includes the following three steps: input item (document) parsing, parts of speech tagging and concept identification. The parsing stage takes SGML, HTML or generalized XML tagged items as input. We have utilities to convert from a variety of input formats to XML, including proprietary airline safety data and US government patent data. Based on AI techniques outlined in [3] and [4], our parts of speech tagging approach includes the use of both lexical and contextual rules for identifying various parts of speech. After identifying each word's part of speech, we invoke a finite-state machine (pictured in Figure 1) that accepts the language generated by a regular grammar of a subset of the English language [5], [6]. Our enhanced state-machine identifies and extracts concepts consisting of complex noun phrases composed of multiple modifiers, including gerund verb forms. The final result of these three steps is a reformulation of the original collection that includes a summation of the location and number of occurrences of each extracted concept. The next stage of the process receives this reformulated collection.

3.2 Concept co-occurrence matrix formation

Co-occurring defines concepts that occur within the same item. An item can be defined as an intelligently created logical unit of text that is cohesive semantically. Examples include abstracts, titles, web pages, airline safety incident reports, patents, etc. The co-occurrence relation is reflexive and symmetric but not transitive. Given concepts extracted by the above process, we compute concept frequency and co-occurrence matrices. We also compute the frequencies of co-occurrences of concept pairs among all items in the set.

The literature discusses various definitions of co-occurrence [7]. Our approach incorporates metric measures based on proximity as well as measures that dynamically define the extent of sub-items within a given item. Our preliminary results indicate that this latter approach is crucial in dealing with the full text of items. We also reported on research in parallelizing the computation of such semantic relations based on the theory of coalescing loop operators [9].

3.3 Knowledge base creation

Knowledge base creation is a meta-level organizational process. For each concept in each matrix (in each of the time-sensitive collections) we rank co-occurring concepts. This one-to-many mapping associates each concept with a list of related concepts ranked by similarity. Co-occurring concepts are ranked in decreasing order of similarity. More general concepts occur lower in the list. Each concept pair (concept to ranked concept) is weighted, creating asymmetric measures of pairwise similarity

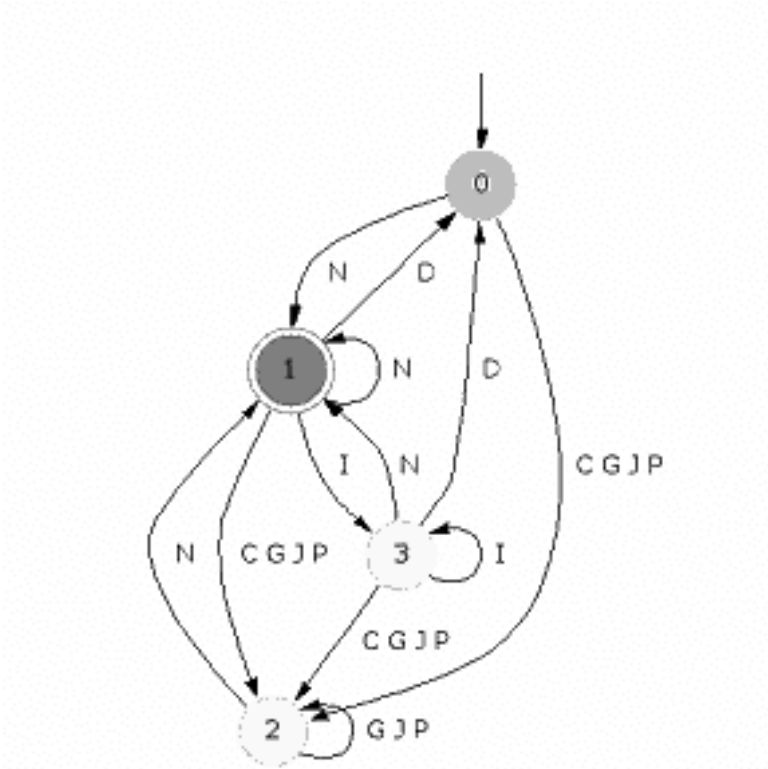

Figure 1. *A Finite State Automaton for Recognizing Complex Noun Phrases*

between concepts. The similarity is a mapping from one concept to another that quantitatively determines how similar they are semantically. We term the resultant mapping a knowledge base[1]. A knowledge base is represented as an asymmetric directed graph in which nodes are concepts and arc weights are similarity measures. The knowledge base can also be visualized as a graph, illustrated by example on the left below, in which vertices represent concepts and edges represent the pair-wise similarity between two concepts.

In [8] the first author implemented techniques that produce a knowledge base using an extension of the statistical model developed in [10] and [11]. Ongoing research at Lehigh University includes enhancement of this cluster function to account for several additional factors including, for example, metrics such as the ratio of commonly used to total words in a concept.

3.4 Identification of regions of semantic locality

The resulting weight assignments from knowledge base creation are context-sensitive. We use these weights to determine regions of semantic locality (i.e., conceptual den-

[1] Note that follow-on work that builds on [10] terms this a *Concept Space*

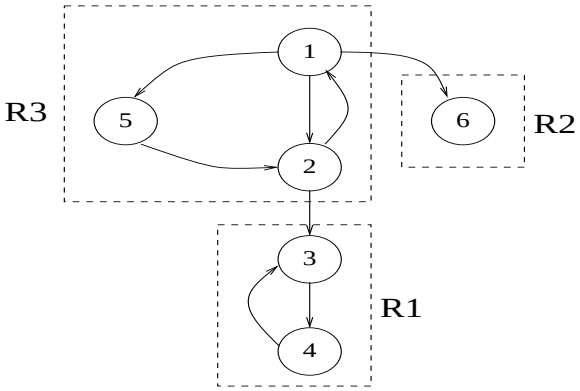

Figure 2. *An Example Application of Tarjan's Algorithm*

sity) within each collection. We thus detect clusters of concepts within a knowledge base [2], [12], [13].

The result is a knowledge base consisting of regions of high-density clusters of concepts: i.e., subtopic regions of semantic locality. These regions consist of clusters of concepts that commonly appear together and collectively create a knowledge neighborhood. Our premise is that we can impute a constrained, contextual transitivity to the co-occurrence relation [2], thereby forming regions of semantic locality. The motivation for the use of the term semantic locality comes from the commonly applied premise that grouping similar concepts together leads to increased effectiveness and efficiency in query search and retrieval [14]. Note however that the similarity relation is by definition not transitive. The theoretical basis for our algorithm, sLoc, is the concept we term contextual transitivity in the similarity relation. In essence, this means that depending on the context (structure and distribution of the similarities in the knowledge base), a threshold is decided upon and transitivity is constrained accordingly. Contextual transitivity extends Schütze's conceptualization of second order co-occurrence [17] by using n-order co-occurrence, where n varies with the underlying semantic structure of the model[2].

The computational core of sLoc is based on an algorithm due to Tarjan [13]. Tarjan's algorithm uses a variant of depth-first search to determine the strongly connected components of a directed graph. This was the first algorithm to solve the problem in linear time. This is an important feature due to the fact that graph clustering is a NP-hard problem and the only heuristics we are aware of are not linear. The theory can be found in [18]. Figure 2 depicts the operation of Tarjan's algorithm as it identifies strongly connected regions (R1, R2, R3) in a simple graph.

Before tackling the sLoc algorithm in detail we must first introduce the following notation:

- Let $\mathcal{N}$ be the set of nodes i in the input graph, and let N be the total number

[2] An interesting aside is thatthe efficacy of LSI can be viewed as the result of an implicit use of constrained, n-order co-occurrence based on contextual transitivity of the co-occurrence relation

of nodes.

- Let $\mathcal{A}$ be the set of arcs in the input graph, A the total number of arcs. An arc $a_{i,j} \in \mathcal{A}$ goes from node i to node j.
- Let $\mathcal{W}$ be the set of arc weights in the graph, $w_{i,j}$ is the weight of the arc going from node i to node j.

Therefore $\mathcal{W} = \{w_{i,j}\}_{(i,j)\in\mathcal{N}^2}$. A knowledge base is an asymmetric graph and thus $w_{i,j}$ may differ from $w_{j,i}$. Moreover, if $a_{i,j} \notin \mathcal{A}$ then $w_{i,j} = 0$; in particular, for all i, $w_{i,i} = 0$. Now, let M be the mean of the arc weights:

$$M = \frac{1}{A} \sum_{(i,j)\in N^2} w_{i,j}$$

We term the standard deviation of the distribution of arc weights SD:

$$SD = \sqrt{\frac{1}{A-1} \sum_{(i,j)\in N^2} (w_{i,j} - M)^2}$$

The sLoc Algorithm Step by Step

Figure 3 depicts the three steps of the sLoc process. Prior to the first step, the weights in the knowledge base are normalized (step 0 in figure 3 below). The first step in sLoc is to statistically prune the input graph. Arcs of weight smaller than a certain threshold τ are virtually pruned. Note that since the similarities are asymmetric, an arc from concept a to concept b can be pruned while the arc back from b to a remains. The second step involves the identification of the clusters within the graph. Tarjan's algorithm is applied to find strongly connected regions. At this stage each strongly connected region is a cluster. The size of a given cluster is the number of nodes (concepts) it contains. During the third and final step, clusters of size smaller than parameter s are discarded (they are assumed to be outliers). We interpret the remaining clusters as regions of semantic locality in the knowledge base.

The greater τ, the more arcs are cut off, and therefore the smaller in size the strongly connected regions. Thus the greater τ the smaller in size and the more focused will be the regions of semantic locality. Our premise is that the optimum τ can be determined statistically as a function of the mean M, the standard deviation SD and other knowledge base dependent variables (e.g., size, maximum weight, etc.). We have conducted some preliminary experiments to study the distribution of weights in various knowledge bases. Our preliminary results indicate that the distribution of weights is quite consistent across both subject domain and collection size. Figure 4 represents one such common distribution. This was computed from the MED gold standard information retrieval collection [16].

It should be stressed that we have yet to compute a knowledge base that does not exhibit a distribution of this nature. Given this fact, we have developed the following heuristic for the threshold τ: $\tau(\alpha) = max(w) - \alpha * SD$. In this equation,

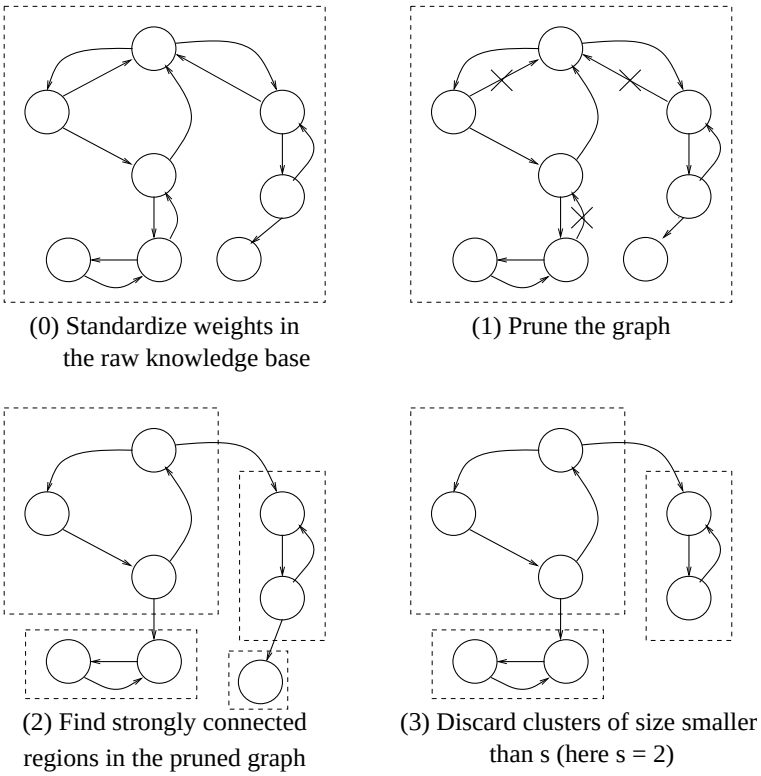

Figure 3. *The sLoc Process*

τ is the cut-off weight used to prune the graph and α is the number of standard deviations. For example, $\tau(1/2)$ is the threshold corresponding to the maximum weight in the graph minus half of the standard deviation of the distribution of arc weights.

We have also conducted experiments that indicate the optimal range of α lies in the range $[1.5, 2.0]$. Obtaining a scalable definition for the threshold is ongoing research.

3.5 Detection of Emerging Conceptual Content

Our fundamental premise is that computer algorithms can automatically detect emerging content by tracing changes over time in concept frequency and association. By taking a snapshot of the statistical state of a collection at multiple points in time, we expect to trace the emergence of hot topics.

We have identified at least two important features that an emerging concept should possess. First, in order to classify a concept as emerging, it should be semantically richer at a later time than it was at an earlier time. Second, an emerging concept should occur more frequently as an increasing number of items (documents)

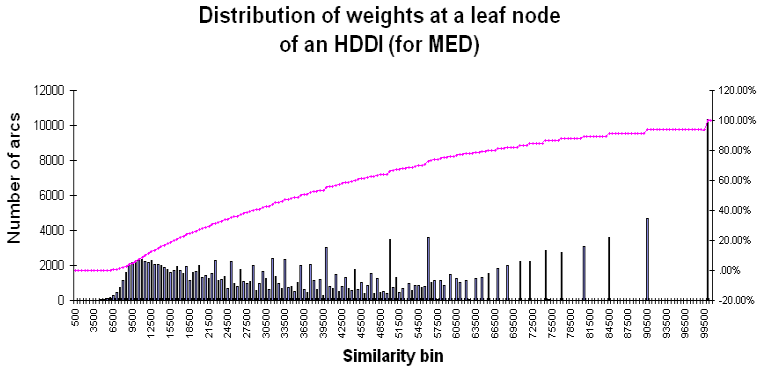

Figure 4. *The Distribution of Weights for the MED Collection*

reference it. We can approximate the semantic richness of a concept by the number of other concepts that are in the same region of semantic locality. To be an emerging concept, we maintain that the number of occurrences of a particular concept should exhibit an accelerating occurrence in a large corpus. In addition, if occurrences are artificially high they fall into a class of redundant concepts. Combining these constraints, we have automatically identified emerging content given a statistically significant sample of items from the domain of interest. In preliminary experiments we have achieved a precision of 34.8% and a recall of 24.1% [15].

We employ a cluster-based approach in which individual clusters of concepts represent regions of semantic locality that encompass portions of one or more items. Item-based approaches attempt to measure deltas in semantics between items. Based on our research in cluster-based retrieval mechanisms, however, we maintain that clusters more accurately capture the dynamics of the semantics between collections of items being compared across time.

Our approach hinges on developing a machine-learning model (e.g., an artificial neural network) to learn the function that maps from the statistical domain to the semantic domain. The input features we experimented with included the following:

(i) Number of occurrences of the concept in the trial year
(ii) Number of occurrences of the concept in the year before the trial year
(iii) Number of occurrences of the concept in the year two years before the trial year
(iv) Number of total occurrences of the concept before the trial year
(v) Number of concepts in the cluster containing the concept in the trial year
(vi) Number of concepts in the cluster containing the concept in the year right before the trial year
(vii) Number of words with length at least four in the concept

The first through the fourth features describe the occurrence frequency of concepts. The fifth and sixth features are cluster related features and describe

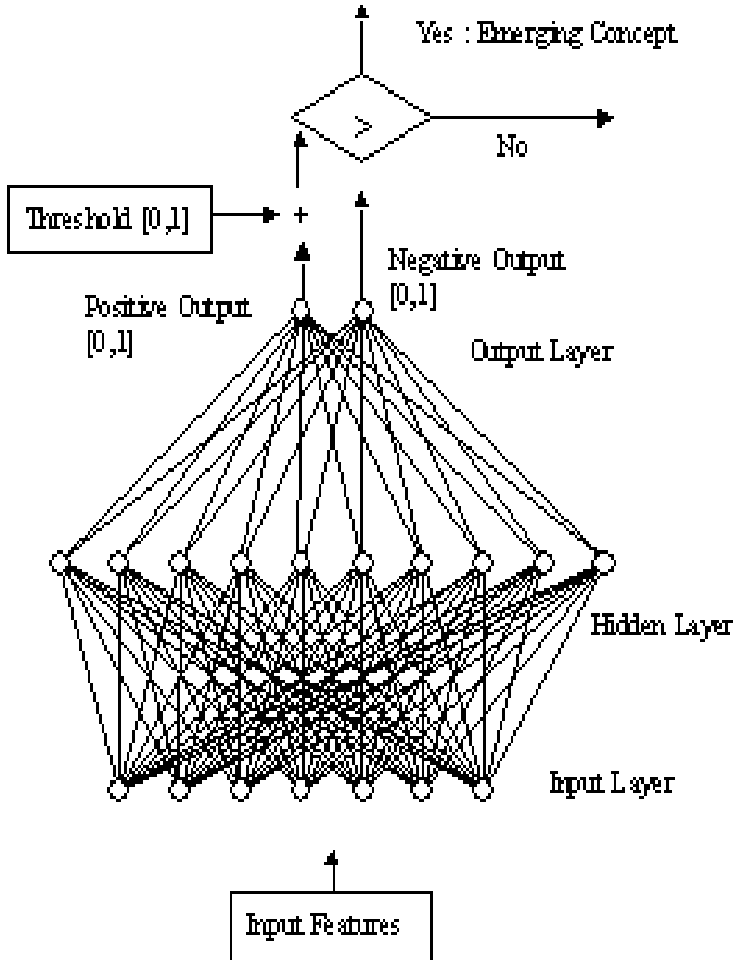

Figure 5. *A 7x10x2 neural network model for learning emerging concepts*

clusters and their change in size over time. The last feature is used to describe concept length and is a heuristic measure of the potential importance of concepts.

Figure 5 depicts a model of our initial approach to classify emerging concepts based on the winner take all strategy [19] employing a 7x10x2 neural network. When evaluating a test example, we employed a variable threshold between 0 and 1 added to the positive output. This new combined value is compared to the negative output. If the combined value is higher, the concept is identified as an emerging concept (positive), otherwise, it is identified as non-emerging concept (negative). The purpose of this approach is to develop a learning model for achieving better recall (just as a radar system depends on good recall) because domain experts will do the final filtering. By adjusting the threshold, we maintain that the domain expert can achieve a suitable balance of precision and recall.

4 Results

There are four databases we used to acquire the data for this research. These databases are the USPTO patent database, IBM patent database, INSPEC database and Compendex database. The first two databases store patent documents submitted to each repository. The last two databases store research publication abstracts. However, there is some overlap of the information stored in these two research publication databases, and potentially overlap of the two patent databases. Four major collections were used in this research. These are the IPP 6999 collection, CPP 6999 collection, BPP 7599 collection and UPP 7698 collection. The first letter in each collection title represents the source database; for example, IPP 6999 collection is set of papers from the INSPEC Database. The following characters represent the abbreviation of the subject domain, for example, PP stands for Processor and Pipeline. The last four digits represent the time frame over which the documents were extracted, for example, 6999 means the documents in this collection were submitted to the database from 1969 to 1999. Each collection contains only the titles and abstracts from the documents. The CPP 6999 collection contains all the titles and abstracts from the papers from 1969 to 1999 that contain both the noun phrases pipeline and processor. The BPP 7599 collection contains all the titles and abstracts from the patent documents from 1975 to 1999 that contain the same two noun phrases. The UPP 7698 collection contains all the titles and abstracts from the patents from 1976 to 1998 that contain the same two noun phrases. The difference in the input years is due to the fact that each database stores different collections of documents. For example, the USPTO database contains patent information from 1976 to 1998; the IBM database contains patent information from 1975 to 1999; the INSPEC and Compendex databases contain publications from 1969 to 1999.

Our results are depicted in terms of an evaluation metric based on the following definitions of precision and recall:

$$precision = \frac{pp}{pp + np} \tag{1}$$

$$recall = \frac{pp}{pp + pn} \tag{2}$$

In these formulas, pp is the number of positive examples identified as positive, pn is the number of positive examples identified as negative (false negatives), and np is the number of negative examples identified as positive (false positives).

When the threshold added is one, all concepts are identified as emerging concepts. In this scenario, the precision is the ratio of detected to total concepts and the recall is 100%. There are no false negatives. If the threshold is zero, the network is trained to minimize the total error in the training set, and the error of the testing set is usually not minimal. By changing the threshold, we can improve either the recall or the precision based on our need.

As noted, because the initial focus of our work is to achieve good recall, we have employed another metric termed F_β in the evaluation of performance:

$$F_{\beta=\sqrt{0.5}} = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision \times recall} \tag{3}$$

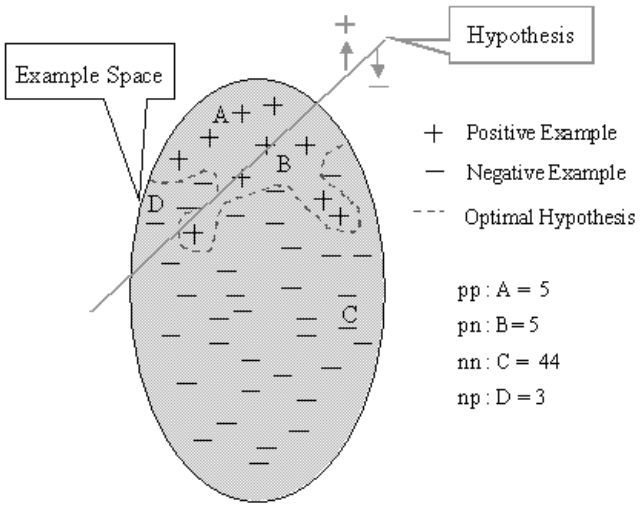

Figure 6. *The concept of positive examples, negative examples and hypothesis*

This metric is a weighted average of precision and recall and depends on the definition of β as the precision/recall ratio at which a user is prepared to trade a given increment in recall for an equal loss of precision [20].

Figure 6 exemplifies how precision and recall have been computed within this framework where precision, recall, and F_β are:

$$precision = \frac{A}{A+D} = \frac{5}{5+3} = 0.625 \tag{4}$$

$$recall = \frac{A}{A+B} = \frac{5}{5+5} = 0.5 \tag{5}$$

$$F_{\beta=\sqrt{0.5}} = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision \times recall} = \frac{(1+0.5) \times 0.625 \times 0.5}{0.5 \times 0.625 \times 0.5} = 0.577 \tag{6}$$

In our examples, three different neural networks were applied for learning emerging concepts. The first one uses 10 hidden neurons and runs 10000 epochs. The second network uses 20 hidden neurons and runs 10000 epochs, and the last network uses 40 hidden neurons and runs 50000 epochs. The reason we chose these three network settings is that these three networks were computationally feasible

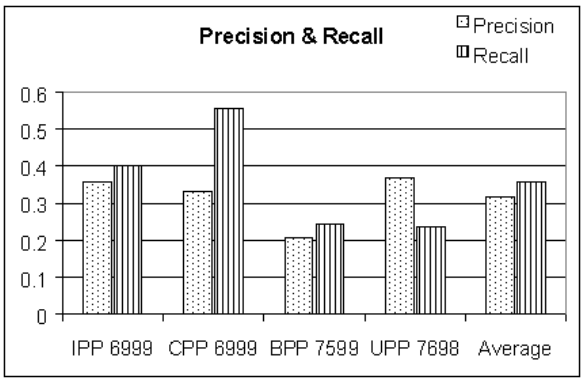

Figure 7. *Precision and Recall for Test Sets*

under current hardware resources. With the SUNW Ultra-Enterprise-10000 processor, it took two hours to train 10 hidden neurons with 10000 epochs, and it takes about 11 hours to train 40 hidden neurons with 50000 epochs. It is possible to use more hidden neurons and more epochs, however, it also takes more computational resources.

Figure 7 shows the precision and recall for the four testing sets using a neural network with 40 hidden neurons and 50000 epochs. The results indicate that the prediction performance of the model ranged from good for CPP 6999 to relatively poor for BPP 7598. The average precision and recall is 0.317 and 0.359 respectively. These results confirm that emerging concepts are learnable under the current framework. Compared with a baseline precision of 0.0686, this is an improvement of a factor of 4.62.

5 Related Work

Several research projects are exploring solutions to the detection of changes in topics. The Topic Detection and Tracking Pilot Study (TDT) project, for example, segments streams of data into distinct stories and identifies new events occurring in news stories [22], [21]. The TDT problem consists of three major tasks: (1) segmenting a stream of data, especially recognized speech, into distinct stories, (2) identifying those stories that are the first to discuss a new event occurring in the news, and (3) finding all stories in the stream given a small number of sample news stories about an event.

Major methods for new event detection in text data mining research come from work at Carnegie Mellon University (CMU), University of Massachusetts (UMass) and Dragon. The CMU approach clusters stories in a bottom-up fashion based on their lexical similarity and proximity in time. The UMass approach uses a variant of single-link clustering and builds up cluster groups of related stories to present events. The UMass method focuses on rapid changes by monitoring sudden changes in term distribution over time. The Dragon approach, based on observations of term fre-

quencies, uses adaptive language models from speech recognition. It hypothesizes a novel event when prediction accuracy of the adapted language models drops relative to the background models. The results show that the CMU incremental approach achieves 62%/67% in precision/recall, the CMU group average clustering top-level approach reaches 83%/43%, the Dragon approach reaches 61%/69%, the UMass 100T approach reaches 34%/53% and the UMass 10T reaches 33%/16%.

Kumar and other researchers in the IBM Almaden Research Center use a graph-theoretic approach to identify emerging communities in cyberspace [23]. The concept is that competitive websites in the same emerging community do not reference one another. Additionally, they may choose not to reference each other because they do not share the same points of view. Noncompetitive sites and those with similar points of view do link to these non-mutual-referencing sites. Thus, websites in the same community become a strongly connected bipartite graph. [23] proposed an efficient and effective algorithm to find the strongly connected cyberspace bipartite sub-graphs and cores. A core is a complete bipartite graph. An initial crawl found that 56% of the sampled communities were not in Yahoo! while a crawl 18 months later found 29% were not in Yahoo!. [23] interprets this finding as a measure of reliability of the trawling process, namely, that many communities that they identified as emerging 18 months ago did later emerge. The average level of these communities in the Yahoo! hierarchy was 4.5.

The Envision system at Virginia Tech is a digital library of computer science literature. It allows users to explore trends in digital library metadata [24]. Envision visually displays information search results as a matrix of icons with layout semantics that users control. The system gives users access to complete bibliographic information, abstracts and full content. It graphically presents a variety of document characteristics and supports an extensive range of user tasks. By using the Envision system, users can browse topics and trends graphically to identify emerging concepts.

ThemeRiver is a prototype (mock-up) that visualizes thematic variations over time across a collection of documents [25]. 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. This enables users to visualize trends and detect emerging themes. However, the proposed ThemeRiver system and the Envision system rely on human expertise to identify topics they do not provide a fully automatic approach to identify emerging topics in collections as does our approach.

The goal of many of these research projects is essentially to detect changes in topics disruptive events exhibiting discontinuities in semantics. Our research [15], [2], focuses on integrative or non-disruptive emergence of topics that build on previously existing topics. There is a subtle but important difference between these two approaches, and based on our research to date, we maintain that an integrative, cluster-based approach is necessary to identify emerging conceptual content with high precision.

6 Conclusion

We have constructed an artificial neural network classification model to classify emerging concepts. Three different networks were used to compare their performance. By changing an output threshold, we provided a method to improve the recall while maintaining an acceptable precision. The results also show that the performance is far better than the baseline precision.

In conclusion, our model has been successfully employed to recognize emerging concepts. There are many reasons we believe this model is well suited for textual data mining. These include such results like (1) cluster space is higher resolution than document space, (2) the features provided by cluster structures reflect the semantic relation well, and (3) the features provide sufficient information for classification. We believe this approach will result in significant advances in the fields of data mining and machine learning based on statistical approaches to semantic analysis.

Bibliography

[1] H. D. White and K. W. McCain, *Bibliometrics*, Annual Review of Information Science and Technology, Elsevier, 1989.

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

[3] G. Cooke, *SemanTag*, www.rt66.com/gcooke/.

[4] E. Brill, *A Simple Rule-based Part of Speech Tagger*, Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, 1992.

[5] 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*, www.eecs.lehigh.edu/~billp/pubs/HDDICB.doc, 2000.

[6] L. Karttunen, *Directed Replacement*, In the Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, California, 1996.

[7] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Eds., *Modern Information Retrieval*, ACM Press, New York, 1999.

[8] William Morton Pottenger, *Theory, Techniques, and Experiments in Solving Recurrences in Computer Programs*, Ph.D. thesis, Center for Supercomputing Research and Development in the Department of Computer Science at the University of Illinois at Urbana-Champaign, 1997.

[9] William M. Pottenger, *The Role of Associativity and Commutativity in the Detection and Transformation of Loop-Level Parallelism*, Proceedings of the 12th International Conference on Supercomputing, Melbourne, Australia, 1998.

[10] H. Chen and K. J. Lynch, *Automatic Construction of Networks of Concepts Characterizing Document Databases*, IEEE Transactions on Systems, Man and Cybernetics, 22(5):885-902, 1992.

[11] H. Chen, J. Martinez, T. Ng and B. R. Schatz, *A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System*, Journal of the American Society for Information Science, Volume 48, Number 1, 1997.

[12] F. D. Bouskila, *The Role of Semantic Locality in Hierarchical Distributed Dynamic Indexing and Information Retrieval*, M.S. Thesis, Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, 1999.

[13] R. E. Tarjan, *Depth first search and linear graph algorithms*, SIAM Journal of Computing, 1:146-160, 1972.

[14] K. Sparck-Jones, *Automatic Keyword Classification for Information Retrieval*, Butterworths, London, 1971.

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

[16] G. Salton, *Automatic Text Processing*, Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

[17] H. Schütze, *Automatic Word Sense Discrimination*, Computational Linguistics, vol. 24, no. 1, pp. 97-124, 1998.

[18] V. Aho, J. E. Hopcroft, and J. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[19] W. Maass, *Neural Computation with Winner-Take-All as the only Nonlinear Operation*, Advances in Neural Information Processing Systems 1999, vol. 12, MIT Press, Cambridge, 2000.

[20] N. Jardine and C. J. van Rijsbergen, *The USe of Hierarchic Clustering in Information Retrieval*, Information Storage and Retrieval, 7, 217-240, 1971.

[21] 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), pp32-43, 1999.

[22] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, *Topic Detection and Tracking Pilot Study: Final Report*, Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[23] R. Kumar, P. Raghavan., S. Rajagopalan and A. Tomkins, *Trawling Emerging Cyber-Communities Automatically*, Eighth International World-Wide-Web Conference, 1999.

[24] L. T. Nowell, R.K. France, D. Hix, L.S. Heath, and E.A. Fox, *Visualizing Search Results: Some Alternatives to Query-Document Similarity*, Proceedings of SIGIR '96, Zurich, 1996.

[25] S. Havre, B. Hetzler, and L. Nowell., *ThemeRiver: In Search of Trends, Patterns, and Relationships*, Presented at IEEE Symposium on Information Visualization, San Francisco, 1999.