Topic Detection, Tracking and Trend Analysis Using Self-organizing Neural Networks

K. Rajaraman* and Ah-Hwee Tan
Kent Ridge Digital Labs
21 Heng Mui Keng Terrace
Singapore 119613
email: {kanagasa,ahhwee}@krdl.org.sg

Abstract. We address the problem of Topic Detection and Tracking (TDT) and subsequently detecting trends from a stream of text documents. Formulating TDT as a clustering problem in a class of self-organizing neural networks, called the Adaptive Resonance Theory (ART) networks, we propose an incremental algorithm to solve this clustering problem. From the topics being detected and tracked, we show how trends can be identified. From our experimental studies, we find that our algorithm has been able to detect hot topics automatically and track them to a good accuracy. The method is also observed to enable discovering interesting trends that are not directly mentioned in the individual documents but deducible only from reading all relevant documents.

Keywords: Topic detection, topic tracking, trend analysis, text mining, document clustering

1 Introduction

In this paper, we address the problem of analyzing trends from a stream of text documents, using an approach based on the Topic Detection and Tracking initiative. Topic Detection and Tracking (TDT) [4] research is a DARPA-sponsored effort that has been pursued since 1997. TDT [1, 2, 3] refers to tasks on analyzing time-ordered information sources, e.g news wires.

Topic detection is the task of detecting topics that are previously unknown to the system[18]. Topic here is an abstraction of a cluster of stories that discuss the same event. Tracking refers to associating incoming stories with topics (i.e. respective clusters) known to the system[18]. The topic detection and tracking formalism together with the time ordering of the documents provides a nice setup for tracing the evloution of a topic. In this paper, we show how this setup can be exploited for analyzing trends.

Topic detection, tracking and trend analysis, the three tasks being performed on incoming stream of documents, necessitate solutions based on incremental algorithms. A class of models that enable incremental solutions are the Adaptive Resonance Theory (ART) networks[6], which we shall adopt in this paper. ART networks are a class of self-organizing neural networks that have found applications in several fields (See, for example, [7, 10]).

The rest of the paper is organized as follows. In Section 2, we describe our document representation and feature selection method. We provide a summary of ART networks in Section 3. In Section 4, we propose our topic detection and tracking algorithm and also describe our trend analysis method. Section 5 reports our experimental results and we conclude the paper with Section 6.

* Corresponding author

2 Document Representation

We adopt the traditional vector space model[14] for representing the documents, i.e. each document is represented by a set of keyword features. We use an in-house morphological analyzer to identify the part-of-speech and the root form of each word. To reduce complexity, only the root forms of the noun and verb terms are extracted for further processing. The following feature selection rules are then used to further refine the feature set:

- All words appearing in less than 5% of the collection are removed.
- From each document, only the top n number of features based on $tf.idf$ ranking are picked.

Let M be the number of keyword features selected through this process. With these features, each document is converted into a keyword weight vector

$$\mathbf{c} = (c_1, c_2, \ldots, c_M) \tag{1}$$

where c_j is the number of occurrence of the keyword w_j in the keyword feature list. The keyword weight vector is then normalized to produce the keyword feature vector

$$\mathbf{a} = \mathbf{c}/c_I \quad \text{where} \quad c_I = max_{i=1}^{M} c_i. \tag{2}$$

We assume that text streams are provided as document collections ordered over time. The collections must be disjoint sets but could have been collected over unequal time periods. We shall call these time-ordered collections as *segments*.

3 ART Networks

ART stands for Adaptive Resonance Theory and was introduced by Grossberg in 1976. ART networks are a class of self-organizing neural networks. There are several varieties of ART networks proposed in the literature[6, 5, 8], of which we shall adopt the fuzzy ART networks[9] .

3.1 Fuzzy ART

Fuzzy ART incorporates computations from fuzzy set theory into ART networks. The crisp (nonfuzzy) intersection operator ($\cap$) that describes ART 1 dynamics [6] is replaced by the fuzzy AND operator ($\wedge$) of fuzzy set theory in the choice, search, and learning laws of ART 1. By replacing the crisp logical operations of ART 1 with their fuzzy counterparts, fuzzy ART can learn stable categories in response to either analog or binary patterns.

Each fuzzy ART system (Figure 1) includes a field, F_0, of nodes that represents a current input vector; a field F_1 that receives both bottom-up input from F_0 and top-down input from a field, F_2, that represents the active code or category. The F_0 activity vector is denoted $\mathbf{I}$. The F_1 activity vector is denoted $\mathbf{x}$. The F_2 activity vector is denoted $\mathbf{y}$.

Due to space constraints, we skip the description of fuzzy ART learning algorithm. The interested reader may refer to [9] for details.

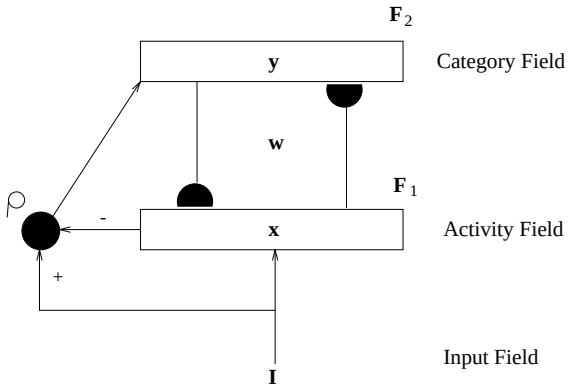

Fig. 1. Adaptive Resonance Theory network architecture.

4 Topic Detection, Tracking and Trend Analysis

As defined in Section 1, Topic detection is the task of identifying topics from document streams, that are previously unknown to the system. Tracking refers to associating incoming stories with topics known to the system.

4.1 Topic Detection Algorithm

As described in Section 3, ART formulates recognition categories of input patterns by encoding each input pattern into a category node in an unsupervised manner. Thus each category node in F_2 field encodes a cluster of patterns. In other words, each node represents a topic. Hence, identification of new topics translates to the method of creation of new categories in the F_2 field as more patterns are presented.

Using this idea, we derive the topic detection algorithm in Table I.

Step 1. Initialize network and parameters.
Step 2. Load previous network and cluster structure, if any.
Step 3. Repeat
- present the document vectors
- train the net using fuzzy ART Learning Algorithm
until convergence
Step 4. Prune the network to remove low confidence category nodes
Step 5. Save the net and cluster structure.

Table I. Topic Detection Algorithm.

Initialization is done in Step 1. In Step 2, the previous network state is loaded if the incremental topic detection is desired. The actual topic detection is done in the iterative Step 3. At every iteration, the document vectors are presented to the network and the network is trained using the ART learning algorithm. This is continued until the network converges, i.e. resonance is achieved for all document vectors. In Step 4, the inputs vectors

are run through the net to detect low confidence nodes and prune them. This prevents topic proliferation. The algorithm ends by saving the network state and topic structure in Step 5.

4.2 Topic Tracking Algorithm

For tracking new documents, the latest topic structure is loaded before processing the documents. For an incoming document, the activities at the F_2 field are checked to select the winning node, i.e. the one receiving maximum input. The document is then assigned to the corresponding topic. This is the idea behind the tracking algorithm presented in Table II.

Step 1. Initialize network and parameters.
Step 2. Load previous network and cluster structure, if any.
Step 3. Present the document to be tracked, to the net
Step 4. Assign the document to the topic corresponding to the winning category node, i.e. category node that receives maximum input.

Table II. Topic Tracking Algorithm.

The algorithm is easy to understand. It can be noted that, by using a threshold on the input to the category nodes, multiple winning nodes could be selected. Hence, the document could be assigned to multiple topics. This means that overlapping topics could be tracked with our ART network model.

4.3 Trend Analysis

The topic detection and tracking setup together with the time ordering of the documents provides a natural way for topic-wise focussed trend analysis. In particular, for every topic, suppose we plot the number of documents per segment versus time. This plot can be thought of as a trace of the evolution of a topic. The 'ups' and 'downs' in the graph can be used to deduce the trends for this topic. For more specific details on the trends, one can zoom in and view documents on this topic segment-wise. This process is illustrated in the following section on our experimental studies.

5 Experiments

For our experiments, we have used news items from CNET and ZDNet. We have made use of a crawler to grab daily news from these sites and grouped the news items into weekly segments. Starting from 1st week of September 2000 up till 4th week of October 2000, we collected 8 segments in all. Totally there were 1468 documents at an average of about 180 documents in each segment. Documents in each segment are processed using the in-house morphological analyzer and converted into feature vectors as described in Section 2. We then applied our topic detection and tracking and performed trend analysis. Some qualitative results are presented below:

5.1 Topic Detection and Tracking

Typically we observed 10 to 15 new topics being identified per segment when choice parameter $\alpha = 0.1$ and vigilance parameter $\rho = 0.01$ (ignoring small clusters with 1 or 2 documents only).

A sample list of some of the frequently mentioned hot topics that have been identified by the topic detection algorithm can be viewed at ??? [website URL]. The tracking results can be viewed at ??? [website URL]. We skip the details due to space constraints.

5.2 Trend Analysis

The evolution graphs for some selected topics are shown below. It may be recalled that the graph plots the number of documents/segment over time. Time is represented through the segment ID which takes values 1,$\cdots$,8. ID=1 corresponds to Sep 1st week, ID=2 corresponds to Sep 2nd week and so on.

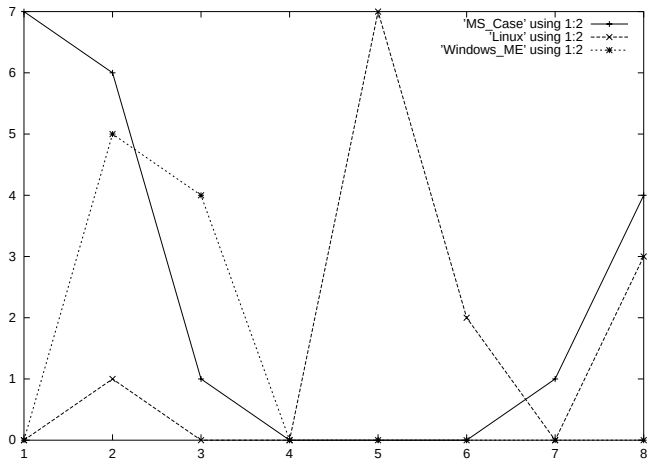

Fig. 2. Trends for 'MS Case', 'Linux' and 'Windows ME'.

The topics 'MS Case' (i.e. Microsoft Case), 'Linux' and 'Windows ME' have been plotted in Fig 2. The 'MS Case' topic shows an initial up trend early September. An examination of the documents under this topic reveals the reason to be Bristol Technology ruling against Microsoft. Similarly the topic on 'Linux' shows a peak for early October when the Open source conference was held. 'Windows ME' graph peaks during September 2nd week coinciding with Win ME release.

The topics 'Apple' and 'Hackings' have been plotted in Fig 3. The 'Apple' topic shows an up trend during mid September when Apple Expo was on. The Microsoft hack-in can be seen to have lead to the sudden peak in 'Hackings' topic around late October.

The above study thus shows that our method can be used to detect hot topics automatically and track the evolution of detected topics. The method also serves to quickly spot emerging trends on topics of interest and make decisions suitably.

6 Related Work

TDT research has been predominantly 'pure IR' based and can be categorized into one of the following types:

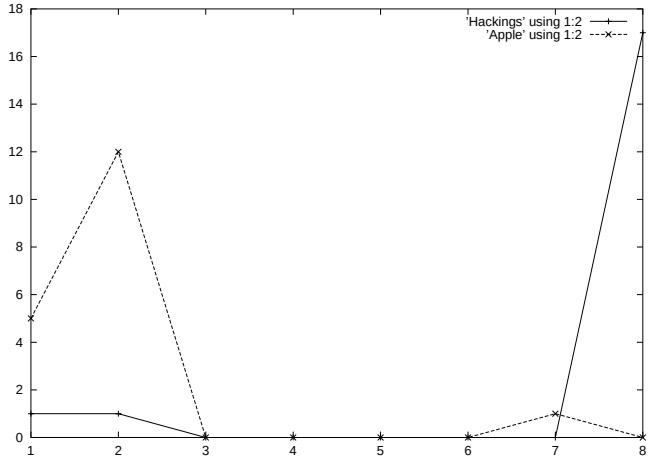

Fig. 3. Trends for 'Apple' and 'Hackings'.

- based on incremental clustering (e.g. [17])
- based on routing-queries (e.g. [13])

One notable exception is the tracking method by Dragon Systems[19]. Their method is based on language-modelling techniques (unigram statistics). Incremental clustering based methods come the closest to our work, but we use ART networks for document processing in contrast to the traditional document similarity measures. ART networks enable truly incremental algorithms and are biologically plausible.

Trend analysis for numerical data has been well investigated. For free-text, where the challenge is tough, we are aware of only very few papers. Incidentally, all of them use the approach of finding trends in terms of emerging or diminishing keywords/phrases.

[11] define concept distributions and propose a trend analysis method by comparing distributions of old and new data. Typically, the trends discovered are of the type "keyword 'napster' appeared x% more now than in old data", "keyword 'divx' appeared y% less now than in old data", etc.

[12] uses the popular a-priori algorithm employed in association-rule learning, for finding interesting phrases. Trend analysis is done by applying a shape based query language on the identified phrases. Queries like 'Up' or 'BigDown' could be used to identify upward and strong downward trends respectively, in terms of phrases. However, there could be potentially large number of candidate phrases that could make this method inefficient.

[15] proposes a method for automatically identifying topics from a newsfeed that are getting greater coverage. The method is based on a statistic estimated from term frequencies. However, this method assumes each news article is annotated manually with relevant keywords.

In contrast, our trend analysis method being based on topic detection and tracking enables finding specific, topic-wise trends. The formulation under the TDT framework offers several advantages. The topic detection and tracking step enables the trend analysis be focussed and more meaningful. Since the documents under each topic are relatively small, the analysis can be done efficiently. (On a related note, the ART learning algorithm can be implemented parallelly and this implies potential further speedup.)

7 Conclusion

We have addressed the problem of topic detection and tracking and thereafter detecting trends, from a stream of text documents. First we have formulated TDT as a clustering problem in ART networks and proposed an incremental algorithm to solve the problem. From the topics being tracked, we have shown how trends can be identified. From our experimental studies, we have found our algorithm has been able to detect hot topics and track them to a good accuracy. Also our method has enabled discovering interesting trends that are not directly mentioned in the individual documents but deducible only from reading all relevant documents.

Our work points to several directions for further investigation. Getting quantitative results under the proposed trend detection formulation is important. Building a benchmarking setup similar to TDT would be a worthwhile to direction to pursue. Currently we are working on addressing this issue. On another direction, ART type networks have recently been shown to facilitate domain knowledge integration[16]. It would be interesting to investigate how domain knowledge can be integrated into our model and whether it can lead to more meaningful trend analysis. This will form part of our future work.

References

1. Proceedings of the TDT-98 workshop. `http://www.itl.nist.gov/iad/894.01/tests/tdt/tdt98/index.htm`, 1998.
2. Proceedings of the TDT-99 workshop. `http://www.itl.nist.gov/iad/894.01/tests/tdt/tdt99/index.htm`, 1999.
3. Proceedings of the TDT-2000 workshop. `http://www.itl.nist.gov/iad/894.01/tests/tdt/tdt2000/index.htm`, 2000.
4. TDT homepage. `http://www.itl.nist.gov/iad/894.01/tests/tdt/index.htm`, 2000.
5. G. A. Carpenter and S. Grossberg. ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26:4919–4930, 1987.
6. G. A. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37:54–115, 1987.
7. G. A. Carpenter and S. Grossberg. Integrating symbolic and neural processing in a self-organizing architecture for pattern recognition and prediction. Technical Report CAS/CNS-TR-93-002, Boston, MA: Boston University, 1993. To appear in: V. Honavar and L. Uhr (eds.) *Symbol Processors and Connectionist Networks in Artificial Intelligence and Cognitive Modeling: Steps Towards Principled Integration*. New York, NY: Academic Press.
8. G. A. Carpenter, S. Grossberg, and J. H. Reynolds. ARTMAP: Supervised real time learning and classification by a self-organizing neural network. *Neural Networks*, 4:565–588, 1991.
9. G. A. Carpenter, S. Grossberg, and D. B. Rosen. ART 2-A: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4:493–504, 1991.
10. G. A. Carpenter and W. D. Ross. ART-EMAP: A neural network architecture for object recognition by evidence accumulation network. In *World Congress on Neural Networks, Portland, OR*, volume III, pages 649–656. Hillsdale, NJ: Lawrence Erlbaum Associates, July 1993.
11. R. Feldman and I. Dagan. Knowledge discovery in textual databases (KDT). In *Proceedings of KDD-95*, 1995.
12. Brian Lent, Rakesh Agrawal, and R. Srikant. Discovering trends in text databases. In *Proceedings of KDD-97*, 1997.

13. Ron Papka, James Allan, and Victor Lavrenko. UMASS approaches to detection and tracking at TDT2. In *Proceedings of the TDT-99 workshop*. NIST, 1999.
14. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.
15. Mark Shewhart and Mark Wasson. Monitoring newsfeed for hot topics. In *Proceedings of KDD-99*, pages 85–92?, 1999.
16. A.H. Tan and F.L. Lai. Text categorization, supervised learning and domain knowledge integration. In *Proceedings of KDD-2000: Workshop on Text Mining*, pages 113–114. ACM, Aug 2000.
17. Fredrick Walls, Hubert Jin, Sreenivasa Sista, and Richard Schwartz. Topic detection in broadcast news. In *Proceedings of the TDT-99 workshop*. NIST, 1999.
18. Charles Wayne. Overview of TDT. `http://www.itl.nist.gov/iaui/894.01/tdt98/doc/tdtslides/sld001.htm`, 1998.
19. J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. Topic tracking in a news stream. In *Proceedings of the TDT-99 workshop*. NIST, 1999.