Self-Organizing Maps for Topic Trend Discovery

Richard Rzeszutek, *Student Member, IEEE*, Dimitrios Androutsos, *Senior Member, IEEE*, and Matthew Kyan, *Member, IEEE*

***Abstract*—The large volume of data on the Internet makes it extremely difficult to extract high-level information, such as recurring or time-varying trends in document content. Dimensionality reduction techniques can be applied to simplify the analysis process but the amount of data is still quite large. If the analysis is restricted to just text documents then Latent Dirichlet Allocation (LDA) can be used to quantify semantic, or topical, groupings in the data set. This paper proposes a method that combines LDA with the visualization capabilities of Self-Organizing Maps to track topic trends over time. By examining the response of a map over time, it is possible to build a detailed picture of how the contents of a dataset change.**

***Index Terms*—Data processing, self-organizing feature maps, statistical analysis, time-series analysis, topic trending.**

I. Introduction

ONE way to describe the Internet is that it is a large collection of interrelated documents. Search engines exploit these relationships [1] to build indexes for fast document retrieval. One important problem, however, is classifying and tracking document content over time based on its semantic content [2]. Techniques exist [3], [4] that can extract semantic relationships from documents. LDA is a particularly powerful method since it is able to infer the underlying statistical model used to generate the observed data. This produces descriptors for each document that then can be used for classification since semantically similar documents will have similar descriptors.

In order to track changes in topics over time, simple time-series techniques have been applied to a corpus analyzed using LDA [5]. For instance, in [5], the authors show how scientific papers on global warming gradually increase in popularity over a ten year period. Unfortunately it is not uncommon to have $100+$ topics (i.e., dimensions) in a descriptor which makes this analysis difficult for any more than two topics.

Therefore, we propose to use a method similar to the WEBSOM [6] and ProbMap [7] algorithms to perform the trend analysis. These algorithms use Kohonen's Self Organizing Maps [8] to nonlinearly project a high dimensional feature space onto a low-dimensional output space. Our method merges the idea behind WEBSOM and ProbMap with the work done in [5] to show how a document corpus can change with time.

Manuscript received March 01, 2010; revised April 11, 2010. Date of publication May 03, 2010. Date of current version May 07, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jen-Tzung Chien.

The authors are with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3 Canada (e-mail: rrzeszut@ee.ryerson.ca; dimitri@ee.ryerson.ca; mkyan@ee.ryerson.ca).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2010.2048940

II. Corpus Modelling

A text corpus is simultaneously a collection of words and a collection of documents. For instance, IEEE Xplore (http://ieeexplore.ieee.org) is a good example of what constitutes a corpus. Each of the journal publications, conferences papers, etc. are all documents which, in turn, contain the collection of words that made up that particular document. Therefore, given a corpus, it is possible to define a word-document co-occurance matrix, $\mathbf{C}$, that relates the number of times a given word occurs for a given document. It is desirable to find a way to decompose $\mathbf{C}$ because a corpus may contain millions of documents and thousands of words.

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation probabilistically decomposes $\mathbf{C}$ such that

$$\mathbf{C} \mapsto \mathbf{\Phi\Theta} \tag{1}$$

where $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are matrices that express how words, topics and documents are related. $\mathbf{\Phi}$ contains the topic-word likelihoods or, put another way, how likely any given word w is to appear in topic k. $\mathbf{\Theta}$ is a collection of document-topic likelihoods which relates how likely a document d is to contain topic k. Therefore, for each topic k, there is a vector, $\vec{\phi}_k$, that contains the word distribution for that topic. Similarly, there exists a vector $\vec{\theta}_d$ for each document d that contains the topic distribution for that document.

We ask the reader to refer to [4] and [5] for a full derivation of LDA and its properties.

B. Document Space

The topic-distribution matrix, $\mathbf{\Theta}$, acts as a natural descriptor for all of the documents in the corpus. For any document, d, its associated vector $\vec{\theta}_d$ describes its location in a **document space**. Finding similar documents then simply reduces to a *k*-nearest neighbor search. LDA ensures that documents that are semantically similar (i.e., share many words that are in the same topics) will be close to one another in the document space. Because $\vec{\theta}_d$ is actually a probability mass function (PMF), it has to satisfy the constraint

$$\sum_{k=0}^{N_T-1} \theta_d^k = 1 \tag{2}$$

where N_T is the number of topics found through LDA.

As a result of this, measuring distance in document space is the same as measuring the dissimilarity between two PMFs. There exists a family of divergences [9] that can be used to measure dissimilarity without having to resort to the generic Euclidean distance. A common choice of divergence in this sort of

situation is the Kullback–Liebler (KL) divergence and it is defined for two D-dimensional PMFs, $\vec{A}$ and $\vec{B}$ as

$$\mathrm{KL}(\vec{A},\vec{B}) = \sum_{i=0}^{D-1} A_i \log\left(\frac{A_i}{B_i}\right). \tag{3}$$

The KL divergence is *not* a distance measure (i.e., metric) so it does not satisfy the triangle inequality and $\mathrm{KL}(\vec{A},\vec{B}) \neq \mathrm{KL}(\vec{B},\vec{A})$. For convenience, it is useful to use a symmetric KL divergence so that the order of the arguments is not important. The symmetric KL measure is defined as

$$\mathrm{sKL}(\vec{A},\vec{B}) = \mathrm{KL}(\vec{A},\vec{B}) + \mathrm{KL}(\vec{B},\vec{A}). \tag{4}$$

III. Time-Varying Corpora

In most cases, the content of a corpora will change over time as new documents are added to it. This is especially true for much of the content on the Internet, such as RSS feeds and blogs, where the *time* that the content was generated can be just as important as the content itself. Applying LDA onto a corpus with N_D documents produces a set of document descriptors $\{\vec{\theta}_d : \vec{\theta} \in \mathbf{\Theta}, 0 \leq d \leq N_D - 1\}$. But, because the documents can be added to the corpus at different times, each descriptor also includes a time component t so that they are really $\vec{\theta}_d(t) = [\vec{\theta}_d\, t]$. That makes it possible to examine how $\mathbf{\Theta}$ varies over time.

A. Dataset

For this paper, a corpus was constructed of 25 754 documents that were collected over a three-month period starting in late May 2009 and ending in late August 2009. The documents were article summaries from RSS feeds from sports news and opinion websites such as ESPN.com and TSN.ca. Prior to running LDA, common words such as "a" and "the" were removed since they do not contribute anything useful to the analysis (words such as these are likely to occur equally for each topic). The values of the hyperparameters were taken from [5] and $N_T = 40$ since we found that this number of topics provided the best tradeoff between descriptor length and the ability to effectively describe the corpus.

B. Arrival Times

Ideally, analyzing the change in topic distribution should be a simple matter of plotting, in one way or another, the values of $\vec{\theta}_d(t)$. Unfortunately there is no guarantee that the documents will arrive in any sort of a consistent manner. Consider Fig. 1. It shows the length of time, in hours, between when two successive documents, d and $d+1$, were added to the corpus. The time between arrivals is more or less random and can even occur out of order. This complicates analyzing the corpus since most time-series methods assume a uniform sampling of the data.

To deal with the non-uniform arrival times, we analyze the documents descriptors by using a sliding window with a fixed step size. The width of each window, T_{wnd}, is defined to be

$$T_{\mathrm{wnd}} = T_{\mathrm{end}} - T_{\mathrm{start}} \tag{5}$$

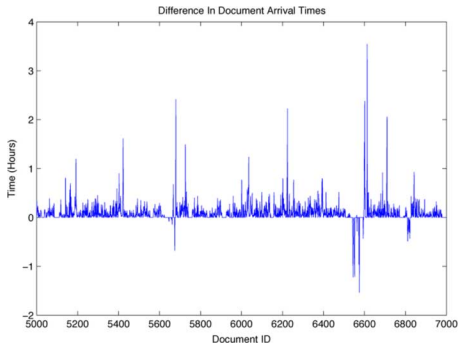

Fig. 1. Differences in document arrival times. Negative values indicate *out-of-order* arrival.

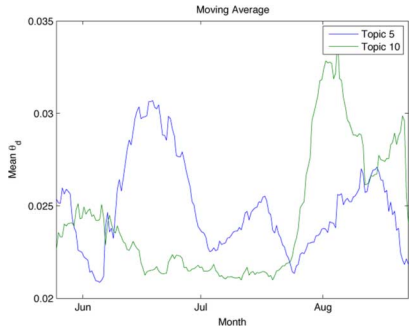

Fig. 2. Moving average topic trend analysis for two topics.

where T_{start} and T_{end} are the start and end times of the window. To ensure a relative degree of smoothness between windows, we keep the step size less than $T_{\mathrm{wnd}}/2$.

IV. Trend Analysis

Given the descriptors, there are a number of ways that they can be examined. The simplest way is to use a moving average to produce an "average" document for each time window, t_w. By looking at the value for a topic, k, as it changes over time, it is possible to infer how popular that topic is, as is done in [5]. A more nuanced approach would be to use a method capable of capturing the distribution of documents in the document space for that particular time window.

A. Moving Average

The moving average approach simply produces a document descriptor that is the average descriptor for any particular time window. For each window, t_w, we obtain a descriptor, $\bar{\theta}(t_w)$, such that

$$\bar{\theta}(t_w) = \frac{1}{N_D(t_w)} \sum_{i=0}^{N_D(t_w)-1} \vec{\theta}_i(t_w) \tag{6}$$

where $N_D(t_w)$ is the number of documents inside of the time window at time t_w. $\bar{\theta}(t_w)$ then represents the central tendency of the documents inside of that time window.

When analyzing only a couple of topics, this method is actually quite useful since it clearly shows how they vary over time and with respect to one another. Fig. 2 shows the results for topics 5 and 10, as returned by the LDA stage. The top ten most likely words for each topic are given in Table I.

TABLE I
TOP TEN WORDS FOR TOPICS 5 AND 10. THESE GROUPS CORRELATE TO THE PGA CHAMPIONSHIP (GOLF) AND THE PRE-SEASON TRAINING FOR THE KANSAS CITY CHIEFS (NFL)

Topic 5	open, championship, major, round, south, ap, reuters, pga, british, golf
Topic 10	camp, chiefs, training, end, practice, thomas, defensive, time, first, reciever

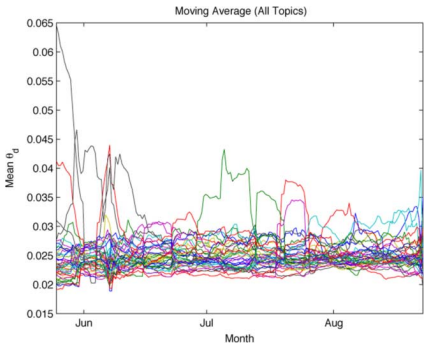

Fig. 3. Moving average for *all* topics.

From the plot, it is clear that topic 5 remained popular for quite some time before topic 10 began to overtake it in August. Topic 5 attained a maximum value on June 18 while Topic 10 attained its maximum on August 4. Notice that topic 5 correlates with golfing and topic 10 correlated with American football. Since sports news is time-sensitive, this is reflected in the value of $\bar{\theta}(t)$. As the golfing season concludes, so do articles with topics relating to golfing and as the NFL preseason approaches, stories relating to the NFL become more popular.

Unfortunately, it is very difficult to extend this sort of analysis to more than just two or three topics. Consider Fig. 3. All of the topics are shown on the same plot and it is extremely difficult to visually see any underlying patterns. More importantly, this type of trending only shows the most *dominant* document types at any point in time. This will not, for instance, show if there are multiple document clusters for the same time period.

B. Self-Organizing Map

The Self-Organizing Map (SOM), as proposed by Kohonen [8], maps a high-dimensional feature space onto a lower dimensional representation (usually one or two dimensions). The map itself is a regular lattice, typically either a hexagonal or rectangular grid. Each node, j, has an associated weight vector, $\vec{n}_j$, that specifies its position in the feature space. This allows the map to perform a nonlinear dimensionality reduction on a dataset. However, it can also be used as a classifier, which is what we do in this paper. By examining how many data points each node classifies, it is possible to map complex structures in the feature space onto the lattice.

This mapping property of SOMs can be interpreted as an "intelligent" moving average where the map can now show the actual distribution of documents in a given time window, rather than just the central tendency. We use the SOM to perform the trend analysis in a two-stage process.

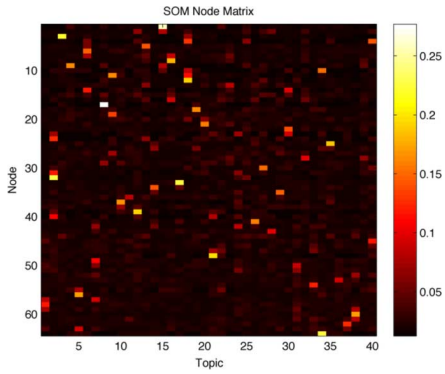

Fig. 4. Visualization of the document vectors for each node in an 8×8 SOM after training. The node ID is represented by the vertical axis while the topic ID is represented by the horizontal axis. The more white the color, the greater the likelihood of that topic being associated with that node.

The first stage trains the SOM while the second stage filters the data through the trained SOM to produce the response maps. The first stage trains the SOM on a small, random subset of the original input data. For the dataset used in this paper, that subset is 500 randomly selected documents. This is done since we do not need the SOM to accurately model the *entire* dataset, just loosely resemble it. This has the added benefit of reducing the computational burden when training the SOM, especially for very large datasets. After training, the SOM will now resemble $\boldsymbol{\Theta}$ such that each node is actually the representation of a cluster of similar documents (Fig. 4). As discussed in Section II-B, KL-divergence is used for the "distance" measure since it well suited to describing the dissimilarity between the document probability vectors.

The second stage filters, or processes, the dataset through the SOM using the sliding window method described in Section III-B. We use the trained SOM as a classifier to determine how many documents in the window are classified by each node. Each node has an associated classification count, N_j, or the number of documents that are associated with node j. A document is associated with a node if $\vec{n}_j$ is the closest node to that document vector. We define a classification density, $D(\vec{n}_j)$, such that

$$D(\vec{n}_j) = \frac{N_j - N_{\min}}{N_{\max} - N_{\min}} \tag{7}$$

where $N_{\min}$ and $N_{\max}$ are the maximum and minimum count values in the window. This ensures that the map is normalized to be the range of [0,1] so that different windows can be compared.

This process produces a response map for each time window. Fig. 5 shows the response maps, or "slices" over a two-week period in July. The more red the color, the more "active" the node, indicating how the documents are distributed in that slice. Please note that the maps have been upsampled using bicubic interpolation from 8×8 pixel images to 64×864 pixel images for clarity.

Notice that over time, the SOM is reporting what *groups* of documents are becoming more or less popular. This is different than the moving average, which can only report which document group is most popular. Fig. 6 shows the SOM slice for July 18th.

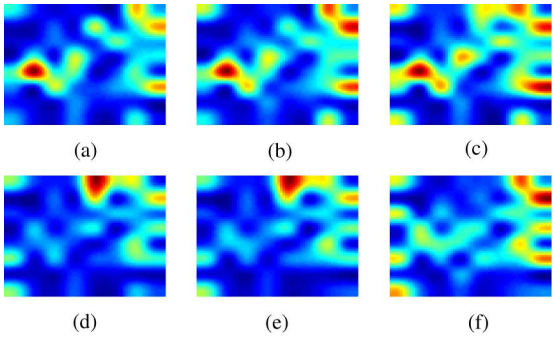

Fig. 5. SOM slices for a roughly two-week period in July 2009. The time difference between the images is two and a half days. (a) July 13, (b) July 16, (c) July 18, (d) July 21, (e) July 23, and (f) July 26.

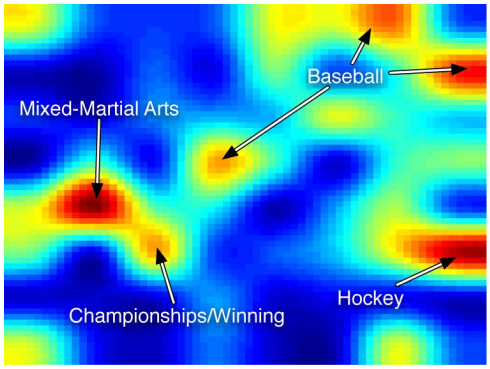

Fig. 6. SOM slice for July 18. The different nodes have been labelled based on the semantic grouping that they best reflect.

The SOM is detecting *several* events occurring simultaneously. First, the strong response to mixed-martial arts (UFC) implies that a UFC tournament is about to occur or has just occurred. Baseball has multiple response since July is in the middle of the baseball season so games are occurring along with other related news. An interesting feature of this slice is the strong response to hockey-related news. Because the response is far from the node associated with championships and winning, this implies that news is not related to any games. In fact, the NHL entry draft occurs around this time, resulting in hockey-related news in the middle of the summer.

C. Topic Volume

As the map responses change over time, it defines a 3-D volume (Fig. 7). This volume describes *how* the map responds over time, as opposed to just observing the response of the SOM at any particular time. As before, the clustering properties of the SOM makes it possible to determine how the document distribution itself changes.

For example, if a document type is very popular over a long period of time, this will be reflected in the volume as a sustained response. This corresponding 3-D structure will the give an indication of how long the topic was popular for. Therefore, it is possible to not only rank topics by popularity, but also for the duration of their popularity.

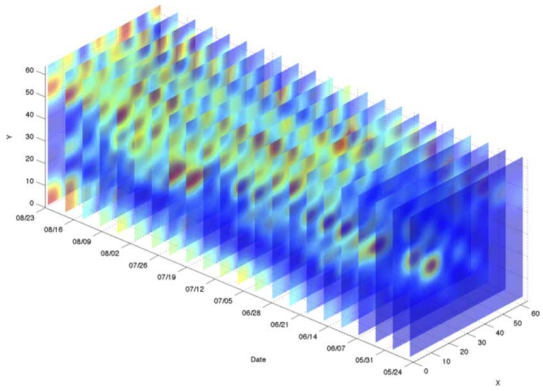

Fig. 7. SOM response volume built up from the responses over time.

V. Conclusion

The Self-Organizing Map is a powerful method for visualizing and examining a complex, time-varying corpus. Latent Dirichlet Allocation provides a means of *describing* individual documents, but it does not provide any information on how the documents change. By combing LDA with SOMs, it becomes possible to observe complex changes in a corpus that would not have been possible otherwise. Furthermore, by using a SOM to examine snapshots of the corpus in time, the slices generated by the SOM visualize the composition of the corpus at different points in time.

Acknowledgment

The authors would like to acknowledge M. Rose and J. Moeinifar at WhoThaMan Media Company for their collaboration on examining topic trends in web content and for use of their news article database. They would also like to acknowledge the MITACS ACCELERATE program for making this collaboration possible.

References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, The Pagerank Citation Ranking: Bringing Order to the Web Stanford University, , Tech. Rep., 1998.
[2] C. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York: Cambridge Univ. Press, 2008.
[3] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, New York, 1999, pp. 50–57, ACM.
[4] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[5] T. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci.*, vol. 101, pp. 5228–, 2004, Suppl. 1.
[6] K. Lagus, S. Kaski, and T. Kohonen, "Mining massive document collections by the WEBSOM method," *Inform. Sci.*, vol. 163, no. 1–3, pp. 135–156, 2004.
[7] T. Hofmann, "Probmap—A probabilistic approach for mapping large document collections," *Intell. Data Anal.*, vol. 4, no. 2, pp. 149–164, 2000.
[8] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Berlin, Germany: Springer, 1995.
[9] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.