Finding topics in dynamical text: application to chat line discussions

Ella Bingham
Neural Networks Res. Centre
Helsinki Univ. of Technology
P.O.Box 5400, 02015 HUT, Finland
ella.bingham@hut.fi

Ata Kabán
Neural Networks Res. Centre
Helsinki Univ. of Technology
P.O.Box 5400, 02015 HUT, Finland
kaba-ci0@paisley.ac.uk

Mark Girolami
Neural Networks Res. Centre
Helsinki Univ. of Technology
P.O.Box 5400, 02015 HUT, Finland
giro-ci0@paisley.ac.uk

ABSTRACT

The problem of analysing dynamically evolving textual data has recently arisen. An example of such data is the discussion appearing in Internet chat lines. In this paper a recently introduced method, termed *complexity pursuit*, is used to extract the topics of a dynamical chat line discussion. Experimental results demonstrate that meaningful topics can be found and also suggest the applicability of the method to query-based retrieval from a temporally changing text stream.

Keywords

chat line discussion, dynamical text, topic identification, complexity pursuit, data mining

1. INTRODUCTION

In times of huge information flow in the Internet, there is a strong need for automatic textual data analysis tools. Algorithms developed for text mining from *static* text collections have been presented e.g. in [1, 2, 7]. Our emphasis is in the recently arisen issue of analyzing *dynamically evolving* textual data; investigating appropriate tools for this task is of practical importance. An example of such data is found in the Internet relay chat rooms: the topic of interest changes after participants' contributions. The online text stream can thus be seen as a time series, and methods of time series processing may be used to extract the topics of the discussion.

We present results of applying a recently introduced powerful method, *complexity pursuit* [3], to topic extraction in a dynamically evolving discussion. Complexity pursuit uses both information-theoretic measures and time-correlations of the data, which makes it more powerful than methods using only one of these — the latter kind of methods include [5, 6, 8, 9].

2. CHAT LINE DATA

The discussion found in chat lines on the Internet is an ongoing stream of text generated by the chat participants and the chat line moderator. To analyze it using data mining methods a convenient technique is to split the stream into windows that may be overlapping if desired. Each such window can now be viewed as one document. We represent the documents using the vector space model [11]: each document forms one T-dimensional vector where T is the number of distinct terms in the vocabulary. The i-th element of the vector indicates (some function of) the frequency of the i-th vocabulary term in the document. The term by document matrix $\mathbf{X}$ contains the document vectors as its columns and is of size $T \times N$ where N is the number of documents.

As a preprocessing step we compute the LSI [2] of the data matrix $\mathbf{X}$, thus acquiring a lower dimensional projection of the the high-dimensional data. We denote the new data matrix as $\mathbf{Z} = (\mathbf{z}(t))$. The topics of the discussion can be found by projecting $\mathbf{Z}$ to the directions $\mathbf{W} = (\mathbf{w}_1 \cdots \mathbf{w}_M)$ given by the algorithm described in the following Section. M, the number of estimated minimum complexity projections, may be smaller than K, the dimension of the LSI projection.

3. THE ALGORITHM

Complexity pursuit [3] is a recently developed, computationally simple algorithm for separating interesting components from time series. The interestingness is measured as a low coding complexity of the distribution of the projection of the data. We model the topics as probability distributions on terms, and the distributions having minimum complexity are assumed to best represent the distinct topics.

We assume that the observations $\mathbf{z}(t)$ are linear mixtures of some latent topics $\mathbf{s}$. Both the latent topics and the mixing process are unknown. The topics are estimated by $s(t) = \mathbf{w}^T\mathbf{z}(t)$ where the projection directions $\mathbf{w}$ are to be found. A separate autoregressive (AR) model, here $\hat{s}(t) = \alpha s(t-1)$, is assumed to model each topic s. At every step of the algorithm, the AR constant α is first estimated. Then the gradient update of $\mathbf{w}$ that minimizes the approximate Kolmogoroff complexity [3] of the AR residuals $s(t) - \hat{s}(t)$ is the following:

$$\mathbf{w} \leftarrow \mathbf{w} - \mu E\{(\mathbf{z}(t) - \alpha\mathbf{z}(t-1)) \cdot \operatorname{sign}(\mathbf{w}^T(\mathbf{z}(t) - \alpha\mathbf{z}(t-1)))\} \quad (1)$$

$$\mathbf{w} \leftarrow \mathbf{w}/||\mathbf{w}|| \quad (2)$$

Details of the algorithm can be found in [3].

4. EXPERIMENTS ON CHAT LINE DATA

The chat line data was collected from the CNN Newsroom chat line[1]. A contiguous stream of almost 24 hours of discussion of 3200 chat participants, contributing 25 000 comment lines, was recorded on January 18th, 2001. The data was cleaned by omitting all user names and non-user generated text. The remaining text stream was split into

[1] `http://www.cnn.com/chat/channel/cnn_newsroom`

overlapping windows of about 750 characters. From these windows a term histogram was generated using the Bow toolkit[2], resulting in a term by document matrix $\mathbf{X}$ of size $T \times N$, that is, 4743×7430. The LSI of order $K = 50$ was computed as a preprocessing step. The choice of the number of estimated topics M is relatively flexible, and in this abstract we present results on $M = 7$.

Figure 1 shows how different topics are activated at different times. We can see that the topics are autocorrelated in time. Some topics show short peaks of contribution whereas others are active for longer periods.

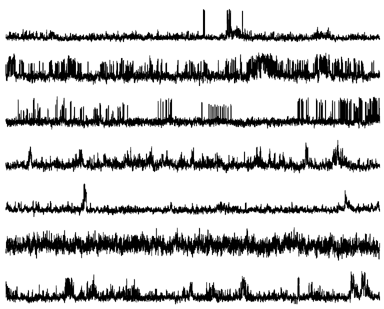

Figure 1: Activity of topics (vertical axis) in each chat window (horizontal axis).

The validity of the identified topics is easy to evaluate using the most representative terms associated with each topic. These are obtained by projecting the term data $\mathbf{Z}_{term}$ (which is the document by term matrix $\mathbf{X}^T$ projected into the LSI space) into the minimum complexity directions $\mathbf{w}_i$ found earlier. By listing the terms corresponding to the highest peaks in the projection $\mathbf{w}_i^T \mathbf{Z}_{term}$ we get a list of keywords for the i-th topic. In Table 1 it is seen that each keyword list indeed characterizes one distinct topic quite clearly. Topic 1 deals with the values of the politicians in the US, topic 2 is a religious discussion and topic 3 corresponds to comments given by the chat line moderator. Topic 4 involves the presidential election in the US and especially the vote recounting in Florida. Topic 5 deals with the problems of the youth: violence, drug abuse etc. In topic 6 the new US president Bush is discussed in general. Topic 7 is about the energy shortage in California in mid-January 2001. Some of the topics display similarity to other topics, whereas some topics are clearly distinct. Indeed, estimating e.g. 10 topics in the same data set has in our experiments brought out topics that resemble topics 1 and 6 to some degree. Also, estimating less than 7 topics gives the most clearly formed topics (such as topics 4, 5 and 7) similarly to what is seen here, in addition to a mixture of the other chat contributions.

We also run experiments with some more traditional methods such as LSI and ICA [5, 6, 9] and noticed that the presented method outperformed the other methods.

5. CONCLUSIONS AND FUTURE WORK

Minimum complexity projections of a dynamically evolving textual data identify some underlying topics of the data. As an example of such dynamical data we used chat

[2] `http://www.cs.cmu.edu/~mccallum/bow/`

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
conserv	god	www	won	violenc	peopl	power
polit	religion	http	vote	gun	kennedi	california
religion	thing	html	count	report	elect	electr
liber	jesu	index	gore	school	cnn	don
govern	bibl	cnn	bush	youth	dai	blackout
mind	doesn	time	stop	children	live	plant
free	white	world	hand	point	call	energi
opinion	don	thing	recount	home	bush	deregul
form	work	stori	court	drug	back	state
life	kill	good	florida	famili	man	problem
philosophi	talk	make	don	major	senat	crisi
establish	good	put	win	gener	presid	build
independ	time	januari	dade	parent	vote	compani

Table 1: The keywords of each topic

line discussions. In our experiments distinct and meaningful topics of the discussion were found, outperforming some more traditional methods. The results suggest that our method could serve in queries on a temporally changing text stream. Also, as some topics of discussion are more distinct than others, natural extensions of the work include recursive binary partitioning [10] of the data, or finding the topographic structure [4] of the topics.

6. REFERENCES

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, New York, 1999.

[2] S. Deerwester et al. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[3] A. Hyvärinen. Complexity pursuit: separating interesting components from time-series. *Neural Computation.* To appear.

[4] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation.* To appear.

[5] C. L. Isbell and P. Viola. Restucturing sparse high dimensional data for effective retrieval. In *Advances in Neural Information Processing Systems 11*, pages 480–486, 1998.

[6] A. Kabán and M. Girolami. Unsupervised topic separation and keyword identification in document collections: a projection approach. Tech. Rep. 10, Dept. of Computing and Information Systems, University of Paisley, August 2000.

[7] T. Kohonen et al. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, May 2000. Special Issue on Neural Networks for Data Mining and Knowledge Discovery.

[8] T. Kolenda and L. K. Hansen. Dynamical components of chat. Tech. rep., Technical University of Denmark, 2000. Available at `http://eivind.imm.dtu.dk/staff/thko/kolenda.hansen.tr2000.zip`.

[9] T. Kolenda, L. K. Hansen, and S. Sigurdsson. Independent components in text. In M. Girolami, editor, *Advances in Independent Component Analysis*, chapter 13, pages 235–256. Springer-Verlag, 2000.

[10] P. Pajunen and M. Girolami. Implementing decisions in binary decision trees using independent component analysis. In P. Pajunen and E. Oja, editors, *Proc. of ICA2000*, pages 477–481.

[11] G. Salton and M. McGill. *Introduction to modern information retrieval.* McGraw-Hill, New York, 1983.