Discovery in Sequential Text Mining

Haiyun Bian
Lawrence J. Mazlack
Computer Science
University of Cincinnati
Cincinnati OH
{bianh, mazlack}@uc.edu

Abstract

Text Mining has been a popular research topic since its introduction in the early 1990s. Several different kinds of patterns may be discovered in the corpus of unstructured textual information. Sequential data mining, targeted at discovering sequential patterns from transaction data, comes to exist at almost the same time.

This paper tries to identify the possibility to de sequential text mining. Both general and restricted applications for sequential text mining are described. This paper aims to evoke more interest both from researchers and practitioners in sequential text mining research area.

1. Introduction

Data mining is a relatively new area of computational discovery. The first workshop on data mining was held in 1993 and the first conference in 1994. Agrawal's paper [1993] is a good marker for the effective start of the area.

Data mining seeks to discover patterns in collections of transaction data. Initially, discovery focused on data that was static. Efforts at recognizing sequences began shortly thereafter. Perhaps the nest known early applications of sequence recognition dealt with telecommunication networks [Hatonen 1996].

Text mining is the process of finding interesting or useful patterns in a corpus of unstructured textual information [Dixon 1997]. Dixon [1997] describes the general text mining process a sequence of steps, including information retrieve, information extraction, information mining, and interpretation. This constitutes the general definition on text mining.

More recently, a narrower view on text mining was proposed, which limits the scope of text mining to those tasks which discover new information [Hearst 1999]. This information retrieve and information extraction do not belong to text mining process, and some information mining tasks such as text categorization are considered out of text mining process because they do not extract new information.

Sequential mining [Agrawal 1995] is defined as finding sequential patterns from ordered list of transactions. An example of such a pattern is that customers typically rent "Star Wars", then "Empire Strikes Back", and then "Return to Jedi". So when a customer who has already had the previous two will most probably rent the third one.

An interesting topic is trying to combine text mining and sequential learning, what we call in the following text as sequential text mining. This paper tries to identify the prospect and potential problems when doing this combination.

The following paper is organized as the following sections. First comes the summary on sequential mining. Secondly some problems when doing sequential text mining are introduced.

2. Sequential mining

This section provides a generalized view of the current state of the sequential analysis. Sequential learning is receiving more and more attention recently both in the neuroscience literature and in computational literature. Different sequential learning problems can be formulated into subcategories according to different criteria. These categorization can give us better view on what sequential mining can do.

2.1 Task-driven perspective

According the different tasks performed by the sequential analysis, [Sun 2000] formulized sequential problems into the following subtasks:

a. Sequence prediction/sequence generation
b. Sequence recognition
c. Sequence decision-making

2.2 Data-driven perspective

Categorical sequential domains, including text, DNA sequences, web usage data, multi-player games, and plan execution traces, forms the first category whose data is mainly symbolic or categorical data [Lesh 2000].

Another category here refers to those data which can be modeled as random variables, either discrete or continuous. In the following, we will call it non-categorical data.

Actually, this taxonomy is not crisp, since some data can be modeled in both ways.

2.3 Technique-driven perspective

Following the data-driven taxonomy, techniques dealing with categorical sequential data mainly include AprioriAll [Agrawal 1995], GSP [Srikant 1996], SPADE [Zaki 1998].

On the other hand, neural networks, hidden Markov models and reinforcement learning are adopted to handle non-categorical data domain [Sun 2000].

2.4 Application-driven perspective

Sequential learning has been proved to be very important domains ranging

from robotics [Sebastiani 1999] to biology [Frezza-Buet 1998] and market analysis [Agrawal 1995]. It has been also successfully applied in pattern recognition [Grossberg 2000], and planning and context based spelling correction [Lesh 2000].

One common characteristic of all the sequential analysis models is that they require that the data be marked with time stamp, which satisfy the precondition to build causality model on this data too. Since causality model can better solve the limitation of the extensibility of the associative sequential analysis [Glymour 2000], it is promising to build the linkage between the causality model and the sequential analysis methods. The most common studied Bayesian causality model will be used in this proposal [Pearl 2000].

3. Sequential text mining

Current text related sequence mining is done using techniques handling categorical data. So the following comments are proposed based on these subdirectory. However, it does not mean that text sequence mining cannot deploy techniques such as HMM and concurrent neural networks [Sun 2000].

According to the general definition for text mining [Dixon 1997], some successful applications on sequential text mining have been reported since 1998. [Allan 1998]. These applications mainly refer to topic detection and task tracking problem.

However, according to the recent definition of text mining [Hearst 1999], current research on text mining [Allan 1998] actually did not do the "mining" job. In other words, they do not discover previously unknown knowledge from the text data, since the knowledge discovered is not applied into prediction of future events. So a complete sequential text mining process can be roughly divided into two main stages.

- The first stage is to discover sequences embedded in the large volume of text data, either from newswire or from web.
- The second stage is to apply these sequences to aid the decision, for example, for prediction or monitoring purpose.

Current researches in text sequence mining mainly focus on the first stage. However, little is done on how to deploy the founded sequences to better further decision.

Lesh [2000] created a good start on the more complete sequential mining model. In his paper, sequential patterns that are found in the first stage are extended to do feature selection and rule pruning job, which is ordinary to data mining process, but new to sequential mining research.

Inspired from the work by Lesh [2000], future text sequential mining research can be carried on according to the general model of data mining process [Fayyad 1996]. The following discussion presents some simple ideas starting from this consideration.

Based on the characteristics of text data, one concern to text sequential mining is to how to handle large volume of data. Scalablity is the very important criteria when choosing text sequence generation algorithms. Another concern is on the

feature selection step. If the mining is done meta-data, then how to select the best summarization for the data should be very important.

We have to decide which technique to using in prediction the sequences. Lesh [2000] used Naive Bayes to do the prediction, but he suggested no criteria about which technique is better for sequential text prediction, considering the characteristics of both text data and sequence data. Another concern is the uncertainty handling method, which has never been mentioned in the previous sequential text mining research. For instance, since Fuzzy sets theory has been proved great success in many data mining areas, it might be promising if introduce it into the sequential text mining model.

We cab see that there are still a lot of researches that are needed to build a whole model for sequential text mining. This paper aims at evoking the interests of both the researchers and practitioners, so that more effort could be devoted to this research area.

4. Conclusions

In this paper, we provide a summary of the current research in sequential mining, and we try to extend the sequential mining techniques into unstructured text data. Some preliminary research in sequential text mining is also introduced. Potential problems in doing sequential text mining are also proposed. In summary, sequential text mining is an attractive are that deserves more research effort.

Reference

[Allan 1998] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and T. Yang, "Topic detection and tracking pilot study final report," *Proceeding of DARPA Broadcast News Transcription and Understanding Workshop*, 1998

[Agrawal 1993] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," *Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD-93),* P. Buneman, S. Jajodia (eds.), Washington, DC, May 1993, 207-216

[Agrawal 1995] R. Agrawal, and R. Srikant, "Mining sequential patterns," *Proceedings of 11th International Conference on Data Engineering*, 1995

[Dixon 1997] M. Dixon, "An overview of document mining technology", *http://citeseer.nj.nec.com/dixon97overview.html*

[Fayyad 1996] Fayyad U, Shapiro G. P, and Smyth P, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, 39(11), November 1996, 27-34

[Frezza-Buet 1998] H. Frezza-Buet, and F. Alexandre, "Multimodal sequential learning with a cortically-inspired model, in *JCIS98 Proceedings*, volume 2, 1998, 24-27

[Grossbery 2000] S. Grossberg, and R. Paine, "A neural model of corticocerebellar interactions during attentive crimination and predictive learning of sequential handwriting movements," *Neural Networks*, 2000

[Hatonen 1996] K. Hatonen, M. Klemettinen, H. Mannila, P. Ronkainen, and H. Toivonen, "Knowledge discovery from telecommunication network alarm databases," *Proceedings of Conference on Data Engineering*, New Orleans, 1996, 115-122

[Hearst 1999] M. A. Hearst, "Untangling text data mining," *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, June 20-26, 1999

[Lesh 2000] N. Lesh, M. J. Zaki, and M. Ogihara, "Scalable feature mining for sequential data," *IEEE Transactions on Intelligent Systems and Their Applications, special issue on Data Minin*g, 15(2), 2000, 48-56

[Sebastiani 1999] P. Sebastiani, M. Ramoni, P, Cohen, J. Warwick, and J. Davis, "Discovering dynamics using Bayesian clustering," *Proceedings of the 3rd International Symposium on Intelligent Data Analysis*, Springer, New York, 1999, 199-209

[Srikant 1996] R. Srikant and R. Agrawal, "Mining sequential patterns: generalization and performance improvements," *Proceedings of the 5th International Conference on Extending Database Technology*, 1996

[Sun 2000] R. Sun, and C. L. Giles, **Sequential Learning: Paradigms, Algorithms, and Applications**, Springer, 2000

[Zadeh 1965] L. A. Zadeh, "Fuzzy sets", *Information and Control*, vol. 8, 1965, 338-353

[Zaki 1998] M. J. Zaki, N. Lesh, and M. Ogihara, "PLANMINE: sequential mining for plan failures," *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining*, 1998