MINING DATA STREAMS APPROACH FOR TEXT-TREND ANALYSIS

Michal Šebek
Doctoral Degree Programme (1), FIT BUT
E-mail: xsebek00@stud.fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@fit.vutbr.cz

ABSTRACT

This paper deals with a trend recognition in textural data. We summarize approaches to mining in data streams as an approach for time-varying analysis and methods for document corpora description. This paper also introduces a system for recognition of phrase time-trend in research papers and discusses recognition results.

1 INTRODUCTION

In science it is very important to know actual trends and ways of research. An actual focus of researchers can be found in conference papers. Human readers know how to find necessary trend information intuitively. But it is difficult to find required information about the actual trend manually because of the high number of conferences.

We suppose that scientific trends are reflected in topics of research papers. Research papers can be viewed as a time-ordered stream of documents (e.g. order by the publish date). In the stream time-trends can be recognized. The problem of research papers analysis was mentioned in [5] where the method based on citation records of papers was presented. These records formed a graph and clustering algorithm was applied on this graph. Then clusters of related papers were composed. Temporal trends were recognized by regression analysis of a frequency of citations per a cluster.

In this paper we propose a system for trend analysis in research papers. We combine classical text analysis methods with some basic approaches from stream data mining for the text trend analysis.

The organization of this paper is following. In section 2 basic concepts of stream data mining and document analysis is described. Then in section 3 we introduce ideas and th design of our trend analysis system. The experimental results are presented in section 4. A possible future work is described in section 5 and conclusions are presented in section 6.

2 CONCEPTS

This section introduces basic concepts of data mining in data streams and discusses methods for a document corpora analysis.

2.1 DATA STREAMS

Unlike traditional data-sets stored in databases, data streams have some special properties. What is significant for a data stream? The **data stream** is defined as potentially infinite, continuously arriving data set. This means that the data stream is mostly massive, very fast and cannot be stored as whole. Also time-order on the data stream can be naturally defined because data arrive in different times. We can imagine data streams as e.g. data continuously generated by sensor network, thermometer or some video data. In section 3 we will discuss a specific type of stream based on the document set. Data streams are discussed in [3].

2.2 MINING DATA STREAMS APPROACHES

Due to properties of the data stream, common data mining algorithms cannot be used for analysis of these streams. Data mining stream algorithms have to be *faster* (we cannot lose data), *single-pass* (whole data stream cannot be saved). Typical data mining algorithms use at least $O(N)$ space, but we require at most $O(log^k N)$ where N is a number of data elements and $k \in \mathbb{N}$ is a constant. Also algorithms must deal with the changing concept of data called *concept drift*. This means that information in the stream can change during the time. E.g. imagine the difference between mining video data in the daylight and in the night.

Current algorithms and approaches for data mining in data streams are described in [2]. Data mining approaches are divided into *Data-based techniques* and *Task-based techniques*.

Data-based techniques use only the subset of the data stream or create summarizing information of the dataset. Some techniques of this category are following:

- **Sampling** – is natural technique based on selecting the subset of elements with a specified probability. The main problem of this method is that it cannot check anomalies because an incoming anomaly element can be skipped.
- **Sketching** – method evaluates frequency moments of possible stream elements. When the available memory is less than required memory, algorithm creates *synopses* (summary values) of frequency moments known as *sketches*.

Task-based techniques mean modification of current algorithms or new algorithms for mining in data streams.

- **Approximation algorithms** – are a group of new algorithms which construct the data mining model with user-defined bounded error. These algorithms use statistical theorems such as Chebyshev's inequality, Chernoff bound or Hoeffding bound to guarantee a maximal devation of the result.
- **Sliding Window** – this method uses a set of most recent elements of the data stream. If we have a sliding window of size w, it contains elements since time $(t-w)$ until t, where t is an actual timestamp. A history can be stored by summarizing old data.

2.3 TOPIC MODELING OF DOCUMENT CORPORA

In this subsection we introduce methods of a text corpora modeling. The main goal is mostly to find important words describing some document or topic. *Information Retrieval* techniques are used for this. Well-known technique is *tf-idf* which evaluates importance of the word for each document of a document corpora.

In this paper we will focus on another technique called *Latent Dirichlet Allocation* (LDA) proposed in [1]. LDA is a generative document model. This model supposes:

- each document is based on one or more topics,
- a probability distribution of possible words differs for each topic.

Formally, let's have a document $\mathbf{w}$ of N words $\mathbf{w} = \langle w_1, \ldots, w_N \rangle$ and a set of k topics. The generative process is parameterized by parameters $\alpha = \langle \alpha_1, \ldots, \alpha_k \rangle$ and a matrix $\beta = k \times |V|$ which determines probability dependence between the set of all possible words V and k topics. The generative process of the document is on Fig. 1. First the distribution of topics in the document θ is evaluated as a *Dirichlet distribution* parametrized by α. Then the topic of each word z_n is $Multinomial(\theta)$ distribution. Finally the next word w_n is generated as $p(w_n|z_n, \beta)$. The procedure is repeated for each word in the document $\mathbf{w}$ of document corpora M.

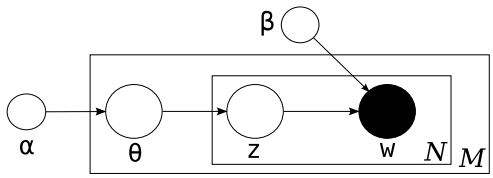

Figure 1: LDA generative model of the document corpora (adapted from [1]).

3 DESIGN OF THE TREND ANALYSIS SYSTEM

In this section a system for discovering trends in research text papers is proposed. The system is divided into two main parts. The first is a text classification, and second one is a statistical trend analysis. A scheme of the system is on Fig. 2.

The first step is **the classification of an incoming document**. For the classification the LDA method is used. The document isn't classified by LDA directly, but a set of keywords for tracked topics is given. The document is classified by keywords of the topic in the document. It requires a training phase when sets of keywords are extracted for topics. When the longer time-period of conference is tracked, the change of keywords of topics is possible. This problem is similar to the problem of the concept drift (mentioned in section 2.2). It is expectable that the change of keyword set is slow so the method of sliding window can be used when solving this problem. Initially the set of keywords is trained with the window of size s, and then the window moves. Keywords are recognized again by LDA in the new document window, and previous keywords of the topic are replaced.

The second step is a **statistical trend analysis**. Initially the user specifies an ontology of tracked phrases. The ontology is necessary when the fast detection of phrase trend-change is required. Theoretically the ontology can be generated automatically, but the dramatically larger number

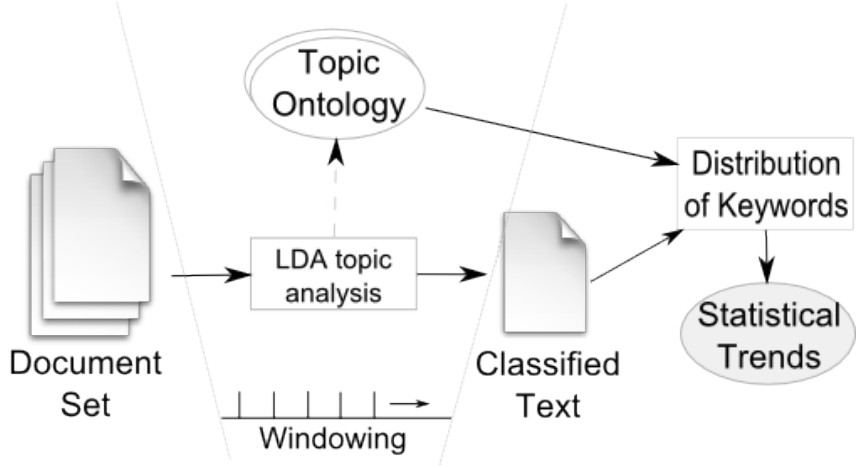

Figure 2: The trend-analysis system based on LDA.

of analyzed papers is required. The trend analysis is based on a frequency counting of the specified ontology phrases and a statistical test for a detecting significant changes. We expect that frequency of keywords have an uniform distribution over the time when there is no trend. When a trend appears, uniform distribution is broken. *Kolmogorov-Smirnov test* can be used for discovery of changes in the uniform distribution.

4 EXPERIMENTAL RESULTS

For experiments we used an available implementation of the LDA method in Mallet framework [4] and ACM SIGKDD Conference papers in odd years from 2001 to 2009 as the document corpora (50 randomly selected papers per a year). The topic of these papers is the Knowledge Discovery in Databases (KDD). Because of the document classification test, reserch papers about grammar systems was added into the set of documents.

The Table 1 shows **top-rated keywords** extracted from the set of papers. LDA was applied in the sliding window mode, so there are simmilar keyword sets of two-year periods for KDD Topic in first columns. Because differences between year periods are small, concept of sliding window with initial training and online modifications can be used. In addition there are keywords about grammar systems in the last column of the table.

Important trend states in the ontology of KDD are in the Table 2. Numbers in rows are **phrases frequencies** in the document corpora per year. A tested hypothesis is that frequency counts have the uniform distribution. The hypothesis was tested by Kolmogorov-Smirnov test on a significance level $\alpha = 0.95$. There are the result of test in the last column of the table. "No" sign in the row means that change in the trend occurred with the probability α (frequency distribution of the phrase has not the uniform distribution). E.g. phrase *SVM* has the increasing trend between years 2001 and 2009.

5 FUTURE WORK

A weak point of the method is a necessity of the ontology. The ontology is actually static and must be defined by the user initially. A next work is to modify the ontology phrases automati-

#	KDD Topic		Grammar
	2001-05	2005-09	Topic
1	data	data	systems
2	time	minus	grammar
3	set	algorithm	language
4	algorithm	model	type
5	number	set	context
6	clustering	time	free

Table 1: Top rated keywords of topics.

Phrase	Year					Unif.
	'01	'03	'05	'07	'09	
neural network	7	7	5	11	3	**No**
decision tree	8	8	6	10	4	Yes
SVM	2	3	9	12	11	**No**
boosting	5	5	6	5	5	Yes
HMM	0	1	4	2	7	**No**
stream	8	14	9	15	14	**No**

Table 2: Frequency counts of phrases per years

cally or semi-automatically. It should allow to recognize trends in the actual research without specifying the ontology.

This method is completely based on frequency counts and statistics. Another way how to recognize trends is to focus on a semantic information in documents, e.g. to find semantic relations between terms and to try to analyze documents more precisely.

6 CONCLUSION

In this paper, we presented an approach for the trend analysis of research papers. We combine the method for the data stream analysis, the method for the document corpora description and the statistic test. In this work we used innovative approach based on the topic modeling instead of *rf-idf scores* or citation records. Experimental results show that the presented statistical approach can recognize important changes of the real research focus.

ACKNOWLEDGEMENT

This work was partially supported by the BUT FIT grant FIT-S-10-2 and the research plan MSM0021630528.

REFERENCES

[1] Blei, D. M.; Ng, A. Y.; Jordan, M. I.: Latent dirichlet allocation. *The Journal of Machine Learning Research*, Vol. 3, 2003: p. 993–1022, ISSN 1532-4435.

[2] Gaber, M. M.; Zaslavsky, A.; Krishnaswamy, S.: Mining data streams: A review. *SIGMOD Rec.*, Vol 34, No. 2, 2005: p. 18–26, ISSN 0163-5808.

[3] Han, J.; Kamber, M.: *Data Mining: Concepts and Techniques*. Elsevier Inc., second edition, 2006, ISBN 978-1-55860-901-3.

[4] McCallum, A. K.: MALLET: A Machine Learning for Language Toolkit, 2002, URL http://mallet.cs.umass.edu.

[5] Popescul, A.; Ungar, L. H.; Flake, G. W.; et al.: Clustering and Identifying Temporal Trends in Document Databases. *Advances in Digital Libraries Conference, IEEE*, Vol. 0, 2000: p. 173-183.