A Survey of Advances in Topic Detection and Tracking: Postprint

Authors: Zhang Jin, Yang Sen, Wang Xiaozong, Luo Wei, Du Pan, Cheng Xueqi

Date: 2017-03-09T00:00:00+00:00

Abstract

With the exponential growth of Internet information, technologies such as information retrieval and topic detection have received widespread attention in recent years to improve the efficiency of information mining. This paper first reviews the developmental history of topic detection and tracking technologies; building upon the introduction of traditional topic detection methods, it conducts an in-depth exploration from two aspects: burst detection and social network-based topic detection and tracking methods; analyzes evaluation methodologies for topic detection and tracking; and finally, discusses the future development trends of topic detection and tracking methods.

Full Text

Preamble

Vol. 8 No. 4
Information Technology Letters

A Survey of Research Advances in Topic Detection and Tracking
Zhang Jin, Yang Sen, Wang Xiaozong, Luo Wei, Du Pan, Cheng Xueqi

Abstract

With the exponential growth of Internet information, technologies such as information retrieval and topic detection have garnered widespread attention in recent years to improve the efficiency of information mining. This paper first reviews the history of topic detection and tracking technology development. Building upon an introduction to traditional topic detection methods, we delve into two key aspects: burst detection and social network-based topic detection and tracking methods. We analyze evaluation methodologies for topic detection and

tracking, and conclude with a prospect for future development trends in this field.

Keywords: Topic detection and tracking, burst detection, social networks

1 Introduction

With the vigorous development and widespread adoption of Internet technology, the volume of online information has grown exponentially, creating an increasingly prominent contradiction between information overload and knowledge scarcity. The vast ocean of web data far exceeds human capacity to manage. Consequently, how to effectively organize and present web data to improve knowledge acquisition efficiency has long been a hot research topic. Topic detection and tracking technologies can help organize information by subject, intelligently push the most active topics to users within specific time periods, and track the dynamic evolution of topics according to user needs, thereby greatly facilitating users' ability to grasp social trends and major events. Applications focused on hot topics and sudden topics have received particularly extensive attention.

Meanwhile, with the application and development of Web 2.0, social networks have become increasingly pervasive. Unlike traditional news media, social networks emphasize user participation. If we can effectively detect and track topics automatically on social networks, it would undoubtedly help users find and comprehensively understand events or topics of interest. However, since social network data is primarily generated by ordinary users, the quality of this data—whether in terms of word choice, format, or specific content—is uneven, posing significant challenges for topic retrieval. Notably, widespread user participation also provides new data information that can be leveraged for topic detection and tracking. Topic detection on social networks is no longer limited to textual information; non-textual information can also be utilized. These new characteristics have made research on social network-oriented topic detection and tracking methods a key focus in recent years.

In this paper, we first review the history of topic detection and tracking. Building upon an introduction to traditional topic detection methods and combining our research findings, we explore in depth two aspects: burst detection and social network-based topic detection and tracking methods. We analyze current evaluation methods for topic detection and tracking, and conclude with a prospect for future development trends.

2 Research Status

Topic detection and tracking research has been conducted for over a decade. In existing studies, a topic is defined as an event or activity and all related events or activities, while an event is defined as something unique that occurs at a specific time or place [**?**]. In previous research, the distinction between events

and topics was minimal and they were often interchangeable. Topic detection can be divided into two relatively independent subtasks: retrospective topic detection and online topic detection. Retrospective topic detection refers to detecting all implicit topics within a dataset after all detection data is known. Online topic detection refers to situations where detection data is only partially known and new data continuously arrives in an online fashion, requiring the system to immediately judge whether newly arrived documents represent new topics or belong to existing historical topics. The topic tracking task involves, for a pre-specified topic (presented in some form), determining in an online data output mode whether current documents belong to that specified topic before new data arrives.

In Topic Detection and Tracking (TDT) evaluations, the corpus used for topic detection and tracking consists of news data, including news texts and transcribed materials, typically arranged chronologically with target events manually annotated. In these evaluations, topic detection and tracking research [**?**, **?**, **?**, **?**, **?**] is further divided into three subtasks: story segmentation, event detection, and event tracking. Story segmentation is defined as partitioning continuous text streams based on report content to correctly identify boundaries between adjacent reports. Event detection can be further divided into Retrospective Event Detection (RED) and Online New Event Detection (NED) [**?**]. RED aims to find all implicit events within a given collection of reports, essentially clustering the target dataset where each resulting cluster represents an event. NED aims to identify new events in a stream of reports in an online fashion. When a new report arrives, NED methods must analyze it and determine before the next report arrives whether it discusses a new event. Event tracking involves finding all reports related to a known event among newly arrived reports.

Since our research focuses primarily on topic detection and tracking, we will mainly analyze existing event detection and tracking methods while ignoring story segmentation research. Topic detection and tracking research can be methodologically divided into two categories. The first seeks new clustering algorithms suitable for topic detection and tracking or modifies existing clustering algorithms. The second focuses on mining new topic features to improve detection and tracking effectiveness. Notably, in some studies, such as [**?**], this distinction is not always clear. For simplicity, we will not provide strict categorization for each method.

3 Main Methods

The primary task of topic detection and tracking systems is to accurately detect topics and track their dynamic evolution, with the most critical issue being how to perform topic detection. James Allan et al. [**?**] divided topic detection into two branches: retrospective topic detection, which reorganizes documents in a corpus by topic and is essentially an unsupervised classification problem that groups documents discussing the same topic together; and online new event detection, which processes incremental online document streams sequentially

to determine whether each document belongs to an already labeled topic or discusses a new topic. The main difference between online and retrospective topic mining is that online topic mining faces incremental document streams, while retrospective topic mining deals with the entire document corpus.

3.1 Topic Representation and Similarity Measurement

"Representation" refers to abstracting documents and topics into models that computers can compute and compare. Similarity measurement includes calculating similarity between documents, between documents and topics, and between topics. These two issues are highly related, as each representation model corresponds to one or more similarity calculation methods.

Commonly used topic representation models are primarily the Vector Space Model, Probability Retrieval Model, and Language Model.

Vector Space Model (VSM): This model represents both documents and topics as vectors, where each dimension represents a word. The entire dictionary constitutes all dimensions of the space, and each document and topic becomes a vector (point) in this space.

A natural similarity measure corresponding to VSM is cosine similarity: $(cid:71)(cid:71)(cid:71)(cid:71)$

Probability Retrieval Model: This model also represents documents and topics as word sets, treating similarity as a probability value—the probability that a document is relevant to a given query. A commonly used similarity measure corresponding to this model is the BM25 formula: $kD, qfavgdl$ where $1k$ and b are free parameters, iq is a word in query Q, $(D, qf$ represents the term frequency of word iq in document D, D represents the length of document D, and $avgdl$ represents the average document length in the corpus.

Language Model: Statistical language models consider language as a probability distribution over an alphabet, specifically forming a distribution of words in a document D, called a language model. In language models, each document is treated as a language model, and the entire corpus is also a language model. In language models, calculating $\cdots = 1$ The relevance between query and document is defined as the probability of generating one language model from another. The Kullback-Leibler distance is used, calculated as: $(D, QPKLQ, D\ Pwlog$ Due to the sparsity of language models, zero-probability words may occur, so the problem of smoothing zero-probability words must be addressed.

3.2 Topic Detection Methods

Research on topic detection and tracking can be divided into two categories: the first aims to find new clustering algorithms suitable for topic detection and tracking or modify existing clustering algorithms; the second focuses on mining new topic features to improve detection and tracking effectiveness. This section

briefly describes existing topic detection and tracking methods from these two categories.

3.2.1 Improved Clustering Algorithms From the definition of topic detection, it bears considerable similarity to clustering algorithm research. Therefore, researchers have sought clustering algorithms more suitable for topic detection and tracking.

In 1998, Yiming Yang et al. proposed a retrospective event detection method based on Group Average Clustering (GAC) [**?**]. GAC is an agglomerative clustering algorithm that aims to maximize the average similarity of text pairs within resulting clusters. Building on GAC, Yang et al. proposed a split-and-recluster method. This method makes full use of event clustering characteristics by splitting data—reports on the same event tend to cluster in a relatively small time region—and minimizes the impact of initial partition boundaries on clustering results during reclustering.

James Allan et al. proposed a multiple K-Means method for retrospective event detection based on the K-Means clustering algorithm [**?**]. The basic idea of multiple K-Means is: at each given time point, with k known clusters, for each report, find its nearest cluster. If the distance is below a certain threshold, assign the report to that cluster; if the distance exceeds the threshold, create a new cluster from that document, then perform conventional K-Means clustering.

Additionally, Li Zhiwei et al. proposed a probabilistic model-based method for retrospective news event detection in 2005 [**?**]. This method uses a probabilistic generative model combining content and temporal information for retrospective event detection. Each known event is represented by a probabilistic generative model. For each document, the event represented by the generative model with the highest probability of generating that document is identified, and the document is assigned to that event. This method also considers that reports on the same event often appear across multiple news sources, so it incorporates reports from different data sources at similar times to help with event detection.

Reference [**?**] proposed a multi-strategy optimized divide-and-conquer clustering algorithm. This algorithm first divides all data into groups with certain similarities, then clusters each group separately to obtain "micro-clusters" within each group. On this basis, it clusters all micro-clusters to obtain final topic results. During clustering, the method employs multiple optimization strategies to improve clustering effectiveness.

It should be noted that although the methods introduced above were initially proposed for either retrospective event detection or online topic detection, retrospective event detection can be decomposed into online topic detection to some extent. Therefore, most online topic detection methods can be applied to retrospective event detection tasks. In the next section, we will no longer distinguish which specific detection task each method targets.

3.2.2 Mining Topic Features Another research approach is to mine inherent topic features to improve topic detection and tracking effectiveness. Topic features include temporal clustering characteristics, topic-specific words, topic life evolution features, and topic named entities.

A widely adopted approach uses various topic features to find appropriate ways to control topic thresholds during detection and tracking, aiming to develop threshold-setting methods with relatively broad applicability. Research in this area includes the time penalty strategy proposed by Allan et al. in [**?**], the incremental Probabilistic Latent Semantic Indexing (PLSI) online event detection algorithm proposed by Tzu-Chuan Chou et al. in [**?**], and the event life feature identification method based on Hidden Markov Models proposed by Chien Chin Chen et al. in [**?**].

Allan et al. primarily used the Single Pass clustering algorithm with a new threshold control model for online new event detection. The basic idea of this threshold control strategy is: two reports far apart in time must have greater similarity to be classified as the same event, while documents close in time require less similarity to be grouped into the same event. The incremental PLSI model implemented by Chou et al. aims to expand the effective range of detection threshold settings. Compared to vector space model-based text and topic representation, probabilistic latent semantic indexing can represent topics more effectively and thus accommodate a wider threshold range. Chen et al.' s life feature identification method posits that event development follows specific patterns—emergence, development, growth, and decline—so HMMs can be trained on known event evolution patterns, then used to predict behavior patterns for new events. By assigning different detection thresholds to different event evolution stages, dynamic topic threshold strategies can improve detection effectiveness of existing methods. Reference [**?**] in 2004 proposed a method for classifying feature words in text into categories such as location, name, time, and general feature words, then comparing text content within each category. Reference [**?**] in 2004 proposed improving new event detection effectiveness through text classification and named entities. This paper classifies texts, assigns different similarity thresholds to different categories, and uses multiple text representation methods—representing a text as three parts: representation by all feature words, representation by named entities, and representation by non-named entities—to improve text content similarity calculation.

In Chinese research, reference [**?**] first classifies text features, dividing all word features into person names, place names, and topic information, assigning different similarity comparison coefficients to each category. On this basis, the weight of each feature word is defined as the product of its term frequency and its category' s similarity comparison coefficient. By assigning different weight calculation coefficients to different categories of feature words, this method strengthens the weight of specific categories in text similarity calculation, thereby improving topic detection accuracy. Additionally, reference [**?**] proposed a method for calculating named entity similarity by constructing a geographic tree. Since

different place names in geographic expressions may share some degree of similarity, introducing a predefined geographic tree can effectively solve the problem of shared similarity between different place names. However, this method is limited to place name comparison and has limited applicability for other parts of speech.

3.3 Sudden Event Detection

In recent years, research on sudden events has attracted increasing attention. Burst features refer to certain features closely related to an event—such as documents or words—that exhibit abnormal explosive growth accompanying the event' s occurrence. Sudden events are events with burst features. Current research on sudden events mainly focuses on finding all burst words related to an event from datasets, then combining these burst words to form features describing the sudden event. The goal of such research differs somewhat from event detection in traditional topic detection and tracking. Sudden event research aims to identify a sudden event through a set of burst words, whereas in topic detection and tracking, events are represented by document sets. Sudden event research is no longer limited to news data but also includes query logs, emails, blogs, and other corpora. Similarly, sudden event detection is also divided into two categories: retrospective sudden event detection and online sudden event detection.

In 2002, reference [**?**] proposed a simple yet powerful automaton model for burst detection in text streams. This automaton model simulates the states of feature words and transitions between states through an automaton. Different states represent different occurrence frequencies of words, while transitions between states represent the emergence or disappearance of bursts. By assigning penalties to state transitions, the automaton model can effectively prevent false detection of excessive non-burst word frequency changes. The authors applied the automaton model to email collections and news collections, demonstrating its effectiveness for burst detection.

Reference [**?**] proposed a parameter-free sudden event detection method. Unlike the automaton model, this method can automatically detect sudden events in text streams without requiring users to specify any parameters. Similarly, this method also aims to find burst word sets, with each burst word set representing a sudden event. Specifically, a binomial distribution is used to represent the likelihood of a word appearing in text. Through this distribution, a sudden increase in frequency of a word with very low probability is considered a burst of that word. Experimental analysis on news data demonstrated the method' s effectiveness.

Reference [**?**] proposed a method that uses Discrete Fourier Transform to decompose time-series signals into a series of sine and cosine signals, analyzing the behavior of signals with maximum energy values (Fourier coefficients) to identify non-periodic and periodic burst words. This method can also distin-

guish between weak burst words and strong burst words. It handles periodic sudden event identification well and can effectively identify weak sudden events.

Reference [**?**] transformed the automaton model to obtain an online sudden event detection method. This method can effectively detect sudden events online in query logs. Since dynamic programming is used to solve for the current optimal state, only a small cost is required at each moment in memory to maintain the state values from the previous time point.

Reference [**?**] further applied sudden event detection ideas to identifying burst user groups in blogs.

In summary, sudden event detection research contains two basic steps: burst word identification and burst word merging. Burst word identification aims to detect all feature words with burst characteristics in datasets. Burst word merging uses these burst feature words to construct final sudden event features. The process of sudden event detection is shown in Figure 1 [Figure 1: see original paper].

3.4 Topic Detection and Tracking Based on Social Networks

Unlike traditional news media, social networks emphasize user participation. Moreover, since social networks provide users with convenient information exchange platforms, various specific forms of online media have developed significantly, such as blogs, online forums, social networking sites, video sharing websites, and the recently emerged microblogs. If we can effectively detect and track topics automatically on social networks, it would undoubtedly help users find and comprehensively understand events or topics of interest.

Previous topic research, particularly topic detection and tracking, focused on news data. News data is formally rigorous, uses precise wording, and has concrete content—all of which differ greatly from social network data characteristics. Since social network data is primarily generated by ordinary users, the quality of this data—whether in word choice, format, or specific content—is not guaranteed. Additionally, widespread user participation provides new data information that can be leveraged for topic detection and tracking. In other words, topic detection on social networks is not limited to textual information; non-textual information can also be utilized. These two characteristics necessitate finding new methods more suitable for social network topic detection and tracking.

Although research on social network topic detection and tracking is valuable, obtaining effective algorithms is not easy due to uneven data quality. Moreover, different social network forms significantly impact topic detection and tracking methods. Related research covers almost all forms of social network data, including query logs, blogs, online forums, and video sharing platforms. As various new social network application platforms emerge, topic detection and tracking methods require continuous improvement.

In 2008, MingLiang Zhu et al. provided a method for detecting and tracking top-

ics in Threaded Discussion Communities [**?**]. This research focused on designing effective methods to eliminate potential noise effects and improving thread similarity calculation by introducing user similarity.

Lu Liu et al. studied video topic detection methods [**?**], forming a bipartite graph through videos and annotation words, then performing topic detection and tracking on this bipartite graph using a Co-clustering algorithm [**?**]. Experimental analysis on YouTube showed that this method can effectively detect and track topics on video web pages. In 2007, Nilesh Bansal et al. detected events on text data streams through user query analysis. The basic idea is: first, identify bursty query words from query logs; then, construct events using the query results of these burst words. Reference [**?**] detected events on User Generated Content (UGC) streams through user query analysis. The basic idea is: first, identify bursty query words from query logs; then, construct events using the query results of these burst words. This paper conducted experiments using query logs and blog data and achieved good event detection results.

Candidate Event Generation
Burst Word Set
Local Event Set
Local Event Set
d3 d4 NL_SW

Figure 2 [Figure 2: see original paper]. Sudden Event Detection Framework with Noise Filtering

Recently, we have conducted in-depth research on topic detection in forums. Reference [**?**] proposed a method for detecting bursty topics in forums, first using burst characteristics to filter bursty feature words and users, then constructing bursty topics through burst word combinations, and further verifying detected bursty topics through bursty user groups. Additionally, we noted that post quality in forums is difficult to guarantee: on one hand, the quality of user-generated text content itself is uneven; on the other hand, forums contain large amounts of non-event text. In actual topic detection, this noise significantly impacts results. To filter such forum noise, we combined text content similarity and event burstiness for topic detection in forums [**?**], and proposed a sudden event detection framework based on noisy data, as shown in Figure 2 [Figure 2: see original paper].

We manually annotated collected Tencent forum data and used this dataset to evaluate our method. Experiments showed that our proposed sudden event detection method with noise filtering can effectively improve detection performance of existing methods on noisy data.

3.5 Evaluation Metrics for Topic Detection and Tracking

Topic detection and tracking evaluations generally employ multiple assessment criteria, including: Precision (p), Recall (r), F_1 value, False Alarm Rate ($False$),

Miss Rate ($Miss$), Normalized Detection Cost ($detC$), and corresponding macro-averages and micro-averages [**?**, **?**].

According to existing topic detection and tracking evaluation methods, the evaluation of topic detection algorithms works as follows: for any number of result topics detected by an algorithm, evaluation focuses only on several topics preselected and manually annotated. For annotated topics, we find the result topic with the largest common document set with the standard topic as the corresponding detection result for that annotated topic [**?**]. The common document set refers to the shared document set between an evaluation topic and a result topic. This evaluation method is suitable for retrospective topic detection performance evaluation and can also be applied to online topic detection performance evaluation.

Specific evaluation metrics are obtained based on a contingency matrix, where each entry represents the number of documents satisfying that requirement, as shown in Table 1 .

Table 1 . Contingency Matrix for Topic Detection Results

	In Result Topic	Not in Result Topic
In Annotated Topic		
Not in Standard Topic		

Based on this contingency matrix, Precision, Recall, F_1 value, False Alarm Rate, and Miss Rate are defined as: $(cid:122)(cid:122)(cid:122)(cid:122)(cid:122)$. Precision is the proportion of documents in detected result events that truly belong to that event. Recall is the ratio of documents in detected events to documents in standard events. Since Precision and Recall often have an inverse relationship—high Precision usually sacrifices Recall, and high Recall usually sacrifices Precision—using only one metric may lead to incorrect evaluation conclusions. A better approach is to consider both metrics together, commonly using the F_1 value.

The False Alarm Rate is the proportion of documents in detected result events that do not belong to annotated topics among all documents not belonging to annotated topics. The Miss Rate is the proportion of documents in annotated topics that fail to be detected. Similar to the relationship between Precision and Recall, False Alarm Rate and Miss Rate also have an inverse relationship. Therefore, a better metric that combines these two evaluation indicators is needed: detection cost.

Detection cost combines False Alarm Rate and Miss Rate, defined as: $falsefalse$ where $missP$ and $falseP$ represent conditional probabilities of miss and false alarm respectively, and is a prior probability. Smaller $detC$ indicates better algorithm detection performance. However, since $detC$ definition depends on

prior probability, to better represent detection algorithm performance, the normalized value of $detC$ is more commonly used in topic detection, defined as [**?**]: $detC)falsedetC)$ value does not exceed 1. Similarly, is set to 0.02. In our experiments, smaller $detC)$ indicates better detection algorithm performance.

Since topic detection evaluation generally uses more than one evaluation event, to describe the comprehensive detection effect of a detection algorithm on each evaluation event, we need to use macro-average and micro-average. Macro-average refers to weighted averaging of evaluation metrics directly across evaluation events, while micro-average refers to first summing the contingency matrices of evaluation events, then calculating overall evaluation metrics on the total contingency matrix [**?**].

4 Conclusion

Although topic detection and tracking research has been conducted for many years, current research primarily focuses on news data due to difficulties arising from diverse Internet data sources and uncertain feature extraction. Research on social network topic detection remains relatively limited. With the rise of social networks, especially the widespread application of forums and microblogs, topic detection for specific needs such as sudden events and emerging applications oriented toward social network data has become increasingly important. We believe that with the integrated application of feature selection methods for social network data and methods mining correlations between user behavior and text content, research and application of topic detection and tracking technology will see further development.

References

Y. Yang, T. Pierce, and J. Carbonell, A study on retrospective and online event detection. In proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval, pages 28-36, 1998.

Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma, A probabilistic model for retrospective news event detection. In proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, pages 106-113, 2005.

Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. Simple Semantics in Topic Detection and Tracking. Kluwer Academic Publishers, pages 347-368, 2004.

Jame Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study: Final Report. In proceedings of DARPA broadcast news transcription and understanding workshop, pages 194-218, 1998.

Ricardo A. Baeza-Yates, Berthier Ribeiro-Neto. Modern information retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1999.

Tzu-Chuan Chou, Meng Chang Chen. Using Incremental PLSI for Threshold-Resilient Online Event Analysis, IEEE Trans. Know. Data Eng. 20, 3, pages 289-299, 2008.

Chien Chin Chen, Meng Chang Chen, Mimg-syan Chen. An Adaptive Threshold Framework for Event Detection Using HMM-Based Life Profiles. ACM Transactions on Information Systems, 2009.

Giridhar Kumaran, and James Allan. Text classification and named entities for new event detection. In proceedings of the seventeenth annual international ACM SIGIR conference on research and development in information retrieval, pages 297-304, 2004.

J. Allan, V. Lavrenko, and R Swan. Explorations within topic tracking and detection. Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic: Massachusetts, pages 197-224, 2002.

J. Kleinberg, Bursty and hierarchical structure in streams. In proceedings of 8th ACM SIGKDD international conference on knowledge discovery and data mining, pages 373-397, 2002.

Gabriel Pui Cheong Fung, Jeffery Xu Yu, Philips S. Yu, Hongjun Lu. Parameter free bursty events detection in text streams. In proceedings of the 31st international conference on very large data bases, pages 181-192, 2005.

Qi He, Kuiyu Chang, and Ee-Peng Lim. Analyzing feature trajectories for event detection. In proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, pages 207-214, 2007.

Nish Parikh, Neel Sundaresan. Scalable and near real-time burst detection from eCommmerce queries. In proceedings of 14th ACM SIGKDD international conference on knowledge discovery and data mining, pages 972-980, 2008.

Nilesh Bansal, Nick Koudas. BlogScope: a system for online analysis of high volume text streams. In proceedings of the 33rd international conference on very large data bases, September 23-27, 2007, Vienna, Austria.

Meishan Hu, Aixin Sun, and En-Peng Lim. Event detection with common user interests. In proceeding of the 10th ACM workshop on Web information and data management, pages 1-8, 2008.

MingLiang Zhu, Weiming Hu, Ou Wu. Topic detection and tracking for threaded discussion communities. In proceedings of the 2008 IEEE/WIC/ACM international conference on Web intelligence and intelligent agent technology, pages 77-83. 2008.

Lu Liu, Lifeng Sun, Yong Rui. Web video topic discovery and tracking via bipartite graph reinforcement model. In proceeding of the 17th international

conference on World Wide Web, pages 1009-1018, 2008.

Inderjit S. Dhilon, Subramanyam Mallela, Dharmendra S. Modha. Information-theoretic co-clustering. In proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, pages 89-98, 2003.

You Chen, Sen Yang, Xueqi Cheng. Bursty topics extraction for web forums. In proceeding of the eleventh international workshop on Web information and data management, pages 55-58, 2009.

Sen Yang, Xueqi Cheng, You Chen, Gaolin Fang, Jin Zhang, Hongbo Xu, Detect Events on Noisy Textual Datasets, In Proceedings of the 12th International Asia-Pacific Web Conference, Busan, Korea, April 2010.

宋丹、卫东、陈英. 基于改进向量空间模型的话题识别跟踪. 计算机技术与发展, 2006. (Continued on page 51)

Note: Figure translations are in progress. See original paper for figures.