An Event Graph Model for Discovering Trends from Text Streams

Chengli Zhao, Xue Zhang, and Dongyun Yi

Department of Mathematic and Systems Science, College of Science,
National University of Defense Technology, 410073 Changsha, Hunan, P.R. China
{chenglizhao,xuezhang,dongyunyi}@nudt.edu.cn

Abstract. In this paper, we formally define and study the event graph model based on set theory and multi-relations theory, and discuss the methods of modeling event and event relations in detail. The event graph model is mainly designed to extract the potential events and the relationships between events from massive text streams, and further discover the trends embodied in the contents in text streams. We also study the connectivity of the event graph model, and give the equivalent conditions to determine the connectivity of event graph.

Keywords: event graph, trend, text stream.

1 Introduction

With the rapid development of network and information technology, people always encounter a lot of text stream data, such as instant messaging chats, e-mail and online news and more. In such stream text data, there often exist interesting events and graph structures constructed by events and their relations. How to mine the useful information from these massive text streams has become an important problem.

In this paper, based on set theory and multi-relations theory, we formally define and study the event graph model to discover, extract and summarize the potential events and event graph structures from massive texts, and discuss the method how to model the node and edge of the event graph with the examples of probabilistic mixture model and relative entropy respectively. We also study the connectivity of the event graph model, and give the equivalent conditions to determine the connectivity of event graph. Connectivity of the graph shows us the way to find the trends in text streams.

The rest of the paper is organized as follows. In Section 2, we discuss the related work in detail. In Section 3, we formally introduce the event graph model. In Section 4, we discuss the connectivity of the event graph and give some proof. We draw the conclusions to summary the paper in Section 5. Finally, Section 6 is acknowledgment.

2 Related Work

General Systems is a common approach that abstracts and considers a system as a set of independent and interacting parts, and provide the general theory and methods of

J. Suzuki and T. Nakano (Eds.): BIONETICS 2010, LNICST 87, pp. 357–362, 2012.

study the systems. In the domain of general systems, Lin studies the concept of general systems mathematically. In his book [1], he defines a system as a set of objects and a set of relations between the objects. Also he introduces a model of general systems based on the methods of set theory, and studies some basic global properties of systems in detail, including layer structures, centralized systems, relations between systems, and a characterization of centralizable systems. The theory and method described in the book presents a general framework to study different kinds of systems.

In text mining, Zhai proposes a generative probabilistic mixture model [2] which models the semantic content in texts in order to perform cross-collection clustering. This model can be estimated using the Expectation Maximization algorithm, so it is a very useful tool in text mining. Based on the above probabilistic mixture model, Mei studies the evolutionary patterns of themes in text streams and explores the methods how to discover and summarize them, and an evolution graph of themes on the Asian tsunami disaster is constructed in his paper [3]. Because there usually exists an evolution relationship between two themes or topics with time stamps, it is necessary to find the evolution graph structure.

Lots of work to predict trends in texts has been presented in the past few years. Based on WWW, Gloor extracts and predicts long-term trends on the popularity of relevant concepts such as brands, movies, and politicians by calculation of betweenness of these concepts. He introduces a novel set of social network analysis based algorithms [4] for mining the Web, blogs, and online forums to identify trends and also find who launch these them. The Algorithms he used include the temporal computation of network centrality measures, mining and analyzing large amounts of text based on social network analysis, and sentiment analysis and information filtering methods.

3 Event Graph Model

Given a text stream, the objective of event graph model is to discover the trends from it through extract the event graph in the text stream. This graph model will give us a specific insight to the massive text stream. In order to explain the event graph model, we firstly define the following concepts.

Definition 1 (Event). An event $e = (\theta, t)$ is an ordered pair of sets $e = (\theta, t)$, such that θ is a semantically topic which normally is represented as a probabilistic distribution of words, and t the time when the event e happens, defined by a time function $t = t(e)$.

Definition 2 (Event Graph). An event graph is an ordered pair of sets $G = (E, R)$, such that E is the set of all vertices (events) of G, and R a set of edges (relations) defined on E. The sets E and R are respectively called the vertex (event) set and the edge (relation) set of the event graph G.

The event graph $G = (E, R)$ is trivial if $E = \varnothing$. It is easy to see that only nontrivial event graph is meaningful to study in practical application.

3.1 Event Modeling

Let $e_i = (\theta_i, t_i)(i = 1, 2, ..., n)$ be n events in a text stream, and $C = \{d_1, d_2, ..., d_T\}$ a text stream, where d_j refers to a text with time stamp $j(j = 1, 2, ...T)$. Probabilistic mixture model [2] can be introduced to describe and extract events in text stream C. In this model, words are regarded as data drawn from a mixture model with component models for the topic word distributions and a background word distribution. Words in the same text share the same mixing weights. The model can be estimated using the Expectation Maximization (EM) algorithm to obtain the topic word distributions.

Each text could be seen as a sequence of words from a vocabulary set $V = \{w_1, w_2, ..., w_{|V|}\}$. The topic θ in the event $e = (\theta, t)$ can be defined by a unigram language model, such as a word distribution $\{p(w \mid \theta)\}_{w \in V}$. Naturally we have $\sum_{w \in V} p(w \mid \theta) = 1$.

Let $\theta_1, \theta_2, ..., \theta_k$ be k unigram language models and θ_B be a background model for the whole text stream C. A text d is regarded as a sample of the following mixture model:

$$p(w : d) = \lambda_B p(w \mid \theta_B) + (1 - \lambda_B) \sum_{i=1}^{k} [\pi_{d,j} p(w \mid \theta_i)] \quad (1)$$

Where w is a word in text d, $\pi_{d,i}$ is the mixing weight for text d for choosing the event θ_i such that $\sum_{i=1}^{k} \pi_{d,i} = 1$, and λ_B is the mixing weight for θ_B. The purpose of using a background model θ_B is to make the event models more discriminative. More details to estimate parameters in (1) can be found in [3]

3.2 Event Relation Modeling

Definition 3 (Event Evolution). Let $e_1 = (\theta_1, t_1)$ and $e_2 = (\theta_2, t_2)$ be two events. If $t_1 \le t_2$ and the similarity between events e_1 and e_2 is above a give threshold, we say that there is an evolutionary transition from e_1 to e_2, which we denote by $e_1 \to e_2$. We also say that θ_2 is evolved from θ_1, or θ_1 evolves to θ_2.

Given two probability mass functions $p(x)$ and $q(x)$, the Kullback-Leibler divergence (or relative entropy) between p and q is defined as

$$D(p \parallel q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \quad (2)$$

$D(p \parallel q)$ is always non-negative and is zero if and only if $p = q$. Because it is not symmetric and does not satisfy the triangle inequality, it is not a true distance between two distributions, but it is still useful to apply the KL-divergence to measure the distance between any two distributions [5].

Based on the property of KL-divergence, we use the KL-divergence to model the relation between any two events and measure their similarity. Let $e_1 = (\theta_1, t_1)$ and $e_2 = (\theta_2, t_2)$ be two events where $t_1 \le t_2$. We assume that e_2 has a smaller evolution distance to e_1 if their unigram language models θ_2 and θ_1 are closer to each other. Since the KL-divergence $D(\theta_2 \| \theta_1)$ can model the additional new information in θ_2 as compared to θ_1, it appears to be a natural measure of evolution distance between two events.

$$D(\theta_2 \| \theta_1) = \sum_{i=1}^{|V|} p(w_i \mid \theta_2) \log \frac{p(w_i \mid \theta_2)}{p(w_i \mid \theta_1)} \tag{3}$$

Formula (3) quantifies the relation between any two events in text streams.

4 Property of Model

In this part we will discuss the connectivity of event graph and give a complete proof in mathematics based on set theory and multi-relations theory [1]. Connectivity of the graph shows us the way to find the trends in text streams.

Let $G = (E, R)$ be an event graph and $r \in R$ a relation. The support of r, denoted $\mathrm{Supp}(r)$, is defined by

$$\mathrm{Supp}(r) = \{e \in E : e \text{ belongs to the event set on which } r \text{ is defined}\} \tag{4}$$

The event graph $G = (E, R)$ is connected if it cannot be represented in the form

$$G = G_1 \oplus G_2 \triangleq (E_1 \cup E_2, R_1 \cup R_2) \tag{5}$$

Where $G_i = (E_i, R_i)(i = 1, 2)$ such that $E_i \subset E$ and $R_i \subset R$ and $E_1 \cap E_2 = \varnothing$.

Theorem 1. An event graph G is connected if and only if for any two events $x, y \in E$, there exists n relations $r_i \in R$ such that

$$x \in \mathrm{Supp}(r_1) \text{ and } y \in \mathrm{Supp}(r_n) \text{ and } \mathrm{Supp}(r_i) \cap \mathrm{Supp}(r_{i+1}) \neq \varnothing,\ i = 1, 2, ..., n-1 \tag{6}$$

Proof: Necessity. We prove by contradiction. Suppose that the event graph G is connected and there exist two events $x, y \in E$ such that there do not exist n relations $r_i \in R, i = 1, 2, ..., n$, for any natural number $n \ge 1$, such that

$$x \in \mathrm{Supp}(r_1) \text{ and } y \in \mathrm{Supp}(r_n) \text{ and } \mathrm{Supp}(r_i) \cap \mathrm{Supp}(r_{i+1}) \neq \varnothing,\ i = 1, 2, ..., n-1$$

From the hypothesis that G is connected, it follows that there must be relations $r_1, s_1 \in R$ such that $x \in \mathrm{Supp}(r_1)$ and $y \in \mathrm{Supp}(s_1)$. Then our hypothesis implies that

$$\mathrm{Supp}(r_1) \cap \mathrm{Supp}(s_1) = \varnothing \tag{7}$$

Let U_0=Supp(r_1) and V_0=Supp(s_1), and for each natural number $n \in \mathbb{N}$ let

$$U_n = \bigcup \{\mathrm{Supp}(r) : r \in R \text{ and } \mathrm{Supp}(r) \cap U_{n-1} \neq \varnothing\} \tag{8}$$

$$V_n = \bigcup \{\mathrm{Supp}(s) : s \in R \text{ and } \mathrm{Supp}(s) \cap V_{n-1} \neq \varnothing\} \tag{9}$$

Then $U_0 \subseteq U_1 \subseteq \cdots \subseteq U_n \subseteq \cdots, V_0 \subseteq V_1 \subseteq \cdots \subseteq V_n \subseteq \cdots$, and $U_n \cap V_m = \varnothing$ hold for all natural numbers $n, m \in \mathbb{N}$.

We now define two subsystems of $G : G_i = (E_i, R_i), (i = 1, 2)$ such that

$$E_1 = \bigcup_{n=0}^{\infty} U_n \text{ and } E_2 = E - E_1 \tag{10}$$

$R_1 = \{r \in R : \mathrm{Supp}(r) \cap E_1 \neq \varnothing\}$ and $R_2 = \{r \in R : \mathrm{Supp}(r) \subseteq E_2\}$

Then we have $R_1 \cup R_2 = R$ and $R_1 \cap R_2 = \varnothing$. In fact, for each relation $r \in R$, if $r \notin R_1$, then $\mathrm{Supp}(r) \cap E_1 = \varnothing$ and so $\mathrm{Supp}(r) \subseteq E_2$, thus $r \in R_2$.

Therefore, $G = (E, R) = (E_1 \cup E_2, R_1 \cup R_2) = G_1 \oplus G_2$, contradiction.

Sufficiency. The proof is again by contradiction. Suppose condition (ii) holds and G is disconnected. Thus, there exist nontrivial subgraphs G_1 and G_2 of G such that $G = G_1 \oplus G_2$.

Suppose that $G_i = (E_i, R_i), (i = 1, 2)$. Pick an event $e_i \in E_i, (i = 1, 2)$. Then there are no relations $r_j \in R, j = 1, 2, ..., n$, for any fixed $n \in \mathbb{N}$, such that

$$e_1 \in \mathrm{Supp}(r_1) \text{ and } e_2 \in \mathrm{Supp}(r_n) \text{ and } \mathrm{Supp}(r_i) \cap \mathrm{Supp}(r_{i+1}) \neq \varnothing,\ i = 1, 2, ..., n-1 \tag{11}$$

Contradiction.

Generally, a connected event graph indicates that some relative stable trends are forming. The event graph can tell us much more than a single event itself does.

5 Conclusions

In this paper, based on set theory and multi-relations theory, we formally define and study the event graph model to discover, extract and summarize the potential events and event graph structures from massive texts, and further discover the trends embodied in the contents in text streams. We also study the connectivity of the event graph model, and give the equivalent conditions to determine the connectivity of event graph. The event graph model can be applied to many fields, such as online news, classification of text data collection, and online text data organization and topic detection of the text streams. This model is still preliminary, but it can serves as a foundation for the analysis of relations between events in massive text streams.

Acknowledgments. At first, we would like to thank Professor Yi Lin for his supervision in set theory and multi-relation theory. Also, we thank the reviewers for their comments that help improve this paper.

References

1. Lin, Y.: General Systems Theory: A Mathematical Approach. Kluwer Academic and Plenum Publishers, New York (1999)
2. Zhai, C., Velivelli, A., Yu, B.: A cross-collection mixture model for comparative text mining. In: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 743–748 (2004)
3. Mei, Q., Zhai, C.: Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining. In: Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2005)
4. Gloor, P., Krauss, J., Nann, S., Fischbach, K., Schoder, D.: Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. In: IEEE Conference on Social Computing (SocialCom 2009), Vancouver, August 29-31 (2009)
5. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley (1991)
6. Zhai, C., Lafferty, J.: Model-based feedback in the KL-divergence retrieval model. In: Tenth International Conference on Information and Knowledge Management (CIKM 2001), pp. 403–410 (2001)