Knowledge-Based Systems 124 (2017) 164–175

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A survey on trends of cross-media topic evolution map

Houkui Zhou[a,c,d], Huimin Yu[a,b,*], Roland Hu[a], Junguo Hu[c,d]

[a] Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, PR China
[b] State Key Laboratory of CAD & CG, Hangzhou 310027, PR China
[c] School of Information Engineering, ZheJiang A&F University, Linan 311300, PR China
[d] Zhejiang Provincial Key Laboratory of Forestry Intelligent Monitoring and Information Technology, Linan 311300, PR China

ARTICLE INFO

Article history:
Received 19 October 2016
Revised 22 February 2017
Accepted 7 March 2017
Available online 9 March 2017

Keywords:
Cross-media
Topic evolution
Topic map
Probabilistic generative model

ABSTRACT

Rapid advancements in internet and social media technologies have made "information overload" a rampant and widespread problem. Complex subjects, histories, or issues break down into branches, side stories, and intertwining narratives; a "topic evolution map" can assist in joining together and clarifying these disparate parts of an unfamiliar territory. This paper reviews the extant research on topic evolution map based on text and cross-media corpora over the past decade. We first define a series of necessary terms, then go on to describe the traditional topic evolution map per 1) topic evolution over time, based on the probabilistic generative model, and 2) topic evolution from a non-probabilistic perspective. Next, we discuss the current state of research on topic evolution map based on the cross-media corpus, including some open questions and possible future research directions. The main contribution of this review is in its construction of an evolution map that can be used to visualize and integrate the extant studies on topic modeling – specifically in regards to cross-media research.

1. Introduction

The extensive development of modern information technology has created an information redundancy problem. There currently exists a wealth – an excess, even – of cross-media data having originated from heterogeneous and homogeneous media with multiple sources, such as news media websites, social media websites, photo/video sharing websites, mobile phones, video surveillance servers, and the Internet of Things. Extracting useful knowledge from cross-media big data on the network space is a challenging endeavor. Search engines are generally relied upon for accessing information, and efforts have even been made to create specialized search and retrieval tools (e.g., academic search and news search).

Search engines are indeed effective in retrieving knowledge, but output lists of search results that are unstructured and potentially unhelpful. Establishing a comprehensive and accurate understanding of the "big picture" when it comes to a given topic based on a large mass of text, speech, image, and video information can be nearly impossible. The "five Ws" (who, what, when, where, and why) are a common starting point in building connections between bits of information on a certain topic, as well as the topic's "evolution". Each of the first four W's is necessary to tell the whole "story" of a given topic or event; these answers can be readily extracted from the data available online, but what about the fifth W – the "why" of the topic?

The underlying cause (i.e., the "why") of a given topic is typically considered its most interesting characteristic; it is also the topic's most elusive aspect in terms of completing the topic evolution map. For this reason, the "why" tends to be relegated to an internal cause between different sub-topics in the map. The quantitative relation among sub-topics is highly appealing to the individual seeking the "cause and effect" of the topic. Knowledge is delivered by an array of data modalities (as opposed to a single medium) which represent the same semantics; this is known as "cross-media" information delivery [1,2]. For example: An image is usually accompanied with a text description on a web page. Although they are in different modalities, the incorporated semantics are consistent between the image and its related text. Cross-media data can represent different aspects of the real world simultaneously, and help to comprehensively document the evolution of certain topics within the real world.

Cross-media-based topic evolution map can be employed to leverage different types of data across multiple sources to strengthen one's knowledge of a given topic; this type of information map represents better solution to the "five W's" than traditional research. Although there is still no formal, academic definition of "cross-media", there has been a substantial amount of research on this approach to topic evolution map. The 2013 ACM International Conference on Multimedia hosted a panel entitled

* Corresponding author.
E-mail addresses: zhouhk@zju.edu.cn (H. Zhou), yhm2005@zju.edu.cn (H. Yu).

http://dx.doi.org/10.1016/j.knosys.2017.03.009
0950-7051/

Cross-media Analysis and Mining, for instance, and the IEEE ICME 2014 held a special session on Cross-media Computing.

Topic evolution map research began with a focus on text data, then shifted to multimedia data, and now centers around cross-media data. Similar work on cross-media topic evolution map dates back to Topic Detection and Tracking (TDT) [3], a DARPA-sponsored initiative to investigate techniques for following news events within a stream of broadcast news stories. The pilot phase started in 1998 [4] and the final phase of the project ended in 2004. TDT-related research has mainly focused on the line structure evolution of text streaming; relevant work includes timeline analysis of news events [5–9] and storyline generation in text information [10–12]. We also summarized some typical topic evolution models with linear structrue in Ref. [13]. These studies provide a general description of the topic evolution map with linear structure from a text corpus for simple topics, but do not describe topic evolution maps for more complex or non-linear subjects.

Complex subjects exhibit a very non-linear structure – they break down into branches, side stories, dead ends, and intertwining narratives. To explore these subjects/stories, the user needs a map or other organized structure to guide them through the unfamiliar territory. Available topic evolution map structures can be roughly divided into two categories. The first includes maps developed from a probabilistic topic model angle, e.g., TOT (topic over time) [8] and TTM (temporal text mining) [7]. The second category includes maps built from a non-probabilistic topic model perspective, such as "metro maps" of science or news articles [14,15] in which the "topic" refers to a single text article or a cluster of text articles. Research on topic evolution map represents an innovative approach to resolving information overload, and is especially helpful in regards to better understanding complex events or topics.

This paper provides a review of the extant research on cross-media topic evolution map. First, the definitions of relevant terms are introduced in Section 2. Next, studies on the traditional topic evolution map based on text corpora are discussed in Section 3, followed by a discussion on topic evolution map based on cross-media corpora in Section 4. Section 5 provides a brief summary and conclusion. Open questions and future research directions are presented in Section 6.

2. Introduction of some terminologies in the survey

There are numerous technical terms utilized in this paper including "topic", "event", "storyline", and "topic map". Terms like "topic" and "topic model" have no standard definition in the topic evolution map research field. For the sake of clarify, we attempt here to establish and standardize the definitions of these important concepts per their usage in the literature.

Topic (Definition 1): A topic z starts with a novel seminal event $\boldsymbol{e_1}$, followed by other related events $\{\boldsymbol{e}_2, \boldsymbol{e}_3, \ldots\}$. Each event $\boldsymbol{e}$ contains a set of stories $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s_i}, \ldots\}$, whereas a topic consists of a collection of events $\{e_1, e_2, \ldots, e_i, \ldots\}$. This definition of topic follows the traditional, non-probabilistic approach of extracting an abstract concept from a collection of documents.

Topic (Definition 2): A semantically coherent topic in a text collection C is represented by a topic model θ, which forms a group of semantically related words. The probabilistic distribution of these words is conditioned on topic z: $\{w, p(w \mid z, \beta)\}_{w \in V}$, where β is a prior parameter. It follows that $\Sigma_{w \in V}\, p(w \mid z, \beta) = 1$ [16].

These two definitions of "topic" are commonly used in studies on topic evolution map. They notably differ by the viewpoint from which they are respectively established: The first definition is non-probabilistic, as the "topic" is viewed as collection of events; the second definition is probabilistic, as the "topic" is viewed as a series of latent variables and documents comprised of sub-topics. Unfortunately, LDA (Latent Dirichlet Allocation) [16] is not well-suited to use in news articles as it offers little guidance for aggregating the articles into "stories".

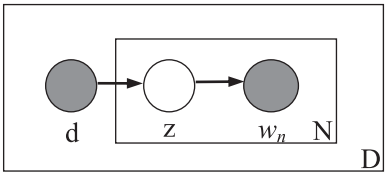

Fig. 1. Graphical model representation of PLSA which is taken from Fig. 3(c) of Ref. [16].

In this survey, these two definitions of "topic" are applied to different research directions for topic evolution map. The first definition is used to describe topics as they relate to event evolution similar to TDT. The second definition is used to describe topics on evolution maps built via probabilistic generative model.

Event: An event is defined as the triplet $\{e, t, l\}$, which represents a particular thing e that happens at a specific time t and place l. Each event e is assigned to a timestamp t, so that event evolution over time becomes a well-defined research problem.

Here an "event" is a smaller concept than the first definition of "topic", which can be viewed as collection of events. These two concepts cannot be strictly distinguished, however – an event can be considered a topic with small granularity or a sub-topic, so in certain conditions, "event evolution" and "topic evolution" are interchangeable.

Storyline: A storyline S in a news topic Q is a chain of events that characterize a certain aspect of Q and involve the same set of actors and places. Here, the definition of "storyline" falls under the first definition of "topic".

Topic map: A topic map (topic graph or topic network) is viewed as a directed acyclic graph (DAG): $G = <V, E>$, where V consists of topics in a collection of vertexes in map G and E represents the relationship between adjacent topics is a collection of edges in map G.

The above definition of "topic map" comes with a few important considerations: 1) The direction of edges follows the timestamp sequence of the topic; 2) there are one or several sources and end vertexes in the topic map; 3) a possible path from a source vertex to an end vertex indicates a storyline along the topic evolution in the map; 4) the points of intersection of different paths in the map represent key issues in the topic evolution process. This map structure is appropriate for research on topic evolution and inference. In this survey, we establish a unified viewpoint of topic evolution structure by using the topic map format. Here, the "topic evolution map" refers to the topic evolution process plotted onto a topic map.

3. Traditional topic evolution map based on text corpora

3.1. The topic evolution map based on the probabilistic generative model

3.1.1. Basic LDA and PLSA models

The topic evolution map based on the probabilistic generative model (PGM) can be considered an extension of the topic model based on LDA [16] or probabilistic latent semantic analysis (PLSA) [17].

The PLSA model, also known as the "aspect model", represents each word in a document as a sample from a larger, mixed model. The developers of the PLSA model made assumptions that each document is characterized by a mixture of different topics, and that each topic is composed of different words. The PLSA model can also be considered a probability generation process (Fig. 1 which is taken from Fig. 3(c) of Ref. [16]). It is expressed as fol-

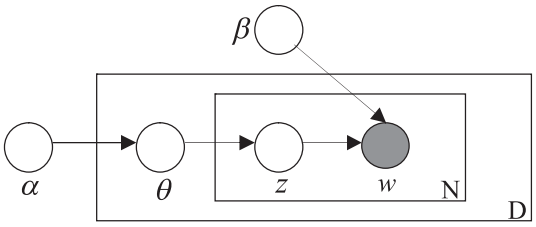

Fig. 2. Graphical model representation of LDA which is taken from Fig. 5 (Left) of Ref. [16].

lows:

$$p(d, wn) = p(d)\sum p\sum p(z|d)p(wn|z) \quad (1)$$

LDA can be considered a generative probabilistic model, as depicted in Fig. 2 which is taken from Fig. 5 (Left) of Ref. [16]. Its main assumption is that documents are composed of latent topics and each topic is composed of a distribution of words. For each document d in corpus D, the generative process of LDA can be summarized as follows [16]:

(1) Choose $N \sim Poisson(\xi)$. (ξ is the paramter of Poisson distribution)
(2) Choose $\theta \sim Dirichlet(\alpha)$.
(3) For each of the N words w_n,
(a) Choose a topic $z_n \sim Multinomial\ (\theta)$;
(b) Choose a word w_n from $p(w_n \mid z_n, \beta)$ and a multinomial probability conditioned on the topic z_n.

Given the parameters α and β (β is prior distribution of words over topics), the joint probability of θ, z, and w is given by [16]:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \quad (2)$$

3.1.2. Typical topic evolution map based on the PGM

"Topic" as discussed in this subsection falls under the second definition provided above, i.e., the probabilistic perspective. PGM (Probabilistic Generative Model) –based topic evolution maps can be split into two categories based on the specific algorithm adopted. In the first category, the whole text corpus is divided into sub-collections by chronology, then the PGM is applied into each sub-collection to generate topics and ultimately establish the topic evolution between each sub-collection. In the second category, the chronological information is integrated into the topic evolution model. The first category includes the Temporal Text Mining (TTM) [7], Dynamic Topic Model (DTM) [18], Temporal Dirichlet Process Mixture Model (TDPM) [19], infinite Dynamic Topic Model (iDTM) [20], and Topic-User-Trend (TUT) model [21].

The earliest work on this subject can be dated back to Mei and Zhai's work on the TTM [7]. They proposed a PGM for mining meaningful temporal theme structures in two types of text datasets: News articles and science papers. The TTM model, which is similar to the PLSA model, can be applied to a background topic θ_B or K common topics θ_k (k = 1, …, K) to distinguish the background words and topic-related words. This model considers the temporal theme evolution using clustering models in the predefined time slice for a fixed number of topics.

There are four main issues with this model corresponding to four potential improvements to be made. First, the model parameters of each sub-collection are predefined and determined independently of each other – identifying parameters automatically for different datasets would substantially improve the model. Second, the model is not suited to online text stream data. Third, the division of sub-collection is data-dependent, leaving no common method for different datasets. Finally, the number of topics is fixed in this model, so it does not reflect new topics in the collection that appear or disappear over time.

As a response to the first problem, Blei and Lafferty [18] proposed the DTM model which considers the time evolution factor between neighboring time slices per sub-collection. DTM divides the whole text collection into time slices similar to TTM and imposes a static topic model similar to LDA in each slice; the parameters of the model in adjacent slices are dependent. DTM uses a variational inference algorithm of Kalman filters and wavelet regression for sequential data. It can provide a highly accurate predictive model and reasonable results for science article datasets, but suffers the same drawbacks as TTM (apart from the first one in the list above). Bolelli's [22] model, the Segmented Author-Topic Model (S-ATM) is similar except that the authorship information of documents is added into the generative model. S-ATM is based on the Author-Topic Model [23,24] and can effectively detect distinct topics and their evolution over time. We also proposed a similar topic evolution model which simultaneously utilized citations and words of papers [25].

Topic evolution models based on LDA or PLSA, as mentioned above, belong to the framework of finite Bayesian mixture models. They are inadequate with respect to modeling sequential data due to the full exchangeability assumption they employ. Ahmed and Xing [19] proposed the temporal Dirichlet process mixture model (TDPM) as a framework for modeling complex longitudinal data. In the TDPM, the data is divided into epochs within which it is non-exchangeable. In the context of the text-stream model, the number of topics in each epoch is unbounded. The generated topics can be retained, die out, or emerge over time, and the actual parameterization of each topic evolves over time in a Markovian fashion. Compared to finite Bayesian mixture models, the advantages of these models are that the number of topics in each epoch is not fixed and the parameterization of each topic evolves over time.

The TDPM builds topics based on the hypothesis that each document is sampled from a single topic. This assumption is simpler than those used by mainstream topic models in which each document is generated from a mixture of topics (e.g., LDA). Ahmed and Xing [20] proposed extending the simple non-parametric dynamic topic model into a full-fledged topic model, the iDTM. The main difference between iDTM and TDPM models is that the former can accommodate the evolution of topic popularity, topic word distribution, and the number of topics in a full-fledged admixture setting; iDTM does not limit the number of topics. Topics can emerge and disappear at any epoch, while the word distributions of topics evolve based on a first-order state space model.

Tang and Yang [21] proposed the TUT model to simulate the generation process of user-generated web contents. Compared to TTM and DTM models, one distinct feature of TUT model is that the number of trends is not fixed over time and does not need to be predefined by leveraging a Recurrent Chinese Restaurant Process (RCRP). This feature is, to some extent, similar to those of TDPM and iDTM models. Another distinct feature of the TUT model is that the participation of multiple users is modeled simultaneously during the generation process. It is critical to model the evolution of topics and users' interests over time in a streaming fashion considering the rapid development of social media sites such as Twitter and Facebook.

The topic evolution models discussed above are all based on LDA and can only be used for offline text datasets. Analyzing a stream of text data at the time of its arrival better suits the modern (i.e., online) data environment, however. AlSumait et al. [26] first proposed an OLDA (on-line LDA) model for mining text streams in TDT applications. Their approach proposes an empirical Bayes method that incrementally builds an up-to-date model

when a new document (or a set of documents) appears. Ahmed et al. [27,28] proposed a unified framework using clustering and topic models to group incoming news articles into temporary but tightly-focused storylines, to identify prevalent topics and key entities within these stories, and to reveal the temporal structure of stories as they evolve for streaming news. Gohr and Hinneburg [29] used latent variables to index new words while deleting outdated words within a sliding window for a stream of documents. Those indexed new words are used to portray topic changes in the information retrieval domain. Kim and Oh [30] are responsible for the latest work on topic evolution for streaming text; they proposed a computational framework based on LDA that can be used to analyze topic chains to identify general topics and short-term issues in online news articles. The topic chain-based model can capture how topics newly emerge and disappear as well as change over time.

The discrete time topic evolution models described above are valuable in regards to the extant research on this subject, however, they are problematic in that the division of sub-collections is data-dependent. Researchers have also attempted to model text and time stamps jointly in newer topic evolution models including the Topics Over Time (TOT) [8], continuous Time Dynamic Topic Model (cDTM) [31], Trend Analysis Model (TAM) [32], nonparametric Topics Over Time (npTOT) [33], Topic-based Information Diffusion and Evolution (TIDE) [34], and other similar models.

TOT [8] models the text and time stamp of a document jointly, assuming that latent topics generate time stamps according to a Beta distribution. Unlike TTM and DTM, this model does not discretize time, so the selection of time slice size is not a problem. Unlike the DTM, this model does not require the assumption that the state between adjacent slices has a Markovian relationship, but instead treats time as an observed continuous variable. TOT can be considered an adaptation of LDA; the main difference is that for each word in a document, the timestamp is drawn from Beta distribution. TOT can yield a topic distribution profile over time, topic co-occurrences over time, and topic evolution over time, but exhibits non-Markovian variations in topic probabilities. Further, the beta distribution used to model the time-varying probability is unimodal and thus limits the available patterns of topic temporal variation. This limitation precludes prediction outside of the bounded time-frame or extension to higher dimensionalities. Another main shortcoming, similar to TTM and DTM, is that the number of topics must be defined a priori.

The cDTM [31] is an extension of DTM in which time is considered to be continuous to resolve the discretization problem. The model proposed here replaces the discrete state space model of the DTM by using Brownian motion to model continuous-time topic evolution. However, this seemingly more complicated model (which generally introduces more latent variables than the DTM) is actually simpler and more efficient to fit.

Kawamae [32] proposed the Trend Analysis Model (TAM) which extends the TOT model and introduces a latent trend class variable into each document. TAM offers two characteristics for modeling trends in documents and realizes the functionalities of both DTM and TOT simultaneously. The TAM model includes trend classes that are responsible for generating both observed timestamps and topic sets as well as words, where each topic is responsible for generating words. The switch variable included in the model can distinguish these different types of words in each token of each document. To this effect, TAM places documents that have similar topic distributions over time into the same trend class. The proposed model can capture interpretable low dimensionality sets of topic classes, and is useful for analyzing the evolution of trends on various datasets.

According to above drawbacks of TOT, which requires a fixed number of topics, the npTOT [33] model was proposed as a non-parametric extension to the TOT model. This model allows an unbounded number of topics, each of which can peak in popularity an unbounded number of times. Related topics trend similarly in npTOT, which induces correlations between the temporal variations in topic popularity. Similar to TOT, npTOT is a joint model of both text and time that includes a Gibbs sampling scheme to make use of tractable exchangeable distributions. These topic models can be applied to image as well as text data [35]. This characteristic suggests that topic models based on npTOT may benefit topic evolution research on cross-media corpora consisting of text, speech, image, and video metadata in the future.

The models described above are only applicable to regular text corpora such as science papers or news reports. Advancements in internet technology (especially social media) have made user-generated text information more and more important in terms of topic evolution map research. These data typically include a great deal of noise information due to irregular language expression and complex structural information.

There have been several new topic evolution models tailored to social media data. Lin et al. [36], for example, defined the popular events tracking (PET) problem in a social community that consists of both a stream of text information and a stream of network structures. PET models the popularity of events over time by taking into consideration the "burstiness" of user interest, the interplay between information diffusion and network evolution, and the evolution of textual topics. The novelty of PET is that it can model the diffusion of interest and evolution of topics together to track the popular events in a social community per the interplay between textual content and social networks. The same authors [36] proposed topic-based information diffusion and evolution (TIDE) [34], a novel probabilistic model for the joint inference of topic diffusion and evolution within social networks. TIDE integrates the generation of text, the evolution of topics, and the social network structure under a unified model. Given the primitive form of any arbitrary topic, TIDE can effectively track topic snapshots as they evolve over time and reveal the latent diffusion paths of the topic.

Yin et al. [37] proposed the mixture model for stable and temporal topics named EUTB model, which can distinguish stable topic and bursty topic evolution over time in a social network under a unified PLSA-based model. In EUTB, stable topics represent regular events occurring periodically and bursty topics represent "hot" events occurring suddenly within a certain period. EUBT was the first model to represents topic distribution temporally by assuming that topics are generated by users or time periods. Hu et al. [38] proposed the Group Specific Topics-over-Time (GrosToT) model, which can simultaneously identify groups and topics from social media streams; GrosToT can model multimodal variation and thus is highly flexible for capturing topic temporal dynamics. Hu et al. [39] also proposed the Community Level Diffusion (COLT) model, a generative model joining text, time, and network information. The main contribution of COLT is that it can discover latent communities and topics, and infers community-level dynamics in a unified latent framework. The common features of these three models are that text is combined with other information (e.g., user participation or social network structure) to model topic evolution. Of course, traditional topic evolution models based on text information alone are not appropriate for social media data. The increasing prevalence (and crucial importance) of social media data necessitates the further development of appropriate topic evolution models, however.

3.1.3. Comparison of topic evolution map models

Table 1 provides a comparison of typical topic evolution map models based on PGM in regards to the way that they respectively divide temporal data and generate topics, as well as their number

Table 1
Comparison among typical PGM-based topic evolution models.

Typical models	Proposed time	The division of time	Basic topic model	Evolution type	Evolution structure	Number of topics	Data type	Online or not
TTM	2005	Discrete time	PLSA	Content	Non-linear	Fixed	News/paper	Not
DTM	2006	Discrete time	LDA	Content/strength	Linear	Fixed	Paper	Not
TDPM	2008	Discrete time	RCRP	Content/strength	Linear	Not fixed	Paper	Not
iDTM	2010	Discrete time	RCRF	Content/strength	Linear	Not fixed	Paper	Not
Topic chain	2011	Discrete time	LDA	Content	Linear	Fixed	Web news	Online
TUT	2012	Discrete time	LDA	Content/strength	Linear	Fixed	Paper/social media	Not
OLDA	2008	Stream text	LDA	Content/strength	Linear	Fixed	Paper/text	Online
Gohr's model	2008	Stream text	PLSA	Content/strength	Linear	Fixed	Paper	Online
PET	2010	Stream text	PLSA	Content/strength	Linear	Fixed	Paper/social media	Online
Ahmed's model	2011	Stream text	LDA	Content/strength	Linear	Fixed	News	Online
TOT	2006	Continuous time	LDA	Content	Linear	Fixed	Paper/text	Not
cDTM	2008	Continuous time	LDA	Content/strength	Linear	Fixed	News	Not
TAM	2011	Continuous time	LDA	Content	Linear	Fixed	Paper/text	Not
TIDE	2011	Continuous time	PLSA	Content/strength	Linear	Fixed	Paper/social media	Not
npTOT	2012	Continuous time	TOT, CRF	Strength	Linear	Not fixed	Paper/social media	Not
EUTB	2013	Continuous time	PLSA	Content/strength	Linear	Fixed	Social media	Not
GrosToT	2014	Continuous time	LDA	Content/strength	Linear	Fixed	Social media	Not
COLT	2015	Continuous time	LDA	Content	Linear	Fixed	Social media	Not

of topics, type and structure of topic evolution, research data type, and whether they are suited to online or social media data.

The topic evolution maps described in Table 1 can be split into three categories per the way they divide temporal data: Discrete time division, continuous time division, and online stream text division. Models adopting the discrete time division method are suitable for science paper and news datasets, which can be readily divided into subsections by time information. Models adopting the continuous time division method are suitable for modeling the topic and time of text data simultaneously. Models adopting the online stream text method are suitable for analyzing massive document collections that are organized in stream form.

Of all the models listed in Table 1, TDPM, iDTM, and npTOT (which adopts non-parametric Bayesian methodology) cannot predefine the number of topics. The majority of models (except PET, TIDE, TUT, and npTOT) could be used only for science papers or news text before 2013. PET, TIDE, TUT, and npTOT can be used for traditional media and social media simultaneously, however, social media posts per user in a time interval are aggregated into a single document similar to a traditional body of text. EUTB, GrosToT, and COLT models can also be applied to social media data, as they consider a post as a single document.

EUTB can detect both stable topics and temporal topics via a user-temporal mixture model. GrosToT places both user and text information in a unified probabilistic model to infer latent user groups and temporal topics simultaneously. COLT can exploit the user relationship and network structure in social media data to model topics and communities under a unified latent framework.

Perplexity [16], a common metric in the topic modeling field, is utilized here to evaluate the performance of each model shown in Table 1. Generally speaking, for a test set D_{test}, perplexity is defined as follows:

$$Perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{3}$$

where N_d represents the amount of words in document d and w_d=(w_{1d}, w_{2d},...,w_{nd}) is the vector form of document d. We used the NIPS dataset as our experimental dataset [11,20,33]. It consists of the full text of the 17 years of proceedings from the 1987 to 2003 Neural Information Processing Systems (NIPS) Conferences, containing 2484 documents with 2865 authors. We removed stopwords, numbers, and the words appearing less than five times in the corpus resulting in a 14,036-word vocabulary.

In our experiment, each document's timestamp was determined by the year of the proceedings. For the first type of topic evolution model based on PGM, we divided the time of collection into 17 segments (i.e., one year per segment). We split each entire dataset into a 50% portion for training and left the remaining 50% for testing, where the training and test documents were selected uniformly across the dataset. To ensure fairness, the parameters of each model in our experiments were the same as the numerical value adopted in the original paper [40] ($\alpha = 50/K$, where K is the number of topics, $\beta = 0.01$) for the LDA model.

Again, the perplexity of different PGM-based topic evolution models are listed in Table 2. Four of these models (TUT, EUTB, GrosToT and COLT) use the author's information of the NIPS dataset. The number of topics was set to 10, 50, or 100 and the perplexity of all models was caculated for each topic values. All topic evolution models based on the non-parameter Bayesian model (TDPM, iDTM, and npToT) have the same perplexity regardless of the number of topics, while the perplexity of other models decreases as topic number K increases. The selection of K for each models refers to model selection problem, which is disscussed in Ref. [40]. The suitable K value is when the lower perplextiy value

Table 2
The comparison exprimental results of typical topic evolution map models based on PGM.

Model	K = 10	K = 50	K = 100
TTM	2763	2214	2174
Topic chain	2089	1632	1587
OLDA	2087	1628	1585
DTM	1685	1427	1424
TDPM	1321	1321	1321
iDTM	1267	1267	1267
cDTM	1328	724	720
TOT	1014	607	605
npToT	587	587	587
TAM	560	554	550
EUTB	735	545	550
COLT	758	535	542
GrosTOT	637	390	487
TUT	372	356	365

is acquired using that K value. For these models, the suitable K value is 50 from Table 2. Discrete time topic evolution models have the higher perplexity than continuous time topic evolution models, likely because the latter consider the time variable during topic modeling and calculate perplexity across the whole dataset. Table 2 also shows where the perplexity of the six discrete time topic evolution models fell into descending order as TTM, Topic Chain, OLDA, DTM, TDPM, and iDTM. The perplexity of the eight continuous time topic evolution models at K = 50 fell into descending order as cDTM, TOT, npToT, TAM, EUTB, COLT, GrosTOT, and TUT. TUT and EUTB utilize the author's information in the modeling process, while TAM, GrosToT, and COLT utilize additional variables such as trend, group, and community in the modeling process. Thus, the latter models have better modeling power with lower perplexity.

Again, the topic evolution maps described above are all based on PGM, which accepts the same assumption as LDA that the topic is a "bag of words" (i.e., distribution of words) and that document timestamps are incorporated into the model. In effect, then, these models can be used to both detect a topic and trace its evolution.

3.2. Non-probabilistic topic (event) evolution map

Non-probabilistic approaches to topic evolution modeling markedly differ from the probabilistic approaches described in Section 3.1. Here, "topic" refers to a cluster of events; an "event" may be a cluster of stories which is extracted from one or several documents. These models can be split into two categories: Topic evolution (where the topic is composed of a cluster of events) and event evolution (where an event is composed of one or several documents).

3.2.1. Generic modeling process

In this type of topic evolution map, "topic" is an abstract concept with a non-probabilistic definition. According to the summarization of existing topic evoluiton map models, the generic modeling process for this type of topic evolution model (as shown in Fig. 3) can be summarized as follows.

(1) Detect typical topics from the text corpus and identify their time range according to the timestamp of the topical documents.
(2) Establish the relationship among topics according to a specific algorithm.
(3) Construct a topic evolution map according to the relationship and timestamp between topics.

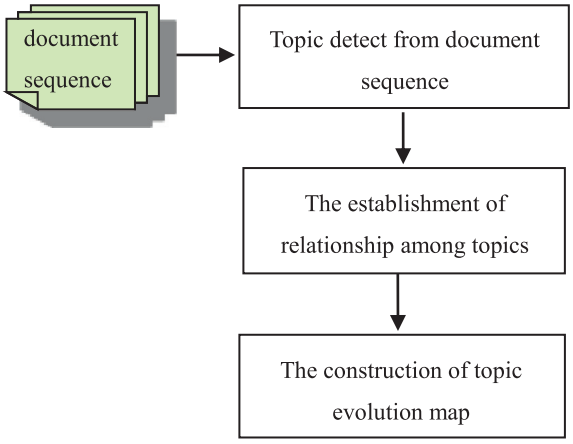

Fig. 3. Generic modeling process.

3.2.2. Event evolution models

This category includes the evolution pattern discovery technique [41], Topic Summarization and Content Anatomy (TSCAN) [42], storyline mining [11], mixture-event-aspect (MEA) [43], topic evolution based on social media [44], and bursty event detection [45,46]. The research corpora for these models are built from text data to social media data and from static data to online streaming data.

Wei and Chang [41] proposed the concept of evolution patterns (Eps) which refers to temporal relationships of event episodes that occur frequently in sequences of documents; they effectively built an EP discovery technique for mining Eps from document sets. Chen and Meng [42] proposed the TSCAN model, which can be used to summarize and integrate core parts of a topic graphically to make it easily understandable. Wei and Chang also proposed an event evolution graph that presents the underlying structure of events for the efficient browsing and extraction of information [47]. These models represent very useful information visualization tools, but do not support users in conducting interactive browsing of their topics of interest among all the relevant (and less relevant) events; they are also prevented from extracting summaries automatically for a specific path in the evolution graph.

Ref. [48] put Wei and Chang's technology [41,47] into practice by building an application for tracing the event evolution of terror attacks. They conducted a case study on "Chechen Terrorists Seizing Beslan School" in which the proposed model yielded favorable results.

Storyline generation can also be utilized to model event evolution. The Story-line based Topic Retrospection (StoRe) [11] system is a typical example; it can be used to identify events from news reports via self-organizing maps and to compose a storyline summary which portrays event evolution within a certain topic by assigning weights to different events based on the similarity between them. This system can generate storyline-based summaries using term weighting schemes, but does not take the temporal information into account. Taiwan electronic toll collection (ETC) news topic retrospection was used to demonstrate and evaluate StoRe performance. The event evolution structure of the system is linear, however, so StoRe is suited only to the simple event evolution of a topic.

Huang et al. [43] proposed the Mixture-Event-Aspect (MEA) model, which represents different types of sub-events per local and global aspects based on their part-whole relationship to the major events. The event storyline is generated by selecting the representative sentences of each aspect based on the aspect sequence for the given event subject. As a departure from other storyline generation and summarization methods, the MEA model introduces a

"local/global" property to distinguish different part-whole relationships between the sub-events and the major event to improve the quality of the storyline.

Event evolution based on social media is more complex than event evolution based on traditional text data, as posts (e.g., tweets) are usually written in an informal manner rife with abbreviations, misspellings, grammatical errors, and noise information. Tracking the evolution of an event through streaming social media posts is challenging in terms of both event extraction and precise, incremental event determination. Lee et al. [44] used a sliding time window in an attempt to accomplish this, wherein they approached event pattern detection and tracking from social streams in terms of birth/death, growth/decay, and merge/split characteristics.

There is a great deal of structure information attached to social posts (e.g., users, tags) to consider in addition to their textual content and time stamp. Bursty event detection from social media content is an important research subject in the event evolution map field. Yao et al. [45] proposed a bursty event detection model which comprehensively explores multiple sources of collaborative contexts based on tag, user, and post data in tandem. Users' tagging actions over time reflect their changing interests. Monitoring and analyzing the temporal patterns of tags can provide useful insight into certain "hot topics" being discussed online. Yao et al. [46] also proposed a bursty tagging event detection model which exploits the co-occurrence and relationships among bursty tags. This model can be divided into three steps: Bursty feature extraction, bursty event clustering, and bursty event evolution map construction.

3.2.3. Other event evolution models

When a single document can be viewed as an event, the event evolution is tracked with a "small granularity" . This type of event evolution modeling includes Events Timeline Analysis (ETA) [49], Evolutionary Timeline Summarization (ETS) [50], connect-the-dots [51], metro map [14,15], and storyline generation.

The timeline temporally summarizes evolutionary news as a series of individual but correlated component summaries. It is beneficial to automatically generate high-quality timelines from a collection of various news sources. ETA [49] can discover event evolution relationships by automatically organizing news events in chronological time order and identifying dependencies between events. ETA helps the user to conveniently and quickly browse news event evolution in graph form and incrementally updates the event evolution process so as to better fit reflect the streaming features of news on the internet.

Yan et al. [50] presented a novel framework for the web mining problem called ETS in which a events are organized solely by their time stamps along a timeline. Neighboring events in a timeline may not be coherent, however, so it is difficult for the user to identify the relationship of events from event threads. Event storyline generation involves summarizing a collection of web documents by extracting representative information based on all the sub-events in order to generate a global summary.

The event timelines generated via the above methods are linear in structure and suited only to simple event evolution tracking. Complex events, conversely, break down into branches and side stories and necessitate similarly complex evolution maps.

Shahaf et al. [14,15,51,52,53] addressed the information overload problem by providing an easy way to navigate within a new topic by creating structured summaries of information called "metro maps". These maps generate useful story chains of complex topics that are coherent, comprehensive, and with high connectivity.

The "connect-the-dots" concept [51] was first proposed for discovering coherent story chains from a series of news articles for a specific topic given the two terminals of the chain. Details regarded this process were provided by Ref. [52], including a relevant optimization algorithm and specific experimental results. The drawbacks to Shahaf et al.'s model [51] include 1) rather low efficiency (where it takes 10 min to create a chain 6 or 7 in length even with a speed-up algorithm) and 2) a failure to address the redundancy problem caused by including multiple articles for a single event in the story chain.

To address these drawbacks, Zhu et al. [54] improved the system for discovering story chains using four criteria: Relevance, coherence, coverage, and redundancy. Shahaf et al. later extended the algorithm for story chain identification by providing the two endpoints to structured summaries of information (news and scientific literature) in metro maps [14,15]. The metro map visualizes the progress of a news topic by organizing both key events and side stories into several event threads. Given a query, the metro map algorithm generates a concise structured set of documents which maximizes the influence, coverage, and connectivity of salient pieces of information. The largest difference between metro maps for news articles and scientific literature is that the side information provided by the citation graph (which is suited only to the latter) is utilized to define coherence and coverage. Shahaf et al.'s latest work [53] extended metro map notion to information cartography to establish the so-called "zoomable metro map", in which each level of zoom shows finer details and interactions.

Metro maps can provide very useful information efficiently and approach the topic evolution map process with fine granularity of individual articles or papers. They can be extended, however, to pursue richer forms of input, output, and interaction; it may also be possible to incorporate higher-level semantic relations into these frameworks. Hu et al. [55], for example, proposed a novel approach to discovering salient storyline interactions to form a clear, global picture of news topics. Storyline interactions that are crucial to characterize the connections among topics can reveal the latent connections among various aspects of the topics.

The event evolution modeling methods described above work based individual documents and attempt only to track event evolution via regular text objects. Tracking event evolution on social media, as discussed above, is a more challenging endeavor. Ref. [56] proposed the innovative generating storylines from microblogs (GESM) method for user input queries to provide both better user experience and deeper understanding of real-time events. The GESM model builds a real-time storyline of the event through a two-level solution. The first level is used to retrieve the related events per the language used to refer to them with dynamic pseudo relevance feedback; the second level generates the storyline via graph optimization. GESM processes the social media in a static form, while Lee's model [44] tracks the evolution of events from online social post streams (such as Twitter timelines and forum discussions) in stream format. The main drawback of GESM is that it generates storylines based on multi-view post graphs in a manner that is neither concise nor coherent compared to connect-the-dots [51]. GESM is relatively new in regards to event evolution research tailored to social media, so it is reasonable to assume that there will be substantial improvements made to the model in the near future.

3.2.4. Comparison among event evolution map models

Table 3 provides a comparison among the typical topic (event) evolution models described in this subsection. We compare the difference of the model from the type of event, the structure of event evolution, the research data type and the characteristic of the models. The first six models listed in Table 3 cluster documents into individual events; the other seven models consider a single document as an event. All models except for Lee's model [44], Qiu's model [49], Yan's model [50], and GESM are only suit-

Table 3
Comparison among typical cross-media topic evolution models.

Models	Event type	Data type	Evolution structure	Characteristic of the models
EP discovery	Cluster	News article	Non-linear,graph in 2D	Discovery event Eps from document sequences
TSCAN	Cluster	News article	Non-linear,graph in 2D	Form evolution graph by semantic and temporal relationships
StoRe	Cluster	News article	Linear,event storyline	Generates storyline from the representative sentences
MEA	Cluster	News article	Linear,event storyline	Generate storyline by distinguish different part-whole relationship between the sub-events and the major event.
Lee's model [44]	Cluster	Tweets data	Non-linear,graph in 2D	Track event evolution patterns(birth/death,growth/decay, merge/split) from social streams,suit for online data
Yao's model [45]	Cluster	Social media	Linear	Comprehensively explores multiple sources of collaborative context to detect bursty event
Yao's model [46]	Cluster	Social media	Linear	Exploits the bursty tag co-ocurrence and correlationship to detect bursty tagging event
ETA	Single	Web news	Non-linear,graph in 2D	Automatically organize news events by time order and dependencies between events,suit for online data
ETS	Single	Web news	Linear,event timeline	Track event evolution along the timeline,relevance,coverage, coherence and diversity,suit for online and offline data
CTD[51]	Single	News article	Linear,story chain	Connect two news paper using a chain of coherence
Metro map	Single	News article science article	Non-linear,graph in 2D	Visualize the progress of a news topic which maximizes influence,coverage and connectivity of information
Zhu's model [54]	Single	News article	Linear,story chain	Discovery story chains using four criteria:relevance, coherence,coverage and redundancy,fast algorithm
Hu's model [55]	Single	News article	Non-linear,graph in 2D	Discover salient storyline interactions to track event evolution
GESM	Single	Tweets data	Linear,event storyline	Generate storylines from microblogs for user input queries

able for traditional (text-based) news. EP discovery, TSCAN, ETA, metro map, and Lee's model [44], Hu's model [55] can discover non-linear evolution structures of complex events. ETA, ETS, and Lee's model [44] can be used for online data. Qiu's model [49] and Yan's model [50] can employ multiple sources of information from social media for event evolution mining.

The topic evolution map methods described above all work based on single types of media, mostly text – however, the rapid proliferation of internet and social media (e.g., Twitter, Facebook, YouTube) technologies has resulted in a wealth of cross-media data which represents the same semantics in reference to the real world. Effective cross-media topic analysis techniques are urgently necessary to facilitate comprehensive and useful knowledge discovery across multiple types of media. Topic evolution map based on cross-media datasets has played an increasingly important role in the research community's attempt to solve the information overload problem.

4. Topic evolution maps based on cross-media datasets

There has already been a great deal of research on modeling topic evolution within cross-media data. At present, representative modeling methods are mainly focused on text and video/image modalities. Some methods are similar to those utilized for text-only data, but because cross-media data has completely different characteristics, modeling topic evolution within it calls for an entirely unique approach. Cross-media topic or event detection still involves topic evolution map construction, but the process differs from that used for text data.

The earliest work on topic evolution map based on cross-media data dates back to Neo et al.'s work [57]. They proposed an improvement to previous news video searching techniques by combining multimodal event information extracted from video data, web news articles, and news blogs to support event evolution analysis and perform question answering (QA). The primary difference between these models and the models described in Section 3 are that the former first extract the event using multi-modal information from news videos and external news articles simultaneously. In other words, their definition of events is based on text features in addition to high-level features such as visual concepts.

Kim et al. [58] proposed a nonparametric approach to the modeling and analysis of temporal topic evolution within web image collections; theirs is the seminal work which considers the timestamps associated with images as a main research subject to uncover dynamic behavior in the images. Different media can also be cross-hyperlinked to visualize and explore video search results. To this end,

Tan et al. [59] investigated the synchronization of multiple media content in the physical form of hyperlinking them. They utilized techniques including content mining and selection from web videos, space-time alignment of multiple media, and augmenting of search results with "when" and "what" information to develop three types of browsing systems: 1) Timeline-based visualization, 2) wiki add-ons, and 3) galaxy browsing.

The dramatic growth of social media has made the effective browsing and searching of videos a challenging task. Wu et al. [60] extended TDT research to cross-media data by mining event evolution structures from web video search results. They explored event discovery and structure construction from a long list of videos by automatically providing a concise structure showing the evolution of events associated with the representative text keywords and visual imagery. Within their system framework, the event structure is constructed from the web videos returned from a search engine.

Unlike traditional text summarization and timeline generation systems [44,49], Wang et al. [61] introduced storyline-based sum-

maries to reflect the evolution of the given topic by integrating text, image, and temporal information taken from online sources. The generated storylines achieve both temporal continuity and content coherence, yielding richer information and better result representation to the user. They employ the Steiner trees algorithm to generate storylines, in which neighboring items are selected and assigned to a spanning tree. Steiner trees do not guarantee a global theme to the storyline, however, so there is no global notion of coherence in this type of model.

Shan et al. [62] proposed the EventSearch system, which can be used to extract cross-media event evolution from news-related data, i.e., web news articles, newspapers, TV news programs, and Weibo posts. This system can be applied to not only offline event extraction but also online event discovery. The event evolution maps yielded by the EventSearch model have linear structure.

Timelines represent information on a linear axis, thus allowing the user to easily follow the evolution of an event or to distinguish between different events while providing a global view. However, the one-dimensional timeline only works for simple stories that are linear in nature. Borrowing the idea of recommendation in heterogeneous network into the cross-modal news summarization. Xu et al. [63] tackle this task and bring out a 1-D cross-media timeline generation framework. Sahuguet et. al. [64] propose an approach to automatically generate a timeline of popular events related to a given topic by mining search behavior and video information. Xu et al. [65] proposed a novel solution to the extraction and reconstruction of storylines for non-linear, complex event summaries based on cross-media datasets. Instead of using the whole document as a story node, the story map is extracted at the sentence level making the process quite different from the metro map extraction process [14,15]. Experiments on four separate datasets indicated that the approach is successful.

The cross-media topic evolution maps described above apply to single media. It is important to consider the difference between cross-media data and single-media (e.g., text) data, however. As discussed at length above, cross-media topic evolution map is considerably more challenging than traditional, text-based topic evolution map. Table 4 provides a comparison between several such map methods in regards to data type, evolution type, evolution structure, and other characteristics.

There are many types of cross-media data, such as crossing text and image modalities or text and audio modalities. Further, traditional cross-media data and social cross-media data have distinguishing features. Topic evolution research based on cross-media data remains a highly challenging endeavor. To date, this research has mostly involved simply extending text media topic extraction methods to cross-media data – with mixed results. There remains much work to be done in terms of unique and effective cross-media topic evolution map.

Here, we summarize the 41 mainstream topic evolution models from regular text data to social media data to cross-media data. Fig. 4 shows a representative evolution map of the models described in Tables 1, 3 and 4 (Sections 3 and 4); the blue, orange, and red railway lines in the figure represent the extant research on certain types of models (PGM-based topic evolution, topic evolution map based on event evolution, and cross-media topic evolution map models, respectively). Green nodes in the map represent the models suited to processing social media, while purple nodes represent models suited to online streaming data. There are nine research directions marked in blue, six in orange, and six in red. The intersections of different lines of the same color represent the relationships among different research methodologies in the same direction, while the intersections of different lines in different colors represent the relationships among different research methodologies in different directions.

Table 4
Comparison among typical cross-media topic evolution models.

Models	Cross-media data type	Evolution structure	Characteristic
Ref. [57]	News video, web news	Linear, 1D map	Topic evolution browsing and QA from both news video and news article.
Ref. [58]	Web image	Linear, 1D map	The first work for image topic evolution,capture the subtopic evolution in the form of the similarity network of the image set.
Ref. [59]	Web video, news article	Linear, 1D map	Topic evolution tracking with rich media information mined from various knowledge sources.
Ref. [60]	Web video	Non-linear, 2D map	Performed the event structure mining on the basis of text co-occurrence and visual feature trajectory.
Ref. [61]	Images, text	Non-linear, 2D map	Summarize the evolution of topics in a pictorial and structural way using both text and image information.
Ref. [62]	Web news, newspaper, TV program, Weibo	Linear, 1D map	EventSearch, a system for event extraction and retrieval on four types of news-related historical data.
Ref. [63]	Web news with image	Linear, 1D map	Automatically generating cross-media timeline.
Ref. [64]	Web video	Linear, 1D map	Automatically generate a timeline of popular events related to a give topic.
Ref. [65]	Web news with image	Non-linear, 2D map	Storyline extraction and reconstruction for non-linear complex event summarization.

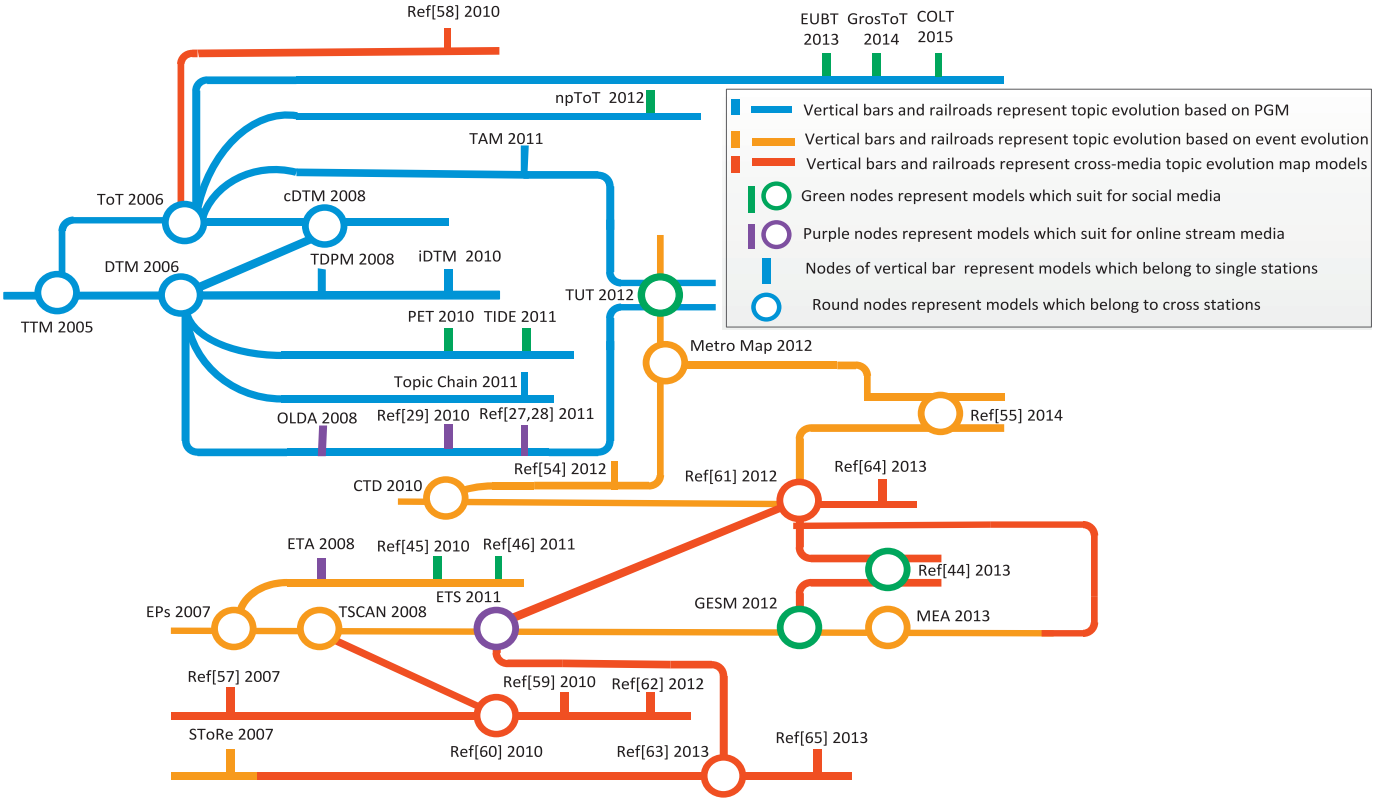

Fig. 4. Evolution map of models described in Tables 1, 3 and 4. The blue roads represent the research directions of topic evolution based on PGM; the orange roads represent the research directions of topic evolution based on event evolution; the red roads represent the research directions of cross-media topic evolution models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Conclusion

In this paper, we discussed the problematic aspects of extant topic evolution maps from two aspects: First, within maps based on text datasets; and second, within maps based on cross-media and social media datasets. The former includes two sub-categories: 1) Evolution over time based on PGMs such as TTM [7], DTM [18], and subsequent models based on them which approach topic evolution with large granularity and linear storyline structures; and 2) evolution with relatively small granularity and simple structures such as timelines, storylines, and connect-the-dots as well as complex structures such as metro maps. The latter category is relatively new and unsophisticated. It includes topic evolution map visualization based on cross-media corpora such as text combined with images and text combined with videos. Researchers in this area tend to utilize the NDK of video and other computer image mining technology as opposed to text-mining technologies. Although research on topic evolution map in cross-media data has yielded notable achievements, there still much work to be done.

6. Open questions and future work

6.1. Open questions

The information overload problem, as described above, is a popular research subject often approached through topic evolution mapping. Topic evolution is nonlinear in nature but is derived from linear structures such as timelines and storylines. Topic evolution models have gradually been extended from single media (mainly text data) to a variety of media data and cross-media objects. Topic evolution map research on cross-media data has yielded notable results, but there is much work to be done. The persistent problems with existing models can be summarized as follows.

(1) There is no generally accepted model checking or evaluation benchmark for cross-media topic evolution models. Perplexity is generally used to evaluate PGM-based topic evolution models, where low scores indicate favorable hold-out performance. There is no coherence between model perplexity and the perceived semantic importance of topic evolution results, however. Topic evolution models that show better held-out qualities may actually represent less semantically meaningful topic evolution results [66], so comparison experiments among various models are difficult to conduct (and potentially meaningless).
(2) There is no unified modeling framework appropriate for topic evolution map across different types of data. The PGM-based topic model, which is based on media or cross-media the “bag of words” assumption, is not suitable for social media or cross-media corpora. PGM-based models do not perform well on short and user-generated text such as Twitter or Weibo posts. There is no universal modeling framework for cross-media topic modeling which can balance the semantic gap between different data modalities.
(3) Current research on topic evolution map is limited to qualitative analysis, mainly based on the correlation and chronological order between topics or events. Quantitative analysis of the evolution of mapped topics would provide better insight into the precise interactions in individual or multiple topics, as well as the probability of topic inference from the topic evolution map.
(4) Cross-media topic evolution map application is still limited to event timelines, video searches, and event summarization.

It remains unclear if there is any other application for cross-media topic evolution maps.

6.2. Future work

There are several potential research directions worth pursuing in the future.

(1) To answer the first open question mentioned above, new model checking and evaluation algorithms for cross-media topic evolution maps should be developed. It is especially necessary to design algorithms with a balance between held-out model performance and semantically meaningful topic evolution results. An evaluation benchmark for cross-media topic evolution map models must be established to facilitate accurate performance evaluation among various models, as well.
(2) To answer the second open question mentioned above, it is necessary to establish a unified cross-media data representation for the same semantic concept from different media. A "law" that governs the topic evolution map over time in different media is necessary to accomplish this. A unified modeling framework for cross-media topic evolution map necessitates new data mining techniques (e.g., deep learning, transferring learning). The "bag of words" topic model does not consider the word order or semantic relationships between words in a sentence and context, though this information may be very useful. With the development of Natural Language Processing (NLP) technology, especailly NLP applied in data mining and knowledge discovery, for instance, the Journal of Knowledge-Based Systems held a special issue on New Avenues in Knowledge Bases for Natural Language Processing in 2016 [67]. In the future, more NLP related technology could be applied to topic evolution modeling in order to produce more robust and versatile findings.
(3) To answer the third open question mentioned above, research on topic evolution relationships should be conducted from qualitative analysis to quantitative analysis and probability inference. One possible approach to this is to treat the topic map as a DAG and exploit Bayesian network theory for topic probability inference and topic cause-effect inference.
(4) Several new applications for cross-media topic evolution research are necessary to answer the fourth open question mentioned above. Applying cross-media data to build smart cities is a particularly interesting and potentially significant application area. Concrete examples of this application include intelligent public security, public opinion analysis systems, and intelligent traffic systems.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions, which significantly contributed to improving the manuscript. This work was supported by the National Key Basic Research Project of China (973 Program No. 2012CB316400), the National Nature Science Foundation of China (No.61471321, No.61202400, No.31300539, No.31570629), the Zhejiang Provincial Natural Science Foundation of China (No.LY15C140005, No.LY16F010004), Science and Technology Department of Zhejiang Province Public Welfare Project (No.2016C31G2010057, No.2015C31004), Fundamental Research Funds for the Central Universities (No.172210261) and the Zhejiang Provincial Key Laboratory of Forestry Intelligent Monitoring and Information Technology Research.

References

[1] Y. Wei, Y. Zhao, Z. Zhu, Y. Xiao, S. Wei, Learning a mid-level feature space for cross-media regularization, in: Multimedia and Expo (ICME), 2014 IEEE International Conference on, IEEE, 2014, pp. 1–6.
[2] S. Wang, Z. Wang, S. Jiang, Q. Huang, Cross media topic analytics based on synergetic content and user behavior modeling, in: 2014 IEEE International Conference on Multimedia and Expo (ICME), IEEE Computer Society, 2014, pp. 1–6.
[3] J. Allan, Introduction to topic detection and tracking, Inf. Retrieval 12 (2002) 1–16.
[4] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, J.A. Umass, et al., Topic detection and tracking pilot study final report, in: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998, p. 194–218.
[5] R. Nallapati, A. Feng, F. Peng, J. Allan, Event threading within news topics, in: Proceedings of the Thirteenth ACM International Conference on Information And Knowledge Management, 2004, pp. 446–453.
[6] S. Morinaga, K. Yamanishi, Tracking dynamics of topic trends using a finite mixture model, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 811–816.
[7] Q. Mei, C.X. Zhai, Discovering evolutionary theme patterns from text: an exploration of temporal text mining, in: Proceedings of KDD '05, 2005, pp. 198–207.
[8] X. Wang, A. Mccallum, Topics over time: a non-markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 424–433.
[9] Q Mei, C.X. Zhai, A mixture model for contextual text mining, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 649–655.
[10] R. Kumar, U. Mahadevan, D. Sivakumar, Research track paper a graph-theoretic approach to extract storylines from search results, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 216–225.
[11] F.R. Lin, F.M. Huang, C.H. Liang, Individualized storyline-based news topic retrospection, in: PACIS 2007 Proceedings, 2007.
[12] A. Ahmed, Q. Ho, C.H. Teo, J. Eisenstein, A.J. Smola, E.P. Xing, Online inference for the infinite topic-cluster model: storylines from streaming text, in: International Conference on Artificial Intelligence and Statistics, 2011, pp. 101–109.
[13] Houkui Zhou, Huimin Yu, Roland Hu, Topic discovery and evolution for scientific literature based on content and citations, Front. Inf. Technol. Electron. Eng.[J] (2016) accept.
[14] D. Shahaf, C. Guestrin, E. Horvitz, Trains of thought: generating information maps, in: Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 899–908.
[15] D. Shahaf, C. Guestrin, E. Horvitz, Metro maps of science, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 1122–1130.
[16] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[17] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn.. 42 (1) (2001) 177–196.
[18] M. David, Dynamic topic models, in: International Conference on Machine Learning, 2006, pp. 113–120.
[19] A. Ahmed, E. Xing, Dynamic non-parametric mixture models and the recurrent chinese restaurant process, in: SDM, 2008, pp. 219–230.
[20] A. Ahmed, E.P. Xing, Timeline: a dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream, in: Proceedings of the 26th International Conference on Conference on Uncertainty in Artificial Intelligence (UAI), 2010.
[21] X. Tang, C.C. Yang, TUT:a statistical model for detecting trends,topics and user interests in social media, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 972–981.
[22] L. Bolelli, Ş. Ertekin, CL. Giles, in: Topic and Trend Detection in Text Collections Using Latent Dirichlet allocation. Advances in Information Retrieval, Springer, Berlin Heidelberg, 2009, pp. 776–780.
[23] M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths, Probabilistic author-topic models for information discovery, in: Tenth ACM Sigkdd International Conference on Knowledge Discovery & Data Mining, 2004, pp. 306–315.
[24] MichalRosen-Zvi Uc, Michal Rosen-zvi, Thomas Griffiths, The author-topic model for authors and documents, in: Conference on Uncertainty in Artificial Intelligence, 2004, pp. 487–494.
[25] Houkui Zhou, Huimin Yu, Roland Hu, Topic evolution based on the probabilistic topic model:a review, Front. Comput. Sci.[J] (2016) accept, doi:10.1007/s11704-016-5442-5.
[26] L. AlSumait, D. Barbará, C. Domeniconi, On-line lda: adaptive topic models for mining text streams with applications to topic detection and tracking, in: Data Mining,2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008, pp. 3–12.
[27] A. Ahmed, Q. Ho, J. Eisenstein, Unified analysis of streaming news, in: Proceedings of the 20th International Conference On World Wide Web, 2011, pp. 267–276.
[28] A. Ahmed, Q. Ho, C.H. Teo, J. Eisenstein, A.J. Smola, E.P. Xing, Online inference for the infinite topic-cluster model: storylines from streaming text, in: International Conference on Artificial Intelligence and Statistics, 2011, pp. 101–109.
[29] A. Gohr, A. Hinneburg, R. Schult, M. Spiliopoulou, Topic evolution in a stream of documents, SDM 9 (2009) 859–872.

[30] D. Kim, A. Oh, Topic chains for understanding a news corpus, in: Computational Linguistics and Intelligent Text Processing, Springer, Berlin Heidelberg, 2011, pp. 163–176.
[31] C. Wang, D. Blei, D. Heckerman, Continuous time dynamic topic models, The 23rd Conference on Uncertainty in Artificial Intelligence, 2008.
[32] N. Kawamae, Trend analysis model: trend consists of temporal words, topics, and timestamps, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 2011, pp. 317–326.
[33] A. Dubey, A. Hefny, S. Williamson, E.P. Xing, A non-parametric mixture model for topic modeling over time, SDM (2013) 530–538.
[34] C.X. Lin, Q. Mei, J. Han, Y. Jiang, M. Danilevsky, The joint inference of topic diffusion and evolution in social communities, in: 2011 11th IEEE International Conference on Data Mining, IEEE Computer Society, 2011, pp. 378–387.
[35] Fei-Fei L., Perona P. A Bayesian hierarchical model for learning natural scene categories. (2005). IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2: 524–531.
[36] C.X. Lin, B. Zhao, Q. Mei, J. Han, PET: a statistical model for popular events tracking in social communities, in: ACM Sigkdd International Conference on Knowledge Discovery & Data Mining, 2010, pp. 929–938.
[37] H. Yin, B. Cui, H. Lu, Y. Huang, J. Yao, A unified model for stable and temporal topic detection from social media data, in: IEEE, International Conference on Data Engineering, 48, 2013, pp. 661–672.
[38] Z. Hu, J. Yao, B. Cui, User group oriented temporal dynamics exploration, Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Press, 2014.
[39] Z. Hu, J. Yao, B. Cui, E. Xing, Community level diffusion extraction, in: ACM SIGMOD International Conference on Management of Data, 2015, pp. 1555–1569.
[40] T.L. Griffiths, M. Steyvers, Finding scientific topics, in: Proceedings of the National Academy of Sciences of the United States of America, 1, 2004, p. 5228.
[41] C.P. Wei, Y.H. Chang, Discovering event evolution patterns from document sequences, Syst. Man Cybern. Part A 37 (2) (2007) 273–283.
[42] C.C. Chen, M.C. Chen, TSCAN: a novel method for topic summarization and content anatomy, in: Proceedings of the 31st Annual International Acm Sigir Conference on Research and Development in Information Retrieval, 2008, pp. 579–586.
[43] L. Huang, Lian'en Huang, Optimized event storyline generation based on mixture-event-aspect model, EMNLP (2013) 726–735.
[44] P. Lee, L.V. Lakshmanan, E.E. Milios, Event Evolution Tracking from Streaming Social Posts, 2013 arXiv preprint arXiv:1311.5978.
[45] J. Yao, B. Cui, Y. Huang, X. Jin, Temporal and social context based burst detection from folksonomies, Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI, 2010 2010July.
[46] J. Yao, B. Cui, Y. Huang, Y. Zhou, Bursty event detection from collaborative tags, World Wide Web 15 (2) (2012) 171–195.
[47] C.C. Yang, X. Shi, C.P. Wei, Discovering event evolution graphs from news corpora, Syst. Man Cybern. Part A 39 (4) (2009) 850–863.
[48] C.C. Yang, X. Shi, C.P. Wei, Tracing the event evolution of terror attacks from on-line news, in: Intelligence and Security Informatics, Springer, Berlin Heidelberg, 2006, pp. 343–354.
[49] J. Qiu, C. Li, S. Qiao, T. Li, J. Zhu, Timeline analysis of web news events, Adv. Data Min. Appl. 5139 (2008) 123–134.
[50] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, Y. Zhang, Evolutionary timeline summarization: a balanced optimization framework via iterative substitution, in: Proc Sigir, 2011, pp. 745–754.
[51] Dafna Shahaf, Carlos Guestrin, Connecting the dots between news articles, in: ACM Sigkdd International Conference on Knowledge Discovery & Data Mining, 2010, pp. 623–632.
[52] D. Shahaf, C. Guestrin, Connecting two (or less) dots: discovering structure in news articles, ACM Trans. Knowl. Discov. Data 5 (4) (2012) 74-74.
[53] D. Shahaf, J. Yang, C. Suen, J. Jacobs, H. Wang, J. Leskovec, Information cartography: creating zoomable, large-scale maps of information, in: ACM Conference on Knowledge Discovery and Data Mining (KDD, 2013, pp. 1097–1105.
[54] X. Zhu, T. Oates, Finding story chains in newswire articles. information reuse and integration (IRI), in: 2012 IEEE 13th International Conference on, 330, IEEE, 2012, pp. 93–100.
[55] P. Hu, M.L. Huang, X.Y. Zhu, Exploring the interactions of storylines from informative news events, J. Comput. Sci. Technol. 29 (3) (2014) 502–518.
[56] C. Lin, C. Lin, J. Li, D. Wang, Y. Chen, T. Li, Generating event storylines from microblogs, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, 2012, pp. 175–184.
[57] S.Y. Neo, Y. Ran, H.K. Goh, et al., The use of topic evolution to help users browse and find answers in news video corpus, in: Proceedings of the 15th International Conference on Multimedia, 2007, pp. 198–207.
[58] G. Kim, E.P. Xing, A. Torralba, in: Modeling and Analysis of Dynamic Behaviors of Web Image Collections, 6315, 2010, pp. 85–98.
[59] S. Tan, C.W. Ngo, H.K. Tan, L. Pang, Cross media hyperlinking for search topic browsing, in: Proceedings of the 19th ACM International Conference on Multimedia, ACM, 2011, pp. 243–252.
[60] X. Wu, Y.J. Lu, Q. Peng, C.W. Ngo, Mining event structures from web videos, Multimedia IEEE 18 (1) (2011) 38–51.
[61] D Wang, T Li, M Ogihara, Generating Pictorial Storylines Via Minimum-Weight Connected Dominating Set Approximation in Multi-View Graphs, AAAI, 2012.
[62] D. Shan, W.X. Zhao, R. Chen, B. Shu, Z. Wang, J. Yao, et al., EventSearch: a system for event discovery and retrieval on multi-type historical data, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 1564–1567.
[63] S. Xu, L. Kong, Y. Zhang, A cross-media evolutionary timeline generation framework based on iterative recommendation, in: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, 2013, pp. 73–80.
[64] M. Sahuguet, B. Huet, Socially motivated multimedia topic timeline summarization, in: Proceedings of the 2nd International Workshop on Socially-Aware Multimedia, 2013, pp. 19–24.
[65] S. Xu, S. Wang, Y. Zhang, Summarizing complex events:a cross-modal solution of storylines extraction and reconstruction, in: EMNLP, 2013, pp. 1281–1291.
[66] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D. Blei, Reading tea leaves: how humans interpret topic models, in: Proceedings of the Advances in Neural Information Processing Systems, 2009, pp. 288–296.
[67] E. Cambria, B. Schuller, Y. Xia, B. White, New avenues in knowledge bases for natural language processing, Knowl.-Based Syst. 108 (2016) 1–4.