www.ijecs.in

International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 11 Nov. 2016, Page No. 18987-18992

Social Event Storyboard Generation from Image Click-A Survey

Sreelekshmi.U[1] *Gopu Darsan* [2]

[1]lekshmishibin@gmail.com , [2] gops601@gmail.com

Department of Computer Science and Engineering in Sreebuddha College of Engineering, Pattoor, Alappuzha, kerala

Abstract

Traditional websites were driven by human-edited events which lead to huge web search traffic. This paper is a survey conducted for identifying the various event detection methods which are useful for event mining. Moreover this paper also suggests an automatic system to detect events from search log data and generate storyboards where the events are arranged along a timeline. Image search log is treated as a good data resource for event mining, as search logs directly reflects people’s interests. In order to discover events from log data, an approach known as Smooth Nonnegative Matrix Factorization framework (SNMF) is used. Moreover, time factor is considered as an important element for event detection as different events develop at different time. In addition, to provide a visually appealing storyboard, each event is mapped with a set of relevant images arranged along a timeline. These relevant images are automatically generated from image search results by analyzing both local and global image content feature

Keywords: *Event storyboard, Social media, Click-through data, Non-negative matrix Factorization, Image search.*

1. Introduction

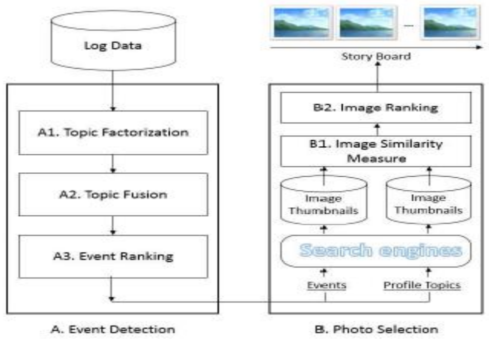
The events are detected from search log data and generate story boards where events are arranged along a time line. It is found that search log data is a good data resource for event detection because: (1) search logs cover a wide variety of real world events

(2) search log directly reflect user’s interests (3) search log respond to real time events.

To discover events from log data, an approach called Smooth Non-negative Matrix Factorization (SNMF) framework is used. There are two basic ideas for SNMF: (1) It promotes event queries (2) It differs events from popular queries. SNMF guarantee weights for each topic to be non-negative and considers time factor for event development. To make event detection easier, relevant images are attached for each event.

There are two phases for the proposed approach: Event detection by SNMF and Event photo selection. In event detection, initially events are searched from log data. Then it discovers groups of queries that have high frequency which is known as topic factorization. Next topics with similar behaviors are merged together along a timeline which is called topic fusion. Event ranking happens in which topics like social events are highlighted. After ranking top topics are called social events and non top topics are called profile topics.

In event photo selection, both the social events and profile topics are sent to search engines like Google or Bing. The search engines generate two sets of image thumbnails which contains relevant images to social events. Image similarity measures occur in which similarity between events and images are measured. Image ranking is done which is sorting of images in the social event image set. Finally all social events together with their images construct a storyboard.



This figure is taken from the paper Automatic Generation of Social Event Storyboard from Image Click-through Data proposed by JunXu et.al [6]. In SNMF topic factorization, the log data is converted into a matrix V of size W x H. Each row in matrix V represents a query and each column indicates one day. Every item Vij represent ith query on jth day. In matrix W, each column represents topic and k indicates number of topics. In matrix H, each column represents decomposed coefficient of topic for a day.

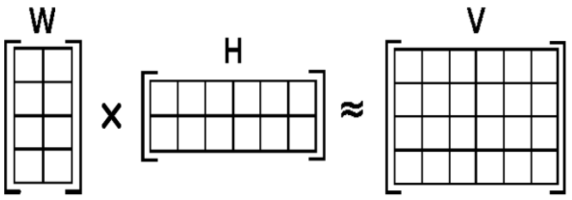


Illustration of approximate non-negative matrix factorization: the matrix *V* is represented by the two smaller matrices *W* and *H*, which, when multiplied, approximately reconstruct *V*.

There is no significant difference between queries from two adjacent days. To achieve this constraint, an approach known as SNMF is introduced. SNMF includes an extra regularization factor, S(H).This factor smoothen two adjacent columns in matrix H. Thus it provides a non-negative weight that adjusts the degree of smoothing. The number of topics, k should be large so that some social events are not missed. If k is large then there is a risk of over-splitting topics. is avoided by topic fusion. In topic fusion, similarity between topics is measured over queries, timeline and search log URLs. Then similar topics are merged in a bottom up way by means of agglomerative hierarchical clustering. Later in event ranking, it distinguishes event related topics from others.

$$Rank_{to} = Score_{ti} \times Score_{q} \times Score_{u}$$

Where $Rank_{to}$ = ranking score of a topic

$Score_{ti}$ = timeline based ranking score

$Score_{q}$ = query based ranking score

$Score_{u}$ = URL based ranking score

To get event related images, directly search image search engines with event queries. But we obtain a lot of irrelevant images. Therefore a better way is needed to collect images for events. Thus there are two steps that identify images that represent the event in question: Image similarity measures and Event photo re-ranking. Image similarity measures consider both local and global image features. Global feature is identified by Block-based intensity histogram and local feature is measured by SIFT (Scale Invariant Feature Transform)

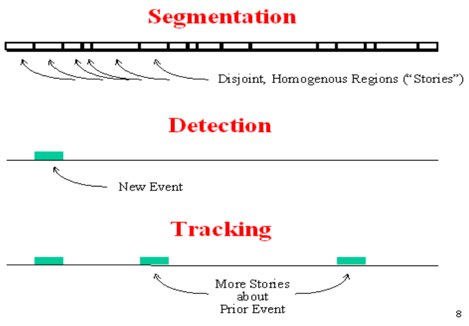
The remaining part of the paper is organized as follows. In Section 2 the survey of different methods will be described. The paper concludes with a brief summary in section 3.

2. Literature Survey

Data mining is the process of semiautomatic ally searching large databases to find patterns that are novel, valid, useful and understandable. The goal of data mining is to extract information from a dataset & transform it into an understandable structure. It is also known as Knowledge Discovery in Databases (KDD).The stages in data mining are: Problem definition, Data gathering and preparation, Model building and evaluation, Knowledge deployment.

I. Topic Detection and Tracking

Topic Detection and Tracking (TDT) [1] is a process which involves the exploration of techniques to detect new topics and track their reappearance and evolution. There are three technical tasks in TDT: Segmentation, Detection and Tracking.


Segmentation
Disjoint, Homogenous Regions ("Stories")
Detection
New Event
Tracking
More Stories about Prior Event
8

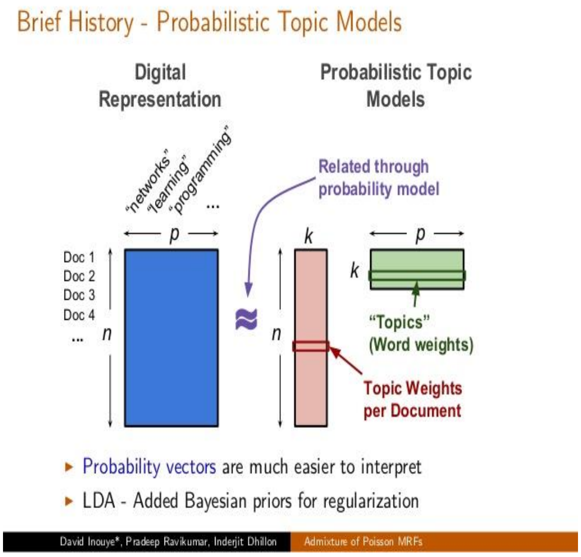
Segmentation is the process of breaking down a continuous stream of text into disjoint, homogenous regions called stories. Detection is the process of identifying new events. Tracking is the process of finding more stories about prior event. There are two types of event detection: Retrospective event detection and online new event detection. In retrospective event detection, stories are grouped into clusters where each cluster represents an event. In online new event detection, it identifies new events in a stream of stories. A decision is made after each story is processed. If the story discusses a new event then it is flagged as YES otherwise NO. This approach is useful for timely information access applications like Yahoo news. Some open issues regarding this approach are: how to choose right level of clusters for users that best fit their information need, how to provide navigation tools for effective and efficient search, how to improve accuracy of on-line detection by introducing limited look-ahead.

II. Event Detection in Twitter

J.Weng et.al proposed Event detection in Twitter [2] which involves Event Detection with Clustering of Wavelet-based signals (EDCoW).The components of EDCoW are: Build signals for individual words, Filter away trivial words and Cluster signals. In order to build signals for individual words, wavelet transformation is used which consists of CWT and DWT. Continuous Wavelet Transformation (CWT) provides a redundant representation of signal. Discrete Wavelet Transformation (DWT) provides a non redundant representation of signals. Then filtering away trivial words is achieved through Auto correlation and Cross correlation. A mathematical tool used to find repeating patterns is called auto correlation. Another tool that searches for a long signal for a shorter known feature is known as cross correlation. Later clustering of signals is achieved by Modularity based graph portioning and Newman algorithm. In Modularity based graph partitioning, it detects events by clustering signals. Newman algorithm detects and removes edges connecting different events. Some advantages of this approach are: Wavelet analysis takes less storage space and EDCoW gives good performance. The disadvantages of this approach are: how to analyze the relationship among users that could contribute to event detection and how to introduce time lag and study the interaction between different words.

III. Introduction to Probabilistic Topic Models

In the paper, Introduction to probabilistic topic models,[3] a topic represents a probability distribution over words. Related words will get high probability in the same topic. In the figure, there are a set of n documents whose digital representation is shown on the left side. These n documents can be related through a probability model as shown on the right side of the figure. In the probabilistic topic model, from the n documents, per document each topic, k is assigned weight and per topic, k each word, p is assigned weight.

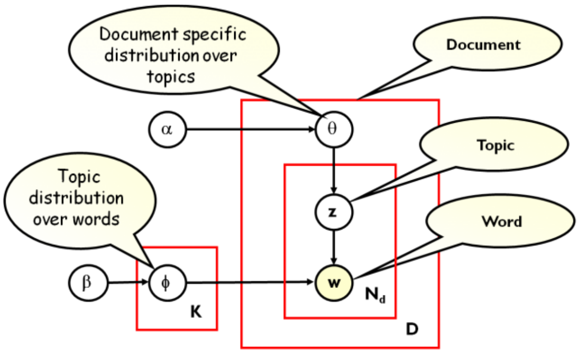


Brief History - Probabilistic Topic Models

- Probability vectors are much easier to interpret
- LDA - Added Bayesian priors for regularization

David Inouye*, Pradeep Ravikumar, Inderjit Dhillon | Admixture of Poisson MRFs

LDA (Latent Dirichlet Allocation) is the simplest topic model. It is a statistical model of document collections. It is defined by statistical assumptions like: Order of words in the document does not matter, Order of documents does not matter & number of topics is assumed known & fixed.

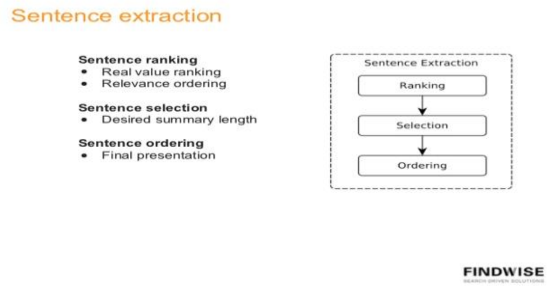


Document specific distribution over topics
Document
α
θ
Topic
Topic distribution over words
z
Word
β
ϕ
w
N_d
K
D

In LDA, it is observed that document D is a probability distribution over topic z and topic is a probability distribution over word w. The advantages of this approach are: LDA can handle ambiguity and helps to organize, summarize and explore large data. Some open issues of this approach are: how to provide evaluation and model checking, how to provide better visualization and user interfaces and to enhance the topic models for data discovery.

IV. Query Based Event Extraction

H.L.Chieu et.al proposed The Query based event extraction [4] along a timeline which describes the extraction of events relevant to a query from a collection of documents and places events along a timeline. The figure shows the framework for a sentence extraction which consists of three steps: Sentence ranking, Sentence selection and Sentence ordering. In sentence ranking, the sentences are ranked or sentences are ordered based on a query. Then sentences are selected based on a desired summary length. Next sentences are ordered along a timeline for final presentation. There are two theoretical measures for ranking sentences: Interest and Burstiness. Interesting sentences are sentences reporting interesting events. Burstiness involves extraction of sentences that are closely related to the date duration of the event. The assumptions for ranking sentences are: any sentence s is relevant to a query q and only one date is attached to each sentence. The advantages of this approach are: it is a more efficient approach and do not require any expensive operations and sentences are treated as better units of information as they allow quick access to the source documents. The drawback of this approach is the difficulty in integrating with a search engine to work on real time system on user queries.



Sentence extraction

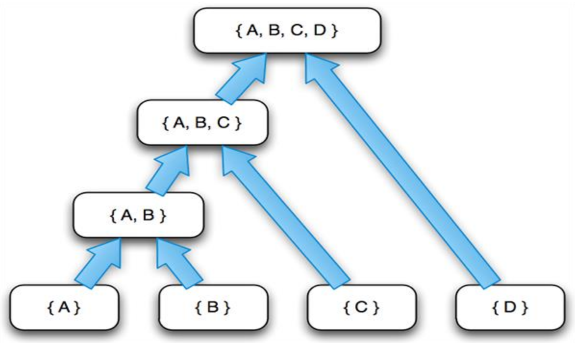
Sentence ranking
- Real value ranking
- Relevance ordering

Sentence selection
- Desired summary length

Sentence ordering
- Final presentation

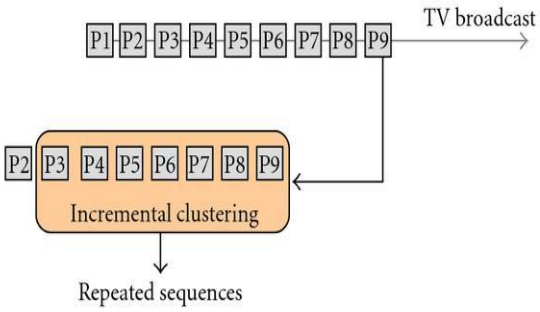
V. A Study on Retrospective & Online Event Detection

The paper, a study on retrospective and online event detection [5] deals with the clustering techniques for event detection. There are two types of clustering methods: Agglomerative or hierarchical or Group Average Clustering (GAC) and Single pass or non hierarchical or Incremental clustering (INCR)



Hierarchical clustering

Hierarchical clustering starts with a single cluster. At each step it joins two closest clusters. In this figure it starts with a cluster A. In the next step it finds a closest cluster B then clusters A and B is joined to form a cluster and so on.GAC is designed for batch processing and is used for retrospective detection.



Non-hierarchical clustering

Non-hierarchical clustering considers only a single event at a time. It works as follows: Assign the first event to a cluster. Then consider next event and either assign event to an existing cluster or create a new cluster. Repeat these steps until all events are clustered. INCR is designed for sequential processing and is used for both retrospective and online detection. The main advantage of this approach is that agglomerative clustering is mainly used for merging topics. On the other hand some disadvantages are: how to make online detection easier and how to improve clustering approach for better accuracy detection.

VI. Automatic Generation of Social Event Story board from Image Click-through data

To discover events from log data, an approach called Smooth Non-negative Matrix Factorization (SNMF) framework [6] is used. There are two basic ideas for SNMF: (1) It promotes event queries (2) It differs events from popular queries. SNMF guarantee weights for each topic to be non-negative and considers time factor for event development. To make event detection easier, relevant images are attached for each event.

There are two phases for the proposed approach: Event detection by SNMF and Event photo selection. In event detection, initially events are searched from log data. Then it discovers groups of queries that have high frequency which is known as topic factorization. Next topics with similar behaviors are merged together along a timeline which is called topic fusion. Event ranking happens in which topics like social events are highlighted. After ranking top topics are called social events and non top topics are called profile topics.

In event photo selection, both the social events and profile topics are sent to search engines like Google or Bing. The search engines generate two sets of image thumbnails which contains relevant images to social events. Image similarity measures occur in which similarity between events and images are measured. Image ranking happens which is sorting of images in the social event image set. Finally all social events together with their images construct a storyboard.

3. Conclusion

This survey has been performed for identifying the various event detection methods which are useful for event mining. It was found that search logs are a good data source for generating an efficient storyboard. SNMF together with time information is emerging as one of the better event detection methods. Moreover it highlights the benefits of mapping events to images along a timeline so as to generate automatically a storyboard. Some advantages of this approach are: there is a large coverage of domains e.g. Entertainment, sports etc., it was found more scalable i.e. it covers large number of topics and it is not at all biased by any editor's interest. Some of the applications of this approach are: monitors social events, creates storyboard and useful for content based news headings.

References

[1]. J.Allan,J.G.Carbonell,G.Doddington,J.Yamron and Y.Yang.Topic detection and tracking pilot study final report.1998.

[2]. J.Weng and B.-S.Lee.Event detection in twitter,*ICWSM*,11:401-408,2011.

[3]. D.M.Blei.Introduction to probabilistic topic models.*Comm.ACM*,55(4):77-84,2012.

[4]. H.L.Chieu and Y.K.Lee.Query based event extraction along a timeline.In *proceedings of the 27th annual international* ACM SIGIR *conference on Research and development in information retrieval*,pages 425-432.ACM,2004.

[5]. Y.Yang,T.Pierce & J.Carbonell.A study of retrospective and online event detection. In *proceedings of the 21st annual international* ACM SIGIR *conference on Research and development in information retrieval,*pages 28-36.ACM,1998.

[6]. JunXu,TaoMei,Seniormember, IEEE,Rui Cai, Member,IEEE,Houqiang Li,Senior Member,IEEE and Yong Rui,Fellow,IEEE.Automatic Generation of Social Event Storyboard from Image Click-through Data, DECEMBER 2015

BIOGRAPHIES

Sreelekshmi.U obtained B. tech. (Computer Science & Engineering) from Sreebuddha College of Engineering, Pattoor, Alappuzha, kerala & pursuing M. Tech. in Computer Science and Engineering from Sreebuddha College of Engineering, Pattoor, Alappuzha,kerala .

Gopu Darsan is currently working as Assistant Professor in the Department of Computer Science and Engineering in

Sreebuddha College of Engineering, Pattoor, Alappuzha, kerala . He obtained his M. Tech. in Computer Vision and Image Processing from Amrita Viswavidyapeetham University.