Design of Improved Social Event Story Board from Image Click-Through Data

Mr. Sahith.R [1], Mr. Ch. Anil Kumar[2], Mrs. A. Venu Madhavi[3]
Senior Assistant Professor, Department of Computer Science &Engineering CVR College of Engineering
Assistant Professor, Department of Computer Science &Engineering BVRIT College of Engineering for Women, Kukatpally, Hyderabad
Assistant Professor, Department of Computer Science & Engineering CVR College of Engineering

***Abstract*-Traditional sites had been pushed by human edited situations which lead to big web search traffic. This paper is actually a survey conducted for identifying the different event detection approaches that are beneficial for event mining.While traditional sites can just present human edited functions, in this particular paper weprovide a novel system to immediately identify events from searchlog data and produce storyboards in which the events are actually arranged chronologically. We selected image search log as the source forevent mining, as search logs could even represent people's interests.In order to find happenings from log data, we show a Smooth Nonnegative Matrix Factorization framework (SNMF) that combinesthe info of query semantics, temporal correlations, research logs as well as time continuity. Additionally, we think about the time factoran essential component since many different events will build in various time tendencies. Additionally, to make a media-rich&visually attractive storyboard, each and every event is actually related with a set of representative pictures set up along a timeline. Theserelevant images are instantly selected from image search engine results by analyzing image content features. Celebrities are used by us as our test domain, that takes a great percent of image search traffics.**

***Index Terms*- Event storyboard, social media, click-throughdata, non-negative matrix factorization, image search.**

I. INTRODUCTION

In recent years, we've witnessed a progression in the acceptance of social media sites, like Flickr, YouTube, and Facebook. These social media web sites offer an interactivesharing platform where huge quantities of unstructured data are actually uploaded every minute. The way we may gain from such richpress is still a challenging and open issue.Events are actually an all natural means of talking about any observableoccurrence grouping activities, times, places, and persons which could be described [5]. Events can also be observable experienceswhich are more and more often documented by individuals through diverse media (e.g., photos and videos). To helpusers grasp events successfully, different event browsing andsearching platforms have been built, that have had good results greatly from social networking event content, e.g., eventful.com,Facebook.com/events, last.fm, and upcoming.org, to name buta few. These services oftentimes have an explicit connectionwith media sharing os's. Usually there's overlap in termsof coverage of upcoming events. Additionally, they offer socialnetwork options to help owners in sharing and deciding upon attending events. Nevertheless, in the Web services, lessattention is given to enhancing the end user experience whenbrowsing and searching content, while the function oflocating target media content to offer vivid data on provided happenings is still missing.In reality, instantly associating social media content with events that are known is actually a difficult issue owing to the noisy and heterogeneous dynamics of the information. Recently, severalworks have been recommended investigating searching event associated media data.

Current search engines typically demonstrate the summaries of prominent individuals as a basic profile. From such a summarization, individuals can smoothly get a celebrity 's essential info as awards, representative works, birthday, nationality, and portrait. Thesearch engine summaries can be regarded as a concentrated model of an individual's bigger related occasion collection. Althoughsuch a brief profile is beneficial for rapidly introducing an individual, it can't gratify people's curiosity for far more detailed and timely info of celebrities. By contrast,some professional sites provide upto-date and comprehensive data on popular persons. Fig.1shows a screenshot of www.people.com, a site famous for celebrityphotos and news. In the marked region of Fig. 1, it showsBritney Spears's the latest news (events) set up along a timeline. This's a really good attribute for fans to trace their idols' pursuits. Nearly all the sites are run by humaneditors, that inevitably leads to a number of limitations.

Fig. 1. Screen shot of www.people.com, a website for celebrity news. The marked region shows recent news of Britney Spears, arranged along timeline.

For starters, thecoverage of man center domains is actually small. Usually, onesite just concentrates on celebrities in a single or perhaps 2 domains(most of them are actually sports and entertainment), also to thebest of the knowledge of ours, there aren't any basic services still for tracing celebrities over different domains. Next, theseexisting services aren't scalable. Even for specific domains,just a couple of best stars are actually covered1, as the editing attempt to discuss a lot more celebrities isn't financially viable. Third, reported eventnews might be biased by editors' interests. With this paper, wegoal to create an unbiased and scalable solution to automaticallydetect social events particularly associated with celebrities along a timeline. This may be an appealing supplement to enrich thepre-existing occasion explanation in search result sites. With this paper, we are going to focus on those events taking place at a particular period favored by owners as the celebrity-related social events of ours.

II. RELATED WORKS

Data mining is the procedure of semiautomatic ally searchinghuge directories to find patterns that are actually understandable, useful, valid, and novel. The objective of data mining is extracting info from a dataset & change it into an easy to understand framework. It's also known as KnowledgeDiscovery in Databases (KDD).The stages in data miningare actually: Preparation, Data gathering, and Problem definition, Model building and evaluation, Knowledge deployment.

Topic Detection and Tracking (TDT) [1] is actually a method whichconsists of the exploration of methods to identify new things and monitor their evolution and reappearance. We will find threespecialized jobs in TDT: Tracking.Segmentation, Detection, and Segmentation is actually the method of breaking down a continuousstream of copy into disjoint, homogenous regions known as stories. Detection is the procedure of determining new events.Tracking will be the method of seeing far more stories about prior occasion. We will find 2 kinds of event detection: Retrospectiveevent detection and online brand new event detection. Inretrospective event detection, stories are actually grouped into clusters in which each cluster belongs to an occasion. In onlinethe latest occasion detection, it identifies brand new incidents in a stream of accounts. A choice is actually made after each story is actually processed. Ifthe story covers a brand new event then it's flagged as Yes overall NO. This strategy is actually beneficial for timelyinfo access applications as Yahoo news. Someopen issues with regards to the method are: how we can choose rightlevel of clusters for owners that best meet their info need,how we can offer navigation resources for efficient and effective search, exactly how to enhance accuracy of on line detection by introducing limited look ahead.

J.Weng et.al proposed Event detection in Twitter [2] whichinvolves Event Detection with Clustering of Wavelet-basedsignals (EDCoW).The components of EDCoW are: Buildsignals for individual words, Filter away trivial words andCluster signals. In order to build signals for individualwords, wavelet transformation is used which consists ofCWT and DWT. Continuous Wavelet Transformation(CWT) provides a redundant representation of signal.Discrete Wavelet Transformation (DWT) provides a nonredundant representation of signals. Then filtering awaytrivial words is achieved through Auto correlation and Crosscorrelation. A mathematical tool used to find repeatingpatterns is called auto correlation. Another tool that searchesfor a long signal for a shorter known feature is known ascross correlation. Later clustering of signals is achieved byModularity based graph portioning and Newman algorithm.In Modularity based graph partitioning, it detects events byclustering signals. Newman algorithm detects and removesedges connecting different events. Some advantages of thisapproach are: Wavelet analysis takes less storage space andEDCoW gives good performance. The disadvantages of thisapproach are: how to analyze the relationship among usersthat could contribute to event detection and how to introducetime lag and study the interaction between different words.

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

LDA (Latent Dirichlet Allocation) is the simplest topicmodel. It is a statistical model of document collections. It isdefined by statistical assumptions like: Order of words in thedocument does not matter, Order of documents does notmatter & number of topics is assumed known & fixed.In LDA, it is observed that document D is a probabilitydistribution over topic z and topic is a probabilitydistribution over word w. The advantages of this approachare: LDA can handle ambiguity and helps to organize,summarize and explore large data. Some open issues of thisapproach are: how to provide evaluation and modelchecking, how to provide better visualization and userinterfaces and to enhance the topic models for datadiscovery.

III. APPROACH

The framework overview of the proposed approach is shownin Fig. 2, which mainly consists of two components: (A) eventdetection and (B) representative event photo selection.To discover events from log data, an approach calledSmooth Non-negative Matrix Factorization (SNMF)framework [6] is used.

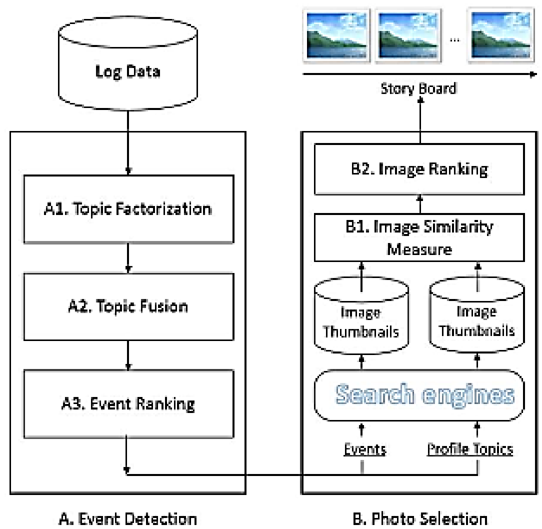

Fig. 2. The overview of the proposed approach, consisting of two main parts:(A) event detection by SNMF and (2) representative event photo selection

There are two basic ideas for SNMF:(1) It promotes event queries (2) It differs events frompopular queries. SNMF guarantee weights for each topic tobe non-negative and considers time factor for eventdevelopment. To make event detection easier, relevantimages are attached for each event.There are two phases for the proposed approach: Eventdetection by SNMF and Event photo selection. In eventdetection, initially events are searched from log data. Then itdiscovers groups of queries that have high frequency whichis known as topic factorization. Next topics with similarbehaviors are merged together along a timeline which iscalled topic fusion. Event ranking happens in which topicslike social events are highlighted. After ranking top topicsare called social events and non top topics are called profiletopics.

For representative event photo selection, top queries fromsocial events and profile topics are first sent to commercialsearch engines (Google or Bing) to collect two sets of imagethumbnails. These two sets are considered the most relevantimages to the social event and the celebrity's background,respectively. However, image search results are very noisy,and sometimes a photo has high-ranking scores in both imagesets. To identify the most representative photos for an event,we propose measuring the content similarity among imagesin these two image sets, using both global and local imagefeatures. The assumption is that event related photos shouldhave similar (duplicate) images in the social event imageset, but should not have similar ones in the profile imageset. Based on this assumption, a simple ranking function isproposed to sort photos in the social event image set. In thisway, we can identify a set of relevant photos to describe eachdetected event. All the social events, together with their photos,construct a story board of that celebrity.

1) SNMF Topic Factorization: In classic topic modeling,the inputs are text documents consisting of words and theoutputs are decompositions of these documents into topics.Here, each topic is a distribution over the word vocabulary.Analogically, we treat one day's log data as a "document"and each query as a "word". The "vocabulary" consists of all the unique queries of a celebrity in his/her log records, i.e.,the set Q defined. Widely used algorithms for topic factorization include probabilistic latent semantic indexing (PLSI) [14], latent Dirichletallocation (LDA) [7], singular value decomposition (SVD) [3],non-negative matrix factorization (NMF) [17], and their variants. In this paper, we choose NMF as it has a nice advantage–data must be decomposed into a sum of additive components. In other words, both the coefficients of "documents'distributions over topics" and the coefficients of "topics'distributions over queries" must be non-negative. This makessense, especially for event modeling, as it is hard to accept theexplanation that we observe a certain query just because someevents didn't happen that day.

2) Topic Fusion: After the factorization step, we have Ktopics $\{t_1,\ldots,t_k\}$ and two matrices W and H. To characterize a topic, the most intuitive clues are its distributions,both over the query vocabulary and over the time line. Thesetwo distributions can be directly obtained from W and H.Another useful clue from the search log data is the set ofsearch log URLs, which have proven to be effective forquery clustering [40]. The assumption is, queries trigging the same URL are very likely to have similar semantics.

3) Event Ranking: The last step is to distinguish eventrelated topics from others. Although this is essentially aclassification problem, collecting enough unbiased trainingdata is quite difficult in practice. Therefore, we treat it as aranking problem, to leverage several heuristics summarizedbased on a number of observations. Similar to the above part, these heuristics are based on the distributions of a topic overthe time-line, over the query vocabulary, and over the searchlog URLs.

4) Event Photo Selection:

People often say that "a picture is worth a thousand words".Without a doubt, interesting events associated with relatedphotos are more attractive to the audience. For each detectedsocial event, it is straightforward to identify a set of mostrelevant queries by inspecting the event's distribution in the query space. The simplest way to get events related photosis to directly search commercial image search engines withthese event queries.

Image Similarity Measures: To measure image similarity, we considered both global and local image features inthis paper. Global features are extracted based on a wholeimage, and are suitable for identifying fully duplicate images.By contrast, local features describe a local image patch, andhave been widely used for recognizing partial duplicates.Supporting partial duplicate detection is quite important inthis step, as many images have been edited (e.g., croppingor stitching) before being published online.

IV. CONCLUSION

In this paper, we use search logs as data source to generatesocial event storyboards automatically. Unlike common textmining, search logs have short, sparse text queries and the datasize is much bigger than some news websites or blogs. It was found that search logs are a good data source forgenerating an efficient storyboard. SNMF together with timeinformation is emerging as one of the better event detectionmethods. Moreover it highlights the benefits of mappingevents to images along a timeline so as to generateautomatically a storyboard. Some advantages of thisapproach are: there is a large coverage of domains e.g.Entertainment, sports etc., it was found more scalable i.e. itcovers large number of topics and it is not at all biased byany editor's interest. Some of the applications of thisapproach are: monitors social events, creates storyboard anduseful for content based news headings.

REFERENCES

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

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

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

[4] H.L.Chieu and Y.K.Lee.Query based eventextraction along a timeline.In proceedings of the27th annual international ACM SIGIR conferenceon Research and development in informationretrieval,pages 425-432.ACM,2004.

[5] U. Westermann and R. Jain. Toward a Common Event Model forMultimedia Applications. IEEE MultiMedia, 14(1):19–29, 2007.

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

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation.theJournal of machine Learning research, 3:993–1022, 2003.

[8] Y.-J. Chang, H.-Y. Lo, M.-S. Huang, and M.-C. Hu.Representativephoto selection for restaurants in food blogs. In Multimedia & ExpoWorkshops (ICMEW), 2015 IEEE International Conference on, pages1–6. IEEE, 2015.

[9] H. L. Chieu and Y. K. Lee. Query based event extraction along atimeline. In Proceedings of the 27th annual international ACM SIGIRconference on Research and development in information retrieval, pages425–432. ACM, 2004.

[10] T.-C. Chou and M. C. Chen.Using incremental plsi for thresholdresilient online event analysis. Knowledge and Data Engineering, IEEETransactions on, 20(3):289–299, 2008.

[11] H. Cui, J.-R. Wen, J.-Y.Nie, and W.-Y. Ma. Probabilistic queryexpansion using query logs. In Proceedings of the 11th internationalconference on World Wide Web, pages 325–332. ACM, 2002.

[12] S. Essid and C. Fevotte. Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. Multimedia, IEEETransactions on, 15(2):415–425, 2013.

[13] G. P. C. Fung, J. X. Yu, H. Liu, and P. S. Yu.Time-dependent eventhierarchy construction. In Proceedings of the 13th ACM SIGKDDinternational conference on Knowledge discovery and data mining,pages 300–309. ACM, 2007.

[14] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings ofthe 22nd annual international ACM SIGIR conference on Research anddevelopment in information retrieval, pages 50–57. ACM, 1999.

[15] T. Joachims. Optimizing search engines using clickthrough data. InProceedings of the eighth ACM SIGKDD international conference onKnowledge discovery and data mining, pages 133–142. ACM, 2002.