Automatic event detection and characterization of social events from Image Click-Through Data

[1]M.Satish Kumar, [2] B.Himagiri
[1] Associate Professor, Department of MCA, Sri Venkateswara College of Engineering and Technology (Autonomous), Chittoor
[2] PG Scholar, Department of MCA, , Sri Venkateswara College of Engineering and Technology (Autonomous), Chittoor ,A.P.

Abstract-

Recent studies have shown that a noticeable percentage of web search traffic is about social events. While traditional websites can only show human-edited events, in this paper we present a novel system to automatically detect events from search log data and generate storyboards where the events are arranged chronologically. We chose image search log as the resource for event mining, as search logs can directly reflect people's interests. To discover events from log data, we present a Smooth Nonnegative Matrix Factorization framework (SNMF) which combines the information of query semantics, temporal correlations, search logs and time continuity. Moreover, we consider the time factor an important element since different events will develop in different time tendencies. In addition, to provide a media-rich and visually appealing storyboard, each event is associated with a set of representative photos arranged along a timeline. These relevant photos are automatically selected from image search results by analyzing image content features. We use celebrities as our test domain, which takes a large percentage of image search traffics. Experiments consisting of web search traffic on 200 celebrities, for a period of six months, show very encouraging results compared with handcrafted editorial storyboards.

Index terms **– Click-through data, event storyboard, image search, nonnegative matrix factorization (NMF), social media.**

I. INTRODUCTION

As social creatures, people are, by nature, curious about others' activities. Information on famous persons have often been of particular interest. This tendency has remained true in the Internet era [35]. Since common search engines, as well as news Web sites, often experience massive search demands about a myriad of current affairs, a great amount of news and events are collected from the Web. However, most social events originate from professional editors. In this case, it is quite meaningful to automatically detect such events for users instead of manual efforts. Current search engines often show the summaries of famous persons as a simple profile. From such a summarization,
people can easily get a celebrity's basic information, such as portrait, nationality, birthday, representative works, and awards. The search engine summaries can be considered a concentrated version of a person's larger relevant event collection.

Although such a short profile is very helpful for quickly introducing a person, it cannot satisfy people's curiosity for more detailed and timely information of celebrities. By contrast, some professional Web sites provide comprehensive and up-to-date information on famous persons. Fig. 1 shows a screenshot of a Web site well known for celebrity news and photos. In the marked region of Fig. 1, it shows Britney Spears's recent news (events) arranged along a timeline. This is a very nice feature for fans to trace their idols' activities. Almost all of these Web sites are powered by human editors, which inevitably lead to several limitations.

First, the coverage of human center domains is small. Typically, one Web site only focuses on celebrities in one or two domains (most of them are entertainment and sports), and to the best of our knowledge, there are no general services yet for tracing celebrities over various domains. Second, these existing services are not scalable. Even for specific domains, only a few top stars are covered,1 as the editing effort to cover more celebrities is not financially viable. Third, reported event news may be biased by editors' interests.

In this paper, we aim to build a scalable and unbiased solution to automatically detect social events especially related to celebrities along a timeline. This could be an attractive supplement to enrich the existing event description in search result pages. In this paper, we will focus on those events happening at a certain time favored by users as our celebrity-related social events.

This paper is organized in five sections. After this introduction, in Section II, literature survey discussed of the paper, section III about the System Analysis, Section IV about System Design, as well as the novel feature of the proposed method. Finally, Sections V and VI provide the simulation results and the conclusions, respectively.

II. LITERATURE SURVEY

1) Learning topic models–going beyond svd

AUTHORS: S. Arora, R. Ge

Topic Modeling is an approach used for automatic comprehension and classification of data in a variety of settings, and perhaps the canonical application is in uncovering thematic structure in a corpus of documents. A number of foundational works both in machine learning and in theory have suggested a probabilistic model for documents, whereby documents arise as a convex combination of (i.e. distribution on) a small number of topic vectors, each topic vector being a distribution on words (i.e. a vector of word-frequencies). Similar models have since been used in a variety of application areas, the Latent Dirichlet Allocation or LDA model of Blei et al. is especially popular. Theoretical studies of topic modeling focus on learning the model's parameters assuming the data is actually generated from it. Existing approaches for the most part rely on Singular Value Decomposition (SVD), and consequently have one of two limitations: these works need to either assume that each document contains only one topic, or else can only recover the {\em span} of the topic vectors instead of the topic vectors themselves. This paper formally justifies Nonnegative Matrix Factorization (NMF) as a main tool in this context, which is an analog of SVD where all vectors are nonnegative. Using this tool we give the first polynomial-time algorithm for learning topic models without the above two limitations. The algorithm uses a fairly mild assumption about the underlying topic matrix called separability, which is usually found to hold in real-life data. Perhaps the most attractive feature of our algorithm is that it generalizes to yet more realistic models that incorporate topic-topic correlations, such as the Correlated Topic Model (CTM) and the Pachinko Allocation Model (PAM). We hope that this paper will motivate further theoretical results that use NMF as a replacement for SVD -- just as NMF has come to replace SVD in many applications.

2) Detecting events from continuous media by intermodal collaboration and knowledge use

AUTHORS: N. Babaguchi, S. Sasamori

We propose an event network, which is a structured representation oriented for the contents of continuous media, as well as present two methods of detecting events as the first step to construct the network. We deal with sports TV programs, considering American football as a case study. The first method is simple intermodal collaboration: linking between visual and linguistic (closed caption) streams. Using domain knowledge about state transitions of football games, the second method attempts to extract specific visual objects including the information about contents. The experimental results indicate that both methods are effective for event detection.

3) Modeling the impact of short-and long-term behavior on search personalization

AUTHORS: P. N. Bennett, R. W. White

User behavior provides many cues to improve the relevance of search results through personalization. One aspect of user behavior that provides especially strong signals for delivering better relevance is an individual's history of queries and clicked documents. Previous studies have explored how short-term behavior or long-term behavior can be predictive of relevance. Ours is the first study to assess how short-term (session) behavior and long-term (historic) behavior interact, and how each may be used in isolation or in combination to optimally contribute to gains in relevance through search personalization. Our key findings include: historic behavior provides substantial benefits at the start of a search session; short-term session behavior contributes the majority of gains in an extended search session; and the combination of session and historic behavior out-performs using either alone. We also characterize how the relative contribution of each model changes throughout the duration of a session. Our findings have implications for the design of search systems that leverage user behavior to personalize the search experience.

III. SYSTEM ANALYSIS

A. Existing System

❖ The most related research topics to this paper are event/topic detection from Web. There have been quite a few works that examine related directions. The most typical data sources for event/topic mining are news articles and weblogs. Various statistical methods have been proposed to group documents sharing the same stories. Temporal analysis has also been involved to recover the development trend of an event.

❖ The representative work for event/topic detection is the DARPA-sponsored research program called TDT (topic detection and tracking), which focus on discovering events from streams of news documents. With the development of Web 2.0, weblogs have become another data source for event detection. Some of these research efforts develop new statistical methods and some others focused on recovering the temporal structure of events.

Disadvantages

First, the coverage of human center domains is small. Typically, one website only focuses on celebrities in one or two domains (most of them are entertainment and sports), and to the best of our knowledge, there are no general services yet for tracing celebrities over various domains.

❖ Second, these existing services are not scalable. Even for specific domains, only a few top stars are covered1, as the editing effort to cover more celebrities is not financially viable.

❖ Third, reported event news may be biased by editors' interests.

❖ Discovering events from a search log is not a trivial task.

❖ Existing work on log event mining mostly focus on merging similar queries into groups, and investigating whether these groups are related to semantic events like "Japan Earthquake" or "American Idol". Basically, their goals are to distinguish salient topics from noisy queries. Directly applying their approaches will fail as the discovered topics are more likely related to vast and common topics, which may be familiar to most users.

B. Proposed System

❖ In this paper, we aim to build a scalable and unbiased solution to automatically detect social events especially related to celebrities along a timeline. This could be an attractive supplement to enrich the existing event description in search result pages.

❖ In this paper, we will focus on those events happening at a certain time favored by users as our celebrity-related social events. we would like to detect those more interesting social events to entertain users and fit their browsing taste, which could be supplementary to some current knowledge bases.

❖ A novel approach is proposed in this paper using Smooth Nonnegative Matrix Factorization (SNMF) for event detection, by fully leveraging information from query semantics, temporal correlations, and search log records. We use the SNMF method rather than the normal NMF method or other MF method to guarantee that the weights for each topic are non-negative and consider the time factor for event development at the same time.

❖ The basic idea is two-fold: 1) promote event queries through by strengthening their connections based on all available features; 2) differentiate events from popular queries according to their temporal characteristics.

Advantages

To provide a comprehensive and vivid storyboard, in this paper, we also introduce an automatic way to attach a set of relevant photos to each piece of event news.

❖ We propose a novel framework to detect interesting events by mining users' search log data. The framework consists of two components, i.e., Smooth Non-Negative Matrix Factorization event detection and representative event related image photo selection

❖ We have conducted comprehensive evaluations on large scale real-world click through data to validate the effectiveness.

IV. SYSTEM DESIGN

A. System Architecture

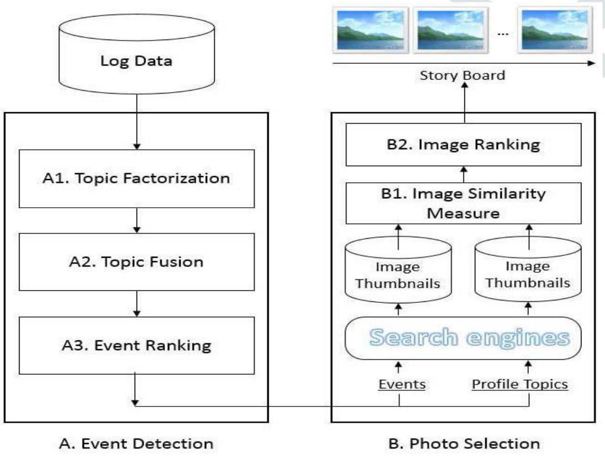


Fig. 1. Overview of the proposed approach, consisting of two main parts. A: Event detection by SNMF. B: Representative event photo selection

The framework overview of the proposed approach is shown in Fig. 3, which mainly consists of two components: 1) event detection and 2) representative event photo selection.

There are three steps for event detection. First, topic factorization methods are adopted to discover groups of queries that have a high co-occurrent frequency. This solves issue with sparsity and random noise in the query set. As we want to detect social events, but not those in the salient topics, we have to keep a relatively large number of topics in the factorization step, and then merge topics with similar behaviors in the second step. To merge correlated topics, we consider topic distributions on both the timeline and the space of click-URLs. Finally, a rank function is introduced to highlight topics, which are very likely to be social events. Again, information on query semantics, temporal correlations, and search log mappings are combined in the ranking process. After ranking, the top topics are referred to as *social events*. Nontop but salient topics are called *profile topics*.

B. Implementation

MODULES:

- System Framework
- User
- Reporter
- Admin

System framework:

In this framework, In this paper, we aim to build a scalable and unbiased solution to automatically detect social events especially related to celebrities along a timeline. This could be an attractive supplement to enrich the existing event description in search result pages. A novel approach is proposed in this paper using Smooth Nonnegative Matrix Factorization (SNMF) for event detection, by fully leveraging information from query semantics, temporal correlations, and search log records. We use the SNMF method rather than the normal NMF method or other MF method to guarantee that the weights for each topic are non-negative and consider the time factor for event development at the same time.

User:

In User module, Initially User must have to register their detail and after login user can view news posted by all reporters. News will be viewed in order of news posted date. Users can search news with search keywords. Users can also search images by entering the keyword in the image search.

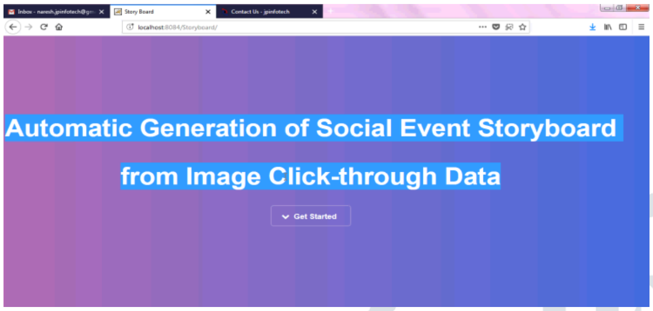
Reporter:

In Reporter module,Reporter can login with Id and Password given by the admin. Reporters can add news and the added news will store in the database. Added news will queued in waiting list for admin approval. Reporters will able to see the added news and overall news posted by other reporters.

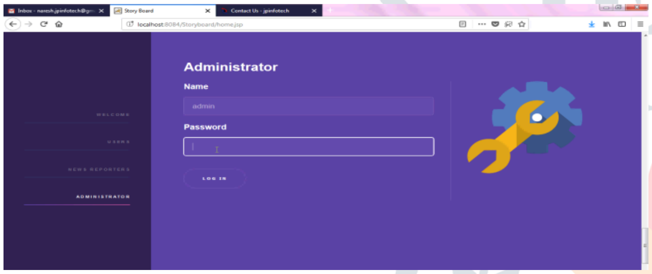
Admin:

In admin module, Admin will add reporters then id and password will be automatically mailed to the reporters email id. Admin will approve the news queued in the waiting list. Admin can view the reporter's details and users details in the database. Admin can view all the news details posted by the all reporters. Admin can view graph analyze of searched keywords, most searched keyword search by all users.

V SIMULATION RESULTS



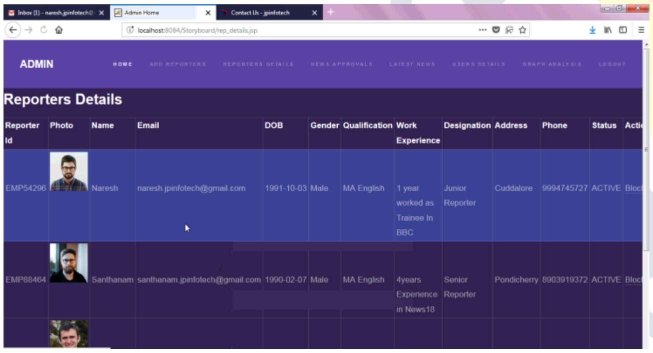
Automatic Generation of Social Event Storyboard from Image Click-through Data



Administrator

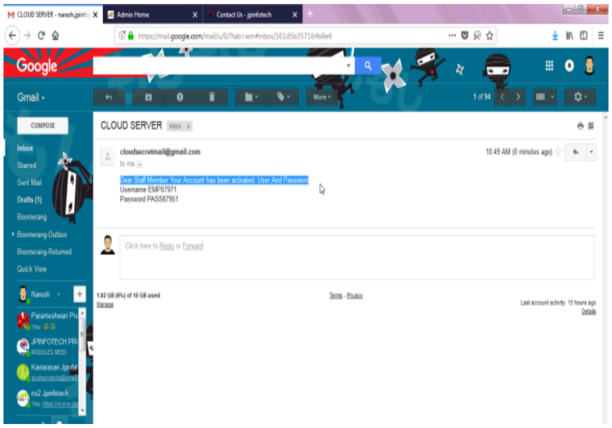
Name: admin

Password:



Reporters Details

Reporter Id	Photo	Name	Email	DOB	Gender	Qualification	Work Experience	Designation	Address	Phone	Status	Acti
EMP54296		Naresh	naresh.jpinfotech@gmail.com	1991-10-03	Male	MA English	1 year worked as Trainee In BBC	Junior Reporter	Cuddalore	9994745727	ACTIVE	Bloc
EMP88464		Santhanam	santhanam.jpinfotech@gmail.com	1990-02-07	Male	MA English	4years Experience in News18	Senior Reporter	Pondicherry	8903919372	ACTIVE	Bloc



CLOUD SERVER Inbox

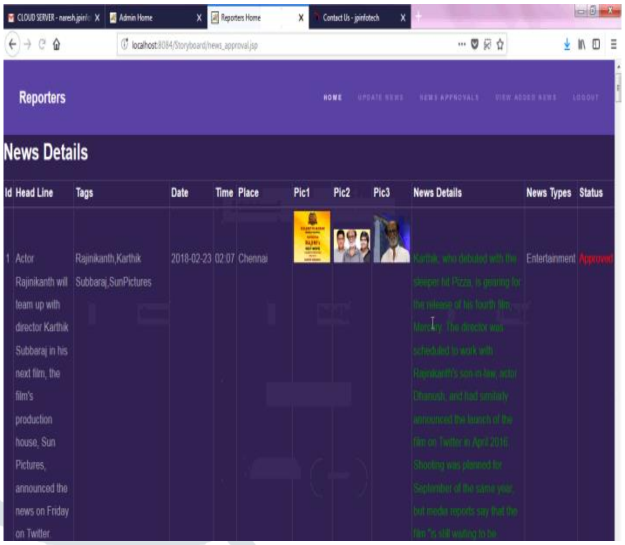
cloudsecretmail@gmail.com 10:49 AM (0 minutes ago)
to me

Dear Staff Member Your Account has been activated. User And Password
Username EMP07971
Password PASS67951

Click here to Reply or Forward

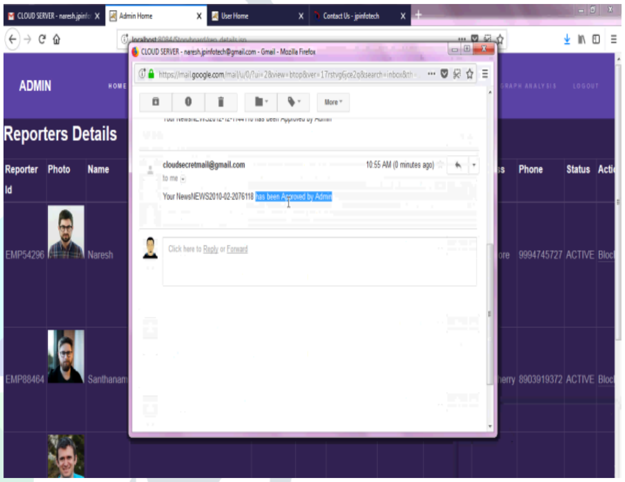
1.02 GB (6%) of 15 GB used
Manage

Last account activity: 15 hours ago



News Details

Id	Head Line	Tags	Date	Time	Place	Pic1	Pic2	Pic3	News Details	News Types	Status
1	Actor Rajinikanth will team up with director Karthik Subbaraj in his next film, the film's production house, Sun Pictures, announced the news on Friday on Twitter.	Rajinikanth,Karthik Subbaraj,SunPictures	2018-02-23	02:07	Chennai				Karthik, who debuted with the sleeper hit Pizza, is gearing for the release of his fourth film, Mercury. The director was scheduled to work with Rajinikanth's son-in-law, actor Dhanush, and had similarly announced the launch of the film on Twitter in April 2016. Shooting was planned for September of the same year, but media reports say that the film "is still waiting to be	Entertainment	Approved


CLOUD SERVER - naresh.jpinf
Admin Home
User Home
Contact Us - jpinfotech
localhost:8084/Storyboard/rep_details.jsp
CLOUD SERVER - naresh.jpinfotech@gmail.com - Gmail - Mozilla Firefox
https://mail.google.com/mail/u/0/?ui=2&view=btop&ver=17rstvg6jce2q&search=inbox&th
More
ADMIN
HOME
GRAPH ANALYSIS
LOGOUT

Reporters Details

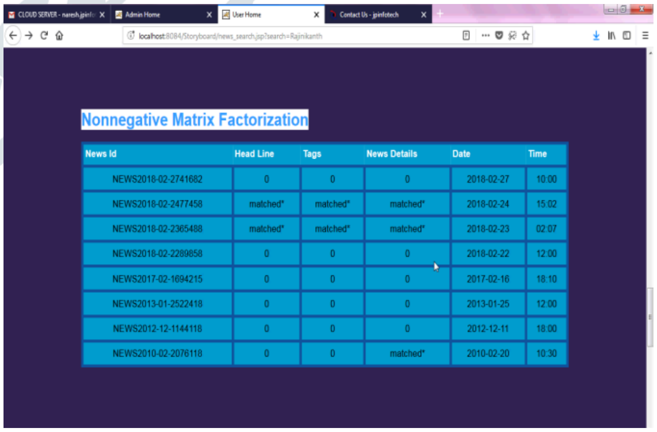
Reporter Id	Photo	Name		Phone	Status	Acti
EMP54296		Naresh	ore	9994745727	ACTIVE	Bloc
EMP88464		Santhanam	herry	8903919372	ACTIVE	Bloc

Your NewsNEWS2012-12-1144118 has been Approved by Admin

cloudsecretmail@gmail.com 10:55 AM (0 minutes ago)
to me

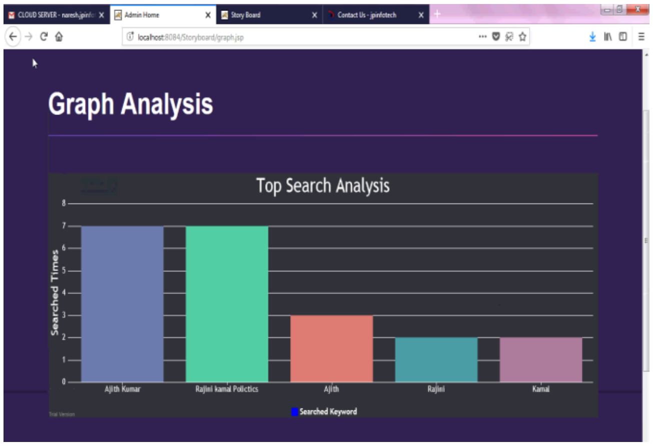
Your NewsNEWS2010-02-2076118 has been Approved by Admin

Click here to Reply or Forward


CLOUD SERVER - naresh.jpinf
Admin Home
User Home
Contact Us - jpinfotech
localhost:8084/Storyboard/news_search.jsp?search=Rajinikanth

Nonnegative Matrix Factorization

News Id	Head Line	Tags	News Details	Date	Time
NEWS2018-02-2741682	0	0	0	2018-02-27	10:00
NEWS2018-02-2477458	matched*	matched*	matched*	2018-02-24	15:02
NEWS2018-02-2365488	matched*	matched*	matched*	2018-02-23	02:07
NEWS2018-02-2289858	0	0	0	2018-02-22	12:00
NEWS2017-02-1694215	0	0	0	2017-02-16	18:10
NEWS2013-01-2522418	0	0	0	2013-01-25	12:00
NEWS2012-12-1144118	0	0	0	2012-12-11	18:00
NEWS2010-02-2076118	0	0	matched*	2010-02-20	10:30



VI CONCLUSION

In this paper, we use search logs as data source to generate social 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. Basedon these features, we do not use the query text informationto do the analysis. Structure and statistic information are usedto get the topics and event detection in our work, which canfit the data well. Furthermore, we add time information in ourapproach to SNMF to make it easier to discover social eventscompared with traditional NMF methods..

REFERENCESS

[1] C. Alexander, B. Fayock, and A. Winebarger. Automatic event detectionand characterization of solar events with iris, sdo/aia and hi-c.InAAS/Solar Physics Division Meeting, volume 47, 2016.

[2] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topicdetection and tracking pilot study final report. 1998.

[3] S. Arora, R. Ge, and A. Moitra.Learning topic models–going beyondsvd. In Foundations of Computer Science (FOCS), 2012 IEEE 53^{rd}Annual Symposium on, pages 1–10. IEEE, 2012.

[4] N. Babaguchi, S. Sasamori, T. Kitahashi, and R. Jain.Detecting eventsfrom continuous media by intermodal collaboration and knowledgeuse.In Multimedia Computing and Systems, 1999. IEEE InternationalConference on, volume 1, pages 782–786. IEEE, 1999.

[5] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk,and X. Cui. Modeling the impact of short-and long-term behavior onsearch personalization. In Proceedings of the 35th international ACMSIGIR conference on Research and development in information retrieval,pages 185–194. ACM, 2012.

[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.