Creating Storyboard of Social Event from Advertorial Data

[1]Bhavana R M, [2]Sangameshwari R K

[1]Assistant Professor, [2]PG Student
Department of Computer Science and Engineering,
Visvesvaraya Technological University Centre for PG Studies, Gulbarga, Karnataka, India

***Abstract*—Recent analyses have shown that searching about social events becoming obvious part of web search activity. Meanwhile existing sites can at most present social events modified by human beings, in this paper we propose a novel framework to discover social events by itself from search log facts and develop storyboards in which events are ordered consecutively. Since search logs can provide the user's interests exactly, we use image search log for detecting various events. To perceive occasions from log facts we propose a smooth nonnegative matrix factorization framework, which blends information for query semantics, temporal relationships, search logs and time durability. Furthermore we take into account the time factor as a key element because different events will generate in different time inclinations. Also to give enriched media and visually attractive storyboards, each event is interrelated with illustrative set of photos ordered chronologically. These suitable photos are spontaneously selected from image search information by examining image content qualities. We take popular persons as our test space that takes huge proportion of image look trade.**

***IndexTerms*—Nonnegative matrix factorization (NMF), temporal relationships, representative photo selection.**

I. INTRODUCTION

We human beings are more eager to know about others' activities especially about celebrities. Because usual search engines and additionally news websites often encounter enormous search requirements about the current issues, a huge amount of news and social events are gathered from web. Almost all social events are provided by professional editors. For this situation, it is very important to consequently recognize such occasions for clients rather than manual endeavors.

The existing search engines usually show information about celebrities just like a basic profile. This cannot fulfill viewers' attentiveness. Still the professional websites provide thorough and advanced information about celebrities. Most websites controlled by human editors, results in several disadvantages: 1) the scope of covering human centered fields is less. 2) Coverage is not extensible. 3) Reported news may be biased by editors' curiosity.

Therefore in this paper our goal is to constructing the social event storyboard inevitably that provides compatible and neutral solution. In this paper, we create storyboards related to celebrities occurring at the specific time.

II. RELATED WORK

S. Arora et al. [1], subject displaying is an approach used for customized comprehension and course of action of data in a combination of settings, and might be the endorsed application is in uncovering topical structure in a corpus of chronicles. Different foundational works both in machine learning and on a basic level have suggested a probabilistic model for reports, hereby records rise as a raised blend of (i.e. movement on) few subject vectors, each point vector being a dispersal on words (i.e. a vector of word-frequencies). Relative models have since been used as a piece of a collection of usage districts; the Latent Dirichlet Allocation or LDA model of Bali et al. is especially pervasive. Theoretical examinations of point exhibiting base on taking in the model's parameters expecting the data is truly made from it. Existing philosophies by and large rely upon Singular Value Decomposition (SVD), and in this way have one of two limitations: these works need to either expect that each report contains only a solitary subject, or else can simply recover the span of the point vectors instead of the point vectors themselves. This paper formally legitimizes Nonnegative Matrix Factorization (NMF) as an essential instrument in this extraordinary circumstance, which is a basic of SVD where all vectors are nonnegative. Using this mechanical assembly S. Arora et al. give the essential polynomial-time figuring for learning point models without the more than two imperatives. The count uses a really smooth assumption about the crucial point organizes called noticeability, which is regularly found to hold, everything considered, data. Perhaps the most charming component of computation is that it totals up to yet more sensible models that wire subject topic connections, for instance, the Correlated Topic Model (CTM) and the Pachinko Allocation Model (PAM).

N. Babaguchi et al. [2], N. Babaguchi et al. propose an event orchestrate, which is a sorted out depiction arranged for the substance of constant media, and moreover show two procedures for perceiving events as the underlying advance to build up the framework. N. Babaguchi et al. oversees sports TV programs, thinking about American football as a logical examination. The essential procedure is fundamental multi-reason composed exertion: associating among visual and phonetic (close engraving) streams. Using region finding out about state advances of football games, the second procedure attempts to evacuate specific visual things including the information about the substance. The exploratory results demonstrate that the two procedures are capable of event acknowledgment.

P. N. Bennett et al. [3], customer direct gives various signs to upgrade the significance of recorded records through personalization. One a player in customer direct that gives especially strong signs to passing on better importance is a man's history of request and clicked reports. Past examinations have researched how without further ado conductor whole deal directs can be farsighted of importance. P. N. Bennett et al. own particular is the fundamental examination to assess how without a moment's hesitation (session)

direct and whole deal (critical) lead interface, and how each may be used as a piece of withdrawal or in the blend to preferably add to grabs in hugeness through request personalization. P. N. Bennett et al. key revelations include: critical lead gives extensive focal points toward the start of an interest session; without a moment's hesitation session direct contributes the lion's offer of augmentations in an extended request session, and the blend of the session and striking behavior out-performs using either alone. P. N. Bennett et al. moreover depicts how the relative duty of each model changes all through the term of a session. P. N. Bennett et al. disclosures have proposals for the layout of chase structures that utilization customer direct to modify the request experience.

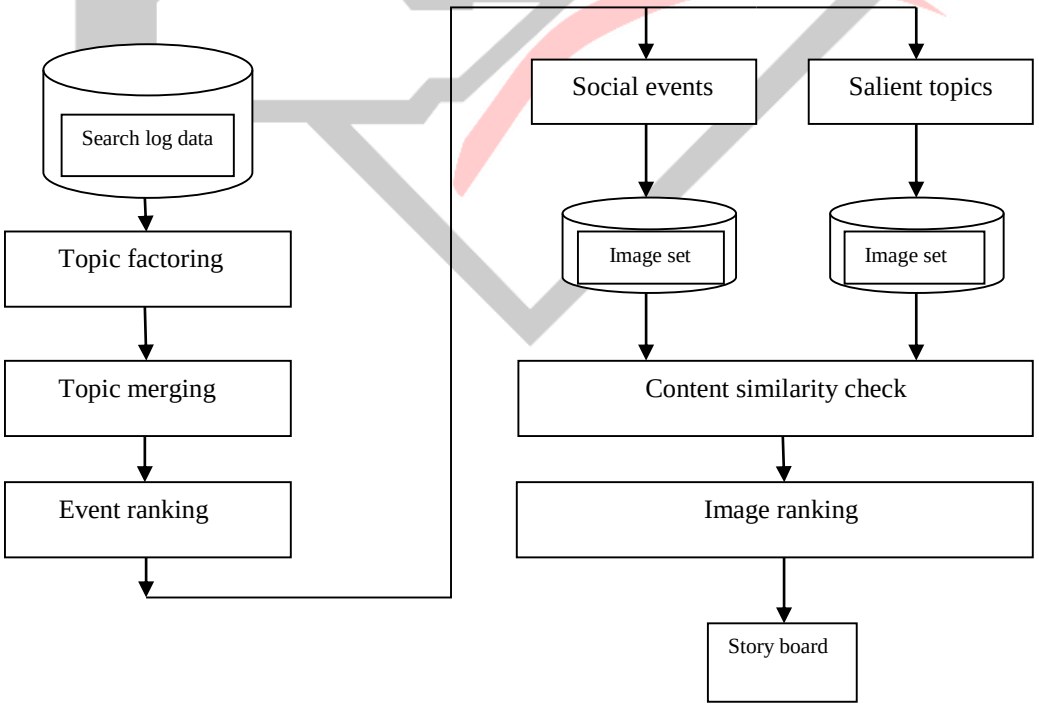
Y.-J. Chang et al. [4], nowadays, people make comments out of restaurants and exchange related photos to support writes in the wake of going there. Working up a versatile application which enables the customer to effectively look diners from data in these online diaries transforms into a creating need. Other than scrutinizing the comments, considerable number individuals will give a gander at food photos of a restaurant and a short time later pick whether to go or what to eat. As needs be, Y.-J. Chang et al. propose a structure to analyze and select delegate photos for each diner in perspective of blog-arrange media. A strong support ID demonstrates is set up to recoup sustenance photos and an a la mode quality evaluation system is utilized to pick specialist photos. In perspective of these operator photos, customers would more have the capacity to easily have the impression of the restaurant and review the blog in a dealt with way. The trial occurs show that structure can make better illustrative photos (i.e. significantly closer to the customers' slants) than existing site stages.

H. L. Chieu et al. [5], in this paper, H. L. Chieu et al. demonstrate a structure and a system that concentrates events relevant to a request from a social occasion C of files, and places such events alongside a course of occasions. Each event is addressed by a sentence removed from C, in light of the assumption that "goal" events are for the most part alluded to in various reports for a time period inside which these events are of interest. In tests, H. L. Chieu et al. used request that is event composes ("tremor") and individual names (e.g. "George Bush"). An appraisal was performed using G8 pioneer names as inquiries: examination made by human evaluators among physically and a system made timetables exhibited that but physically delivered courses of occasions are general all the more best, the structure created timetables are at times judged to be better than physically created ones.

III. METHODOLOGY

The proposed framework is consists of 2 parts: event detection and picture selection. In the event detection, there are three steps. First, in the topic resolving we find the set of queries that are not typical. Second, in the topic merging, these various set of queries having similarities are merged. Third, event numbering is made to highlight the events based on query semantics, temporal relationships and search log figuring. After rating, the top topics are called social events and nontop yet remarkable topics are called profile topics.

In the image selection, top queries from events and salient topics are searched in search engines. It outputs as set of images one for the event and other one for the celebrity's background. Then content similarity check is performed on these sets to get a representative image for an event. Then storyboard with suitable images can be built. The overview of the proposed framework is shown in figure 1.



a) Event detection b) Picture selection

Fig 1: System Overview

Detection of Event:
Here we detect the events that are not typical by applying smooth nonnegative matrix factorization. For example, a famous sport person's marriage. Here we need to determine "how noisy a query is". To handle noise data topic factorization has ended up being a viable approach.
For celebrity with N search log records, each record is illustrated as $r_x = (q_x, d_x, u_x)$ where 1≤ x≤ N and $q_x \in$ Q, $d_x \in$ D, and $u_x \in$ U. The sets Q, D, and U are set of unusual queries, days, and URLs in the celebrity's log information respectively.

1. Topic factorization: Here we use NMF to factorize the events. Both coefficients of records' distribution over topics and coefficients of topics' distribution over queries should be nonnegative. Firstly the log data are changed to matrix L of size |Q| × |D| where each row indicates a query and each column indicates a day. NMF obtains the matrices T and H that should meet

$$L \approx T \times H \quad (1)$$

Where T = [t_1,t_2,…,t_I] in which each column $w_i (1 \le i \le I)$ signifies a topic and I is the predefined number of topics. H = $[h_1, h_2, \dots, h_{|D|}]$ in which each column $h_d (1 \le d \le |D|)$ is the factorized coefficients of topics for the dth day. This factorization problem converts to minimizing the cost function.

$$\arg \min_{T,H} D^{g}_{K\,L}(L||T \times H) \text{ s.t } T \ge 0, H \ge 0. \quad (2)$$

Where $D^{g}_{K\,L}(A||B)$ is the generalized Kullback-Leiber (K-L) divergence of matrices A and B.
Further there must not be any noteworthy dissimilarity between topics from two nearby days. To fix this, SNMF was proposed with an additional regularization term R(H) to cost function.

$$\arg min_{T,H}\{D^{g}_{K\,L}(L||T \times H) + \lambda \times R(H)\} \quad (3)$$
$$\text{Where, } R(H) = \sum_{d=2}^{|D|} ||h_d - h_{d-1}||_2 \text{ s.t } T \ge 0, H \ge 0.$$

Here R(H) behaves as penalty for smoothness between two nearby columns in H, and λ is a nonnegative weight to adjust the degree of smoothing

arg

2. Topic merging: To identify a topic, we have instinctive intimations. They are its distributions and search log URLs. We consolidate these intimations to determine the similarity between two topics and merge the topics in unsupervised way.

a) Topic similarity over queries: Given a topic $t_i (1 \le i \le I)$, its distribution over queries can be estimated by ith column of T. Since T is nonnegative matrix, it is sufficient enough to convert the ith column into a distribution $P_Q(q_x|t_i)$ by normalizing it with the sum of its elements that is, $P_Q(q_x|t_i) = T_{xi} / \sum_{j=1}^{|Q|} T_{ji}$. Also the distance between the topics t_i and t_m is given by the symmetric K-L divergence, as

$$\text{dist}_Q(t_i, t_m) = K\,L_Q(t_i, t_m) = \frac{1}{2}\sum_{x=1}^{|Q|}\left(P_Q(q_x|t_i)\ln\frac{P_Q(q_x|t_i)}{P_Q(q_x|t_m)} + P_Q(q_x|t_m)\ln\frac{P_Q(q_x|t_m)}{P_Q(q_x|t_i)}\right) \quad (4)$$

b) Topic similarity over timeline: A topic t_i's distribution over the timeline can be estimated by normalizing the ith row in H. That is, $P_Q(d_x|t_i) = H_{ix} / \sum_{j=1}^{|D|} H_{ij}$. The distance between t_i and t_m is

$$\text{dist}_Q(t_i, t_m) = \min\{K\,L_D(t_i, t_m; \Delta), \Delta \in \{-1, 0, 1\}\} \quad (5)$$

Where $K\,L_D(t_i, t_m; \Delta)$ is the shift-enabled K-L divergence and Δ is compensate for days.

c) Topic similarity over URLs: From search log, the connections between log URLs and the queries can be portrayed by |U| × |Q| matrix G in which every element G_{xj} indicates the circumstances that URL u_x being clicked given the query q_j. Then a topic t_i's distribution over the log URLs is given as $P_U(u_x|t_i) = (GT)_{xi} / \sum_{j=1}^{|U|} (GT)_{ji}$, and the corresponding distance between t_i and t_m is

$$\text{dist}_U(t_k, t_l) = K\,L_U(t_k, t_l) \quad (6)$$

These three distances are added to get the overall distance between two topics. Further agglomerative hierarchical clustering is used to merge similar topics in bottom-up way.

3. Event ranking: The last step in event detection to recognize the topics relevant to event from others. The ranking scores on the basis of timeline, query and URL are assigned to each topic as follows.

$$rank(t_i) = \text{score}_t(t_i) \times \text{score}_q(t_i) \times \text{score}_u(t_i) \quad (7)$$

$$\text{Where, } \text{score}_t(t_i) = \exp^{-K\,L(P_D(.|t_i)||\text{Gamma}(.|t_i))} \quad (8)$$

$$\text{score}_q(t_i) = 1.0 + \frac{1}{\ln|Q|}\sum_{x=1}^{|Q|}\left(P_Q(q_x|t_i)\ln P_Q(q_x|t_i)\right) \quad (9)$$

$$\text{score}_u(t_i) = 1.0 + \frac{1}{\ln|U|}\sum_{x=1}^{|U|}(P_U(u_x|t_i)\ln P_U(u_x|t_i)) \quad (10)$$

Event Image Selection:

Doubtlessly, events associated with interesting images are very attractive to the users. So to obtain event related photographs is to search image search engines with these event queries. But we get many irrelevant images. Therefore to obtain event related images there are two steps: image similarity measures and image reranking. In image similarity measures both local and global features are admitted. The global feature identifies duplicates images completely where as local feature identifies duplicate images partially. For the local feature based similarity measurements scale invariant feature transform (SIFT) is used. In the image reranking, to support the image which has duplicates in social events' photos weighting score is defined. More the duplicates in events' pictures, the more essential the photo is. Likewise, to oppose the photos as in the profile topics' photos again weighting score is defined. The final reranking score is calculated using these two weighting scores. The top scores are considered to be the most illustrative pictures of that event.

IV. RESULTS AND DISCUSSION

By proposing SNMF framework, it detects the interesting events from web search log and gives set of illustrative photographs identified with occasions for storyboard.

We make use of web search logs rather than news articles and weblogs since these are good sources of data as they may cover diverse realistic events and exactly mirror users' interests. The SNMF is an effective approach that guarantees nonnegative weights for each topic and categorize events from others based on query semantics, temporal characteristics and search log figuring. To give a global and clear storyboard, set of appropriate images generated by our method are attached to each bit of news.

For quantitative performance estimation we construct ground truth table with data set as benchmark. To evaluate the system performance we use classic precision and recall measures. The social events are arranged in the decreasing order as indicated by their scores after event ranking. Thus for each famous person c_i, the precision and recall are calculated according to top topics in the ranking list.

$$\text{prec}_{\text{micro}} = \frac{\sum_{i=1}^{n}\text{matched}(c_i)}{\sum_{i=1}^{n}\text{gt}(c_i)}$$

$$\text{rec}_{\text{micro}} = \frac{\sum_{i=1}^{n}\text{covered}(c_i)}{\sum_{i=1}^{n}\text{gt}(c_i)} \quad (11)$$

$$\text{prec}_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{matched}(c_i)}{\text{gt}(c_i)}$$

$$\text{rec}_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{covered}(c_i)}{\text{gt}(c_i)} \quad (12)$$

Where $gt(c_i)$ is the number of events in the c_i's ground truth list and n is the number of persons.

The micro averages compute the performance at the event level, where as the macro averages estimate the performance at the celebrity level. At last we compare the overall performance of our proposed approach with three different methods. One is the abnormal query-based method, the second one is the query-URL clustering as in [8] and third one is time pattern prediction as in [1]. For the ease of comparison, the number of abnormal queries and the number of clusters are set to ground-truth event $gt(c_i)$. The comparison results are shown in table 1. It is clear that our proposed approach achieves better performance than other three approaches.

Table 1 Comparison of Overall Performance with Other Approaches.

	Micro		Macro	
	Precision	***Recall***	***Precision***	***Recall***
Abnormal query-based solution	0.23	0.16	0.32	0.29
Query-URL clustering	0.43	0.48	0.52	0.56
Time pattern	0.49	0.44	0.58	0.57
Our method	0.58	0.56	0.67	0.61

At last, the storyboard will be produced utilizing the chosen events with important photographs. To guarantee the high caliber of the photographs, low visual quality photographs will be wiped out at first. In addition, time and location settings are other key fact for the storyboard photographs. For each distinguished event, the event time and location will be separated at first. We utilize the surrounding writings of the photographs at Web to distinguish whether the photograph is important to our recognized event. Those photographs with for sure and location distinction contrasted and recognized event will be positioned in the low portion.

V. CONCLUSION

In our model, we make use of search logs to create community event storyboards inevitably. Not at all like regular content mining, search logs have short, limited text requests and the size of information is significantly more noteworthy than some news web sites or blogs. In view of these features, we don't utilize the query content information to do the analysis. Structure and estimation information are used to get the topics and event disclosure in our work, which can fit the information well. Besides, we include time data to SNMF to make it uncomplicated to find social affairs differentiated with existing NMF approaches. Our work performs superior to conventional works around there, since we can recognize the topics in a way that perceive the events which attracts typical users. The related pictures were made up the storyboard in a course of events to display good depiction of the mined events using the photo search features and associations.

REFERENCES

[1] S. Arora, R. Ge, and A. Moitra, "Learning topic models–going beyond SVD," in Proc. IEEE 53rd Ann. Symp. Found. Comput. Sci. (FOCS), Oct. 2012, pp. 1–10.

[2] N. Babaguchi, S. Sasamori, T. Kitahashi, and R. Jains, "Detecting events from continuous media by intermodal collaboration and knowledge use," in Proc. IEEE Int. Conf. Multimedia Comput. Syst., vol. 1. Jul. 1999, pp. 782–786.

[3] P. N. Bennett *et al.*, "Modeling the impact of short- and long-term behavior on search personalization," in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2012, pp. 185–194.

[4] Y.-J. Chang, H.-Y. Lo, M.-S. Huang, and M.-C. Hu, "Representative photo selection for restaurants in food blogs," in Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW), Jun./Jul. 2015, pp. 1–6.

[5] H. L. Chieu and Y. K. Lee, "Query based event extraction along a timeline," in Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2004, pp. 425–432.

[6] C. Alexander, B. Fayock, and A.Winebarger, "Automatic event detection and characterization of solar events with IRIS, SDO/AIA and Hi-C," in Proc. AAS/Solar Phys. Division Meeting, vol. 47. p. 1, 2016.

[7] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Probabilistic query expansion using query logs," in Proc. 11th Int. Conf. World Wide Web, 2002, pp. 325–332.

[8] Q. Zhao, T.-Y. Liu, S. S. Bhowmick, and W.-Y. Ma, "Event detection from evolution of click-through data," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 484–493.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Proc. Adv. Neural Inf. Process. Syst., 2001, pp. 556–562.