Auto Storyboard Generation Using Web Search Log

Mutharaju K S [1], Manukumar N [2], Pradeep Gowda A.B[3] , Prashanth P [4], Smt. Ranjana S Chakrasali[5]

[1,2,3,4]Department of Computer Science, BNM Institute of Technology, Bengaluru- 560-070, Karnataka, India

[5]Assistant Professor, Department of Computer Science, BNM Institute of Technology, Karnataka, India

***Abstract*— Social Events are responsible for the major percentage of web search traffic as shown by recent studies. This paper is a survey conducted for identifying the various event detection methods which are useful for event mining. While traditional Web sites can only show human-edited events, this Paper suggests a System which automatically detects events from search log data and generate a storyboard where the events are arranged along a timeline. Web search log is considered as a good data resource for event mining, since they reflect people's interests directly and wide variety of real world events are covered in it. 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.**

***Keywords*: Event, Storyboard, Social media, Log data, Smooth Non-negative matrix Factorization, Event detection, Photo Selection.**

I. INTRODUCTION

People are curious about other's activities and that is obviously the nature of social creatures. Life style of a celebrityis being the interest of people in today's world and it has been reported to be true in modern years. Engines such as Google, Bing experience massive search demands for current affairs or any hot news and events extracted from search log of search engine. Hot news or events shown on the websites are mostly originated from the Professional editors. Instead of manual efforts, it will be fruitful to extract such eventsfrom search made by people around world [1].

Search engines show the information or summary of celebrity in as a profile containing basic information. Thus, through which people are able to get celebrity's basic information from a summarized view such as their nationality, their age, birthday, their works, social interest and awards.

Study on commercial search engine data through statistics have shown that 30% of search queries aims at searching an informationof real-time events [13].Remaining set of queries are said to be about a celebrities their information and their life style in real-time such as their social work and so on.Thus people are more interested in the above specified domains. The traditional websitesare said to be controlled by human editors known as reporters, which leads to limitations of the website. Primarily, these websites are limited meaning not a scalable websites. Even for a particular domains in which user is interested, information about few celebrity will be provided. Second, the information covered by human controlled websites are limited i.e., reporter of one web site focuses on limited number of domains for a particular celebrity. For example a website with url www.POPSugar.com focuses mainly on singers or on stage performers.Finally, information on these websites are said to be biased by the reporters. Thus to overcome the above limitation we in this paper propose a system that extractsreal-time events from log data automatically and generate anevent story board for each and every event detected.

Log data is the input to the system. Search log data is parsed to detect Events and storyboards are generated where this detected events are sorted based on timestamp. Web search is said to be a better data to be used as a input for the event detection, because :(1) Wide range of real-time events are covered by them. (2) User's interests are directly reflected in it. (3) They respond to real-time events [2]. Proposed approach in this paper contains 2 consecutive stages as shown in figure 1: (1) Real-time event detection by Smooth non Negative Matrix Factorization (SNMF) and (2) Photo selection for detected events. Initially, real-time events are detected from search log data using SNMF. This event detected is considered to have a higher search frequency which process is known as topic factorization[11]. Ranking of this detected events is carried out based on frequency of search and highlighted. Now once the event has been detected, images/photos for each particular event has to be selected, this is done by sending a top query to search engines like Google to download images relevant to social events[5]. The suitable image related the event is chosen for the Storyboard. Storyboard contains a Heading of the event, an image and a short description or summary about the event.

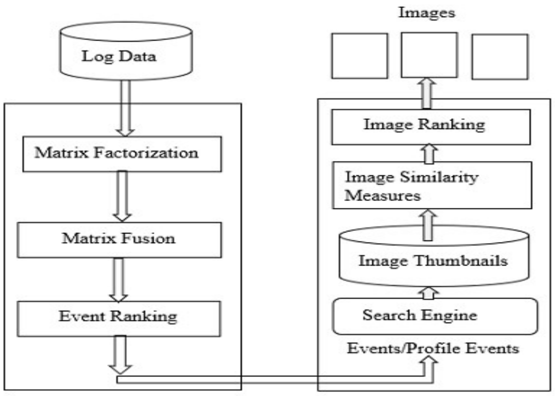


Figure 1: Overview of the Proposed System

II. LITERATURE SURVEY

[1] *Jun Xu et al., "Automatic Generation of Social Event Storyboard From Image Click-Through Data" IEEE Trans. On circuits and system, vol.28, No. 1, January 2018.*

The approach proposed in this paper has two components. 1) Event detection and 2) Representative event photo selection.

In Event detection process, the topics with high frequency or the topics which appear concurrently were discovered using the topic factorization methods.

This method eliminated the least occurred topics in the query set which were considered as the random noise.

This paper focuses on detecting the social events from the query set, Hence in the event detection step large number of topics are considered and the topics which are correlated or similar are merged into a single topic. Social events are detected and highlighted by introducing a rank function.

In the photo selection step, the topic detected in the previous step is sent to the Search engines like Google or Bing. Two sets of Image thumbnails are collected from the search engines which are likely to match the Social event detected. To select a photo which is more appropriate to the Social event, the similarity between the two image sets is measured using the rank function. The photos related to the event, should have duplicate or similar images in the Social image set, but there should not be no similar images in the Profile image set. This is the assumption made in the above paper for selecting a photo for the storyboard. A ranking function is used to sort the photos in the social event image set. Finally a photo with a top rank which is the most appropriate photo for the detected event is chosen.

The above two processes i.e. Topic detection and Representative Event photo selection were considered for generating a storyboard.

[2]Sreelekshmi.U and GopuDarsan, "Social Event Storyboard Generation from Image Click" International Journal of Engineering and Computer Science, Vol 5 Issue 11 Nov. 2016, Page No. 18987-18992.

This paper mainly focuses on search log data stating that, it is a good source for the detection of event for the following reasons

1)People interest are directly reflected in the search log. 2)Varieties of real world events are covered in the search log. 3)They respond to the real time events.

What people search is their interest, mainly trending events, current affairs, information about celebrities are searched which gets appended in the search log.

To detect a topic from the search log data, an approach called Smooth Non-Negative Matrix Factorization (SNMF) is used.

It has two basic ideas. 1) It promotes event queries. 2) Events from the popular queries are differentiated.

SNMF assigns weights to each topic which is non-negative and also considers time factor in the process of event detection, which makes it easier. Relevant photos are attached for the each detected event and is presented as a storyboard.

[3]S. Essidet al., "Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring," IEEE Trans. Multimedia, vol. 15, no. 2, pp. 415–425, Feb. 2013.

Similar to most of topicdetection algorithms, the Non-negative Matrix Factorization (NMF) considered as a standard algorithm also ignores the order in which documents are inputted. In other way, the decomposition resultswould not be affected by the permutation of information in columns of matrix. Somehow, in the log mining process, the temporal order is anedge factor i.e., a difference between queries (and related topics) from any two adjacent days must not be significant.

Thus to embed this constraint of NMF, SNMF was proposed that introduced an extra S(H) as aregularization factor to the cost function. S(H) acts as a regulation factorthat smoothness (reduces distance) between values contained in two adjacent columns of H, and λ is a nonnegative weight that adjusts the smoothing. In general, we choose a relatively large number of topic, thus ensuring that all events has been covered.

[4] H. Liu,et al., "Detecting and tracking topics and events from Web search logs," ACM Trans. Inf. Syst., vol. 30, no. 4, 2012, Art. no. 21.

Provides insight of Topic Detection and Tracking (TDT) as a process that explores the techniques of detecting new topics and track their reappearance (duplication) and evolution. TDT involves three tasks as shown in fig 2 :Segmentation, Detection and Tracking.

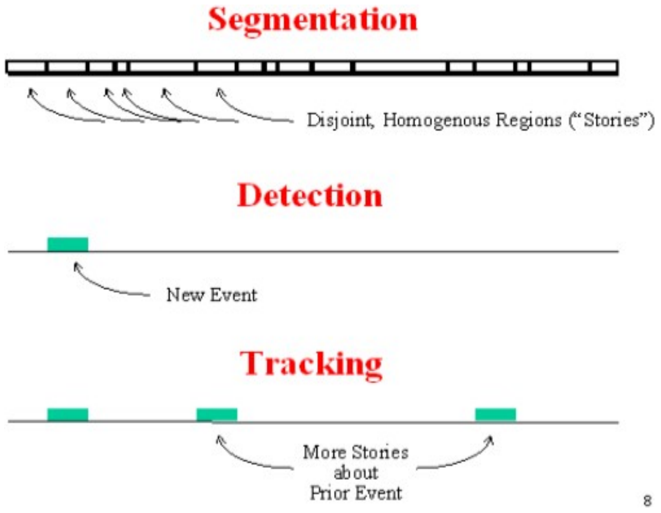


Figure 2: Illustrating Segmentation, Detection and Tracking

Segmentation means to divide the events into parts, or segments, which are definable, accessible, actionable called stories. Detection means the process of detecting or identifying a new event . Tracking is carried out to check for any duplication of prior detected events

[5] S. F. Chang, Y. Rui and T. S. Huang, "Image retrieval: Current techniques, promising directions, and open issues," J. Vis. Commun. Image Represent., vol. 10, no. 1, pp. 39–62.

To select a relevant images, a similarity measure is computed, both global image feature and local image features of a particular image are computed in this paper. Global features are computed to identify duplicatebased on a full image, and are particularly used to identify fully duplicate images. On the other hand, local features are used to identify partial duplicates by computing a local patches on image and thus used for recognizing partial duplicates. partial duplicate of images have been a quite importantstep, since many images in search engine would have been edited.

The global feature of an image is computed by the block-based intensity histogram. Wherein, each image is divided into 64 blocks of 8X8 matrix form, then 256-D intensity histogram is computed for each block which is based on the pixels within that particular block. The similarity between two images Ix and Iyusing global feature is definedas:

$$\text{sim}_{\text{hist}}(I_x, I_y) = \max\left\{1.0 - \frac{1}{64}\sum_{i=1}^{64}\left\|\mathbf{g}_i^{I_x} - \mathbf{g}_i^{I_y}\right\|_2, 0\right\}.$$

To compute the local feature-based similarity between images, a classic scale-invariant feature transform (SIFT) feature is employed for the process of identifying duplicates. This SIFT process considers each pixel of the image and compares it with the same pixel in other image and the value is computed. When the value computed is greater than the specified threshold then the image is considered to be similar.

III. PROPOSED METHODOLOGIES

Event Detection by SNMF

Event detection: As a very first step the real-time events are detected using search-log data. Then a group of queries are discovered having high frequency, this process is termed as topic factorization.

The search log data taken as input in form of a matrix V which is of the size |Q|×|H|. Every single row in Videntifies a query searched by people and every column in V identifies day. Thus every tuple in V(Vij) indicates theith query that was searched by people around world on the jth day. SNMF computes two other nonnegative matrices W and H that satisfies the product as shown fig 3: V=WXH

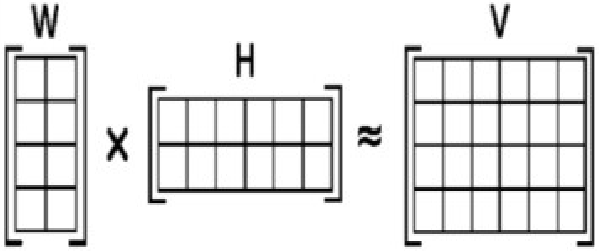


Figure 3: Illustration of approximate SNMF

W where each column identifies a topic. H where each column identifies the decomposition coefficients of topics for that day.

In the log mining process, the temporal order is anedge factor i.e., a difference between queries (and related topics) from any two adjacent days must not be significant.

Thus to embed this constraint of NMF, SNMF was proposed that introduced an extra S(H) as aregularization factor to the cost function. S(H) acts as a regulation factorthat smoothness (reduces distance) between values contained in two adjacent columns of H, and λ is a nonnegative weight that adjusts the smoothing.

Topic Fusion: Once after the detection of events we will be having K number of topics along with the 2 nonnegative matrix W and H.The topic distribution along with words used in query and the timeline plays an important role in characterizing a topic. Here W is used to get the query vocabulary and the H matrix to obtain timeline of a particular event. Log data's are also used as a useful clue that provides a list of URL for effective clustering of query. Meaning the queries triggering the same URL is said to be the query for similar topics. Above 3 clues defined are used to measure the similarity between topics that get fused providing only distinct or unique topics.

(1) Similar topic detection using queries: For any given topic Tk, the kth column of W is used for measuring its distribution over the queries

(2) Similar topic detection using timeline: By normalizing the kth row in H matrix, a Tk topics distribution over the timeline is measured

(3) Similar topic detection using URLs: Log data's are also used as a useful clue that provides a list of URL for effective clustering of query. Meaning the queries triggering the same URL is said to be the query for similar topics.

Event Photo Selection

*(1) Image Similarity Measures:*To select a relevant images, a similarity measure is computed, both global image feature and local image features of a particular image are computed in this paper. Global features are computed to identify duplicatebased on a full image, and are particularly used to identify fully duplicate images. On the other hand, local

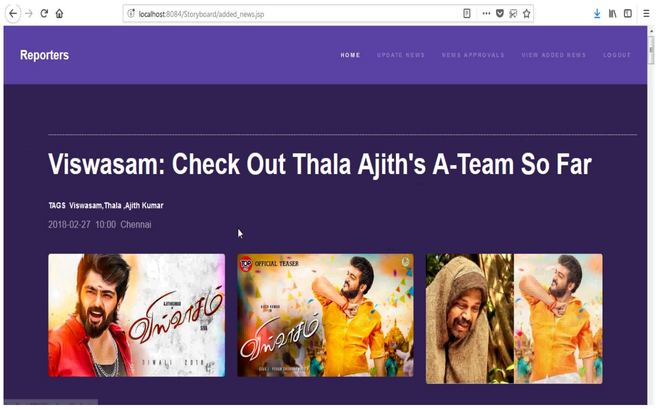
features are used to identify partial duplicates by computing a local patches on image and thus used for recognizing partial duplicates. partial duplicate of images have been a quite importantstep, since many images in search engine would have been edited.

The global feature of an image is computed by the block-based intensity histogram. Wherein, each image is divided into 64 blocks of 8X8 matrix form, then 256-D intensity histogram is computed for each block which is based on the pixels within that particular block.

To compute the local feature-based similarity between images, a classic scale-invariant feature transform (SIFT) feature is employed for the process of identifying duplicates. This SIFT process considers each pixel of the image and compares it with the same pixel in other image and the value is computed. When the value computed is greater than the specified threshold then the image is considered to be similar.

(2) Event Photo Re-ranking: After measuring the similarity between the images each image Ix is ranked based on the similar images found. Higher the duplicate images found higher is its rank.

After Event detection and photo selection the Storyboard is generatedas shown in the following figure 4.


localhost:8084/Storyboard/added_news.jsp

Reporters

Viswasam: Check Out Thala Ajith's A-Team So Far

TAGS Viswasam,Thala ,Ajith Kumar

2018-02-27 10:00 Chennai

Figure 4: Storyboard generated for the automatically detected event.

REFERENCES

[1]. Jun Xu, Tao Mei, RuiCai, Houqiang Li and Yong Rui, “Automatic Generation of Social Event Storyboard From Image Click-Through Data” IEEE Trans. On circuits and system, vol.28, No. 1, January 2018.

[2]. Sreelekshmi.U and GopuDarsan, “Social Event Storyboard Generation from Image Click” International Journal Of Engineering And Computer Science, Vol 5 Issue 11 Nov. 2016, Page No. 18987-18992

[3]. S. Essid and C. Févotte, “Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring,” IEEE Trans. Multimedia, vol. 15, no. 2, pp. 415–425, Feb. 2013.

[4]. H. Liu, J. He, Y. Gu, H. Xiong, and X. Du, “Detecting and tracking topics and events from Web search logs,” ACM Trans. Inf. Syst., vol. 30, no. 4, 2012, Art. no.21.

[5]. S. F. Chang, Y. Rui and T. S. Huang, “Image retrieval: Current techniques, promising directions, and open issues,” J. Vis. Commun. Image Represent., vol. 10, no. 1, pp. 39–62.

[6]. D. M. Blei, “Probabilistic topic models,” *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[7]. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[8]. 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.

[9]. 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.

[10]. T.-C. Chou and M. C. Chen, “Using incremental PLSI for threshold resilient online event analysis,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 3, pp. 289–299, Mar. 2008.

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

[12]. S. Essid and C. Févotte, “Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring,” *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 415–425, Feb. 2013.

[13]. Q. Mei, C. Liu, H. Su, and C. Zhai, “A probabilistic approach tospatiotemporal theme pattern mining on weblogs,” in *Proc. 15th Int.Conf. World Wide Web*, 2006, pp. 533–542.

[14]. T. Hofmann, “Probabilistic latent semantic indexing,” in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1999, pp. 50–57.

[15]. T. Joachims, “Optimizing search engines using clickthrough data,” in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 133–142.