INVOLUNTARY CREATION OF SOCIAL OCCURRENCE STORYBOARD FROM PICTURE CLICK-THROUGH INFORMATION

[1]Kavyashree S, [2] Mrs. ShashiRekha H
[1]Student, [2] Asst. Professor
[1] Department of Computer Science and Engineering
[1]Visvesvaraya Technological University Department of PG Studies, Regional Office
Mysuru, Karnataka, India

Abstract : **— Traditional websites were motivated by human-edited events which lead to huge web search traffic. This project conducted for identifying the a variety of event detection methods which are useful for event mining. Moreover this paper also suggests an involuntary system to detect events from search log data and produce storyboards where the events are approved along a timeline. Image search log is treated as a good data reserve 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 part for event detection as different events develop at different time. In addition, to provide a visually attractive storyboard, each event is mapped with a set of related images arranged along a timeline. These related images are involuntarily generated from image search results by analyzing both local and global image content.**

***Index Terms*- Occurrence storyboard, Social media, Click-through data, Non-negative matrix Factorization, Image search**

I. INTRODUCTION

A ***storyboard*** is a graphic organizer in the form of illustrations or images displayed in sequence for the purpose of previsualizing a motion picture, animation, motion graphic or interactive media sequence. One advantage of using storyboards is that it allows (in film and business) the user to experiment with changes in the storyline to evoke stronger reaction or interest. Flashbacks, for instance, are often the result of sorting storyboards out of chronological order to help build suspense and interest. Another benefit of storyboarding is that the production can plan the movie in advance. In this step, things like type of camera shot, angle, and blocking of characters are decided.

The process of visual thinking and planning allows a group of people to brainstorm together, placing their ideas on storyboards and then arranging the storyboards on the wall. This fosters more ideas and generates consensus inside the group.

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.

Automatic audiovisual document structuring represents a key technological component as part of the global effort to set up efficient multimedia and video indexing tools. Though there seems to be no consensual definition of this process, it is widely accepted that it is one of extracting a temporal organization of an audiovisual document, by arranging it into different sections, or structural units, each conveying a homogeneous (audio/video) type of content (possibly highlighting content repetitions). The definition of a "structural unit" highly depends both on the particular type of content that is processed and the application considered, for which a human-generated ground truth is generally available for a set of manually annotated documents. Then, the structuring problem comes down to automatically recreating the documents temporal-organization ground truth (obviously in view of automatic all y structuring new documents that have not been manually annotated). As such, shot boundary detection or scene segmentation, also referred to as sequence, story unit or logical unit segmentation, etc., can be considered as instances of video structuring problems. Other works consider more specific structuring tasks and rely on expert techniques specifically tailored for the particular structuring scheme that is envisaged. A number of proposals employ supervised approaches exploiting prior knowledge on the general structure of the type of documents to be processed and using domain rules and specific concept or event detectors.

II. RELATED WORK

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.

[1]Topic Detection and Tracking (TDT) 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 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 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.

[2]Event detection in Twitter 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.

[3]Introduction to probabilistic topic models, 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.

[4]The Query based event extraction 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 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.

[5]A study on retrospective and online event detection 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).

[6]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.

II. METHODOLOGY

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 nonnegative 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: 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.

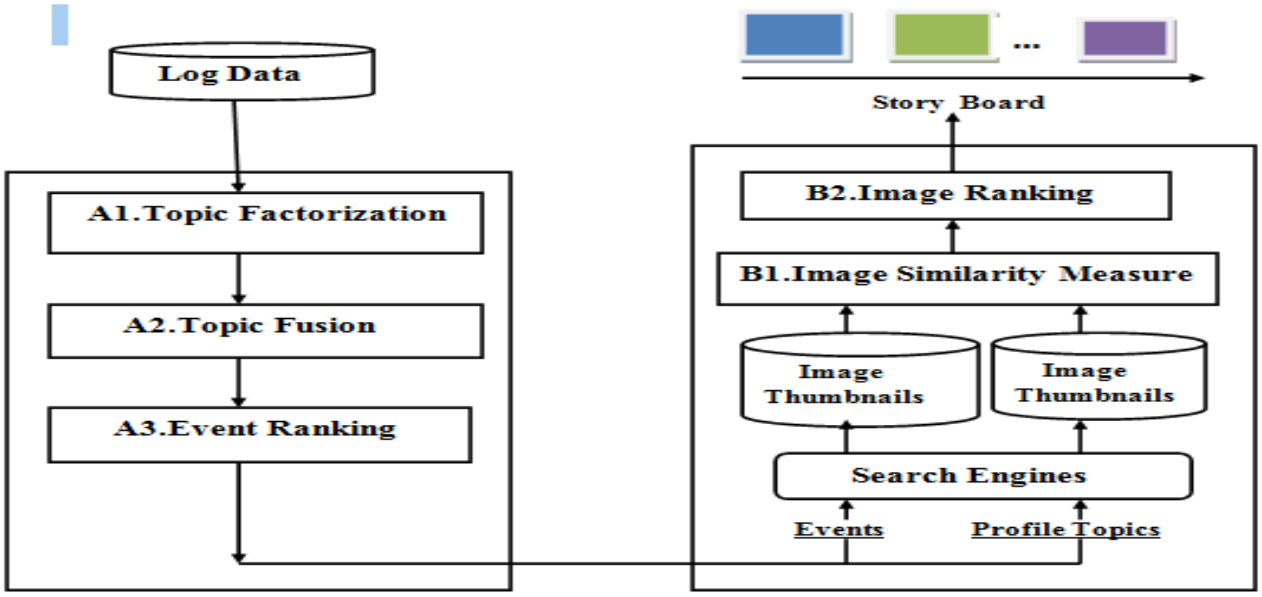

Figure 1: System Architecture

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 V_{ij} represent i^{th} query on j^{th} 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.

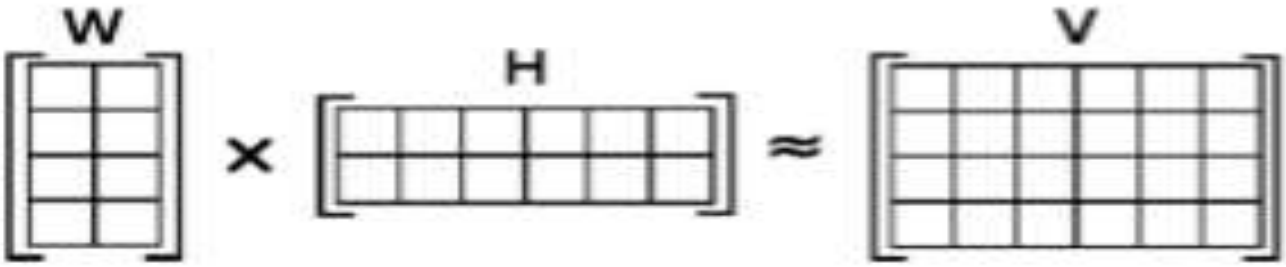

Figure 2: Matrix view

Illustration of approximate non-negative matrix factorization the matrix V is represented by the two smaller matrices W and H. 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 nonnegative 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 proposed algorithm is implemented using bottom up approach of multi agent system (MAS). The MAS are used to define functionality of state awareness. These concepts are demonstrated in simulation environment. An algorithm is an effective method expressed as a finite list of well-defined instructions for calculating a function. Starting from an initial state, the instructions describe a computation that when executed will proceed through a finite number of well defined successive states, eventually producing "output" and terminating at a final state. Failures in the instructions will result in faulty output. A program can hence be viewed as an elaborate algorithm.

SHA-1 Algorithm A cryptographic hash (sometimes called digest) is a kind of signature for a text or a data file. A hash is not an encryption, it cannot be decrypted back to the original text (it is a one-way cryptographic function, and is a fixed size for any size of source text).

SHA-1(160 bit message) Algorithm Framework:

Step 1: Append Padding Bits….

Message is "padded" with a 1 and as many 0's as necessary to bring the message length to 64 bits less than an even multiple of 512.

Step 2: Append Length....

64 bits are appended to the end of the padded message. These bits hold the binary format of 64 bits indicating the length of the original message.

Step 3: Prepare Processing Functions….

SHA1 requires 80 processing functions defined as:

f(t;B,C,D) = (B AND C) OR ((NOT B) AND D) (0 <= t <= 19)
f(t;B,C,D) = B XOR C XOR D (20 <= t <= 39)
f(t;B,C,D) = (B AND C) OR (B AND D) OR (C AND D) (40 <=t<=59)
f(t;B,C,D) = B XOR C XOR D (60 <= t <= 79)

Step 4: Prepare Processing Constants....

SHA1 requires 80 processing constant words defined as:

K(t) = 0x5A827999 (0 <= t <= 19)
K(t) = 0x6ED9EBA1 (20 <= t <= 39)
K(t) = 0x8F1BBCDC (40 <= t <= 59)
K(t) = 0xCA62C1D6 (60 <= t <= 79)

Step 5: Initialize Buffers….

SHA1 requires 160 bits or 5 buffers of words (32 bits):

H0 = 0x67452301
H1 = 0xEFCDAB89
H2 = 0x98BADCFE
H3 = 0x10325476
H4 = 0xC3D2E1F0

Step 6: Processing Message in 512-bit blocks (L blocks in total message)….

This is the main task of SHA1 algorithm which loops through the padded and appended message in 512-bit blocks.

Input and predefined functions:

M[1, 2, ..., L]: Blocks of the padded and appended message
f(0;B,C,D), f(1,B,C,D), ..., f(79,B,C,D): 80 Processing Functions
K(0), K(1), ..., K(79): 80 Processing Constant Words
H0, H1, H2, H3, H4, H5: 5 Word buffers with initial values

Step 7: Pseudo Code….

```
For loop on k = 1 to L
        (W(0),W(1),...,W(15)) = M[k] /* Divide M[k] into 16 words */
        For t = 16 to 79 do:
        W(t) = (W(t-3) XOR W(t-8) XOR W(t-14) XOR W(t-16)) <<< 1
        A = H0, B = H1, C = H2, D = H3, E = H4
        For t = 0 to 79 do:
            TEMP = A<<<5 + f(t;B,C,D) + E + W(t) + K(t) E = D, D = C,
                        C = B<<<30, B = A, A = TEMP
        End of for loop
      H0 = H0 + A, H1 = H1 + B, H2 = H2 + C, H3 = H3 + D, H4 = H4 + E
  End of for loop
```

Output:

H0, H1, H2, H3, H4, H5: Word buffers with final message digest
The message digest of the string:
"This is a test for theory of computation"
4480afca4407400b035d9debeb88bfc402db514f

We have proposed a new PDP scheme (referred to as MB-PMDDP), which supports outsourcing of multi-copy dynamic data, where the data owner is capable of not only archiving and accessing the data copies stored by the CSP, but also updating and scaling these copies on the remote servers. To the best of our knowledge, the proposed scheme is the first to address *multiple* copies of *dynamic* data. The interaction between the authorized users and the CSP is considered in our scheme, where the authorized users can seamlessly access a data copy received from the CSP using a single secret key shared with the data owner. Moreover, the proposed scheme supports public verifiability, enables arbitrary number of auditing, and allows possession-free verification where the verifier has the ability to verify the data integrity even though he neither possesses nor retrieves the file blocks from the server.

Through performance analysis and experimental results, we have demonstrated that the proposed MB-PMDDP scheme outperforms the TB-PMDDP approach derived from a class of dynamic single-copy PDP models. The TB-PMDDP leads to high storage overhead on the remote servers and high computations on both the CSP and the verifier sides. A slight modification can be done on the proposed scheme to support the feature of identifying the indices of corrupted copies. The corrupted data copy can be reconstructed even from a complete damage using duplicated copies on other servers. Through security analysis, we have shown that the proposed scheme is provably secure.

IV. EXPERIMENTAL OVERVIEW

This is the most generalized method for storyboard. This experiment gives the most appropriate method for understanding the storyboard

.

A. Event Detection by SNMF:

The most straightforward way to discover events from search log data is to identify "abnormal" queries. For example, for the well-known singer Adele, the query "Adele pregnant" is somewhat abnormal in comparison to more common queries, such as "Adele lyrics" and "Adele mp3."As a result, the evidence of an event becomes obscure, as we cannot integrate the statistics of correlated queries. Experimental results reported later in this project show the limitations of this simple solution. To deal with noisy and sparse data, topic modeling (or topic factorization) has proved to be an effective approach, especially for text mining. Through topic modeling, high-dimensional sparse data are projected into a low dimensional topic space, in which the correlations among original feature dimensions are embedded. Topic modeling is also good at suppressing random noise.

B. Event Photo Selection by SNMF:

For each detected social event, it is straightforward to identify a set of most relevant queries by inspecting the event's distribution in the query space. The simplest way to get events related photos is to directly search commercial image search engines with these event queries.

V. CONCLUSION

This Project 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 involuntarily 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.

VI. ACKNOWLEDGMENT

I would like to thank Dr.K.Thippeswamy Ph.D., Professor & Chairman, Dept of studies in Computer Science &Engineering and Ms.ShashiRekha H M.Tech., Assistant Professor, Dept of studies in Computer Science &Engineering, VTU Regional office,Mysuru and Anonymous reviewers encouragement and constructive piece of advice that have prompted us for new round of Rethinking of our research, additional experiments and clear presentation of technical content.

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. Wengand B.-S.Lee.Evet 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,RuiCai,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.
[7].C. Alexander, B. Fayock, and A. Winebarger. Automatic event detection and characterization of solar events with iris, sdo/aia and hi-c. In AAS/Solar Physics Division Meeting, volume 47, 2016.
[8].X.Wang and A.McCallum. Topics over time:A non-markov continuous time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 424–433. ACM, 2006.
[9].Q. Zhao,T.-Y. Liu,S. S. Bhowmick, and W.-Y. Ma.Event detection from evolution of click-through data. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 484–493. ACM, 2006.
[10].H. Liu, J. He, Y. Gu, H. Xiong, and X. Du. Detecting and tracking topics and events from web search logs. ACM Transactions on Information Systems (TOIS), 30(4):21, 2012