Proposed Model for Twitter Data Summarization

Dr. Sandeep M. Chaware
Head of the Dept., Department of Computer Science Engineering, Marathwada Mitra Mandal's College of Engineering, Karvenagar, Pune, India
Miss. Avanti Pandit, Miss. Ashwini Pawar, Miss Pratiksha Tambe, Miss Nirmal Thakare.
B. E Students, Department of Computer Science Engineering, Marathwada Mitra Mandal's College of Engineering, Karvenagar, Pune, India

Abstract- Microblogging offerings have revolutionized the way human beings trade facts. Confronted with the ever-growing numbers of tweets with multimedia contents and trending topics, it's far proper to offer visualized summarization to assist users to quickly hold close the essence of topics. While existing works normally attention on text-based strategies best, summarization of a couple of media sorts (e.g., text and image) are scarcely explored. In this paper, proposes a twitter summarization framework to automatically generate visualized summaries for trending topics. Specifically, a novel generative probabilistic model, termed multimodal-LDA (MMLDA), is proposed to find subtopics from microblogs by means of exploring the correlations amongst different media kinds. The contribution work is, to extract microblogs which shows the text, image and video posts. We evaluate the widespread experiments on an actual-international Twitter microblog dataset to illustrate the prevalence of our proposed technique towards the modern-day processes.

Keywords- Twitter, Tweets, Summarization, Trending Topic, Social Media, OCR, MMLDA

I. INTRODUCTION

Users are allowed to share multimedia content on such platforms, such as news, images and video links. With the wide availability of information sources, rapid information propagation and ease of use, twitter has quickly become one of the most important media for sharing, distributing and consuming interesting contents, such as the trending topics. Currently, some microblogging platforms, such as Twitter, offer users the list of (manually created) hot trending topics, together with a set of related microblogs in each trend. Such service offers a potentially useful way to help users to conveniently gain a quick and concise impression of the current hot topics. In addition, users may obtain further understanding of the topics by browsing the related mircoblogs. However, due to the tremendous volume of microblogs and the lack of effective summarization mechanism in existing trending topic services, users are often confronted with incomplete, irrelevant and duplicate information, which makes it difficult for users to capture the essence of a topic. Therefore, it would be of great benefit if an effective mechanism can be provided to automatically mine and summarize subtopics (i.e., divisions of a main topic) from microblogs related to a given topic.

Motivation

1. A twitter summarization method to automatically generate visualized summaries for trending topics.

2. Images can supplement the textual content with additional information, especially in the circumstance of microblogging, where the text lacks sufficient expressive power as aforementioned.
3. Multimedia contents can facilitate subtopic discovery.
4. Incorporating concrete multimedia exemplars into summarization can assist users to gain a more visualized understanding of interesting topics and/or subtopics.

II. RELATED WORK

[1] P. Sinha, S. Mehrotra, and R. Jain, "Summarization of personal photologs using multidimensional content and context," in Proc. 1st ACM Int. Conf. Multimedia Retrieval, 2011, p. 4.

Proposed methods to compute quality, diversity and coverage properties using multidimensional content and context data. The proposed metrics which will evaluate the photo summaries based on their representation of the larger corpus and the ability to satisfy user's information needs. Advantages are: The greedy algorithm for summarization performs better than the baselines. Summaries help in effective sharing and browsing of the personal photos. Disadvantages are: Computation is expensive.

[2] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2010, pp. 912–920.

In multi-document summarization, redundancy is a particularly important issue since textual units from different documents might convey the same information. A high quality (small and meaningful) summary should not only be informative about the remainder but also be compact (non-redundant). Advantages are: The best performance is achieved. Submodular summarization achieves better ROUGE-1 scores. Disadvantages are: The proposed system very expensive to solve.

[3] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi, "Eddi: Interactive topic-based browsing of social status streams," in Proc. 23nd Annu. ACM Symp. User Interface Softw. Technol., 2010, pp. 303–312.

Eddi is a novel interface for browsing Twitter streams that clusters tweets by topics trending within the user's own feed. An algorithm for topic detection and a topic-oriented user interface for social information streams such as Twitter feeds. (1) benchmark TweeTopic against other topic detection approaches, and (2) compare Eddi to a typical chronological interface for consuming Twitter feeds. Advantages are: A simple, novel topic detection algorithm that uses noun-phrase detection and a search engine as an external knowledge base. Eddi is more enjoyable and more efficient to browse than the traditional chronological Twitter interface. Disadvantages are: Users had access to our clients for a limited time, making it difficult to extrapolate conclusions on how the tool might be used longitudinally. Users were viewing the history of their feed rather than tweets they had never seen before, making our task slightly less realistic.

[4] P. Goyal, L. Behera, and T. M. Mcginnity, "A context-based word indexing model for

document summarization," IEEE Transactions on Knowledge & Data Engineering, vol. 25, no. 8, pp. 1693–1705, 2013.

Proposes the novel idea of using the context sensitive document indexing to improve the sentence extraction-based document summarization task. In this paper, proposes a context sensitive document indexing model based on the Bernoulli model of randomness. Advantages are: The new context-based word indexing gives better performance than the baseline models. Disadvantages are: Need to calculate the lexical association over a large corpus.

[5] D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 66–73.

In this paper we argue that for some highly structured and recurring events, such as sports, it is better to use more sophisticated techniques to summarize the relevant tweets. The problem of summarizing event-tweets and give a solution based on learning the underlying hidden state representation of the event via Hidden Markov Models. Advantages are: The advantage of leveraging existing query matching technologies and for simple one-shot events such as earthquakes it works well. The HMM is able to learn differences in language models of sub-events completely automatically. Disadvantages are: The disadvantage that SUMMHMM has to account for tweet words that only occur in some of the events, but not in others.

III. OPEN ISSUES

Text summarization is performed for the purposes of saving users time by reducing the amount of content to read. However, text summarization has also been performed for purposes such as reducing the number of features required for classifying or clustering documents. Some microblogging platforms offer users the list of (manually created) hot trending topics, together with a set of related microblogs in each trend. Such service offers a potentially useful way to help users to conveniently gain a quick and concise impression of the current hot topics. In addition, users may obtain further understanding of the topics by browsing the related mircoblogs. However, due to the tremendous volume of microblogs and the lack of effective summarization mechanism in existing trending topic services, users are often confronted with incomplete, irrelevant and duplicate information, which makes it difficult for users to capture the essence of a topic. Some of the issues are given below:

- The lack of effective summarization mechanism.
- Users are often faces with incomplete, irrelevant and duplicate information due to existing trending topics services.
- It makes difficult for users to capture the essence of a topic.

IV. SYSTEM OVERVIEW

Traditional documents that contain only textual objects, tweets constitute of multiple media types, such as image and text. In this paper proposes a novel framework to summarize tweets for trending

topics. Specifically, first proposes a novel generative probabilistic model, called multimodal-LDA (MMLDA), to partition the microblogs relevant to the same topic into different subtopics. MMLDA model is capable of not merely capturing the intrinsic correlation between visual and textual information of microblogs, but also estimating the general distribution as well as subtopic-specific distribution under a trending topic. For text summarization, specifies three criteria, namely coverage, significance and diversity to measure the summarization quality. For visual summarization, a two-step process is devised to automatically select the most representative images: 1) images within a subtopic are grouped by spectral clustering; and 2) images in each group are ranked by a manifold ranking algorithm and the top-ranked image is selected as representative. The Fig. 1 shows the block diagram of twitter summarization system. The processes of generating textual and visual summaries for each subtopic, by utilizing the reinforced textual/visual distributional information. Then, the textual and visual summaries are aggregated to form a comprehensive multimedia summary.

The contribution work is, to extract microblogs which shows the video posts. These videos are separated in two parts: one is number of images and second is speech transcriptions. We evaluate the widespread experiments on an actual-international Twitter microblog dataset to illustrate the prevalence of our proposed technique towards the modern-day processes.

A. Block Diagram

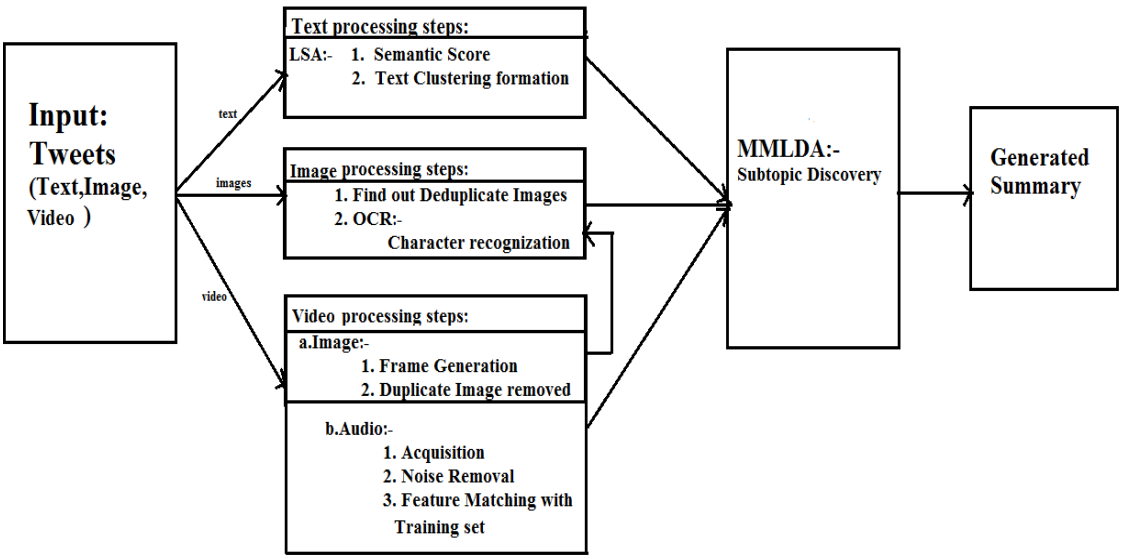

Fig. 1 Block diagram

Advantages

- It provides to automatically mine and summarize subtopics (i.e., divisions of a main topic) from microblogs related to a given topic.

- Microblogs comprise of multiple media types, such as image and text and video.
- Multimedia contents can facilitate subtopic discovery.
- Well organizing the messy tweets into structured subtopics.
- Generating high quality textual summary at subtopic level.
- Selecting images relevant to subtopic that can best represent the textual contents.

B. Mathematical Model:

Consider the textual summary generation of the k-th subtopic from the subset S_k^t. Denote $G_k \subseteq S_k^t$ as the summary set consists of the selected textual samples, and $\widetilde{S_k^t} = S_k^t - G_k$ is the remaining subset. In order to determine which sample is subsequently selected from $\widetilde{S_k^t}$, we calculate a selection score for each sample by considering coverage, significance and diversity as follows.

Coverage:

Given the current summary set G_k, the new sample T_i to be selected should be the one that makes the new summary $(i.e., G_k \cup \{T_i\})$ achieve the best coverage (i.e., minimize the distance between $\Theta_{G_k \cup \{T_i\}}$ *and* φ_K^{TS}). Therefore, the coverage of each candidate T_i could be measured by the following equation:

$$u_C(T_i) = D_{KL}(\Theta_{G_k \cup \{T_i\}} || \varphi_K^{TS}) \qquad (1)$$

Significance: In general, the popularity of a microblog can be revealed from the repost number. A large repost number means that the microblog has gained a lot of attention and interest from other users, and can indirectly represent the quality of this microblog. Therefore, we use the repost number to measure the significance of a candidate:

$$u_S(T_i) = log(RepostNum(T_i) + 1) \qquad (2)$$

Diversity: We take the information redundancy into consideration in sample selection. Consider a candidate Ti, the redundancy it brings to the summary set can be measured by the similarity between this candidate and the previously generated summary, which is:

$$u_D(T_i) = D_{KL}(\Theta_{T_i} || \Theta_{G_k}) \qquad (3)$$

C. Algorithms

1. MMLDA Algorithm

Steps:

1. For the topic T, draw $\varphi^{TG} \sim Dir(\lambda^{TG})$ and $\varphi^{VG} \sim Dir(\lambda^{VG})$ denote the general textual distribution and visual distribution, respectively. $Dir(\cdot)$ is the Dirichlet distribution. Then draw $\phi^Z \sim Dir(\beta^Z)$, which indicates the distribution of subtopics over the microblog collection corresponding to T.
2. For each subtopic, draw $\varphi_K^{TS} \sim Dir(\lambda^{TS})$ and $\varphi_K^{VS} \sim Dir(\lambda^{VS})$, $k \in \{1, 2, \ldots, K\}$, correspond to the specific textual distribution and visual distribution.
3. For each microblog M_i, draw $Z_i \sim Multi(\phi^Z)$, corresponds to the subtopic assignment for Mi. Multi(·) denotes the Multinomial distribution. Then draw $\phi_i^R \sim Dir(\beta^R)$ indicates the general-specific textual word distribution of M_i. Similarly, draw $\phi_i^Q \sim Dir(\beta^Q)$ indicates that for visual words.
4. For each textual word position of M_i, draw a variable $R_{ij} \sim Multi(\phi_i^R)$:

- If R_{ij} indicates General, then draw a word $W_{ij} \sim Multi(\varphi^{TG})$.
- If R_{ij} indicates Specific, draw a word W_{ij} from the Z_i-th specific distribution $W_{ij} \sim Multi(\varphi_{Z_i}^{TS})$

5. The generation of visual words is similarly done as in step 4.

2. Optical character recognition (OCR) Algorithm

Step 1: Image Preprocessing

Step 2: Edge Detection

Step 3: Detection of Text Regions

Step 4: Enhancement and Segmentation of Text Regions

3. Hidden Markov Model (HMM) algorithm for speech recognition:

A HMM is characterized by 3 matrices viz., A, B and PI.

A - Transition Probability matrix $(N \times N)$

B - Observation symbol Probability Distribution matrix $(N \times M)$

PI - Initial State Distribution matrix $(N \times 1)$

Where, N =Number of states in the HMM

M = Number of Observation symbols

After can apply HMM for speech recognition by using following steps:

1. Recursive procedures like Forward and Backward Procedures exist which can compute P (O|L), probability of observation sequence.

Forward Procedure:

Initialization:

$\alpha_1(i) = \pi_i b_i o_1,\ 1 \le i \le N$

Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}),$$

$$1 \le t \le T-1, 1 \le j \le N$$

Termination

$$P(O|\lambda) \sum_{i=1}^{N} \alpha_T(i)$$

Backward Procedure:

Initialization:

$\beta_T(i) = 1,\ 1 \le i \le N$

Induction

$$\beta_T(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1}) \beta_{t+1}(j),$$

$$T-1 \le t \le 1, 1 \le i \le N$$

Termination

$$P(O|\lambda) \sum_{i=1}^{N} \alpha_T(i)$$

2. The state occupation probability $t(sj)$ is the probability of occupying state sj at time t given the sequence of observations

$$O_1, O_2, \dots, O_N.$$

3. Baum-welch algorithm for parameter re-estimation.

V. RESULT AND DISCUSSIONS

Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 7, MySQL 5.1 backend database and Jdk 1.8. The application is web application used tool for design code in Eclipse and execute on Tomcat server. Some functions used in the algorithm are provided by list of jars like Twitter-core and Twitter-stream jars etc.

Tweets are retrieved in a streaming way, and Twitter provides the Streaming API for developers and researchers to access public tweets in real time. The aim of this paper is to bridge the gap by carrying out a performance evaluation, which was from two different aspects of NLP and machine leaning algorithms. Some functions used in the algorithm are provided by list of jars like standford core NLP jar for keywords extraction using POS tagger method. This is very useful for implementation of MMLDA algorithm. The OCR algorithm is used to extract text from image using tesseract training dataset. The resultant outcome is to display the list of tweets with images and videos summary using MMLDA algorithm.

The Micro-blog Summarization Framework is based on twitter posts. Fig. 2 shows that when summarizing micro-blog data which gives better accuracy. The image file applies the OCR algorithm for textual part extraction and generates summary. The Video file first fragment the no. of frames and finally apply on unique images OCR algorithm for text extraction and audio signals convert into digital and HMM model convert into text format.

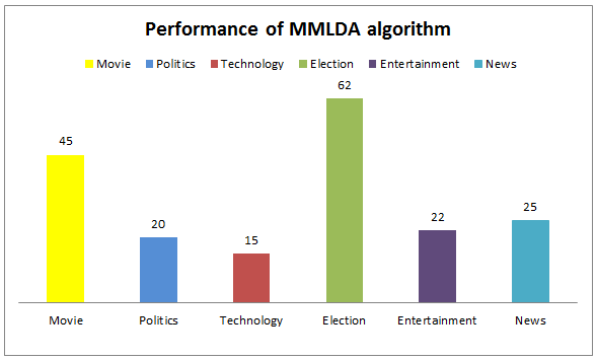

Fig. 2 Performance graph for Twitter Summarization framework

VI. CONCLUSION

In this paper, we proposed a multimedia microblog summarization method to automatically generate visualized summaries for trending topics. Tweet comprise of multiple media types, such as image and text and video. Specifically, a novel multimodal-LDA (MMLDA) model was proposed to discover various subtopics as well as the subtopic content distribution from microblogs, which explores the correlation among different media types. Based on MMLDA, a summarizer is elaborated to generate both textual and visual summaries. Well organizing the messy tweets into structured subtopics. Generating high quality textual summary at subtopic level.

References

[1] P. Sinha, S. Mehrotra, and R. Jain, "Summarization of personal photologs using multidimensional content and context," in Proc. 1st ACM Int. Conf. Multimedia Retrieval, 2011, p. 4.

[2] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2010, pp. 912–920.

[3] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi, "Eddi: Interactive topic-based browsing of social status streams," in Proc. 23nd Annu. ACM Symp. User Interface Softw. Technol., 2010, pp. 303–312.

[4] P. Goyal, L. Behera, and T. M. Mcginnity, "A context-based word indexing model for

document summarization," IEEE Transactions on Knowledge & Data Engineering, vol. 25, no. 8, pp. 1693–1705, 2013.

[5] D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 66–73.