Web Event Topic Analysis by Topic Feature Clustering and Extended LDA Model

Yuanzi Xu
School of Computer Science and Technology, Shandong University, Jinan, China
xu_yuanzi@163.com

Qingzhong Li, Zhongmin Yan and Wei Wang
School of Computer Science and Technology, Shandong University, Jinan, China
lqz@sdu.edu.cn

***Abstract*—To analyze topics of a large number of web events, we proposed an event topic analysis approach by topic feature clustering and extended LDA (latent dirichlet allocation) model. The extended LDA model is dimension LDA (DLDA) which integrates topic probability of LDA model. We represent an event as a multi-dimensions vector and use DLDA model to select topic feature words in events. We aggregate events which have a common topic by topic feature clustering. In clustering process we use dynamic K-means method to automatically select suitable number of clusters. In this paper a topic term generating rule is proposed to compose topic terms by clustered topic feature words. We accurately detect a common topic from lots of different events and analyze topic terms for events. Experiments on dataset results show that the web event topic analysis approach has high accuracy.**

***Index Terms*—Event topic analysis, DLDA model, Topic feature clustering, Topic term generating rule**

I. INTRODUCTION

Topic analysis is highly valuable for numerous different web events and has been widely studied in recent years. Topic analysis of web event can detect a common topic from many different events and provide valuable information for market intelligence. These information help enterprise policymakers understand themselves and know the development trends of other enterprises. Event topic analysis is important in information retrieval, data integration and topic detection and tracking (TDT). According to the large scale of web events, we propose an approach for analyzing topics of lots of different events by topic feature clustering and extending LDA model.

An event is an activity that occurs at a special time and involves participants. Some events have a common topic although time and participants of events are obviously different. Web event topic analysis can find the common topic from different events. Fig. 1 shows different events which are extracted from webpages and microblogs but they belong to a common topic. The first event reports the sales quantity of a new electronic product of Microsoft. The second and the third events report the new electronic product sales quantity of Apple Inc and Xiaomi company. Although these events are obviously different, they have a common topic which is the sale of new products. Event topic is closely related to the activity and participants of an event. In this paper we need to detect common topic from events and give these events a suitable topic term.

Topic: The sale of new products
In 2010,Microsoft sold 8,000,000 kinect cameras in 60 days and kinect became the fastest selling electronic product in the world.
In 2011, Apple Inc sold 4,000,000 iPhone4S in three days, and sold 33,000,000 in 78 days. Its new product sales record has exceeded kinect.
Xiaomi company convened a meeting in 2012, and announced that MI mobile phone since published had been sold 3,520,000 during one year.

Figure 1. Events and their common topic

Aggregating common topic events and giving them suitable topic terms poses multiple interesting technical challenges. First, an event contains syntactic structure and semantic information, this information is useful for topic features selection. We use eight dimensions to present an event and detect latent topic feature words in dimensions so that this method can combine syntactic structure and semantic information of events. Second, many events belong to a common topic but the time and participants of events are different. We need to measure the content of every dimension respectively and compose a final result. Third, according to large scale of web events, we need to aggregate events which belong to the same topic and compose a suitable topic term for every topic cluster.

We propose a dimension latent dirichlet allocation (DLDA) model which integrates dimension information with LDA model. LDA model [1-2] can be used for discovering underlying topics from text documents, but it ignores syntactic structure of events and distributes probability on topic unrelated words. Due to the shortcoming of LDA, some people improved it. We improve it by integrating dimension topic probability with it. In this paper we use dimensions to represent an event and it is conducive to combine syntactic structure information and semantic information for event topic

doi:10.4304/jsw.9.4.977-984

analysis. The dimensions are denoted as {agent, activity, object, time, location, cause, purpose, manner}. We obtain prior topic information about the topic ability on different dimensions and select some topic feature dimensions, such as agent, activity, and object. DLDA model selects latent topic feature words in these dimensions. We aggregate events by topic feature clustering that is computing the similarity of the content of topic feature dimensions. In clustering process we extend K-means clustering method for dynamic cluster quantity selection. We use different similarity measure methods to compute the similarity of the different topic feature dimensions. In this way we can effectively cluster many different events which belong to the same topic. A topic terms generating rule is proposed to compose suitable topic terms for every topic cluster. We obtain topic feature words in every cluster after clustering， use topic feature words to compose topic terms by generating rule and dictionary.

In this paper we proposed an event topic analysis approach. We use it to detect a common topic of many different web events and choose a suitable topic term for these events. We use dimensions to represent an event and present DLDA model to integrate dimension information for latent topic words selection. For lots of events analysis, we use topic feature clustering to aggregate events and propose a topic terms generating rule to compose topic terms for every cluster. Compared to other methods, this approach analyzes a common topic from many events even though the time and participants of these events have obvious differences.

The remainder of this paper is organized as follows. Section 2 presents the related work. We describe event topic analysis problem and explain some concepts in Section 3. We present DLDA model and describe some details in section 4. In section 5, we describe our experimental evaluations and report our results. We conclude our paper and discuss the future work in Section 6.

II. Background

Analyzing topic of events is an important research in information retrieval, data integration and topic detection and tracking, etc. Recent topic analysis research has focused on web events from microblogs and webpages. The web event is a sentence which has latent topic. Latent dirichlet allocation (LDA) model [1-2] can be used for discovering underlying topics from text documents. Some researches use it to analyze the topic of microblogs [3-7]. LDA model has some shortcomings [8] so some people improved the model [9]. For example, it ignores syntactic structure of events and distributes probability to topic unrelated words. D. Ramage [4] presents the content of microblog into dimensions and in this way he can characterize different microblogs to meet users' interest. Daniel did not combine the result of every dimensions but a comprehensive result is important for topic analysis. Dong proposes an extended evidence theory [10] for composing different matchers to get a comprehensive measure result and this method can be used to combine the similarity of dimensions.

For the character of large number of web events, clustering events which have common topic is an effective method for topic analysis [11-12]. Unsupervised and simple clustering method may adapt to the large scale web data, and many researchers use K-means clustering method for event topic clustering [13]. Existing works are interested in identifying the most suitable cluster quantity [14-15]. In this way K-means method eliminates the limit of human determine cluster quantity and it is important for topic analysis. These researches did not recognize topic terms for clusters. Q. Li [16] has presented a key phrase identification program (KIP) to find significant topic terms for a given document. In order to improve the ability of presenting all keywords' meanings, some works use dictionary for topic terms recognition [17-18].

We use eight dimensions [19] to present an event and propose a DLDA model which provides dimension supervision to improve the accuracy of topic feature words selection. We use different similarity matchers to measure the content of different dimensions and compose similarity result. In this way we can detect topic more accurate than other methods for different events. For large scale of different web events, we aggregate them to find a common topic and use cohesion degree measure function [15] to select suitable number of clusters in K-means clustering method. Compared to many researches do not select a topic term for events which belong to a common topic, we propose a topic terms generating rule, and use this rule to compose suitable topic terms according to aggregated topic keywords.

III. Problem Definition

To make a clear presentation and facilitate the following discussions, we describe event topic analysis problem and explain some concepts of our approach in this section.

Event topic analysis needs to detect a common topic of many different events. We propose an event topic analysis approach by topic feature clustering and extended LDA model. DLDA model is a kind of extended LDA model which integrates topic feature information with LDA and it analyzes event topic more accurate than LDA model. There are two concepts of our approach.

- **Event:** An event is an activity that occurs at a special time and involves participants. In this paper we use eight dimensions to represent an event. An event can be denoted as {agent, activity, object, time, location, cause, purpose, manner}.
- **Topic feature dimension:** An event is represented as eight dimensions and the dimension which can provide more topic information than others is topic feature dimension. For example, activity, agent and object. These topic feature dimensions are the basis of topic feature clustering.
- **Event topic:** Event topic is an inductive and general theme of many different events. Common topic is mainly derived from the participant and

activity from many events although time and location of these events are obviously different.

- **Event topic analysis:** Event topic analysis is effectively detecting a common topic of events and giving these events a suitable topic term.

In this paper our goal is to analyze a common topic of many different web events, we cluster events which belong to a topic and choose suitable topic terms for clusters. Topic feature clustering aggregates the similar contents in topic feature dimensions. These events in a common cluster have a same topic. Event topic analysis can denote as $C_T = \{c_{t1},c_{t2},...,c_{ti}\} = \{\{e_{11},e_{12},...,e_{1k}\},\{e_{21},e_{22},...,e_{2p}\},\ldots, \{e_{i1},e_{i2},...,e_{in}\}\}$. C_T is a topic set, c_{ti} is a cluster that contains some events with a common topic, and k, p, n are the number of events in every topic cluster. The process of our approach is described in Fig. 2. We aggregate common topic events in topic clustering stage and we choose suitable topic terms for clusters in topic terms generating stage.

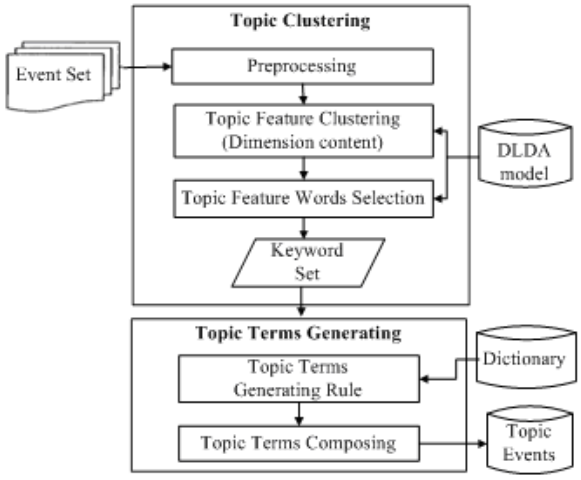

Figure 2. The process of event topic analysis

In Fig. 2, the input of process is an event set which contains many different events. The outputs are some event clusters and events in a cluster belong to a topic. The process of event topic analysis is divided into two stages. In topic clustering stage, we preprocess the event set for representing each event in eight dimensions. We use DLDA model to cluster events by topic feature clustering which aggregates the content of topic feature dimensions and to select topic feature words. Then we obtain a keyword set from clusters. In topic terms generating stage, we present a topic terms generating rule. We use topic terms generating rule and dictionary to compose topic terms by keywords for every topic event cluster.

IV. Event Topic Analysis

In this section we explain DLDA model and some detail of event topic analysis approach such as topic feature clustering, topic feature words selection and topic terms generating.

A. DLDA Model

Latent dirichlet allocation (LDA) model is a generative probabilistic model which is used for topic feature words detection. It models the words under the "bag-of-words" assumption and considers topic is the probability distribution of words in a document. According to this assumption, LDA model ignores the syntactic structure information of documents and distributes topic probability to topic unrelated words. In order to utilize syntactic structure information and improve the accuracy of topic detection for events, we propose DLDA model which integrates topic information of dimensions with LDA model. We represent an event in dimensions, select some topic feature dimensions and use DLDA model to improve the accuracy of topic feature words selection.

We represent an event in eight dimensions and use words to represent the content of dimensions by word segmentation system ICTCLAS 2010. Although each dimension is a mixture of latent topics, the ability of topic providing in every dimension is different. Due to topic prior information, we select some dimensions as topic feature dimensions (i.e., activity, agent and object) and use DLDA model to select latent topic words in topic feature dimensions. DLDA model uses the existence of a set of labels d as dimensions, and each dimension is characterized by a multinomial distribution β over all words. Fig. 3 represents the graphic model of DLDA.

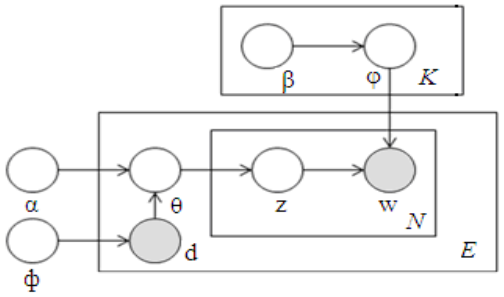

Figure 3. Graphical model representation of DLDA

For each event e in set E, DLDA model first picks a multinomial distribution $\theta_e = [\theta_{e1},\ldots, \theta_{em}]^T$ from the dirichlet distribution $\alpha_e = [\alpha_{e1},\ldots, \alpha_{em}]^T$, and then the model assigns a topic $z_{ie} = k$ to the ith word in the event. We use d to denote a dimension and use ϕ to denote the topic distribution of every dimension. Compared with LDA model which distributes probability to topic unrelated words, DLDA model adds dimension information to model and only distributes probability to topic related dimensions. Some dimensions, like time and location, DLDA model do not distribute topic probability to them. Given the topic $z_{ie} = k$ (k is topic number), the model pick words w from vocabulary of dimensions according to the distribution φ which is generated from the dirichlet distribution β for each topic k. We use maximum likelihood estimation to find the parameters of the dirichlet distribution. Paramenter α can be estimated by prior probability distribution, $z_{ie}|\varphi^{(di)} \sim Dirichlet(\alpha)$. In order to solve the maximum likelihood estimation from a huge dataset, the Gibbs sampling is used to approximate the solution. We use the variant of Gibbs sampling [20] to approximate the solution.

$$P(z_{ie}=k|z_{-ie},w_{ie}) = \frac{\frac{n_{-ie,k}^{(d_i \cdot w_{ie})}+\alpha}{n_{-ie,.}^{(d_i)}+T\alpha}}{\sum_{k=1}^{T}\frac{n_{-ie,k}^{(d_i \cdot w_{ie})}+\alpha}{n_{-ie,.}^{(d_i)}+T\alpha}} \quad (1)$$

In this formula, z_{ie} = k represents distributing w_{ie} to topic k. w_{ie} not only presents word w, but also is associated with the dimension this word from. z_{-ie} is all distribution for $z_{ie} \neq$ k, $n_{-ie,k}^{(d_i \cdot w_{ie})}$ is the number of words allocated to topic k and similar to w_{ie}., $n_{-ie}^{(d_i)}$ is the number of words allocated to topic k in the feature dimension of an event.

B. Topic Feature Clustering

We select activity, agent and object dimensions as topic feature dimensions. Because of dimension content is constructed by words, clustering the content of topic feature dimensions is important for event topic analysis.

We cluster the content of topic feature dimensions to analyze the common topic for events. In clustering process we use K-means clustering method. K-means clustering method is a simple and effective method adapted to large scale data, but it has a weakness for static cluster number k. It is hard for people to predict the most suitable k and we use a cohesion degree measure function [15] which can reflect cohesion degree in clusters and can monotonically change in the iterative process until terminate. The following is definition of this function and we use the function to select suitable k automatically in clustering process.

$$E(C) = \sum_{r=1}^{k} \frac{\sum_{p_i,p_j \in Cr} sim(p_i, p_j)}{n_r} \quad (2)$$

In this function, p_i and p_j are two data points which denote two events, n_r is the number of events and c_r is the number of topic clusters. The number of topic clusters should be as little as possible, so we use a function [15] to punish the large k and make cohesion degree measure function converge to the small direction. The function is shown as follows.

$$E'(C) = \left(1 - \frac{2k}{n}\right) E(C) \quad (3)$$

In the function, k is the number of topic clusters and n is the number of events. The following is topic cluster algorithm with dynamic K-means clustering method.

We propose a topic cluster algorithm for event topic clustering as follow.

In this algorithm we first select the k cluster centers randomly and the number of clusters may dynamic change in algorithm. We compose the k clusters according to the similarity of cluster centers and events. Second we compute the connectivity of clusters and merge two clusters if their connectivity in the range of [T/4, T]. We compute new cluster centers, repeatedly measure similarity and connectivity until E'(C) beyond the designated range. The termination conditions are E'(C) becomes little (less than T/4) or C[] has only two independent clusters. We obtain the events clustering result according to different topics at last.

Algorithm 1. Topic cluster algorithm

Input: The event set E and every e in this set is composed of topic feature dimensions, similarity threshold T
Output: The array C[] which contains events clustering result

```
(1)  k=| √E |          /*use the integer part of square root of E
     as current cluster quantity*/
(2)  Randomly select k events as cluster centers {c1,c2,…,ck},
     C[i] is initial cluster which only has ci as cluster center in
     cluster
(3)    For i=1 to k do;
(4)     for each e ∈ E do
(5)       Simtopic = MeasureSimilarity( ci , e );
(6)       if Simtopic >T
(7)         Put event e into C[i] and compose initial clusters
     /* put event which has common topic of ci in cluster */
(8)       else qcur= computeconnectivity(C[i],C[i+1])
     /*compute E'(C) of current clusters */
(9)         if (qcur>T/4 )&&(qcur< T );
(10)          C[]=Mergecluster(C[i],C[i+1]);
(11)          k-- ;
(12)          Compute new cluster centers and repeatedly
     measure similarity in line 5.
(13)        else unmerge any cluster
(14)      Compute E'(C) and decide whether it reaches
     termination conditions
(15)    End for
(16)   End for
(17) Return C[]
```

Figure 4. Algorithm for event topic clustering

Most of topic feature words are verbs and nouns. The great majority of verbs are from activity dimension and nouns are from agent and object dimensions. To compute the similarity of topic feature words, we use Hownet [21] to compute the semantic similarity for verb words. In the Hownet semantic network architecture, a word is composed of primitives. The activity verb itself may be the primitive or can be deconstructed into primitives. We use primitive S_a={a_i | i=1,2,…,m} to denote verb a, S_b ={b_j | j=1,2,…,n} to denote verb b, and use depth-based semantic matcher to measure the similarity of verbs.

$$Sim_{Verb}(S_a, S_b) = \frac{2 \times depth(a_i, b_j)}{depth(a_i) + depth(b_j)} \quad (4)$$

In this formula depth(a_i ,b_j) is the whole depth of the common ancestor of primitives in Hownet, depth(a_i) and depth(b_j) denote the depth of each primitive with a common ancestor.

We use word element similarity [22] to computer the similarity of nouns and other words, use an extended evidence theory [10] to distribute weight of different words and compose complete similarity. We cluster events which belong to the same topic by topic feature clustering.

C. Topic Feature Words Selection

After clustering events belonged to a common topic, we use DLDA model to select topic feature words. Topic feature words are some words in the content of topic feature dimensions. We use KL dispersion formula [23] to compute the discrete degree between dimension topic

distribution and event topic distribution. Using DLDA model to compute topic distribution probability in dimensions function is shown as follow.

$$P(d \mid t) = \sum_{w_i \in d} P(w_i \mid t) \times P(e \mid t) \quad (5)$$

In this function , $P(w_i \mid t)$ is topic distribution in words which may compute by DLDA, $P(e \mid t)$ is topic distribution in events and we use distribution function [23] to compute them.

We use E_{kl} as dispersion factor to denote the discrete degree between dimension topic distribution and event topic distribution. We can select topic feature words in following function.

$$f_{topic}(w_i) = \gamma \cdot E_{kl}(P(t \mid d) \| P(t \mid e)) \quad (6)$$

In this function γ is -1 by experiment and we select topic feature words which has high probability.

Due to topic clustering and topic feature words selection, we obtain some topic feature words from every cluster.

D. Topic Terms Generating

After clustering events according to common topic, we need to composite topic terms for every topic event cluster. We propose a topic terms generating rule for composing topic terms from clustered topic feature words.

- **Topic terms generating rule:** According to clustered topic feature words and dictionary, use merging, replacing and concluding steps to compose topic terms for every cluster.

We introduce the three steps in generating rule. In the merging step we merge words with same value from keywords. Two keywords A-B and B-C can be merged as A-B-C. For example, two keywords 'product sell' and 'sell quantity' can be merged as 'product sell quantity'. In the replacing step we use an existing keyword which has more entire meaning to replace other similar keywords. For example, we use 'company and product' to replace 'product'. After the two steps we obtain some keywords as candidate topic terms. We compose suitable topic terms according to these candidate topic terms in the concluding step. A topic event cluster C can be presented as a set of candidate topic terms. For example, C= $\{e_1,e_2,\ldots,e_k\}$ =$\{\{w_{11},\ldots,w_{1n}\},\{w_{21},\ldots, w_{2m}\},\ldots,\{w_{k1},\ldots, w_{kp}\}\}$, w_{11} and w_{kp} are candidate topic terms. We can compute probability distribution of topic candidate terms by the merging and replacing times and obtain two kinds of topic candidate terms. (1) Primary topic candidate terms. Terms have higher probability distribution than others by sorting. (2) Secondary topic candidate terms. Terms have lower probability distribution than others by sorting.

We use dictionary to select words which contain the meanings of topic candidate terms. We use a cover degree factor μ to show the ability of topic terms. The higher value of μ indicates the stronger ability to cover the meaning of topic candidate terms.

Web event topic analysis contains detecting common topics from many different events and choosing topic terms for topic clusters. We aggregate events belong to a common topic and compose suitable topic terms for topic clusters by proposed topic terms generating rule.

V. Evaluation and Results

In this paper we proposed an event topic analysis approach which extends LDA model to analyze topic. We use real data and do a series of experiments to evaluate the effectiveness of our approach.

A. Data Set

We extracted 10,625 events from Sina micoblogs and news webpages (i.e., biz.163.com and finance.ifeng.com). These events are collected in food and phone fields from 2010 to 2012. These events constitute two experiment datasets. Before experiment we preprocess events by ICTCLAS segmentation system and represent every event in eight dimensions.

B. Experiment Evaluation

The result of web event topic analysis is some event clusters. These events in a cluster have a common topic. We use an information retrieval evaluation method and divided the cluster results into four sets.

A = True Positives (events that are clustered in a cluster is correct)
B = False Negatives (events that are not clustered in a cluster is incorrect)
C = False Positives (events that are clustered in a cluster is incorrect)
D = True Negatives (events that are not clustered in a cluster is correct)

The precision, recall, and F-measure are calculated as follow.

$$Precision = \frac{|A|}{|A| + |C|} \quad (7)$$

$$Recall = \frac{|A|}{|A| + |B|} \quad (8)$$

$$F - measure = \frac{2 \times \mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \quad (9)$$

These evaluation functions, "precision" evaluates the soundness of the clustering, "recall" evaluates the cover degree of correct clustering, and "F-measure" is the comprehensive evaluation. We test web event topic analysis approach from four aspects in experiments. (1)We compare the accuracy of topic feature clustering method to other methods with similar usage. (2) We compare using DLDA to select topic feature words with other methods. (3) We evaluate the effectiveness of topic terms generating rule. (4) We compare the accuracy of proposed approach with others for event topic analysis.

C. Experiment Results

(1). We compare the accuracy of topic feature clustering method to other methods with similar usage.

TABLE I
MAIN EVENT TOPIC CLUSTERS IN DATASET

Topic	The scale of new product	Stock trading	Develop New function	Product defect	Adjust price	Company meeting
The number of events	1076	993	967	874	651	492

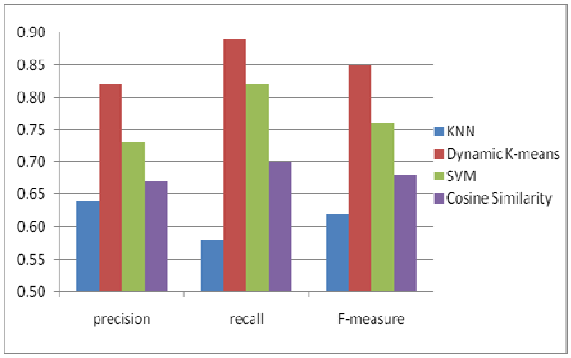

Figure 5. Compare accuracy of clustering methods in food dataset

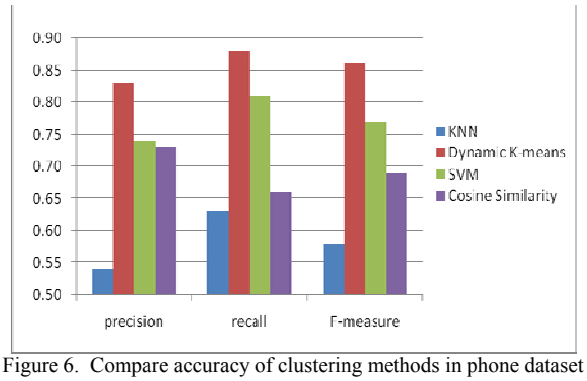

Figure 6. Compare accuracy of clustering methods in phone dataset

In this paper we use dynamic K-means clustering method to aggregate the content of topic feature dimensions as event topic feature clustering. The cluster results are many topic clusters and each cluster contains some events which belong to a common topic. We select main event topic clusters according to the number of events in every cluster and showed them in Tab. I.

Then we use three other clustering methods to compare the accuracy with topic detection. Fig. 5 and Fig. 6 show evaluations of cluster results about KNN, Dynamic K-means (proposed topic clustering method), SVM and cosine similarity clustering method.

Dynamic K-means method is our proposed topic clustering method and it is obviously better than KNN, SVM and cosine similarity method for event topic clustering in Fig.5. We use dimensions to present an event, and use DLDA model to distribute topic probability to related dimensions. We use different similarity matchers to compute the content of topic feature dimensions. Therefore, this method has higher cluster accuracy than other methods. In Fig.6, the experimental result in phone dataset also presents our approach has obvious advantage. Fig. 5 and Fig. 6 show that our proposed method has high cluster accuracy in different datasets. We can get the conclusion that the advantage of our approach for topic clustering is not limited in a certain field.

(2). We compare using DLDA model to select topic feature words with other methods.

In this paper we use DLDA model to select topic feature words. We choose 600 events and label topic feature words by people as correct result. The number of events in topic clusters and their topic terms are shown in Tab. II.

TBALE II
EVENTS IN TOPIC CLUSTERS AND THEIR COMMON TOPIC TERMS

Topic	Product sell	Goods depreciate	Stock trading	Product defect	Quality detection	Company meeting
The number of events	98	101	106	95	107	93

In this experiment we use following function to compute the precision of selecting topic feature words.

$$Precision = \frac{N_{correct}}{N_{total}} \quad (10)$$

$N_{correct}$ is the number of words that selected by method are corresponded to people labeled. N_{total} is the number of all words in test events. We use DLDA model, LDA model, TFIDF [24] and Z-SCORE [24] to find topic feature words. Tab. III shows the precision of these methods for topic feature words selection.

Tab. III shows that the DLDA method has the best topic feature words selection result in four methods. LDA model can find the latent topic in events but does not perform well. DLDA model provides dimension supervision for topic feature words selection and it has better result than LDA. The results of TFIDF and Z-SCORE are both worse than DLDA. This experiment shows that using DLDA model to select topic feature words in events is more accurate than others.

TBALE III
THE PRECISION OF FOUR METHODS FOR TOPIC FEATURE WORDS SELECTION

Topic	Precision			
	DLDA	LDA	TFIDF	Z-SCORE
Product sell	83.4%	78.1%	51.3%	32.6%
Goods depreciate	82.2%	75.8%	54.6%	35.4%
Stock trading	90.5%	86.4%	57.4%	49.8%
Product defect	85.7%	79.6%	63.2%	57.5%
Quality detection	86.9%	82.3%	60.8%	46.2%
Company meeting	91.6%	85.2%	62.7%	39.1%

(3). We evaluate the effectiveness of topic terms generating rule.

We proposed a topic terms generating rule for composing topic terms. We use a cover degree factor μ to evaluate the cover ability of topic terms. We fix the number of topic terms (i.e., three words) and use function 9 to evaluate effectiveness. The correct topic terms are given by people and in this experiment $N_{correct}$ is the number of words that composed by rule are corresponded to people given. N_{total} is the number of all composed topic terms. Tab. Ⅳ shows the precision of different μ for topic terms.

TBALE IV
THE EFFECTIVENESS OF DIFFERENT μ

μ	0.5	0.6	0.7	0.8
F-measure	94.8%	90.3%	85.7%	72.6%

Tab. Ⅳ shows that the F-measure rate decreased with the increased μ. Because fixed the number of topic terms, some meanings of candidate topic terms do not be contained. Although we improve the cover degree may influence effectiveness, the F-measure is also higher than 70%.

(4). We compare the accuracy of proposed approach with others for event topic analysis.

In this paper we use dynamic K-means method as topic clustering method to aggregate same topic events and use DLDA model to select topic feature words. We have evaluated the results of topic clustering and topic feature words selection. Tab. Ⅴ is the accuracy of event topic analysis approach.

TABLE V
THE ACCURACY OF DIFFERENT APPROACHS

Approach	Precision	Recall	F-measure
LDA	78.4%	80.3%	79.4%
LDA+SVM	83.6%	86.9%	85.2%
DLDA+Dynamic K-means	89.7%	92.1%	90.9%

DLDA model provides dimension supervision for topic feature words selection and dynamic K-means clustering method has higher accuracy than other methods. So the comprehensive result has higher accuracy than LDA and LDA + SVM approaches. These experiment results show that our approach can effectively detect a common topic from many different events and give suitable topic terms for topic clusters.

VI. Conclusion

In this paper, we proposed a topic analysis approach for lots of web events by topic feature clustering and extending LDA model. We use dimensions to represent an event and present a DLDA model which integrates topic dimension probability with LDA. DLDA model is important in event clustering and topic feature word selection. We aggregate events which have a common topic in a cluster by topic feature clustering. In clustering process a dynamic K-means clustering method is used to select suitable number of clusters automatically. We propose a topic terms generating rule to compose topic terms by topic feature words for every topic cluster. Compared to other approaches, our approach is more accurate in analyzing common topic from many different events and choosing suitable topic terms for topic clusters. In future work, we intend to optimize DLDA model and analyze the value of events which have a common topic.

Acknowledgment

This work is supported by the National Key Technologies R&Dprogram (No.2012BAH54F01), Shandong Province Independent Innovation Major Special Project (No.2013CXC30201), the Natural Science Foundation of China (No.61303005) and the Shandong Distinguished Middle-aged and Young Scientist Encouragement and Reward Foundation (No.BS2012DX015).

References

[1] Y. Hu, et al., "ET-LDA: Joint Topic Modeling for Aligning Events and their Twitter Feedback," in AAAI, 2012.
[2] L. Lei, et al., "LDA boost classification: boosting by topics," EURASIP Journal on Advances in Signal Processing, vol. 2012, pp. 1-14, 2012.
[3] H. Ma, et al., "A Novel Online Event Analysis Framework for Micro-blog Based on Incremental Topic Modeling," in Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), 2012 13th ACIS International Conference on, 2012, pp. 73-76.
[4] D. Ramage, et al., "Characterizing Microblogs with Topic Models," in ICWSM, 2010.
[5] L. Wang, et al., "Topic Discovery based on LDA_col Model and Topic Significance Re-ranking," Journal of Computers, vol. 6, pp. 1639-1647, 2011.
[6] X. Xu, "A New Sub-topics Clustering Method Based on Semi-supervised Learing," Journal of Computers, vol. 7, pp. 2471-2478, 2012.
[7] M. Xie, et al., "A New Intelligent Topic Extraction Model on Web," Journal of Computers, vol. 6, pp. 466-473, 2011.
[8] V. Krishnan, "Short comings of latent models in supervised settings," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, pp. 625-626.
[9] H. Xia, et al., "Plink-lda: Using link as prior information in topic modeling," in Database Systems for Advanced Applications, 2012, pp. 213-227.
[10] Y.-Q. Dong, et al., "A query interface matching approach based on extended evidence theory for deep web," Journal of Computer Science and Technology, vol. 25, pp. 537-547, 2010.
[11] A. Pons-Porrata, et al., "On-line event and topic detection by using the compact sets clustering algorithm," Journal of Intelligent and Fuzzy Systems, vol. 12, pp. 185-194, 2002.
[12] B. Huang, et al., "Microblog topic detection based on LDA model and single-pass clustering," in Rough Sets and Current Trends in Computing, 2012, pp. 166-171.
[13] S. Li, et al., "The key technology of topic detection based on K-means," in Future Information Technology and Management Engineering (FITME), 2010 International Conference on, 2010, pp. 387-390.
[14] M. Dutta, et al., "QROCK: A quick version of the ROCK algorithm for clustering of categorical data," Pattern Recognition Letters, vol. 26, pp. 2364-2373, 2005.
[15] Y. Jin and W.-L. Zuo, "A clustering algorithm using dynamic nearest neighbors selection model," Jisuanji Xuebao/Chinese Journal of Computers, vol. 30, pp. 756-762, 2007.

[16] Q. Li, et al., "Automatically Finding Significant Topical Terms from Documents," in AMCIS, 2005, p. 120.
[17] H. Wang, et al., "Supervised class-specific dictionary learning for sparse modeling in action recognition," Pattern Recognition, vol. 45, pp. 3902-3911, 2012.
[18] M. Georgescu, et al., "Extracting event-related information from article updates in wikipedia," in Advances in Information Retrieval, ed: Springer, 2013, pp. 254-266.
[19] C.-Y. Zhang, et al., "Extracting web entity activities based on SVM and extended conditional random fields," Ruanjian Xuebao/Journal of Software, vol. 23, pp. 2612-2627, 2012.
[20] S. Jing and L. Wanlong, "Topic Discovery Based on LDA Model with Fast Gibbs Sampling," in Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on, 2009, pp. 91-95.
[21] Y. Bin, et al., "Using Information Content to Evaluate Semantic Similarity on HowNet," in Computational Intelligence and Security (CIS), 2012 Eighth International Conference on, 2012, pp. 142-145.
[22] Z. Yihua, "A Comparison of Two Algorithms for Computer Recognition of Chinese Synonyms [J]," The Journal of The Library Science In China, vol. 4, 2002.
[23] M. Zhang, et al., "An Automatic Summarization Approach based on LDA Topic Feature," Jisuanji Yingyong yu Ruanjian, vol. 28, 2011.
[24] Y. Liu, et al., "Comparison of two schemes for automatic keyword extraction from MEDLINE for functional gene clustering," in Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE, 2004, pp. 394-404.

Yuanzi Xu, Born in 1983, currently a Ph.D candidate in Shandong University of computer science and technology. Her research interests are in the areas of Web information integration, data mining and event detection.

Qingzhong Li, Ph.D, currently a professor in Shandong University. His research interests are in the areas of Web information integration, data mining and large-scale network data management.

Zhongmin Yan, Ph.D, currently an associate professor in Shandong University. Her research interests are in the areas of Web information integration and data mining.

Wei Wang, currently a master candidate in Shandong University of computer science and technology. His research interests are in the areas of data mining and event detection.