© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Performance and Analysis Public Options Based on Using K-Means Algorithm

T. G. Babu*[1], R. Sharmila[2]

*[1]Assistant Professor, PG &Research Department of Computer Science and Science and Applications, Arignar Anna Govt Arts College Arcot Road, Cheyyar, Vellore, Tamil Nadu, India

[2]M.Phil (CS) Research Scholar PG &Research Department of Computer Science and Science and Applications, Arignar Anna Govt Arts College Arcot Road, Cheyyar, Vellore, Tamil Nadu, India

ABSTRACT

The hot spot in the general public are frequently the most ready to be found, shared and remarked by we-media like Facebook or Twitter. Mining hot spot from we-media can assist people to optimize their own investment behavior, assist enterprises to adjust their production and investment strategies to meet market demand, and help government to monitor public opinions and grab the chance to direct the healthy development of public opinions. In this paper, where the authors confronted a need to propose an exact calculation for news bunching that might assemble news into semantically close sets. A two phase way to deal that objective is proposed. First comparability estimation between news messages is performed using semantic similarity metric in view of WordNet. Next, the most reasonable for given data structure bunching calculations is chosen so as to acquire topical news groups and watch their size circulation after some time. Experiments were made on news volumes from several news mass media official pages in Facebook.

Keywords: We-media, Hot spot clustering, K-Means, psychological warfare, ontology, similarity analysis, news clustering, and news analysis.

I. INTRODUCTION

A. General Study

In the recent years of we-media era, there have been many events, such as the boom of mobile red envelopes, the popularity and dilemma of the special car market, and the investigation of China's smog by Chai Jing at her own expense. Behind every public hot spot, there are many we media professional express their opinions. Whether the government departments or the general public, excavating the current social hot spots can optimize their behavior in social life and generate positive economic and social value. Therefore, it is necessary doing hot spot mining to the collected classified information. Data mining results affect the accuracy of the users' obtained information about hot spot directly. At present, there have been many successful studies on hot spots mining. In the article "Breaking news detection and tracking in Twitter", Murata et al. proposed a method of effectively finding sudden hot spots on the Twitter platform through data collection, clustering, sorting and hot spot tracking. Yang Yonghong et al. used data mining techniques in network hot research, applying the methods of association analysis, clustering, classification and prediction in data mining to hot spot mining. Huang Min et al. Proposed a hot-spot mining analysis method using PageRank and Hits algorithm based on complex network.

B. Problem Definition

The rapid development of the Internet exerts a profound impact on the country, society, and individuals and how to eff ectively master mass data and extract the hotspot information there in have been a problem urgently to be solved in the management of internet public opinions. Solving this problem has an extensive application prospect: first, for individuals, it is an important means to promptly and conveniently obtain the hotspot information in current society; second, for enterprises, it can help enterprises master the most cutting-edge information and hot technology in their fields, increase their competitiveness for enterprises through this method; especially for the country, it can provide important clues for relevant departments of the governments to promptly know about the direction of public opinions in current society, be conductive to the governments to analyse and guide the public opinions, actively guide the healthy development of internet public opinions; meanwhile, help the governments to grasp the problems mostly cared by the people in each period as well as the viewpoints and attitudes on these problems, so as to make scientific and correct decision, keep the society stable, and truly reach the aim that the Internet serves for the society and the people. In the past, public opinions workers rely on manual work to sort the contents on the webpage to discover the hotspot information of the society, not only low efficiency in work, but also easy to be subjectively influenced and make the result deviate from the truth.

C. Motivation

At present, search engines, to some extent, meet people's demand on rapidly acquiring information needed among massive and messed information; however, its adoption of simple key words matching to find information causes a great deal of redundant and irrelevant contents in search results, results in redundant information overwhelming the information needed, leads to the incomplete analysis on topics of relevant personnel, and makes it difficult to have to a comprehensive mastery. Te premise for discovering hotspot information by search engines is that analysts know in advance the existence of such topics, so such method is obviously lagging and it is not good for discovering new problems, easy to miss the best timing to solve problems, making the problems spread and difficult to be controlled. Therefore, if the real-time hotspot information in a period is to be obtained and the internet hotspot topics in current society are to be periodically discovered, automatic solutions are becoming a valuable research orientation.

D. Contributions

The focus of this paper is hotspot clustering after the collected text is preprocessed and classified. The main purpose of text clustering is to compare the similarity of a group of textual information and classify the similar textual information into one category. There is no prior agreement on the type of text, so the number of categories is not sure. In the clustering process, the text of the cluster is generated automatically. The commonly used text clustering algorithms include K-Means, K-Medoids, Clarans Hierarchical-based and Density-based algorithms. Among them, the K-Means algorithm has low complexity and high efficiency, making it widely used. In addition, selecting the proper initial centroid is the key step of the basic clustering algorithm. One common technique for choosing an initial centroid is to make multiple runs, using a different set of random initial centroids each time, then select the cluster with the smallest SSE (square sum of errors). This strategy is simple, but may not work well; it depends on the number of data sets and clusters you are looking for. Canopy algorithm can select initial centroid of cluster effectively. Unlike traditional clustering algorithms (such as k-means), the biggest feature of Canopy clustering is that it does not need to specify the K value (the number of clustering) in advance, so it has a great practical value [4].Compared with other clustering algorithms, Canopy has a great advantage though its precision is low. Combining the advantages of K-means and Canopy, this paper improves the basic K-Means algorithm according to the characteristics of hot spot discovery. Firstly, Canopy clustering is used to select the centroid of the cluster, and the data was "coarse" clustering, then K-means is applied for further "fine" clustering after getting k value. Experimental results show that the proposed algorithm makes the purity and F values of the clustering results improved.

II. LITERATURE SURVEY

A. Web-Scale Computer Vision Using Mapreduce For Multimedia Data Mining

In that paper **Brandyn White** says that overview of MapReduce and common design patterns are provided for those with limited MapReduce background. Discussed both the high level theory and the low level implementation for several computer vision algorithms: classifier training, sliding windows, clustering, bag of features, background subtraction and image registration. Experimental results for the k-means clustering and single Gaussian background subtraction algorithms are performed on a 410 node Hadoop cluster.

Drawbacks

1) Developing and maintaining a computer cluster is a costly undertaking with a multitude of considerations: power, cooling, support, physical space, hardware, and software.
2) The "utility computing" model replaces these complexities with a fixed cost for the resources used, reducing the barrier to entry for researchers and small companies.
3) Moreover, vendor support is available for virtualized MapReduce clusters, further increasing their accessibility.
4) The most effective algorithms for natural language disambiguation for large datasets need not be the most effective for small datasets.

B. Identifying opinion sentences and opinion holders in internet public opinion

Zhang yu-feng introduces a methodology for analyzing judgment opinions, and defines a judgment opinion as consisting of a valence, a holder, and a topic. Also decompose the task of opinion analysis into four parts: a) recognizing the opinion; b) identifying the valence; c) identifying the holder; and d) identifying the topic. This paper addresses the first three parts and evaluates our methodology using both intrinsic and extrinsic measures.

Drawbacks

1) It is hard to apply opinion bearing words collected from one domain to an application for another domain.
2) One might therefore need to collect opinion clues within individual domains.
3) In case we cannot simply find training data from existing sources, such as news article analysis, we need to manually annotate data first.
4) In the NIST pilot study, it was apparent that human annotators often disagreed on whether a belief statement was or was not an opinion.

C. Detection And Tracking Technology Research For Chinese Microblogging Hot Topic

Sun shengping explains that with the widespread application of microblog, it has become a platform for people to publish their interest topics, to express personal feelings, to participate in a centralized discussion. Aiming at the tracking problem of hot events in microblogs, this paper proposes a new tracking method of hot events in microblogs based on the analysis of the features of mood information that is the total amount of emotion value and its growth rate. The experiment is carried to verify the feasibility of the method by the real-time capture and analysis of microblogs correlating with the hot events in different periods. Experimental results show that the method based on analyzing the features of mood information can effectively identify hot topic in microblogs.

Drawbacks

1) The information in microblogs mostly emerged fragmentation, immediacy, mobility and other features, and most of the content for the microblog users to express their feelings, so the proportion of information of emotional words in microblogs is larger than that of the traditional text, so the calculation method of emotion value in microblog for detecting hot events is feasible.
2) Based on the analysis of emotional words and especially the features of mood words, this paper proposes a tracking method of hot events in microblogs based on the calculation and analysis of the weight value of mood words.

D. Construction of topic detection and tracking system

Liu xiaodong, proposes that Topic Detection and Tracking (TDT) is a DARPA-sponsored initiative to investigate the state of the art in finding and following new events in a stream of broadcast news stories. The TDT problem consists of three major tasks: (a) segmenting a stream of data, especially recognized speech, into distinct stories; (b) identifying those news stories that are the first to discuss a new event occurring in the news; and (c) given a small number of sample news stories about an event, finding all following stories in the stream. The TDT Pilot Study ran from September 1996 through October 1997. The primary participants were DARPA, Carnegie Mellon University, Dragon Systems, and the University of Massachusetts at Amherst. This report summarizes the findings of the pilot study. The TDT work continues in a new project involving larger training and test corpora, more active participants, and a more broadly defined notion of "topic" than was used in the pilot study. The following individuals participated in the research reported.

Drawbacks

1) Stories that discuss unexpected events will of course follow the event, whereas stories on expected events can both precede and follow the event.
2) Events might be unexpected, such as the eruption of a volcano, or expected, such as a political election.
3) The notion of an event differs from a broader category of events both in spatial/temporal localization and in specificity.

E. Breaking news detection and tracking in Twitter

S. Phuvipadawat & t. Murata said that Twitter has been used as one of the communication channels for spreading breaking news. We propose a method to collect, group, rank and track breaking news in Twitter. Since short length messages make similarity comparison difficult, we boost scores on proper nouns to improve the grouping results. Each group is ranked based on popularity and reliability factors. Current detection method is limited to facts part of messages. We developed an application called "Hotstream" based on the proposed method. Users can discover breaking news from the Twitter timeline. Each story is provided with the information of message originator, story development and activity chart. This provides a convenient way for people to follow breaking news and stay informed with real-time updates.

Drawbacks

1) In tweet based clustering method, first collect tweets using certain queries, then cluster them together and extracts topics from them.
2) But tweet based clustering method may lead to cluster fragmentation problems. In other words, pattern based clustering method,
3) First find out patterns, then cluster them together and find out topics from the pattern clusters.
4) Wrong correlation of keywords is an important disadvantage of pattern based clustering techniques.
5) So the second objective of the proposed system is to introduce a new pattern based topic

detection technique, which is free from wrong correlation of patterns.

III. RESEARCH DESIGN

A. Existing system

Behind every public hot spot, there are many we media professional express their opinions. Whether the government departments or the general public, excavating the current social hot spots can optimize their behavior in social life and generate positive economic and social value. Therefore, it is necessary doing hot spot mining to the collected classified information. Data mining results affect the accuracy of the users' obtained information about hot spot directly. The implementation of this paper is based on Hadoop and R language. First, with Nutch's crawler function, the multithreading crawls data from we-media platform and completes data de-noising. Different strategy is used for different platform: direct crawl strategy for Baidu we-media platform and 360 we-media platform, second crawling strategy of Sougou for WeChat public platform, direct crawl or RSS strategy for other blog lists.

Disadvantages of Existing System:

- ✓ After the data de-noising phase, Html Parser is used to de-noisy data.
- ✓ The interface provides functions such as stopping words, participles, word frequency statistics, and producing word frequency vectors.
- ✓ As commercial ICTCLAS require regular payment to update the license, so the interface used in the paper is a free version of the web API.
- ✓ After the text is preprocessed, the processed text representation information is stored in the database for the next feature extraction

B. Proposed System

We use vector space model (VSM) which is the commonly used document representation, and apply TF-IDF formula to realize the feature weighting of text vector space. Firstly, all the words that are less than three after preprocessing are filtered out by using the DF feature selection algorithm, and then the irrelevant features in the text space are removed by using chi-square feature word selection algorithm. In this way, the original eigenvector is reduced in dimension; the selection of t values can represent the characteristics of the text, the candidate feature field is obtained. Next, the features are weighted by TF-IDF, which is essentially the product of TF and IDF. If a term appears frequently in document of a class, it can represent the characteristics of the text of this class. The term can be assigned higher weights with TFIDF, as the characteristic words of the text to distinguish it from other texts

Advantages of Proposed system:

- ✓ The average similarity of each cluster is calculated compared to the basic K-means algorithm.
- ✓ The time complexity of this process is only O(NKT),
- ✓ The total time complexity of the algorithm is still O(NKT),
- ✓ And there is little impact on efficiency.

C. Limitations Of K-Means Clustering Algorithm

Cluster analysis is an important technology in the data mining technology. Clustering is gathering some things together into one class according to certain attributes. So that similarity between classes becomes as small as possible, similarity within one class as large as possible. Clustering is an unsupervised learning process. Differences between it and classification are: it is necessary to know in advance what is the basis of the data characteristics for classification, and clustering is to find out the data characteristics. Therefore, in many applications, the cluster analysis as a data preprocessing procedure is the basis for further analysis and data processing. For example, in business, cluster analysis can help market analysts find a different customer base from the client library. And buying patterns depict different features of the customer base.

Clustering analysis methods commonly used are: method based on classification, method based on hierarchical, method based on density, method based on grid and so on. The *K*-means clustering algorithm (*K*-means clustering) is one kind of classical clustering algorithm which is proposed by Mac Queen. Realization of this algorithm is simple; the complexity is low. It obtains extremely widespread use, and becomes improvement object or the foundation for many other algorithms. The *K*-means algorithm main steps are as follows:

```
Input: Data set D; the K-means clustering counts K
Output: clustering results C*
Step1： Selecting k points as initial central point
Step 2： Repeat
Assigning each point to the nearest the center,
forming a k-th cluster. Recalculate the center of each cluster.
Until Center point does not change
Step3： Returning clustering results C*
```

K-means algorithm as a classical clustering algorithm although has relatively scalable and efficiency advantages. However, because of the limitations of the algorithm itself, there are still defects as follows:

1) Requiring The User In Advance To Give The Desired Cluster Number K:

Clustering algorithm based on *K*-means clustering algorithm is the most classic and the most commonly used algorithm. The clustering method requires users in the process of cluster analysis to input expected clustering number *k*. For example, when clustering computer audit methods of the basic endowment insurance, users who has many years society guarantee work experience will set the expected clustering number as 5 according to his society guarantees experience and knowledge, anticipated bunch of number *k* supposes will be 5 (i.e. basic old-age insurance synthesis auditing methods, basic old-age insurance collection- pay auditing methods, basic old-age insurance management auditing methods, basic old-age insurance payment auditing methods and basic old-age insurance finance auditing methods), but those who do not have any society guarantees experience and knowledge will stochastically set the expected clustering number *k*. Because the clustering results are extremely sensitive to regarding the input parameter, different input can obtain entirely different clustering results. Consequently, *K*-means algorithm requires the user in advance to give expected clustering number *k*. This deficiency not only increases the users 'burden, but also made it difficult to control the clustering results 'quality.

2) Random Selection Of Initial Cluster Centers:

Selecting the appropriate initial cluster center is a key step in the traditional *K*-means algorithm process. However, it often leads to the final local optimization results that the traditional *K*-means algorithm always randomly selects initial cluster centers. It can be explained by data set in literature [2]. The data set consists of two clusters: that is four types of data composition. Within each cluster is closer, while distance between clusters is larger. If setting two initial centers for each cluster, final cluster centers will not change, even if the two initial centers are assigned to one cluster, with the iteration of the algorithm, the cluster center will re-distribution. If a cluster is only assigned one initial centers, another cluster three, after several iterations, two clusters originally belong to one cluster are divided, and two clusters originally did not belong to one cluster are merged. Thus, in the *K*-means clustering process, different initialized clustering center can produce different results. Furthermore, it can discover: So long as two initial central points of one cluster fall on the internal cluster, no matter which position falls on, the optimal cluster can be obtained. That is because the multiple iteration of algorithm will redistribute the cluster center, and finally each cluster will has one cluster center. However, with clustering objects increase, possibilities that central point of cluster become more or less than two also gradually increase. Because distance between clusters is larger, center point cannot the redistributed. Therefore, only local optimum can be obtained.

3) Researches in Clustering Technology Based On Association Rules:

Along with the development of association rules and clustering- two mining technologies, research in clustering technology based on association rules has also become more and more. Firstly, researchers had many improvements in the similarity computing methods mainly through the association rules technology. Literature [3] has given a new association rule method. It measures the distance between the rules by commodity information classification information. The entire process scanned primitive data sets only once, thus it saves time. Literature proposes one similarity computing algorithm based on the words "relational degree". This algorithm obtains good clustering results. In addition, the frequent item set is the foundation of association rules, so clustering technology based on the frequent item set had many improvements. Literature [5] has improved text clustering method based on frequent item-set in WEB documents through the cross link chart instead of traditional calculating methods obtaining the frequent item-set. To solve the two limitations *K*-means algorithm has, Longhao, Fengjianlin, ET propose R-means algorithm

IV. ALGORITHMS

A. News Similarity Estimation Algorithm

a) Sentence tokenization and stop-words removal. At this stage we represent each text message as token vector $\bar{v} - (v_1 \dots v_n)$ consisting of words. We remove stop-words to avoid additional infelicities.

b) Part-of-speech disambiguation. Each word is tagged by two tags: the first one indicated syntactic role of the word (object, subject e.t.c.) and the second one point at functional role (verb, noun e.t.c.). We estimate similarity between nouns aiming to reveal news similar by discussed theme.

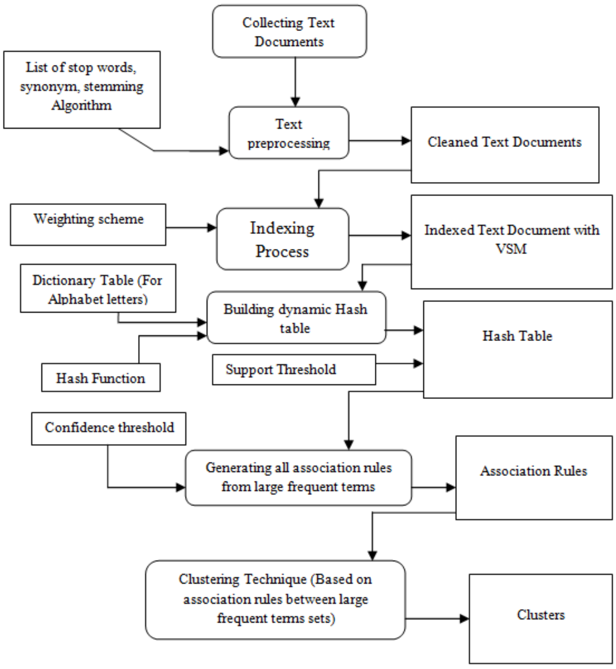

Collecting Text Documents
List of stop words, synonym, stemming Algorithm
Text preprocessing
Cleaned Text Documents
Weighting scheme
Indexing Process
Indexed Text Document with VSM
Dictionary Table (For Alphabet letters)
Building dynamic Hash table
Hash Table
Hash Function
Support Threshold
Confidence threshold
Generating all association rules from large frequent terms
Association Rules
Clustering Technique (Based on association rules between large frequent terms sets)
Clusters

Figure 1. Architecture Diagram

3) Word stemming what means removing the common morphological and inflexional endings of words. This operation is especially useful in the field of information retrieval and increases accuracy.

4) Word Sense Disambiguation. In this stage we investigate which of word senses is more appreciating in current context. Lesk algorithm could be used for that task. Word disambiguation is based on comparing of glossaries containing each word sense. The most probable sense is that one which is concluded in same glossary with the majority of other words in sentence. In adapted version of Lesk algorithm they achieved more accuracy.

5) Compute sentences relatedness. This estimation is based on pair of words similarity according JCN metric. First similarity matrix has to be constructed. The matrix element $R_{i,j}$ is similarity estimation value between could token corresponding to first sentence and token w_j corresponding to second sentence. Similarity matrix could be examined as bipartite graph and sentence similarity computing task as

computing a maximum total matching weight of this bipartite graph. Thus resulting similarity could be computed as average value

$$S = \frac{2 \times Match(N,M)}{|N|+|M|}$$

Where N, M is token vectors and Match (N, M) is token matching computed by Hungarian method. This estimation takes into account the influence of each pair similarity value. After we define sentences similarity computation method, we have to estimate similarity between all collected data and fetch out clusters of related messages.

B. Hot spot Discovery Algorithm

1) Means Algorithm Principle:

Steps for -means clustering algorithm are as follows

- ✓ Select objects as the initial cluster seeds on principle;
- ✓ Reassign each object to the most similar cluster in terms of the value of the cluster seeds;
- ✓ Update the cluster seeds; that is, recompute the mean value of the object in each cluster, and take the mean value points of the objects as new cluster seeds.
- ✓ Repeat (b) and (c) until no change in each cluster.

C. Improved K-Means Algorithm based on association rules technology

To solve the two limitations *K*-means algorithm has, we propose improved *K*-means algorithm based on minimum cover set.

Algorithm 1: Seeking the minimum rules set

Input: Frequent closed item sets FCI; Minimum confidence minconf;

Output: minimum rules set MRS

```
For each item K in FCI
{
Finding all subset s in K;
For each subset s in K{
If (s->K ¢ MRS){
Confidence=support(IK)/SUPPORT(s);
If(Confidence>=minconf)
Add s->K to MRS;
}
}
}
```

V. IMPLEMENTATION MODULES & SCREENSHOTS

A. Modules

1) New Uploads

Website Administrator can process uploaded with the content/text or text and images (News). The information's that were uploaded successfully are moved to a given directory. The class may reject data's that exceed a given size limit. The descriptions are picked from the value of a form text field that is submitted with the file field data. The keyword description and size information are stored in a database with separate name that can be used to retrieve the necessary information to generate pages on which the uploaded data's are displayed.

2) News Acquisition And Preprocessing

Verification data acquisition and preprocessing in this paper mainly include the following steps.

a) Public opinions data acquisition adopts web search technology, traversing the entire Web space within designated scope to collect all kinds of public opinions information, establishing indexes of acquired information through indexer, and save in the index database. Objects of data acquisition are mainly each major web portals, BBS, blogs, and so forth.

b) Word segmentation processing of website text: public opinions information acquired are unstructured data, which shall be preprocessed. Word segmentation study of Chinese language has been mature. This thesis adopts the Chinese Lexical Analysis System of Institute of Computing Technology (ICTCLAS).

c) Text features abstraction: the aim of selecting features is to further filter works with no much amount of information and less influence on the discovery of public opinions hotspots, reaching the effect of dimension reduction of website feature vector, so as to improve the processing efficiency and reduce the complexity of calculation. Form of dimension reduction adopted in this thesis to build evaluation function of webpage theme through statistical methods, evaluating each feature vector and choosing words meeting the preset threshold as the feature item of webpage;

d) Feature representation: this paper adopts vector space model (VSM) to indicate public opinions information; here omit the specific forms

3) News Similarity estimation

We purpose to amplify clustering accuracy by estimating similarity between news data based on ontology. Using ontology could give a better understanding of information spreading and impact. We aim to obtain some news clusters were each cluster contains information concerning one theme or even one point of view regarding this theme.

We use WordNet – lexical database of English. Words in WordNet are united in synsets (sets of cognitive synonyms) which are interlinked by means of conceptual-semantic and lexical relations. This structure could be convenient to estimate words and sentences similarity. There're many measures of semantic similarity based on WordNet. Some measures rely on WordNet structure to produce a numeric score that quantifies the degree to which two concepts are similar. It measures which use information content values along with ontology structure are more accurate and provide greater correlation with human similarity judgments. That is why we use

$$jcn(c_1, c_2) = \frac{1}{IC(c_1)+IC(c_2)-2*res(c_1,c_2)}$$

Where res(c_1,c_2) is Resnik measure of similarity and IC (c) is information content value of concept.

Since we analyze news messages from social networks for the moment, so news usually are presented by one or few sentences. Our first step is to understand which messages are related to same theme. Thus sentences similarity estimation method should be proposed. We provide several steps for that method

(a) Sentence tokenization and stop-words removal. At this stage we represent each text message as token vector $\overline{v} = (v_1, .. v_n)$ consisting of words. We remove stop-words to avoid additional infelicities.

(b) Part-of-speech disambiguation. Each word is tagged by two tags: the first one indicated syntactic role of the word (object, subject e.t.c.) and the second one point at functional role (verb, noun e.t.c.). We estimate similarity between nouns aiming to reveal news similar by discussed theme.

(c) Word stemming what means removing the common morphological and inflexional endings of words. This operation is especially useful in the field of information retrieval and increases accuracy.

(d) Word Sense Disambiguation. In this stage we investigate which of word senses is more appreciating in current context. Lesk algorithm could be used for that task. Word disambiguation is based on comparing of glossaries containing each word sense. The most probable sense is that one which is concluded in same glossary with the majority of other words in sentence. it suggested adapted version of Lesk algorithm where they achieved more accuracy.

(e) Compute sentences relatedness. This estimation is based on pair of words similarity according JCN metric. First similarity matrix has to be constructed. The matrix element $R_{i,j}$ is similarity estimation value between token v_1 corresponding to first sentence and token w_j corresponding to second sentence. Similarity matrix could be examined as bipartite graph and sentence similarity computing task as computing a maximum total matching weight of this bipartite graph. Thus resulting similarity could be computed as average value

$$s = \frac{2X\ Match(N, M)}{|N| + |M|}$$

where N, M are token vectors and Match(N,M) is token matching computed by Hungarian method. This estimation takes into account the influence of each pair similarity value. After we define sentences similarity computation method, we have to estimate similarity between all collected data and fetch out clusters of related messages.

4) Text clustering

Mass media news data from social networks has several features:

- ✓ News is presents as short text in 18 words in average;
- ✓ Text corpora could contain hundreds of thousands of news and more. News set is always replenishing;
- ✓ The number of clusters in unknowns and it could vary in different moments of time;
- ✓ Preprocessing stage of algorithm has a similarity matrix in output which is also represents a bipartite graph.
- ✓ These features make a class of spectral clustering algorithms more suited for application. Since spectral divide & merge clustering algorithm shows high accuracy when similarity matrix is known and doesn't need the number of clusters for input to reveal clusters for news data.

B. Result

1) Screenshots

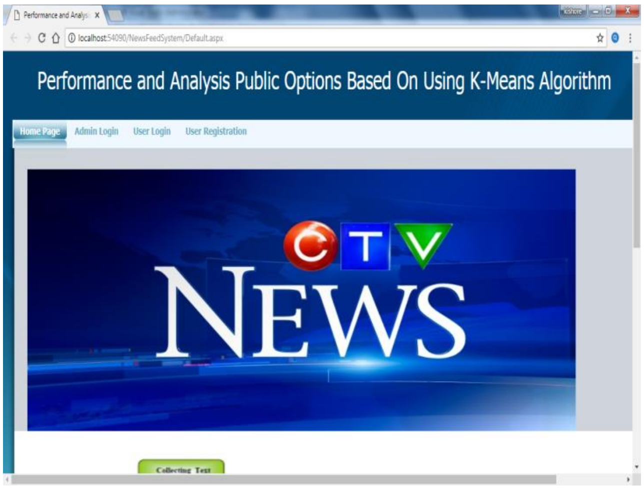
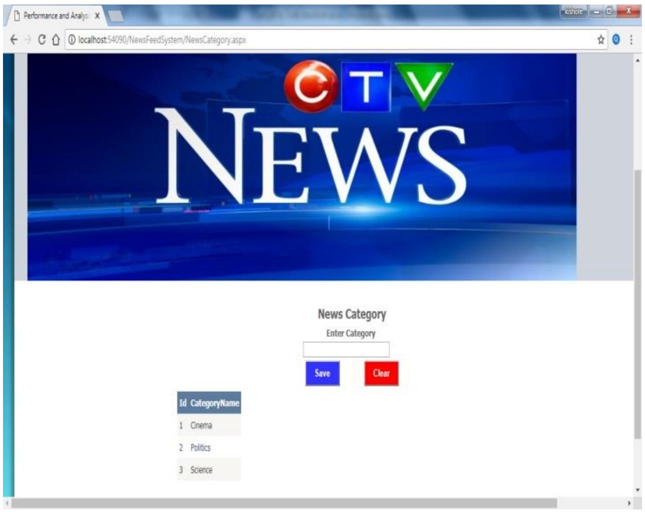
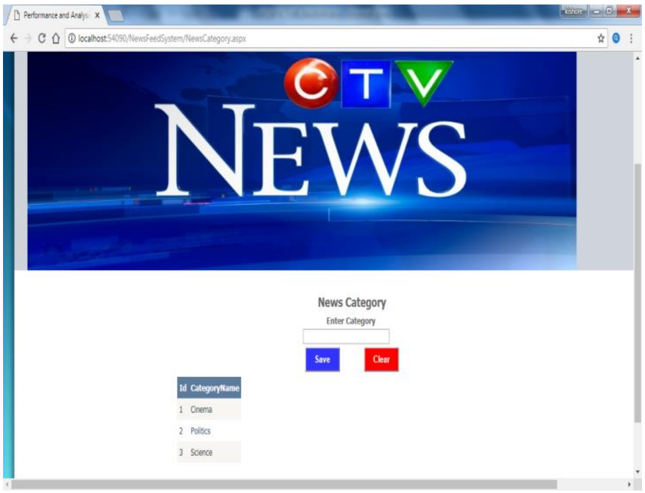

Performance and Analys
localhost:54090/NewsFeedSystem/Default.aspx
Performance and Analysis Public Options Based On Using K-Means Algorithm
Home Page
Admin Login
User Login
User Registration
CTV
NEWS
Collecting Text

Figure 2. Home Page



Id	CategoryName
1	Cinema
2	Politics
3	Science

Figure 3. Admin Login



Id	CategoryName
1	Cinema
2	Politics
3	Science

Figure 4. News Category

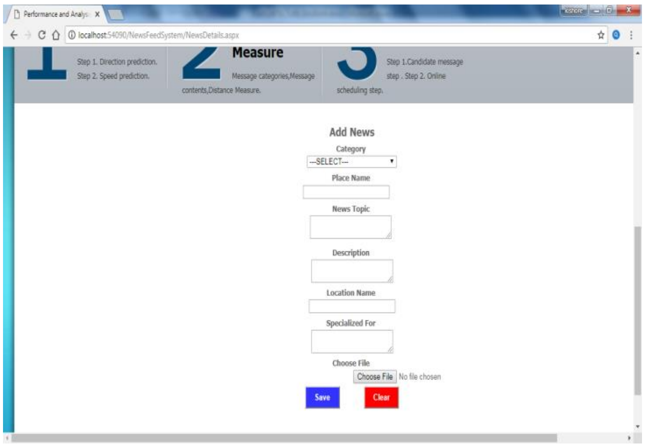
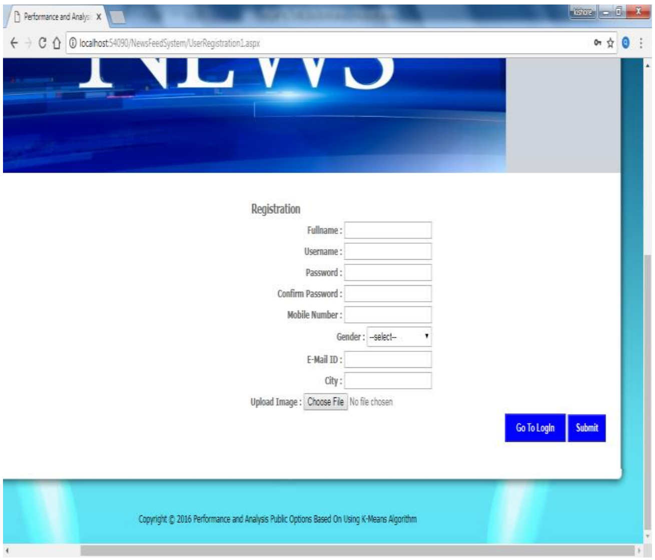

Performance and Analys
localhost:54090/NewsFeedSystem/NewsDetails.aspx
1
Step 1. Direction prediction.
Step 2. Speed prediction.
2
Measure
Message categories,Message contents,Distance Measure.
3
Step 1.Candidate message step . Step 2. Online scheduling step.
Add News
Category
---SELECT---
Place Name
News Topic
Description
Location Name
Specialized For
Choose File
Choose File
No file chosen
Save
Clear

Figure 5. Upload News



localhost:54090/NewsFeedSystem/UserRegistration1.aspx

Registration

Fullname :

Username :

Password :

Confirm Password :

Mobile Number :

Gender : --select--

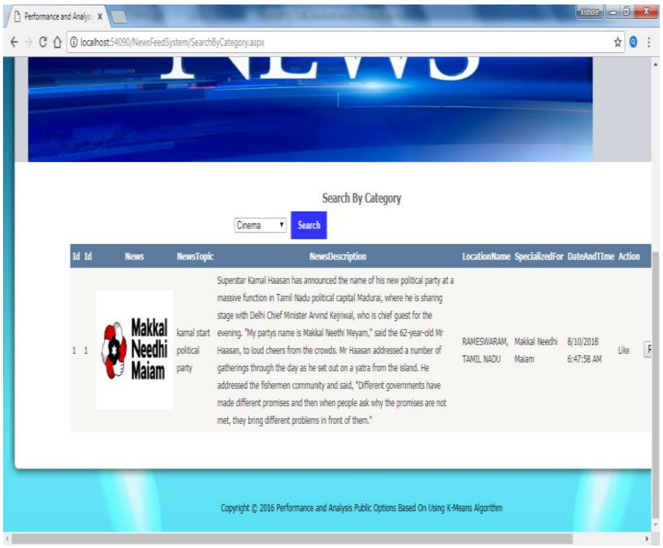
E-Mail ID :

City :

Upload Image : Choose File No file chosen

Copyright © 2016 Performance and Analysis Public Options Based On Using K-Means Algorithm

Figure 6. User registration



localhost:54090/NewsFeedSystem/SearchByCategory.aspx

Search By Category

Cinema

Id	Id	News	NewsTopic	NewsDescription	LocationName	SpecializedFor	DateAndTIme	Action
1	1		kamal start political party	Superstar Kamal Haasan has announced the name of his new political party at a massive function in Tamil Nadu political capital Madurai, where he is sharing stage with Delhi Chief Minister Arvind Kejriwal, who is chief guest for the evening. "My partys name is Makkal Neethi Meyam," said the 62-year-old Mr Haasan, to loud cheers from the crowds. Mr Haasan addressed a number of gatherings through the day as he set out on a yatra from the island. He addressed the fishermen community and said, "Different governments have made different promises and then when people ask why the promises are not met, they bring different problems in front of them."	RAMESWARAM, TAMIL NADU	Makkal Needhi Maiam	8/10/2018 6:47:58 AM	Like

Copyright © 2016 Performance and Analysis Public Options Based On Using K-Means Algorithm

Figure 7. Category Centric Search Result

VI. CONCLUSION

These days, internet is turning into the main channel for individuals to get and release data, the managing part of internet public opinions information is larger and larger. It has excited wide consideration in the business how to carry out public opinions gathering and hotspots discovery on the basis of information acquisition of Internet public opinions as well as track and analyze the hotspots to guarantee the information security. Under such background, this paper, based on analyzing the advantages and disadvantages of all kinds of clustering algorithms, chooses K means clustering as the website text clustering model and puts forward a new discovery algorithm of internet public opinions hotspots through improving its shortcoming of sensitivity to initial number of clusters and initial clustering centers. The test shows the appropriateness and unwavering quality of technique in this paper.

VII. REFERENCES

[1]. Brandyn White, "Web-Scale Computer Vision using Map Reduce for Multimedia Data Mining", MDMKDD '10 Proceedings of the Tenth International Workshop on Multimedia Data Mining. ACM New York, pp.287-292, 2010.

[2]. Zhang Yu-feng, "Identifying Opinion Sentences and Opinion Holders in Internet Public Opinion", Industrial Control and Electronics Engineering (ICICEE), pp.1668- 1671, 2012.

[3]. Sun Shengping, "Detection and tracking technology research for Chinese microblogging hot topic", Beijing Jiaotong University School of Economics and Management, pp.18-48, 2011.

[4]. Liu Xiaodong, "Construction of topic detection and tracking system", Computer Science, Beijing University of Posts and Telecommunications, pp.9-50, 2011.

[5]. S. Phuvipadawat, T. Murata, "Breaking news detection and tracking in Twitter", 2010IEEE/WIC/ACM International Conference on Web Intelligent Agent Technology, pp.120-123, 2010

[6]. H. Liu and J. H. Xu, "Research of internet public opinion hotspot detection," Bulletin of Science and Technology, vol. 27, no. 3, pp. 421–425, 2011.

[7]. G. Hamerly and C. Elkan, "A new algorithm based on K-means and its application in internet public opinion hotspot detection," Pattern Recognition, vol. 32, no. 6, pp. 521–534, 2012.

[8]. L. M. Kristina, "Document clustering in reduced dimension vector space," Journal of Computer Application, vol. 27, no. 10, pp. 37–49, 2011.

[9]. H. J. Andreas, "Research on text document clustering," Com-puter Simulation, vol. 24, no. 7, pp. 84–99, 2010.

[10]. C. D. Wagstaf f and S. S. Rogers, "Constrained K-means clustering with background knowledge," Journal of Computer Engineering and Application, vol. 21, no. 5, pp. 467–479, 2011.

[11]. B. T. Ya, "Research on public opinion hotspot detection based on SVM," Science and Technology Management Research, vol. 25, no. 2, pp. 64–69, 2009.

[12]. P. S. Bradley and L. S. Managasarian, "K-plane clustering," Jour-nal of Global Optimization, vol. 16, pp. 23–32, 2010.

[13]. Y. Tang and Q. S. Rong, "An implementation of clustering alg-orithm based on K-means," Journal of Hubei Institute For Nationalities, vol. 22, no. 1, pp. 69–71, 2011.

[14]. Z. H. Yang and Y. T. Yang, "Document clustering method based on hybrid of SOM and K-means," Computer Application, vol. 27, no. 5, pp. 73–75, 2012.

[15]. Y. F. Zhang and J. L. Mao, "An improved K-means algorithm," Computer Application, vol. 23, no. 8, pp. 31–33, 2009.

[16]. N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," Decision Support Systems, vol. 48, no. 2, pp. 354–368, 2010.

[17]. T. Pedersen "Information Content Measures of Semantic Similarity Perform Better Without Sense-Tagged Text". Proceeding HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, p. 329-332.

[18]. Bach, F., Jordan, M.: Learning spectral clustering. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems 16 (NIPS), pp. 305–312. MIT Press, Cambridge, 2004.