Auto-detection of Hot Topics in Mass Chinese Internet Information

Chun-hui DENG¹, Yun LUO² and Hui-fang DENG²,*

¹School of Computer Engineering, Guangzhou College of South China University of Technology, Guangzhou, China, 510800

²School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, 510006

*Corresponding author

Keywords: Hot topic auto-detection, Text preprocessing, Text clustering, Web structure-based text feature weight calculation, Concept feature extraction, Multi-resolution density clustering, Sub-cluster relation-based clustering, Feature-based news text vector representation, Feature space, Advanced Chinese word segmentation ICTCLAS system.

Abstract. In order to overcome the weakness of the traditional topic detection clustering strategy and realize hot topic auto-discovery, we re-examined density-based clustering algorithms, and then put forward a sub-cluster relation-based and multi-resolution density clustering algorithm (SRB-MRClustering) which considers both adjacency information of sub-clusters and relative density concept. And in the meanwhile, in order to reduce the computational complexity, we proposed a Web structure-based text feature weight calculation method and a concept-feature extraction method and used feature-based news text vector representation method to improve the textual representation and shrink the dimension of feature space. Finally, we used Chinese news corpus of June-July 2012 to verify our algorithm. The experimental results show that the algorithm’s performance and clustering quality are improved to a notable extent.

Introduction

Relative to the TDT (Topic Detection Task), auto-detection of hot topics and the way of presenting topics’ core information are the research of greater applicability. The prevalent problem for current hot topic auto-detection system is the high incorrect classification rate in detecting hot events or topics. Most news portals still have to rely on manual work to classify and release some special news topics or are only able to provide hot topics within a certain period of time (say, a month or a year), lacking in timeliness. So how to automatically discover, identify, and extract hot topics (keywords and summary info) from mass internet information is of a great challenge.

Topic detection process is a kind of incremental clustering; the commonly used clustering strategies are Agglomerate Hierarchical Clustering, and Incremental K-Means Clustering, etc. [1]. And the commonly adopted clustering algorithms include partition clustering algorithm, hierarchical clustering algorithm, density-based clustering algorithm, grid-based clustering algorithm and model-based clustering algorithm etc.. Text clustering process combines a series of natural language processing (NLP) techniques, such as text segmentation, text preprocessing, text features extraction, text feature vectors representation and text similarity calculation. Text clustering process is as shown in Figure 1:
Choosing which kind of text clustering algorithm depends on the available data types and the specific objectives of applications. Different clustering algorithm exhibits a great difference in time complexity, clustering result, and the application area.

In this paper, the research of clustering algorithms is mainly concentrated on the improvements of traditional clustering algorithms. Through deeply analyzing the strengths and weaknesses of the density-based clustering algorithm, we then proposed a sub-cluster relationship-based and multi-resolution density clustering algorithm (SRB-MRClustering), which can improve the efficiency and effectiveness of hot topics automatic discovery to a certain degree.

Internet hot topics detection mainly includes offline topics detection of a set of static news corpus and online topics detection of real-time news corpus collection. But these two methods are both lack of prior knowledge (new topics for reference). Therefore, unsupervised learning clustering algorithms are required to realize hot topics detection in massive Internet information. The process framework is as shown in Figure 2 and the process is described below:

1. Web page information extraction and pre-processing. Transforming the free text into structured or semi-structured information, then a series of text preprocessing operations (such as text segmentation, text features extraction and text feature vectors representation etc.) are applied to the results of Web text information extraction;

2. Hot topics auto-detection. Clustering webpage text feature vector set using SRB-MRClustering algorithm and ending up with webpage text cluster of related topics;


**Density-based Clustering Algorithm and Its Improvements**

The principle of density-based clustering algorithm [2] is shown in Figure 3.
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a typical spatial clustering algorithm based on density[2]. Let us examine it briefly below:

(1) Description of DBSCAN

DBSCAN can handle spatial data and spot any shape clusters in a noised spatial database by dividing the database into clusters with density of high enough.

Given parameters of Eps and MinPts, searching for all \( P \)-density-reachable objects based on breadth-first search from the spatial database to get a cluster. The process of DBSCAN is as shown in the Figure 4:

![Flow chart of DBSCAN](image)

Before clustering, R*-Tree is built for the objects in spatial database to conduct region searching to extract neighborhood objects. Therefore, the algorithm has \( O(n \log n) \) time complexity on average, where \( n \) is the number of objects.

(2) The shortage and current improvements of DBSCAN

DBSCAN is capable of discovering clusters with arbitrary shape and shielding the influence of noise. But the algorithm also has some drawbacks:

1) Clustering results are sensitive to parameters. When the defined density threshold is too large, a cluster maybe divided into a few sub-clusters; or if the density threshold value is too small, several clusters will be merged. One of the methods to estimate the values of Eps and MinPts in DBSCAN algorithm is to draw k-dist graph, by observing the varying trend of the curve to determine the parameters.
2) DBSCAN used fixed global density parameters to identify different clusters, and therefore is lack of flexibility when handling spatial database with uneven density distribution. Few density-based clustering algorithms can automatically process the data with non-uniform density. When a unified criterion is adopted, the natural structure of clusters is likely to be damaged. For example, when relative density (the neighborhood radius is smaller) is larger, non-uniform density distribution space cannot be distinguished; when relative density (neighborhood radius is larger) is smaller, part of the cluster information will be lost.

3) The conflict between maximality and connectedness in the DBSCAN’s definition of density-connectivity and density-reachability may lead to a density-connected cluster be divided into two.

The improvements made on DBSCAN are about how to reduce the sensitivity of the algorithm to the input parameters or to seek more effective methods to determine parameters. For example, Su Zhong et al. proposed a RDBC Algorithm (Recursive Density-based Clustering Algorithm)[3], and the parameters can be modified dynamically and intelligently in this algorithm; Shui-Geng Zhou et al. constructed a data partitioning-based DBSCAN algorithm[4].

Based on current research works, we continue to improve DBSCAN with a view to overcoming its drawbacks to the maximum extent and propose SRB-MRClustering algorithm.

Hot Topic Auto-Detection Based on Srb-Mrclustering Algorithm

The Description of Text Representation Procedure

(1) Using ICTCLAS (Institute Of Computing Technology Chinese Lexical Analysis System) to realize Chinese word segmentation [5, 6];

(2) Text concept feature extraction algorithm. It applies the concept and association expansion to text feature vectors in order to reduce the relevant degree of terms and enhance the ability to represent the text theme. This new method eliminates the drawbacks of traditional IDF(Inverse Document Frequency) method. It can efficiently conduct text feature extraction and thus improve the accuracy of automation text cluster.

(i) The definition of text concept features. In the paper, text concept features refer to a collection of words of identical or similar semantics [7]. The way to determine the semantic similarity of words is to check if the similarity threshold is greater than or equal to threshold value $\alpha$ [8]. In this paper, the value of $\alpha$ is 1.0.

(ii) Text concept feature weights $W$ calculation method: $W$ is calculated as follows:

$$ W = W(t_i, d) \times f_{i,d} \quad (1) $$

Where $W(t_i, d)$ is text feature weight and calculated as:

$$ W(t_i, d) = \frac{\sqrt{\sum TF(t_i, d) \times \log \left( \frac{N}{DF(t_i)} + 0.01 \right)^2}}{\sum \left( \frac{N}{DF(t_i)} + 0.01 \right)^2} \quad (2) $$

Where $TF(t_i, d)$ denotes the frequency at which a given word appears in document $d$:

$$ TF(t_i, d) = \frac{n_i}{\sum n_k} \quad (3) $$

$n_i$ denotes the counts of term $t_i$ appearing in document $d$; $\sum n_k$ is the sum of the counts of all the terms appearing in document $d$.

The Inverse Document Frequency $IDF(t_i, d)$ is used to measure the capacity of a given term $t_i$ distinguishing document $d$: 

584
\[ IDF(t_i, d) = \log \frac{N}{DF(t_i)} \]  
\( N = |d| \), represents the total number of documents in corpus; \( DF(t_i) \) represents the number of documents containing terms \( t_i \) in corpus.

Terms appearing in title and content in a news text will be given different weight coefficients \( f_{t, d} \). The coefficient lies in \([0,1]\). \( \alpha, \beta, \lambda, \) and \( \chi \) respectively represent the feature weight of content, time, location and subject of a news event, and their values are shown in Table 1.

<table>
<thead>
<tr>
<th>Title</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Content</th>
<th>Time</th>
<th>Location</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

(iii) Semantic similarity calculation algorithm. Continuing from the semantic similarity calculation algorithm in reference [9], in this paper, the term similarity means the ability of replacement of two terms in different context without changing their syntax structure.

\[
\text{Sim}(T_i, T_j) = \max_{i=1, j=1} \text{Sim}(C_i, C_j) 
\]

\( C_i \) is the concept feature set of term \( T_i \), \( C_j \) is the concept feature set of term \( T_j \). Similarity between \( T_i \) and \( T_j \) is equivalent to the maximum value of every two concept feature similarity in the concept feature set of them.

(3) Text representation method based on news feature. In reference [10], an event is divided into four parts: time, location, subject and event content. Each part is regarded as one-dimensional vector space. In this paper, we use time dimension, location dimension, subject dimension and content dimension to build text vector space, aiming to improving the traditional text vector space representation model.

(4) Text similarity calculation [11]. Text similarity is calculated using cosine factors as follows:

\[
\text{Sim}(d_1, d_2) = \cos \theta = \frac{\sum_{k=1}^{n} W_{i,k} W_{j,k}}{\sqrt{\sum_{k=1}^{n} W_{i,k}^2} \sqrt{\sum_{k=1}^{n} W_{j,k}^2}} 
\]

The similarity for time dimension, location dimension and subject dimension is calculated using formula below:

\[
\text{Sim}(d_1, d_2) = \begin{cases} 
1 & d_1 \cap d_2 = \emptyset \\
0 & \text{else}
\end{cases} 
\]

The similarity \( \text{Sim}(d_1, d_2) \) between two news events (documents) is the weighted sum of content dimension \( \text{Sim}(d_{c1}, d_{c2}) \), time dimension \( \text{Sim}(d_{t1}, d_{t2}) \), location dimension \( \mathcal{S} m(d_{l1}, d_{l2}) \) and subject dimension \( \mathcal{S} m(d_{s1}, d_{s2}) \):

\[
\text{Sim}(d_1, d_2) = \alpha \times \text{Sim}(d_{c1}, d_{c2}) + \beta \times \text{Sim}(d_{t1}, d_{t2}) + \lambda \times \mathcal{S} m(d_{l1}, d_{l2}) + \chi \times \mathcal{S} m(d_{s1}, d_{s2}) 
\]

Definitions in SRB-MRClustering Algorithm

1) Definition 1 (k-distance of object \( p \) [12]):
In the collection of objects $D$, for any natural number $k$, $k$-distance of object $p$ defined as $(k - \text{distance}(p))$ is the distance between $p$ and an object $o$; $o$ should satisfy certain conditions as follows:

1) At least there exist $k$ objects $o' \in D \setminus \{p\}$ that makes $d(p, o') \leq d(p, o)$.
2) At most there exist $k-1$ objects $o' \in D \setminus \{p\}$ that makes $d(p, o') < d(p, o)$.

2) Definition 2 (Neighborhood of $(k - \text{distance}(p))$ [12]):

\[ N_{k \text{-distance}}(p) = \{q \mid d(p, q) \leq k - \text{distance}(p)\} \] (9)

The neighborhood of $(k - \text{distance}(p))$ contains all the objects $q$ within the distance of $p$ less than or equal to $(k - \text{distance}(p))$.

3) Definition 3 (k-neighborhood average distance [13]):

\[ NAD_{k \text{-distance}}(p) = \frac{\sum_{o \in N_{k \text{-distance}}(p)} d(p, o)}{|N_{k \text{-distance}}(p)|} \] (10)

4) Definition 4 (k-neighborhood relative density):

\[ RD_{k \text{-distance}}(p) = \frac{\min_{q \in N_{k \text{-distance}}(p)} [NAD_{k \text{-distance}}(p), NAD_{k \text{-distance}}(q)]}{\max_{q \in N_{k \text{-distance}}(p)} [NAD_{k \text{-distance}}(p), NAD_{k \text{-distance}}(q)]} \] (11)

This ratio reflects the close degree of object $p$ and its neighborhood domain objects $q$, namely local density can be obtained.

5) Definition 5 (collection of core objects):

The core object $p$ is defined as: in the collection of objects $D$, given threshold $\eta$, $(1 - RD_{k \text{-distance}}(p)) < \eta$ should be satisfied.

The core object set of $p$ is denoted as $CO_{k \text{-distance}}(p)$ which consists of the core object $p$ and all the core objects in neighborhood of $(k - \text{distance}(p))$. If $p$ is not a core object, this definition is meaningless.

6) Definition 6 (shared sub-cluster boundary objects $s$):

In the object space, there are at least two core objects $p_1$ and $p_2$ belonging to different sub clusters, satisfying the following conditions:

1) $p_1$ and $p_2$ are both density-reachable $s$;
2) $s$ is not a core object, but a boundary object on the border of clusters (if $s$ is a core object, $p_1$ and $p_2$ are density-connected through $s$).

7) Definition 7 (sub-cluster connection relationship):

If sub-clusters $C_1$ and $C_2$ share at least a boundary object $s$, it is said that $C_1$ is connected to $C_2$ through $s$.

As shown in Figure 5, cluster 1 composed of solid points and cluster 2 composed of hollow points are connected by $s$ (MinPts = 4).
Description of SRB-MRClustering Algorithm

In the algorithm, we build up two data structures: cluster category data structure and object *Obj* data structure in the data set *D*.

The algorithm mainly consists of two parts: multi-resolution density clustering and sub-clusters fuzzy clustering. Figure 6 is the flow chart of SRB-MRClustering algorithm.

![Flow chart of SRB-MRClustering algorithm.](image)

Experimental Analysis of SRB-MRClustering Algorithm

The experimental analysis of SRB-MRClustering algorithm contains two parts: performance evaluation comparisons of different algorithms and clustering result analysis.

1. We conducted the performance evaluation comparisons experiment of SRB–MRClustering mainly with the traditional density-based clustering algorithm (DBSCAN) and Agglomerative Hierarchical Clustering (AHC).

![Performance evaluation comparisons of different algorithms.](image)

SRB-MRClustering algorithm is based on DBSCAN, but used the relative density concept and fuzzy clustering to obtain the final results by saving adjacency information of sub-clusters. Therefore, SRB-MRClustering algorithm, basically following the same processing steps as
DBSCAN, when R*-tree index is adopted, has the time complexity of $O(n \log n)$. Adjacency information of sub-clusters can be kept in memory (the number of sub clusters is far less than the number of objects in space), so SRB-MRClustering algorithm has no additional I/O overhead and its execution efficiency is similar to DBSACN, but is slightly lower than AHC (Agglomerative Hierarchical Clustering).

(2) The clustering effectiveness comparison analysis of DBSCAN and SRB-MRClustering algorithm.

Figure 8 is the distribution of test data set. Figure 9 and Figure 10 are the clustering results of DBSCAN and SRB-MRClustering under different parameters.

Through the above experimental results, we can see that DBSCAN is sensitive to the input parameters. When the neighborhood radius is smaller, uneven distribution area cannot be distinguished well; When the neighborhood radius is bigger, some clustering information get lost. This is to say when the relative density changes, the clustering effects changed greatly.
From Figure 10, we can see that SRB-MRClustering algorithm is insensitive to the input parameters, the clustering results obtained are relatively stable, and high density area can be identified from low-density area.

Experiments

In this paper, the news corpus are sourced from domestic and international news that happened during the period of June-July 2012 [14]. There is no authoritative data available for comparison with and evaluation of the clustering results. After conducting text clustering for the news corpus, the top ten hot topics are shown in Table 2. The results reflect hot topics during the period of the June-July 2012. They are partially overlapping with domestic and international top ten events voted by authoritative media (Chinese network media news top 10 of 2012 and Xinhua News Agency 2012 annual national and international news top 10), but represents the distribution of hot topics during the period of June-July 2012 to a greater extent.

<table>
<thead>
<tr>
<th>Topic No.</th>
<th>Domestic</th>
<th>International</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Manually controlled rendezvous and docking of Shenzhou-9 with Tiangong-1 has been realized successfully</td>
<td>The American history’s most cutting-edge probe named “Curiosity” landed on Mars in August.</td>
</tr>
<tr>
<td>2</td>
<td>The sea diving trial of manned submersible vehicle named Jiaolong has been crowned with complete success</td>
<td>The persistence fermentation of European debt crisis drags on the global economic recovery</td>
</tr>
<tr>
<td>3</td>
<td>The prefecture-level city of Sansha has been set up</td>
<td>The Egyptian elected President Mohamed Morsi was sworn in</td>
</tr>
<tr>
<td>4</td>
<td>The 12th Meeting of the Council of Shanghai Cooperation Organization (SCO) Heads of States</td>
<td>The 7th G20 summit of leaders opens in Mexico</td>
</tr>
<tr>
<td>5</td>
<td>The hero driver Wubin in Hangzhou</td>
<td>The Japanese government has officially started the acquisition of Diaoyu Islands</td>
</tr>
<tr>
<td>6</td>
<td>China’s economy has entered a period of contraction after continuously fast growth for three decades</td>
<td>The third round of China-EU High-level Strategic Dialogue</td>
</tr>
<tr>
<td>7</td>
<td>Implementation of State Council’s regulation and control measures of real estate market</td>
<td>New drug “bath salts” has started showing widespread trend in Japan</td>
</tr>
<tr>
<td>8</td>
<td>The central bank has cut interest rates again</td>
<td>The U. S. House of Representatives has formally approved a bill to apologize for the Chinese exclusion act in history</td>
</tr>
<tr>
<td>9</td>
<td>Ministry of Education and other relevant Departments strengthen the management of pupil’s nutritional meal standard</td>
<td>The international oil price plummeted</td>
</tr>
<tr>
<td>10</td>
<td>Heavy rain hit Beijing</td>
<td>The vote of “consumption tax increase bill” has gone through by Japan’s House of Representatives</td>
</tr>
</tbody>
</table>

Observing the distribution characteristics of the 2012 annual domestic and international top ten hot events released by the authority media, they are mainly concentrated in the fields of military, finance, science and technology, political events and social events. In the paper, the profile of the hot news topics extracted form news corpus of June-July 2012 exhibits the same fields distribution characteristics. So the experimental results are valid and reliable, as shown in Table 3.
Table 3. Comparison of fields distributions.

<table>
<thead>
<tr>
<th>F (Fields)</th>
<th>Military</th>
<th>Sports</th>
<th>Finance</th>
<th>Entertainment</th>
<th>Technology</th>
<th>Education</th>
<th>House property</th>
<th>Politics</th>
<th>Society</th>
<th>Automobile</th>
</tr>
</thead>
<tbody>
<tr>
<td>The field distribution of Chinese network media news ranking list of 2012</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>The field distribution of Xinhua News Agency 2012 annual national and international</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>The field distribution of hot topics during June-July 2012</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4 and Table 5 show the Precision Rate and Recall Rate of the clustering results of topic 2.

Table 4. Precision rate of topic 2.

<table>
<thead>
<tr>
<th>Assessment Criteria</th>
<th>Clustering Algorithm</th>
<th>Number of text files in topic2 after clustering</th>
<th>Text files really belonging to topic2 after conducting artificial screening for the clustering results</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>DBSCAN</td>
<td>100,296</td>
<td>80,251</td>
<td>80.01%</td>
</tr>
<tr>
<td></td>
<td>SRB-MRClustering</td>
<td>99,768</td>
<td>85,032</td>
<td>85.23%</td>
</tr>
</tbody>
</table>

Table 5. Recall rate of topic 2.

<table>
<thead>
<tr>
<th>Assessment Criteria</th>
<th>Clustering Algorithm</th>
<th>Text files really belonging to topic2 after conducting artificial screening for the clustering results</th>
<th>The actual number of news text files related to topic2 that are contained in experimental corpus</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>DBSCAN</td>
<td>80,251</td>
<td>98,532</td>
<td>81.45%</td>
</tr>
<tr>
<td></td>
<td>SRB-MRClustering</td>
<td>85,032</td>
<td></td>
<td>86.30%</td>
</tr>
</tbody>
</table>

Precision = Number of text files really belonging to category C after clustering / Total number of text files in category C after clustering

Recall = Number of text files belonging to category C both before and after clustering / Total number of text files originally in category C before clustering

Conclusions

In the paper, we constructed a Web structure-based text feature weight calculation method and concept-feature extraction method, and used feature-based news text vector representation method to improve the textual representation, and reduce dimension in feature space, and thus to acquire a better computational efficiency and the clustering quality. In analysis of the drawback of the traditional clustering method in topic detection and on the basis of density-based text clustering method, we put forward a sub-cluster relationship-based and multi-resolution density clustering (SRB-MRClustering) algorithm to achieve the hot topics auto-detection. This new algorithm considered sub-cluster neighbourhood relation and relative local density, and thus, overcame the drawback of the density-based clustering strategy which is very sensitive to the algorithm’s input
parameters to a large extent. Finally, we conducted hot topics auto-discovery experimental verification using the past news corpus, clustering effects’ presentation, and experimental results’ comparison analysis and performance assessment.

Although the experiments adopted the advanced Chinese word segmentation ICTCLAS, the segmentation dictionary is incomplete. Meanwhile, how to build a sound evaluation system (for example, determining the performance evaluation criteria and optimization of parameters) to reduce the impact of subjective factors in algorithm upon experimental results is also deserving further research.

Acknowledgement

This work is partially supported by Department of Education of Guangdong under Special Innovation Programme (Natural Science) with Project No.: 2015KTSCX183.

References


