Research on Hotspot Discovery of Academic Resources based on Hadoop

Li Guo¹, Dejun Chen², Shuoxuan Fang² and Rong Huang²,*

¹Department of Information Engineering, HBEPC Technology Training center
(Wuhan Electric Power Technical College), 430079 Wuhan, China
²School of Information Engineering Wuhan University of Technology,
430070 Wuhan, China

ABSTRACT

In view of the situation that user can not access timely and accurately to the current research hotspot, this paper proposes a hotspot of academic resources discovery method, which aims to combine the map-reduce function of Hadoop with the text clustering technology based on DBSCAN, and to realize hotspot discovery of academic resources in distributed environment. By comparing the efficiency of hotspot discovery of academic resources in the single centralized architecture management environment, the efficiency and rationality of Hadoop-based hotspot discovery of academic resources is proved.

INTRODUCTION

With the continuous development of science and technology, academic resources sharing become more frequent. However, the explosive growth of academic resources makes it difficult for users to get the current hotspot quickly and accurately when facing massive academic resources. How to meet the needs of users, at the same time improve the experience of Value-added logistics service, have become a new problem that affects the rapid development of academic resource sharing platform.

Hotspot discovery is a kind of information processing technology in recent years. It extracts hot topic from the vast amount of complex information with statistical and cluster methods, and automatically shows users the result[1]. However, with the rapid growth of academic resources, information resources sharing is an new trend[2]. However, academic resources management and hotspot discovery using traditional centralized architecture will not meet the need of high-efficiency, high-scalability[3]. Therefore, there is a great demand for seeking hotspot discovery method with distributed environment management to deal with massive academic resources. Hadoop is an open-source and free distributed computing platform with high-scalability, high-availability, high-reliability and many other advantages[4], which all qualify it with the ability to manage the massive academic resources in the distributed environment.

Overall, hotspot of academic resources discovery method based on Hadoop is proposed in this paper, which aims to combine the MapReduce function under the Hadoop architecture with the text clustering technology based on DBSCAN, and to realize the function of hotspot discovery of academic resources in distributed environment, providing a new method to realize hotspot discover of academic resources.
**DBSCAN TEXT CLUSTERING METHOD BASED ON HADOOP**

DBSCAN (Density Based Spatial Clustering of Applications with Noises) is a classical clustering algorithm based on density[5]. The main function of DBSCAN is to gather all the data points with connected density into one class, which is introduced in detail in the literature[6].

Under the Hadoop architecture, the data collected by the system is stored in different memories according to the structure of name-node. It is necessary to firstly divide the data set when clustering analysis with DBSCAN algorithm in Hadoop environment. If the data set is divided into some independent small subsets, it will result in dividing some objects, which belonged to one cluster, into different ones. Hence, the accuracy of cluster will be declined.

As shown in Figure 1, Object \( o \) is density reachable to object \( p,q \). According to the DBSCAN definition, object \( o,p,q \) should be in the same cluster. If the data set is divided into two subsets A and B as Figure 1, obviously the objects that should be in the same cluster is divided into different clusters. The reason for this phenomenon is that there are objects which are related to both the two clusters. When the related object is a core object, the objects from the same cluster would be divided into different ones. In Figure 1, the object \( o \), which is associated with both subsets of A and B, belongs to the core object.

According to the above analysis, DBSCAN algorithm based on Hadoop must ensure that there are overlaps among every data subset, that is to say each subset is associated with its front and latter subset. For simplicity, the width of the overlapping area should not exceed \( 2Eps \) (\( Eps \) is radius threshold). If the overlap region width is too large, the number of related objects, which need to be judged whether it is the core object or not, will increase accordingly, resulting in more time complexity of the algorithm. In this paper, the width of overlapping area is set as \( Eps \). As shown in Figure 2, the subsets of A and B are clustered independently with DBSCAN. If the result shows that the two classes contain the same object, and the object happens to be the core object, they will be combined.

DBSCAN algorithm based on Hadoop: firstly, the map function is used to realize the segmentation of data set. Secondly, the reduce function is used to realize the clustering of the data subset of different nodes. Finally, the clustering results of each
node contained the same object are combined to get the final result.

The pseudo code of data segmentation based on Map function.

```
Map process pseudo code

input: dataset
output: <id, value>  // m indicates the partition number, value indicates
                    // space vector
for each point p in DB DO
  m = value / d;
  n = value % d;
  if p is not in overlapping areas
    output(m, value);
  else if p is in front overlapping areas
    output(m - 1, value);
    output(m, value);
  else if p is in latter overlapping areas
    output(m, value);
    output(m + 1, value);
END FOR
```

(2) The pseudo code of clustering the data sunset in different nodes base on Reduce function.

```
Reduce process pseudo code

input: <id, value>
output: <id, value>
FOR each <id, value> DO
  List ds.add(value);
END FOR
DBSCAN(ds);
output(id, ds);
```

(3) The pseudo code of combining the clustering results contained the same object.

```
Reduce process pseudo code

input: <id, value>
output: <id, value>
FOR each <id, value> DO
  IF an object is located in two clusters, and is the core object
    Merge two clusters
  END IF
END FOR
```

**HOTSPOT DISCOVERY METHOD OF ACADEMIC RESOURCES BASED ON HADOOP**

Hotspot of academic resources discovery process is as shown in Figure 3:
Figure 3. Hotspot discovery process of academic resources.

The flowchart of Hadoop-based hotspot discovery is shown as Figure 4:

As shown in Figure 4, the whole process consists of three stages. The first stage contains feature item extraction, feature set reduction and construction of feature space; the second stage completes feature item weight calculation; the third stage mainly completes clustering result description.

1. **MR-I**: MR-I is mainly to extract feature items, and it consists of one map and two reduce processes.

   **MR-I map process pseudo code**
   ```plaintext```
   input: dataset
   output: <key,1>
   keys ← analysis keywords of academic keywords
   FOR each key in keys DO
     output(key,1)
   END FOR
   ```

   **MR-I reduce process pseudo code**
   ```plaintext```
   input: <key,1>
   output: <key,num>
   FOR each value in values DO
     num++
   END FOR
   output(key,num)
   ```

   **MR-I reduce` process pseudo code**
   ```plaintext```
   input: <key,num>
   output: <key,num>
   FOR each value in values DO
     Priority Queue queue.add(value)
     IF queue.size>N
       queue.remove(0);
     END IF
   END FOR
   END FOR
   FOR each value in queue DO
     Output(key,num)
   END FOR
   ```
(2) MR-II is mainly to compute the weight of feature item.

| MR-II map process pseudo code |
|-------------------------------|
| input: <key,num>             |
| output: <id, values>         |
| FOR each key keys             |
|     W←calculate the feature weight |
| END FOR                       |
| Output(id,W);                 |

TEST RESULTS

The common indicators of text clustering are Precision and Recall[7]. The former indicates the proportion of the number of real objects in clustering result which do belong to a class of objects, and the number of all objects of this class in the result. The latter indicates the proportion of the number of real objects in clustering result which do belong to a class of objects and the number of all real objects which should be in this class.

The paper selects 497 academic resources collected by 2013-2015 in a periodical network operation platform as test samples. The effectiveness of hotspot discovery is verified by comparing the clustering results with the artificial analysis. Through the sample analysis of 497 academic resources, five hotspots, including cloud computing, genetic algorithm, BP neural network, triangular fuzzy number and fuzzy comprehensive evaluation, had been divided.

In this example, the total number of resources in the test samples is 497, the value of MinPts is 4, the value of Eps is 0.7, the number of feature items is 15, the comparison results of clustering and artificial analysis are shown in TABLE I:

Through the observation of feature items, there is a clear inclusion relation between some groups of feature items, such as “BP neural network” and “artificial neural network”, “genetic algorithm” and “hybrid genetic algorithm”. But these algorithms are used as independent feature item, resulting in the same academic resources possibly being divided into different clusters. Merge the feature items with the inclusion relation, then to compare it with artificial analysis. The comparisons are shown in TABLE II:

| Hotspot category                  | Number of clustering results | Artificial classific number | Clustering effect evaluation |
|-----------------------------------|-----------------------------|-----------------------------|------------------------------|
| cloud computing                   | 45                          | 47                          | 1.0                          | 0.9782                        |
| genetic algorithm                 | 35                          | 40                          | 1.0                          | 0.875                         |
| BP neural network                 | 15                          | 19                          | 1.0                          | 0.7895                        |
| triangular fuzzy number           | 15                          | 16                          | 1.0                          | 0.9375                        |
| fuzzy comprehensive evaluation    | 15                          | 13                          | 0.8667                       | 1.0                           |
TABLE II. COMPARISON RESULT OF RESOURCE HOTSPOT DISCOVERY WITH FEATURE MERGING.

| Hotspot category          | Number of clustering results | Artificial classification number | Clustering effect evaluation |
|--------------------------|-----------------------------|----------------------------------|-----------------------------|
|                          |                             |                                  | Precision       | Recall       |
| cloud computing          | 45                          | 47                               | 1.0             | 0.9782       |
| genetic algorithm        | 38                          | 40                               | 1.0             | 0.9500       |
| BP neural network        | 18                          | 19                               | 1.0             | 0.9470       |
| triangular fuzzy number  | 15                          | 16                               | 1.0             | 0.9375       |
| fuzzy comprehensive evaluation | 15                      | 13                               | 0.8667         | 1.0000       |

Comparing TABLE I and TABLE II, the recall of “genetic algorithm” and “neural network” have obvious improvements, because the “genetic algorithm” and “hybrid genetic algorithm” are merged as “genetic algorithm”, which increases the number of academic resources of corresponding hotspot, and proves the effectiveness of the improvements.

The time consuming of hotspot discovery of respectively 10, 50, 1000, 2000 and 3000 academic resources sample were tested in both single and Hadoop Mode. In order to make the comparison fairer and more accurate, the Hadoop Mode is pseudo distributed, which is a distributed environment with five processes on a computer. The comparison results are shown in Figure 5:

![Figure 5. Comparison of time consuming in single and Hadoop mode.](image)

From Figure 5, it can be found that when the number of academic resources is small, Hadoop mode would consume more time than single model, because resource scheduling, load balancing and network communications would take additional time in distributed environment. When the number grows to a thousand, the advantage of Hadoop model begins to show up. The larger the number, the more time the Hadoop model will save, which proves that hotspot discovery method based on Hadoop can greatly improve the efficiency.

CONCLUSIONS

This paper aims at the situation that user is unable to grasp the current research hotspot timely and accurately due to massive academic resources, and proposes a DBSCAN clustering method based on hadoop, which can realize the automatic discovery of resource hotspots. The accuracy of the hotspot discovery is increased by
merging feature items with inclusion relation. It implements the hotspot discovery of academic resources based on Hadoop, the efficiency is proved to be greatly improved through the comparison of test samples.

REFERENCES

1. Strugala V., Avis J., and Johnstone L. M., et al. 2002. Introduction to Topic Detection and Tracking. Springer US, pp:1-16.
2. Bai B., and Gao B. 2013. “Present Status, Feature and Enlightenment of Foreign Library Resources Sharing.” Journal of Library Science in China, 2013, 39(3):108-121.
3. Wang L., Tao J., and Ranjan R., et al. 2013. “G-Hadoop: MapReduce across distributed data centers for data-intensive computing.” Future Generation Computer Systems, 29(3):739-750.
4. White T. 2010. “Hadoop: The Definitive Guide,” O’Reilly Media Inc. Gravenstein Highway North, 215(11):1-4.
5. Li Q., and Yuan J. S. 2012. “Optimal density text clustering algorithm based on DBSCAN,” Computer Engineering and Design in China, 33(4):1409-1413.
6. Yu Y. F., and Zhou A.W. 2011. “An Improved Algorithm of DBSCAN,” Computer Technology and Development in China, 21(2):30-33.
7. Huang C. H., Yin J., and Hou F. 2011. “A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method,” Chinese Journal of Computers, 34(5):856-864.