IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. V (Jan – Feb. 2015), PP 24-27
www.iosrjournals.org

Research on Clustering Method of Related Cases Based On Chinese Text

Gang Zeng[1]
(Police Information Department, Liaoning Police Academy, Dalian, China)

Abstract: *With the movement of population, crimes are showing mobility and professionalism characteristics, the cases which the same suspects committed have the same or similar characteristics, it is helpful to find the related cases. In this article, firstly, analyzes the factors associated with the cases, discusses how to form vector space of Chinese cases, then proposes a method to associate cases by joint use of FKM and Canopy clustering algorithm.*

Keywords: *related cases, Chinese segmentation, Fuzzy K-Means, Canopy.*

I. Introduction

With the development of transportation and movement of population，Crimes are showing mobility，professionalism and other characteristics, the suspects often gather his associates and wander around to commit the crime. These cases have same characteristics, because the same person or gang do it. if the cases are clustered, Investigation will make an unexpected progress.

When the case investigation is in trouble, Investigators often merge cases with the same characteristics and investigate them, to find out the relationship between the cases, to find more clues and speed up the investigation of cases. Cases often be searched in the cases database, to find out related cases with the same characteristics, we call it related cases, after cases clustered, there will be a few hundred records in a class, or even more. Investigators will spend much time and effort. to find out related cases in these cases, In this way it is trivial and inefficient.

Ning Han presented a clustering method for related cases using FKM algorithm(Fuzzy K-means algorithm, also known as the fuzzy C-means algorithm), and analyze the information of the cases. and do the experiment based on 681 records of cases, this analysis method is feasible in the low-dimensional stand-alone environment. But his method does not solve the problem of Chinese word segmentation, Under condition of high-dimensional and big data, the clustering results could not be gotten within an acceptable time using his method, We propose a method, in the Hadoop environment, we can analyze related cases using Canopy and FKM clustering algorithm, this method can overcome the lack of FKM clustering algorithm, and improve the efficiency and quality of analysis.

II. Conditions Of Associated Cases

Conditions of related case includes many factors, such as time and location of the crime, the way into the scene of crime, tool for criminal purpose, criminal activity sequence, modus operandi, the way and means of escape, the way of transportation and sales of stolen goods, and so on. Because a variety of factors, such as the suspect's life experiences, history of crime, level of education, professional features, life skills, mental set, etc. In the course of committing the crime in different places, means and methods of the suspect committing the crime have formed a stable characteristic, we call it crime stereotypes of criminal behavior, in the different cases that the same suspect has made, the factors that mentioned above may are manifested more or less, this shows that the stability of the crime. To find more investigation clues of cases, we must analyze the relationship of all cases recorded in database, to find out the information we need.

1. The Same Or Similar Nature Of The Cases

All kinds of cases , such as criminal cases, security cases, economic cases, etc. the suspects are affected by the above factors, their criminal behavior presents the characteristics of continuity. The suspect who unlock a lock by tools will steal things in a long time, the suspect who murder will still kill someone.

2. Regularity Of The Time Of Occurrence Of The Case

A person or a group of people will commit crimes , at the same time or similar time. The time when the suspect commit crime often shows a certain regularity, for example, a type of cases often occur at a certain time in a day, a certain day in a week, a specific time period in a year. for example , Car theft cases are far more likely to occur from midnight to 6 a.m. Robbery, theft and other usurpation of cases occurred on the eve of Chinese New Year.

3. Target Suspects Violated Is Same Or Similar

Target suspects violated is closely related to the suspect's motive, Because of different psychological needs, the choice of target is in a somewhat tendentious. for example, young single woman often become the object of rape or robbery, Precious industrial raw materials or digital products become the object of theft.

4. Suspect'S Body Features Is Same Or Similar

if suspect's body features is same or similar in different cases, we can consider the cases is related. body features including height, weight, gender, countenance, clothing, accent, and so on.

5. Crime Means, Methods, Processes Are Same Or Similar

Because of the suspect's history of crime, life skills, and other factors, In terms of tools of crime, the characteristics of the crime, the crime procedures, etc. a certain pattern has been formed, And with its own specificity and stability. for example, to destroy grille, some suspect prefer to use the jack, some prefer to use a pipe wrench, some suspect prefer to use stick. to avoid video surveillance, some suspect choose cap, some suspect choose mask, some suspect choose headgear.

6. Vestige Are Same Or Similar

Vestige is defined as a vestige of something is a very small part that still remains of something that was once much larger or more important. we can use it as evidence in court, vestige includes footprints, fingerprint, handprint, shot mark, semen, bloodstain, hair, toxicant, exploder, handwriting, and so on. After comparison, overlapping, anastomosis, checkout, If identified as same or similar things in different cases, We can believe that these cases are related.

Now, police manage cases by RDMS, each field describe a characteristic of the case, for example, the 'time' field describe the time when the case occurred, the 'body features' field describe the suspect's features of his(her) body, including common human characteristics, such as left ring finger broken, yellow hair, myopia and with glasses. The method for the management of the cases made important contributions.

But disadvantages are also obvious, firstly, It is complex to enter data into database, each case must be recorded in the management system, a full-time police is responsible for data entry, because it is very complex to describe the case. this is waste of police. secondly, the quality of the data is uneven. Because of different understanding of the case, different expression level, different degree of familiarity with the management system. much of the data is wrong, useless. thirdly, it is excessive precision, When associated with the cases, we can seldom find related cases, It did not play the desired effect to associate with cases.

For the above reasons, we present that interrogation record of the case is adopted as the basis for association. we can relate cases based on Chinese text, playing a specialty of big data, we can find related cases nationwide.

III. Cluster Analysis Of Cases

Vector space model (VSM) is a common method to describe cases, the case set is described as a vector space. Then we analyze the related cases by studying the similarity of cases. Clustering analysis is an important aspect of the study of the data mining. It is process of division physical and abstracted collection into several classes composed of similar objects. Clustering analysis is the learning process of a free guide. It can divide criminal acts into a number of classes or cluster. In same cluster crime acts enjoy high similarity. On the contrary, in different clusters they have large difference. in this article we will use fuzzy K-means clustering algorithm to associate with cases.

1. The Choice Of Distance Measure

It is very important, how to calculate the distance between the vectors, and it is directly related to the accuracy of clustering. how to calculate is the best way? there are many factors which infects the accuracy, but the most important factor of all factors is the choice of distance measure.

(1) Euclidean Distance Measure

The Euclidean distance is the simplest of all distance measures. It's the most intuitive and matches our normal idea of distance. Euclidean distance between two n-dimensional vectors (a1, a2, ... , an) and (b1,b2, ... , bn) is

$$\mathrm{d} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_i - b_i)^2}$$

(2). Manhattan Distance Measure

The distance between any two points is the sum of the absolute differences of their coordinates. This distance measure takes its name from the grid-like layout of streets in Manhattan. we can't walk from 3th

Avenue and 3th Street to 5thAvenue and 5th Street by walking straight through buildings. The real distance walked is two blocks up and two blocks over. the Manhattan distance between two n-dimensional vectors (a1, a2, ... , an) and (b1,b2, ... , bn) is

$$\mathrm{d} = |a_1 - b_1| + |a_2 - b_2| + \cdots |a_n - b_n|$$

(3). Weighted Distance Measure

Weighted distance measure is based on Euclidean and Manhattan distance measures. A weighted distance measure allows you to give weights to different dimensions in order to either increase or decrease the effect of a dimension on the value of the distance measure. This will affect specific distance measures differently but will in general make the distance value more sensitive to differences in a dimension.

2. Representing Chinese Text Documents As Vectors

The vector space model (VSM) is the common way of vectorizing text documents. All the words in the world form a set of words, each word being assigned a number, the number is the dimension, it'll occupy in document vectors. the dimension of these document vectors can be very large. the maximum number of dimensions possible is the cardinality of the vector. Because the count of all possible words or tokens is unimaginably large, text vectors are usually assumed to have infinite dimensions.

Chinese text processing involves the Chinese segmentation. Chinese is different from English, English words in the document are divided into independent word by a space or punctuation, Chinese Document can cut apart into sentences by Chinese punctuation. but there is no space in sentence. Chinese segmentation is the most important factor in representing Chinese text documents as vectors. The method commonly used in Chinese segmentation is "two-way maximum matching", First, maximally splitting in the positive direction, then maximally splitting in the opposite direction, the combination of the two parts is the final result, during the procedure, an atomic word base is needed.

The more dimensions in vector space, the more complex the calculations become, It is proved by theoretical calculations and practical tests. Representing the case with high-dimensional vector require enormous computing resources, so two questions must be solved to calculate the related cases: firstly, reduce the dimension of vector space, secondly, provide sufficient computing resources.

Atomic word is usually expressed as a natural word, the natural words are usually not standard, the same meaning is expressed in different words, atomic word must be regulatory. the method to establish a thesaurus and stop words dictionary can reduce the dimension of vector space effectively.

Sufficient computing resources in a stand-alone environment is impossible, with the emergence of Hadoop, MapReduce model can break the huge computing task into small tasks, the small tasks can be executed at different nodes, Hadoop provides us with sufficient computing resources. so we can calculate the relevance of the cases in the Hadoop environment.

3. Term Frequency-Inverse Frequency

Term frequency–inverse document frequency (TF-IDF) weighting is a widely used improvement on simple term-frequency weighting. The IDF part is the improvement; instead of simply using term frequency as the value in the vector, this value is multiplied by the inverse of the term's document frequency. That is, its value is reduced more for words used frequently across all the documents in the dataset than for infrequently used words.

the TF-IDF weight Wi for a word wi is:

$$W_i = TF_i \cdot log \frac{N}{DF_i}$$

Where: TFi represents the term frequency (TFi) of word wi, N represents the document count. the document vector will have this value at the dimension for wordi. This is the classic TF-IDF weighting. stop words get a small weight, and terms that occur infrequently get a large weight. The important words, or the topic words, usually have a high TF and a somewhat large IDF, so the product of the two becomes a larger value, thereby giving more importance to these words in the vector produced.

IV. Application Of Clustering Algorithm

1. Fuzzy K-Means Clustering Algorithm

K-means clustering algorithm tries to find the hard clusters (where each point belongs to one cluster), But fuzzy k-means discovers the soft clusters. In a soft cluster, any point can belong to more than one cluster with a certain affinity value towards each. So the theory of the fuzzy clustering is more suitable for the nature of things, and it can reflect objectively the reality. Currently, the fuzzy K-means clustering (FKM) algorithm is the most widely used.

FKM partitions set of n objects $X = (x_1, x_2, \cdots, x_n)$ into K fuzzy clusters with $C = (c_1, c_2, \cdots, c_k)$ cluster centers. In fuzzy matrix $U = (u_{ij},)$, uij is the membership degree of the ith object with the cluster, The characters of uij are as follows：

$$u_{ij} \in [0,1]\ i \in 1,2,\cdots,n; j\epsilon 1,2,\cdots,k$$

$$\sum_{i=1}^{k} u(i,j) = 1,\ \ j = 1,\cdots,n$$

Update uij according to (1):

$$u_{ij} = \frac{1}{\sum_{k=1}^{k} \left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{(m-1)}}} \qquad (1)$$

Where: m>1is fuzziness exponent, cj is the clustering center. $d_{ij} = \|x_i - c_j\|$is the distance between xi and cj. Update cj according to (2)

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m} \qquad (2)$$

The objective function is the equation (3):

$$J(U, c_1, \cdots, c_k) = \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^m\ d_{ij}^2 \qquad (3)$$

The nature of FKM algorithm is to apply the gradient descent method to find out optimal solution, so there is a local optimization problem. And the algorithm convergence speed is greatly influenced by the initial value, especially in the case of large number of clusters. For many real-world clustering problems, the number of clusters isn't known beforehand, A class of techniques known as approximate clustering algorithms can estimate the number of clusters as well as the approximate location of the centroids from a given data set. One such algorithm we can use is called canopy generation.

2. Canopy Clustering Algorithm

Canopy's advantage lies in its ability to create clusters extremely quickly—it can do this with a single pass over the data. But this algorithm may not give accurate and precise clusters. But it can give the optimal number of clusters without even specifying the number of clusters, k, as required by fuzzy k-means.

V. Conclusion

This article proposed a method to associate cases by joint use of FKM and Canopy clustering algorithm, it overcome the difficult problem of clustering Chinese cases, but this method is relatively complex, the results are not very accurate, We hope to improve this method in the future.

Acknowledgment

This work was partly supported by general scientific research project of education department of Liaoning Province of China (No. L2013490).

References

[1]. Ning Han, Wei Chen. Research on clustering analysis on related Cases. Journal of Chinese People's Public Security University(Science and Technology), 2012(1):53-58.
[2]. Sean Owen, Robin Anil, Ted Dunning, Ellen Friedman. Mahout in Action[M]. Manning Publication, Westampton.
[3]. Tony H. Grubesic. On The Application of Fuzzy Clustering for Crime Hot Spot Detection[J]. Journal of Quantitative Criminology, 2006,22(1):77-105.
[4]. [Deguang Wang, Baochang Han, Ming Huang. Application of Fuzzy C-Means Clustering Algorithm Based on Particle Swarm Optimization in Computer Forensics. 2012 International Conference on Applied Physics and Industrial Engineering, 1186-1191.
[5]. Gang Liu. Programing Hadoop[M]. BeiJing:China Machine Press, 2014.
[6]. Hesam Izakian, Ajith Abraham, Václav Snášel. "Fuzzy Clustering Using Hybrid Fuzzy c-means and Fuzzy Particle Swarm Optimization" 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), pp.1690-1694,2009.