2nd International Conference on Education, Social Science, Management and Sports (ICESSMS 2016)

Application of Web Browsing Records in Anti Terrorism

REN Weizheng[1, a]

[1]School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

[a]li_yuxiang@126.com

Keywords: counter-terrorism; accurate identification; potential members; machine learning; KNN; SVM

Abstract. Potential terrorist organizations to develop accurate identification of objects is a direction of counter-terrorism work. Wen expounded based on personal web browsing history, human is used to determine the key way by KNN algorithm for text classification, according to browse the document classification results of using the SVM algorithm to abnormal discriminant method to determine the development potential terrorist organization object. The article also expounds the specific implementation process is important and difficult.

Introduction

In the process of maintaining development, terrorist organizations need to replenish financial and human resources constantly. So they keep recruiting new members through a variety of means during operation. As it stands, new members tend to be adolescents. A necessary step for terrorist organizations to attract new members is to propagandize radical thoughts.

About half of 30 terrorist organizations officially identified by the United States as early as in 1998 had their own websites. The content of these websites not only published their political motives, but also used religion to bewitch potential objects and disseminate all kinds of terrorist information. Some even discussed skills to implement terrorist activities on the websites. Most of terrorist incidents took the form of bombing. Many terrorists talked about the way to prepare and use explosives on a specific website undisguisedly.

From the state of the art, it is already feasible to identify potential objects of terrorist organizations using big data technology. Once a potential object is detected, he/she will be surveilled and guarded closely. If the recruitment of new members by terrorist organizations is limited, it will play a significant role in restricting the development of terrorist organizations.

Overview of Machine Identification

Whoever uses the network will undoubtedly leave certain records, especially when he/she browses webs. Using technical means, counter-terrorists can obtain each user’ s web browsing history and analyze the type and content of articles that he/she has read using a machine. So it is possible to identify whether a user is affected by a terrorist group’ s extreme thoughts and take precautions.

According to the proposed method in this paper, first of all, it is necessary to classify and grade webs and articles in the network and identify according to the distribution of various types of webs in individual browsing history. The advantage of doing this is that it not only ensures the effectiveness of results in primary identification, but also ensures that there are definite dimensions in secondary identification, which is more conducive to the self-evolution of algorithm.

Compared with large-scale counter-terrorist publicity, education and psychological surveys, computer identification is more efficient and targeted. Especially after some unexpected incidents, under artificial intervention, computer identification can provide guidance for counter-terrorism work rapidly and accurately.

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Relative to personalized recommendation services of many websites, the identification scope of the proposed method in this paper is not limited to a particular site’ s searching or browsing records, but a certain user’ s browsing history across the Internet.

The Data Processing of Webs and Articles

A. The source and preprocessing of web data

To analyze a user’ s browsing behavior, we must first classify and grade webs and articles in the network. First of all, a variety of means and tools were used and webs and articles updated on relevant websites were collected in the shortest possible time. The collected webs and articles cannot be used directly. We needed to remove all distractions, such as labels, ads and structures, other than plain texts. But prior to this, the hyperlink contained in articles and source of articles were two important characteristics, too. If necessary, the weights of these two characteristic values can be increased.

B. To determine training sets

When classifying and grading webs and articles using a machine, first of all a certain number of keywords must be identified. This job shall be done manually. After reading a certain number of relevant articles of various types, keywords were detected from the articles as characteristics of the articles. According to these characteristics, webs and articles can be roughly classified and graded, using a certain machine learning algorithm.

Since terrorist groups mostly had an extreme religious background, keywords, such as “holy war (jehad)” and “summon”, were likely to appear at high frequency. Sometimes terrorists would talk about skills to implement terrorist activities undisguisedly, such as the way to prepare and detonate explosives. When identifying keywords, word frequency statistics or correlation analysis can be used as appropriate. After counting a lot of articles of the same category, we were bound to find some high frequency words. These words were so-called keywords. Machine learning also had a corresponding algorithm for correlation analysis. Keywords and their correlation degree constituted characteristics of such kind of webs and articles. Articles and characteristics of the same category constituted training sets for machine learning.

C. To choose a web classification algorithm

After identifying training sets, the next step was to let clusters identify according to training sets. There were many optional algorithms. At present, there are a dozen famous machine learning algorithms.

The analysis of webs and articles belonged to classification analysis and the purpose of analysis was not to predict, so classification and supervised learning algorithms can be applied. Famous classification and supervised learning algorithms included KNN (K-nearest neighbor) algorithm and SVM (support vector machine) algorithm, etc.

SVM algorithm: the essence of SVM was to determine a separating hyperplane to separate data sets. The decision plane had one fewer dimension than the data set. One side of the separating hyperplane was a category. The other side was another category. In fact, the longer geometrical distance from a separated object to decision plane, the higher reliability of the results. What SVM algorithm did was to find the optimal separating hyperplane. It was commendable that the operating efficiency of SVM algorithm on a suitable platform was very high, which can win precious time for imminent counter-terrorism tasks.

KNN algorithm: to put it simply, KNN algorithm was to classify using the position of different characteristic values in their own dimensions. The extraction of characteristics was done manually and the characteristic values were also identified using word frequency statistics. So if N characteristics were identified before, each characteristic and its corresponding value constituted a point in N-dimensional space. KNN algorithm was realized by comparing the geometric distance between the analyzed object and a point in N-dimensional space, whose category had been identified.

For example, after calculation, the position of a given category in the N-dimensional space was Point A, whose coordinate was (A1, A2, A3... An). The position of another category was Point B, whose coordinate was (B1, B2, B3…Bn). The target point was X. The distance from Point X to A and B was calculated using the Pythagorean theorem. If it was very close to A, but far away from B, then it was considered to belong to Category A. Otherwise, it belonged to Category B. If in an example, many predefined keywords recurred frequently and different keywords presented a certain proportion, then they may come from an article of a given category. This relationship was derived by artificial identification, followed by a machine analysis. The resulting multi-dimensional characteristic value was the coordinate of a corresponding point in multi-dimensional space.

SVM algorithm applied to two-type problems only. So to divide articles into multiple types, we must carry out a multiple SVM analysis. Multiple SVMs may cause error accumulation.

The accuracy of KNN algorithm was high and it was not very sensitive to outliers. But the calculating time and space complexity were relatively high. KNN algorithm cannot train a machine, so the correction, update and evolution of characteristic value of a point of a known category must be done manually.

However, counter-terrorist articles had very strong timeliness. Although SVM algorithm can learn by itself and the operation was more reliable, when some incidents happened, perhaps even the dimension of data set would change. SVM algorithm was inferior in convergence rate, that is, its instant adaptability was poor. Besides, the application of multilevel SVM algorithm was in all probability to cause error accumulation. Common algorithms also included decision tree algorithm and Naive Bayes algorithm, but due to the lack of characteristic values to make an assertion, decision tree algorithm was not taken into account. Since the conditional probability in web grading was too complex, it was not taken into account, either. So in the classification of articles, KNN algorithm was more advisable.

The Validation of Web Classification Algorithm

It was not easy to validate the feasibility of KNN algorithm in the classification of articles, for the following reasons: first of all, it was much difficult to get sufficient articles. Secondly, the author knew too little about extreme terrorist thoughts. So the author found 19 articles about an NBA player, Kevin Durant. The reasons for choosing this topic were as follows: first of all, there were sufficient articles about Kevin Durant for classification. Secondly, Kevin Durant had just made a decision to turn to Warriors, which defeated him in the last season’ s playoff and this was a controversial decision. Thirdly, the author was familiar with relevant articles and background.

During the validation, all articles selected by the author came from the network. Most articles were sourced from other NBA players’ comments on Durant’ s turn to Warriors. After reading all of these articles, the authors decided to classify them into three categories: pro, con and uncertain. In the choice of keywords, this article chose 8 keywords.

Since these articles were short, the correlation between keywords was not so clear. So these words would not appear in large quantity simultaneously. So the regularity was not very strong. The keyword frequency statistics were as follows:

Table 1 The Proportions of Keywords in All Samples

		?	LeBron James	Free	Westbrook	Thunder	Know	Decide	No
Pro	1	1	2	1	0	6	1	0	3
	2	0	0	1	0	0	0	3	1
	3	2	1	2	4	13	0	0	28
	4	9	0	0	0	3	0	0	13
	5	1	2	0	0	4	0	1	3
	6	19	6	0	0	0	0	1	25
Con	1	0	0	0	0	0	0	1	1
	2	9	0	0	0	11	1	1	12
	3	2	0	0	3	2	0	0	6
	4	0	0	0	0	3	1	0	1
	5	0	0	0	0	2	4	1	6
	6	1	1	0	3	4	0	1	12

	7	0	0	0	0	0	1	3	5
Uncertain	1	0	0	0	0	0	1	3	2
	2	0	0	1	0	0	3	0	3
	3	3	0	0	0	0	1	1	8
	4	0	0	0	2	0	0	4	5
	5	0	0	0	0	0	2	1	4
	6	0	0	0	1	10	1	2	16
	7	1	0	0	0	0	0	4	6

After the keyword frequency statistics in the articles, to avoid too much influence of the frequency of occurrence on classification results when the length difference was too large, the keyword frequency was divided by the length of article (thousand words). The result was corresponding characteristic value. Below are the characteristic values of all dimensions in pro articles. The last row was means.

Table 2 The Characteristic Values Obtained by All Keywords Divided by Length in Pro Samples

?	LeBron James	Free	Westbrook	Thunder	Know	Decide	No
1.44	2.89	1.44	0	8.7	1.44	0	4.34
0	0	5	0	0	0	15	5
1.78	0.89	1.78	3.57	11.60	0	0	25
20.45	0	0	0	6.81	0	0	29.5
2.63	5.26	0	0	10.52	0	2.63	7.89
12.5	3.947	0	0	0	0	0.65	16.44

After processing three types of articles as above, the average coordinate of three categories were obtained.

Below the proposed algorithm was validated. Due to a small number of articles, the author used previous articles to validate the algorithm. The validation method was to calculate the distance from the coordinate of each article to the average coordinate of three categories. After comparison, the nearest point was taken to get the classification results.

The classification results were as follows:

Pro:	5/6	83%
Con:	0/7	0%
Uncertain:	5/6	83%
Total:	10/19	53%

From the above results, the classification accuracy of pro and uncertain was high, while the classification of con all failed. The reason was as follows: when choosing keywords, the expression of con articles varied, such as satire, abuse and even many ironies and allusions, so it was difficult to choose keywords. Only two keywords were identified in con articles, “Westbrook” and “Thunder”. Eventually, two invalid keywords greatly affected the accuracy of the entire analysis results.

But it can also be seen that when the choice of keywords was appropriate, the reliability of classification results was acceptable. Articles containing extreme religious thoughts and terrorism often had distinctive themes. The complexity of expression was not very high, so professionals observed a lot of samples. When they were sure that the selected keywords were reasonable, the prospect of KNN algorithm for the classification of webs and articles was still promising.

The Screening of Potential Objects of Terrorist Organizations

After classification in the previous step, the dimensions of data sets were determined, too. Besides, the identification of potential objects was a typical two-type operation, so at this point, SVM algorithm was more suitable.

In the processing of webs in the previous step, webs were divided into 10 grades. So in the proposed algorithm, the number of dimensions of data sets was ten and the number of dimensions of separating plane was nine. The browsing history of people that had been identified as potential objects shall be used as training sets as far as possible to train the system.

First of all, the statistics of browsing history was conducted. The occurrence of various kinds of webs was recorded and divided by the total to get their proportions. Then the characteristic values of all dimensions of a given sample were obtained.

As mentioned above, the principle of SVM algorithm was to choose the optimal decision hyperplane. The principle of choosing the optimal hyperplane was actually to maximize the distance from the nearest point to hyperplane. So once the data sets were determined, the optimal hyperplane had a unique solution. The ongoing evolution in the process of machine learning was also to approach the optimal solution with the increase of data. So fundamentally, the learning process of SVM algorithm was a process to adjust hyperplane equation parameters.

Suppose that X_i was the data to be classified. Y_i was a classified category. $Ax+B=0$ was a separating hyperplane. If x_a satisfied $|a \cdot x_a + B| = 1$, then x_a was called support vector and the optimal hyperplane satisfied $y_i \left[A \cdot x_i + B\right] \geq 1$.

On the current popular Spark platform, MLlib had a rich built-in function library, including classification, regression, clustering, association rules, recommendation, dimension reduction, optimization, feature extraction and screening, etc. Moreover, a memory-based operating framework made the computing efficiency of Spark almost one hundred times as high as Hadoop. So in many cases, Spark was an ideal platform for machine learning.

With the help of tools, we can get the final decision hyperplane.

However, SVM algorithm still required a lot of labelled samples as its training sets. This process can only be obtained through potential objects identified previously, even through simulated data created by researchers. The bottleneck of the whole algorithm also lay in this. It was a worldwide problem and also a research hotspot to train a good classifier using a few classified samples and lots of unclassified samples.

Conclusion

Big data has an important status and plays an important role in counter-terrorism work. Whether or not big machine learning technology can achieve desired accuracy in counter-terrorism at the current stage, it is bound to show great power in the future. Except big data and the strong analysis and processing ability of machine learning technology, high technical barriers also make it difficult for terrorist organizations to use them and thereby form an effective resistance. In the process of application, there are still a lot of difficulties, for example, text classification not only requires choosing keywords, but also finding their internal correlation and embodying these correlation and timeliness in the algorithm. It is hard for training sets identified by potential objects to reach a certain scale, so the optimization of classifier will be an emphasis and difficulty in this part of work. Despite so many challenges, it is believed the wonderful prospect of this technology will bring new ways for counter-terrorism work.

References

[1] Zhou, Danqi, Customs Intelligence System in the Era of Big Data, Economics and Management Strategies Journal, 2014(3), pp.76-81.

[2] Jiang, Xinyu, Preliminary Study on the Big Data Analytics and Its Adaptability in Intelligence Studies, Library and Information, 2014(5), pp.15-17.

[3] Lv, Xuemei, Surveying the Crime Analysis in U. S. Prediction Policing from Big Data, Journal of Intelligence, 2015(12), pp. 17-19.

[4] Huang, Xiaobin, On the Innovation and Development of Enterprises Competitive Intelligence Analysis in the Big-data Era, Library and Information, 2012(6), pp. 10-14.

[5] He, Qing, A Survey of Machine Learning Algorithms for Big Data, Pattern Recognition and Artificial Intelligence, 2014(4), pp. 328-334.

[6] Li, Benxian, Application of Big Data in Counter-terrorism Intelligence, Journal of Intelligence, 2014(12), pp. 2-12.

[7] Li, Wenlian, Business Model Innovation Based on “Big Data”, Economics and Management Strategies Journal, 2014(3), pp. 85-90.

[8] Yu, Ning, Customs Intelligence System in the Era of Big Data, master’s thesis. Chongqing: Chongqing University, pp. 6-10.

[9] Hadoop: The Definitive Guide, Zeng, Dadan (Trans), Tsinghua University Press, 2010.

[10]Nick Pentreath, Spark Machine Learning, Beijing: Posts and Telecom Press, 2015.

[11] Cai, Wen, Fundamentals of Extension Logic, Beijing: Science Press, 2004.

[12] Cai, Wen, Extension Engineering, Beijing: Posts and Telecom Press, 2007.