TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.1, January 2014, pp. 379 ~ 387
DOI: http://dx.doi.org/10.11591/telkomnika.v12i1.4141

An Improved AC-BM Algorithm for Monitoring Watch List

Tang Jun, Mo Yiwen*
Zhongnan University of Economics and Law
Wuhan, PR China (430074)
e-mail: evenmomo59@gmail.com

Abstract

With the expanding of database of the watch list of anti-money laundering, improving the speed in matching between the watch list and the database of account holders and clients' transaction is especially important. This paper proposes an improved AC-BM Algorithm, a matching algorithm of subsection, to improve the speed of matching. Experiment results show the time performance of the improved algorithm is better than traditional BM algorithm, AC algorithm and the AC-BM algorithm. It can improve the efficiency of on-line monitoring of anti-money laundering.

Keywords*: watch list, matching algorithm, anti-money laundering*

1. Introduction

Some international organizations, including the UN, and some related departments of anti-money continuously release the organizational and personal watch list of suspicious money laundering and terrorist financing, which require close monitoring by countries' financial institutions. Because the data of monitoring system updates constantly and the size of data expands day by day, the financial institutions have to spend much time in matching the database of account holders and clients' transaction, which affects the normal transaction and the on-line monitoring of anti-money laundering.

The earliest algorithm of matching single pattern is the algorithm of Brute-Force (the algorithm of BF), which is algorithm of matching in order and the efficiency is low. In 1977, Knuth D.E, Morris J.H and Pratt V.R proposed an algorithm of matching single pattern, the algorithm of KMP, which eliminate the problem of the comparison of backtracking. In the same year, Boyer and Moore proposed the algorithm of BM [1], which can skip by the rule of bad character and good postfix. The algorithm of BM performs efficiently in the algorithm of matching single pattern. In 1975, Aho and Corasick proposed the algorithm of AC [2], which uses the finite state machine to match strings and can match all the pattern strings by scanning text strings once. However, the algorithm of matching single pattern scans text strings once, and it only can match one pattern string. In 1993, Fan Jang-Jong proposed the algorithm of AC-BM [3], which uses the idea of the finite state machine of AC in preprocessing and uses the idea of skipping of BM in scanning. The algorithm highly improves the traditional the algorithm of AC in the part of scanning and matching.

In order to improve the speed of matching in watch list, this paper compares and analyses some kinds of essential pattern matching algorithms, such as BM algorithm, AC algorithm and AC-BM algorithm, and proposes an improved AC-BM Algorithm, a matching algorithm of subsection, which not only combines with the advantages of BM algorithm and AC algorithm, but also performances better than the traditional AC-BM algorithm. The speed of matching is highly improved, so that it can satisfy the demand of on-line monitoring of anti-money laundering. At last, this paper designs a model of the monitoring system, and structures the module of matching, which is one of the submodules in the system, with the algorithm of the improved AC-BM in detail. The structure can perfect the whole system of on-line monitoring of anti-money laundering well.

Received June 25, 2013; Revised August 26, 2013; Accepted September 19, 2013

2. Watch List and Matching Process

Watch list is a list about the staffs, organizations and classes of business who may lead financial institutions to face the risk and the financial institutions need to pay close attention to. The information of watch list comes from the international organization of anti-money laundering (such as Financial Action Task Force on Money Laundering, Basel Committee on Banking Supervision, Egmont Group, International Criminal Police Organization and so on), public and media, the report of the financial internal monitoring system, the risk information of government departments, while some professional institutes also collect the related information and release it regularly. It is a legal duty of financial institutes to implement the on-line monitoring of watch list. Figure 1 gives an example of watch list comes from the Bank of England, which shows some important persons of Taliban. In addition, the watch list also includes senior officials of government departments, corrupt officials, and tax evasion personnel and so on.

4383	Last Updat	2007-9-20									
4384	Name 6	Name 1	Name 2	Name 3	Name 4	Name 5	Title	DOB	Town of Birth	Country of	Nationality
4385	KARADZIC	Radovan						1945-6-19	Petnjica, Savnik, Mor	Serbia and	Bosnia and
4386	KARAJ NUCLEAR RESEARCH CENTRE										
4387	KARAM	Nabil	Victor					00/00/1954			Lebanese
4388	KARAM	Nabil	Victor					00/00/1954			Lebanese
4389	KARIM	Yves	Andoul					1973-8-20	Bunia		Congolese
4390	KARIMAN	David	Ishemunyc	Godi				1947-5-25			
4391	KARIMI SA	Javad									
4392	KARPENK	Ihar	Vasilievich					1964-4-28	Novokuznetsk	Russia	
4393	KARPENK	Igor	Vasilievich					1964-4-28	Novokuznetsk	Russia	
4394	KASKAR	Daud	Hasan	Shaikh	Ibrahim			1955-12-26	(1) Bombai (2) Ratnag	India	Indian
4395	KASKAR	Daud	Hasan	Shaikh	Ibrahim			1955-12-26	(1) Bombai (2) Ratnag	India	Indian
4396	KASKAR	Daud	Ibrahim	Memon				1955-12-26	(1) Bombai (2) Ratnag	India	Indian
4397	KASKAR	Daud	Ibrahim	Memon				1955-12-26	(1) Bombai (2) Ratnag	India	Indian
4398	KASKAR	Dawood	Hasan	Ibrahim				1955-12-26	(1) Bombai (2) Ratnag	India	Indian
4399	KASKAR	Dawood	Hasan	Ibrahim				1955-12-26	(1) Bombai (2) Ratnag	India	Indian
4400	KASKAR	Dawood	Ibrahim				Sheikh/Sh	1955-12-26	(1) Bombai (2) Ratnag	India	Indian
4401	KASKAR	Dawood	Ibrahim				Sheikh/Sh	1955-12-26	(1) Bombai (2) Ratnag	India	Indian
4402	KASMURI	Abdul	Manaf					1955-5-28	Selangor	Malaysia	Malaysian
4403	KASTSIAN	Siarhiei	Ivanavich					1941-1-15	Usokhi, Mogilev district		
4404	KASUKUV	Saviour						1970-10-23			
4405	KATANGA	Germain									Congolese
4406	KATHIM	Rashid	Taan								Iraq
4407	KAUKONE	Ray						1963-3-4			
4408	KAVOSHYAR COMPANY										
4409	KAYANI	Waseem						1977-4-28			
4410	KAZAKH JAMA'AT										

Figure 1. The watch list of Bank of England

The matching process of watch list is a matching between the watch list and the database of account holders and clients' transaction. Usually this process costs lots of time because there are plenty of personal properties of the clients requiring matched and the database is keeping expanding. Consequently, decreasing the consumed time of the matching process is especially important to financial institutes when they are doing normal businesses and the on-line monitoring of money laundering.

The watch list that this paper discusses is the basic data of the system monitor of finance. Because of the large system of finance, a lot of subsystems will use the personal information to confirm some businesses, such as opening an account, transaction, transfer and so on. Because of the large volume of business, it is easy to cover the facts of crime. The system of watch list achieves the goal of anti-money laundering by paying close attention to the staffs that are in the high risk. When a person, who is in the watch list, opens an account or trades with the personal information in the system of finance, the system monitor will sound the alarm. Like this, the suspicious criminals will be discovered in time. Obviously, it simplifies the process of anti-money laundering and it is convenient to supervise the money laundering.

The watch list mainly involves the data module of personal information. The systems of account holders and clients' transaction have close relations with it. When the operators help clients open accounts or trade, the behaviors will involve the personal information. So the watch list that this paper discusses is mainly about the systems of account holders and clients'

transaction. Figure 2 shows the structure of matching process of watch list. Supervise the money laundering by matching the information of the watch list. Judge the person whether is suspicious by matching the client's personal information with the watch list.Figure 1. Effects of selecting different switching under dynamic condition

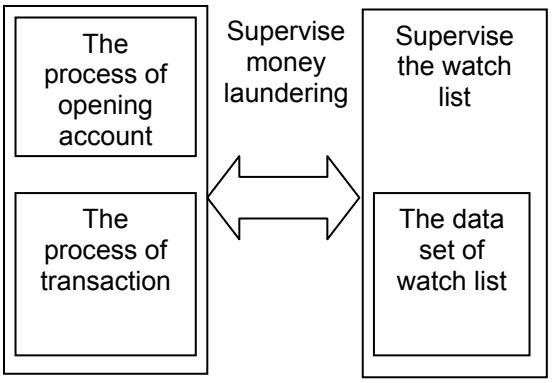

Figure 2. The structure of matching process of watch list

3. Study of Matching Algorithm

According to the requirement of filtering of the watch list's content, this paper mainly discusses the matching algorithm. On the basis of the number of pattern string, it can be classified as algorithm of matching single pattern and algorithm of matching multiple patterns. The former one matches only one string every time, while the latter one matches several strings every time. The algorithm of matching single-pattern is represented by the BM algorithm; on the other hand, the algorithm of matching multiple patterns is represented by the AC algorithm.

We assume P is a string with size m, P[0…m-1], then the P[i…j] is a substring of P, P[i…j] is a prefix of P, and P[i…m-1] is a postfix of P. ($0 \le i \le j \le m-1$).

Let T is a text string with size n and P is a pattern string. The characters of the P and the T belong to a limited character set Σ; Therefore, the goal of the matching is to find out the equal substring between P and T.

3.1. BM Algorithm

BM algorithm (Boyer-Moore) [1] is a kind of matching algorithm based on postfix. Matching backwards from right side to left side is a distinctive characteristic of BM algorithm. When there is a character that fails to match in text string T, T[i] =c, it will skip in the text string to speed up the pattern string moving. If c is a string in P, finding out the rightmost c in P and then moving P to make T[i] align it; if c isn't a string in P, moving P to make P[0] align T[i+1].

The fundamental of BM algorithm: matching one character at a certain distance in T, and then determining whether skip to the right side and the distance of skipping according to the character, which is matched at present, whether appears in P or appears in which location. If it doesn't skip, it will compare with P from the right to the left in the present location. Otherwise it will skip to the next location of P according to the skipping distance that has been computed in advance.

The algorithm also includes two parts, preprocessing and scanning. In the part of preprocessing, it just consider the pattern string p and the set of strings Σ. We introduce a function of shift (c). It shows the situation that every character c appears in pattern string P, and the distance that the pattern string moves. The function value of shift is also saved in an array. The following is a method of computing the array of shift. If c appears in the pattern string P, find out the location of c in the rightmost of P, and the location is index. The value of function is the distance between the c of rightmost and the rightmost of P, it is m-1-index; if c doesn't appear in the pattern string P, the value is the length of P.

For example, Σ={a,b,c,d} and the pattern string P = abacab, we can compute it the Shift(a)=1, Shift(b)=0, Shift(c)=2, Shift(d)=6. And the time complexity of shift is O(m+|Σ|).

Although the average efficiency of BM algorithm is high, it doesn't performance well in the worst case. The complexity of the worst case is O(mn), such as, a text string T=aaa…a, a pattern string P=baa…a.

BM algorithm is designed as a kind of algorithm that matches the single pattern string in text. Among the algorithms of matching single-pattern, BM algorithm is proved to performance best. However, when there are various kinds of key words to match in the filtering and matching of content, the BM algorithm has to match every kind of pattern. The time complexity of the BM algorithm is O (n) when matching the single pattern, but it is O (kn) when matching the multiple patterns.

3.2. AC Algorithm

AC algorithm (Aho-Corasick) [2] is a kind of algorithm that based on finite state machine. It preprocesses the set of pattern strings to form a state machine in a tree before start matching. It just scans the text string T once, and then it can find out the all patterns that match with T in P.

The algorithm also includes two parts, preprocessing and scanning. It generates three functions: the transition function: goto, failure function: failure and output function: output. The following is the matching process of AC algorithm: starting from state zero, picking up a character from the text string, and then going to the next state with goto function and the failure function. When the output function of some state is not a null, it means that it successes to find out the pattern string.

It structures the finite state machine in the part of preprocessing, and the time complexity of structuring transition function is O(M), M is the total length of all patterns. The time complexity of structuring the failure function is O (M), too. In the part of scanning, it scans every character of the text T on the basis of the finite state machine that has structured before. Every character just has one transition function, and the time complexity of scanning is O (n). So the total time complexity of the algorithm that based on the finite state machine is O (M+n). The time complexity is related to the length of text and pattern, but is not related to the content of text and pattern. It means that the average time of scanning in the best case or the worst case is the same, O(M+n).

The disadvantage of AC algorithm is that the demand of space is large. Too many matching patterns will occupy plenty of space; even maybe make the system crash. So AC algorithm can satisfy the demand and performances well in the case of a few patterns. However, it doesn't skip when it scans the text. It inputs the text in order, which means it can't skip the unnecessary comparison. Obviously AC algorithm is not the best matching algorithm in practical process of matching.

3.3. An Improved AC-BM algorithm

3.3.1. AC-BM Algorithm

AC-BM algorithm [3] combines the AC algorithm and BM algorithm. It includes two parts, preprocessing and scanning. The part of preprocessing structures a finite state machine likes AC algorithm. The part of scanning based on the idea of skipping likes BM algorithm, and matches the text string with the finite state machine. So it improves the speed of AC algorithm efficiently in the part of scanning.

3.3.2. An Improved AC-BM Algorithm

On basis of the traditional AC-BM algorithm, this paper proposes an improved AC-BM algorithm, a matching algorithm of subsection. Figure 3 shows the flow chart of the algorithm.

a. Let P align the left side of T, and start matching with BM algorithm. If matching successfully once, we call it match completely, and then return the result. When Pi≠Ti, pick up the Pi as L1, so L1={P1…Pi }.

b. Continue to match Pi+1 and make subsection when they are inequality.

c. Do step a and step b repeatedly, we get L1,L2…Lk.

d. Divide T into t1,t2, …tm, and { t1,t2, …tm }∈Σ.

e. Let { L1,L2…Lk } match { t1,t2, …tm } with the technology of the multi-thread AC matching, and then return the result.

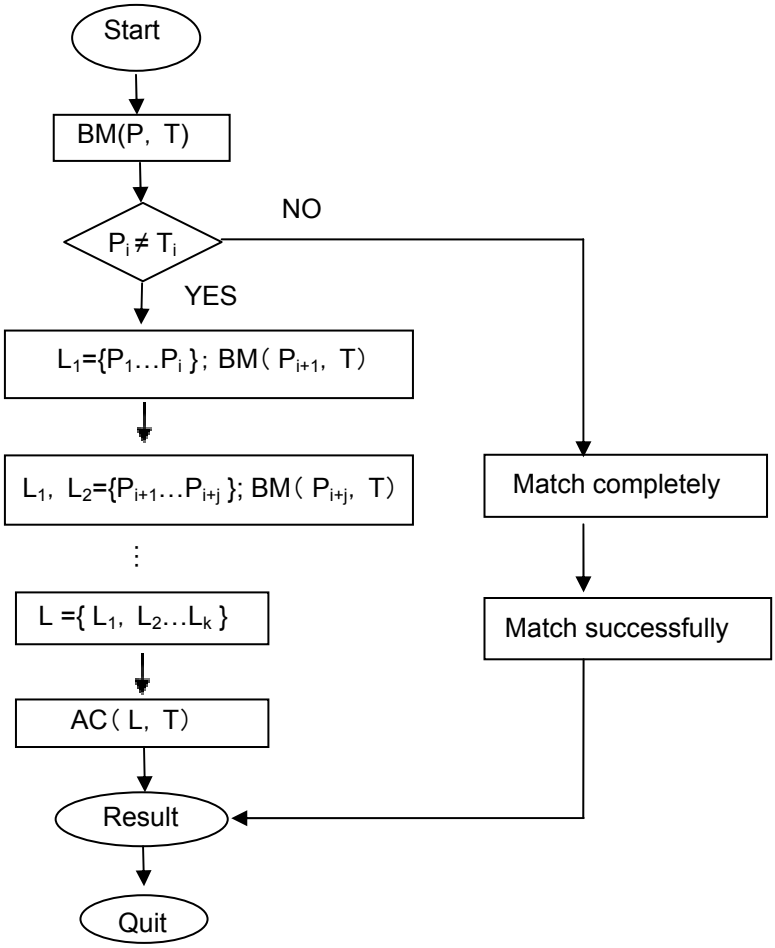

Figure 3. The flow chart of the improved algorithm

The improved AC-BM algorithm above uses the technology of subsection, so the length of string can't affect it when it is matching the single pattern. However the length of the string it is, it affects the speed of matching less. In the matching of multiple patterns, the improved AC-BM algorithm is not only able to match multiple patterns, which likes AC algorithm, but also skip the unnecessary comparison, which likes BM algorithm. In addition, it improves the traditional AC-BM algorithm and the improvement makes it better than the traditional one in the speed of matching.

4. Experiment

In order to test the efficiency of the algorithms we mentioned above, we extract the 100MB data from the informational database of clients. And we extract randomly the pattern strings from the database of watch list to test the algorithms. The system configuration is 2.4GHz Intel CPU and 2G RAM. We generate two sets of experimental data, and the size of the data is 100MB. The first data set has no repetitive strings, and the second one has plenty of repetitive strings. The length of pattern strings includes 10, 20, 30, 40, 50, 100, 500, 800, 1000. Figure 4 shows the time of the first data set that every algorithm takes. Figure 5 shows the time of the second data set that every algorithm takes.

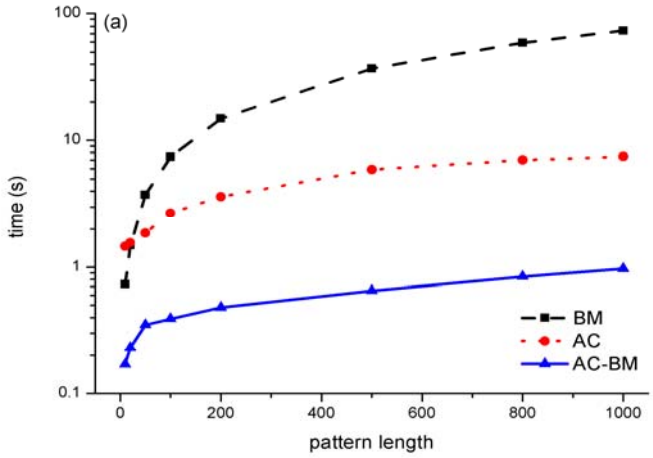

Figure 4. The result of the first data set

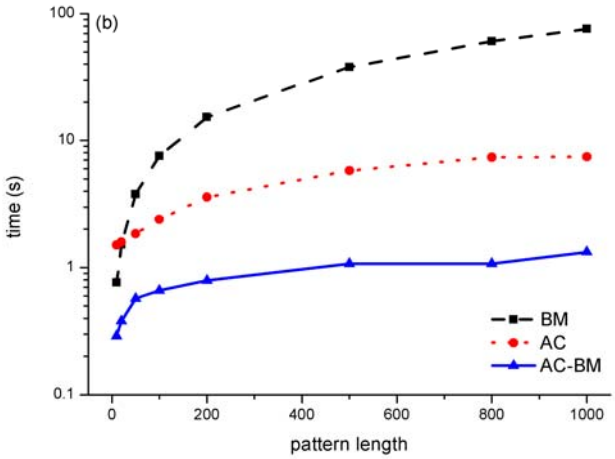

Figure 5. The result of the first data set

It shows that the improved AC-BM algorithm performs best. When the number of pattern sets increasing, the time of all the algorithms increase. Because the finite state machine is larger and it has to take more time to transform; meanwhile, the times of matching increase. Using the SMA algorithm, it can match by 100MB bytes per second and 1000 kinds of patterns. So it improves the performance of the matching algorithm of key words well. In addition, as the number of pattern strings increasing, the improved AC-BM algorithm is affected least. It means that it's convenient to add key words, and we don't need to worry about that it may affect the speed of matching. Obviously it is one of advantages of the improved AC-BM algorithm.

5. A Model of the Monitoring Watch List

5.1. The Structure of the Whole Monitoring System

The whole monitoring system should include the management and control center, the module of preprocessing, the module of matching and the module of on-line monitoring and analyzing. Figure 6 shows the structure of the system.

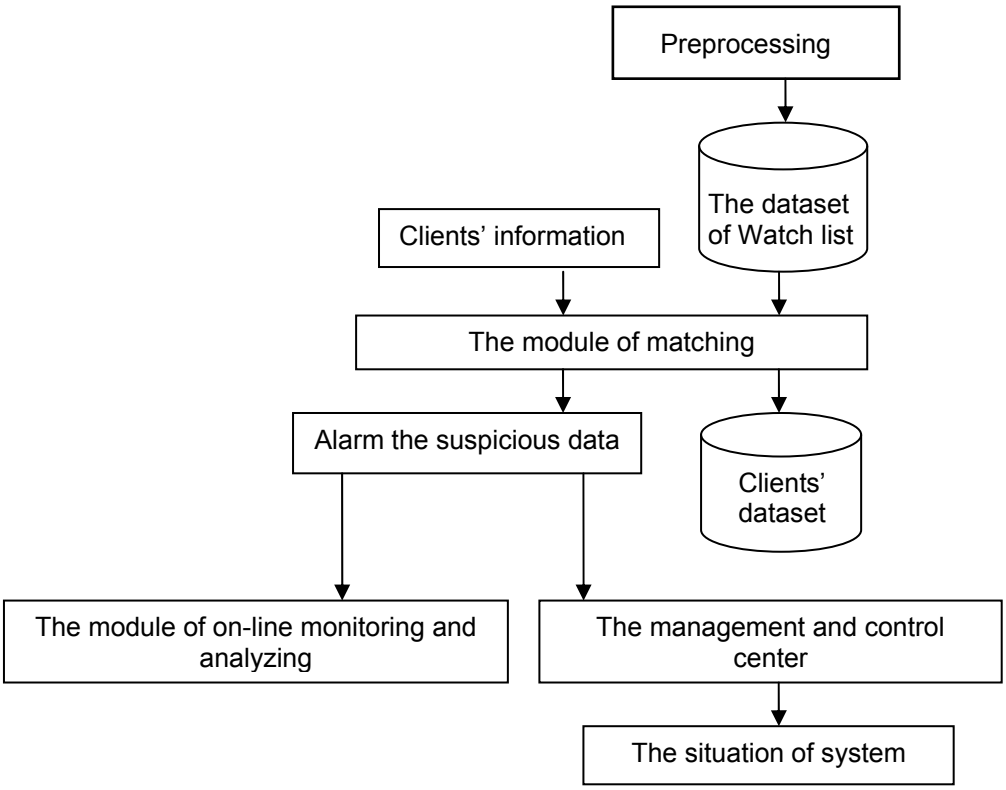

Figure 6. The structure of the monitoring systemt

The management and control center has a nice user interface, and shows the risk index of the list that is detected at present. The module of preprocessing deals with the information of watch list, and makes it standardization. The module of matching is used to match clients' information with the watch list. The module of on-line monitoring and analyzing is used to monitor the data of transactions in time.

5.2. The Structure of the Module of Matching

This paper mainly discusses the structuring of the module matching. Improving the efficiency of this module will improve the operating efficiency of the whole system and help it achieve the goal of on-line monitoring. The module of matching uses the algorithm of improved AC-BM that this paper proposes to match and detect the clients' information. If the module discovers it there is someone suspicious or someone who maybe have some relationship with suspicious organization after matching accurately, the system of monitoring will sound alarm. If the module discovers it there isn't any suspicious person or organization, it will save the clients' information to the clients' dataset. Figure 7 shows the workflow chart of this module.

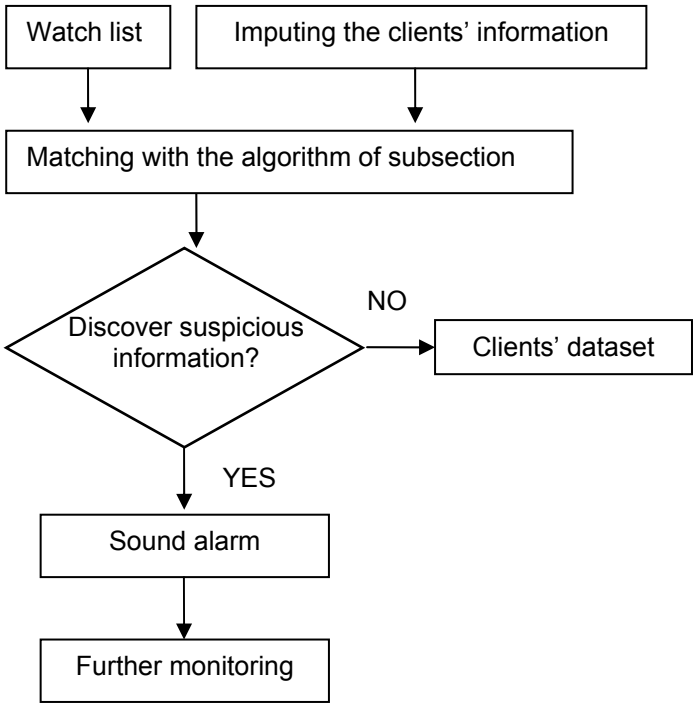

Figure 7. The workflow chart of matching module

6. Conclusion

This paper focuses on the problems that anti-money laundering meets, and deliver related research about the algorithms of matching pattern strings. On basis of BM algorithm, AC algorithm and traditional AC-BM algorithm, we propose an improved algorithm, the matching algorithm of subsection. The experiment shows that application of the algorithm to the matching of watch list of anti-money laundering, could improve the speed of matching significantly. Even the database of the watch list is expanding and the number of the key words is increasing, the speed of the matching algorithm is affected least. These advantages can increase the speed of matching watch list and meet the demand of on-line monitoring of anti-money laundering. In addition, this paper designs a model of the monitoring system of anti-money laundering. Applying the algorithm of the improved AC-BM to the module of matching in the monitoring system will highly improve the efficiency of the whole monitoring system of anti-money laundering.

Acknowledgment

This work is supported by National Social Science Foundation of China Project "Studies on Suspicious Financial Transaction Monitoring System based on Behavior Pattern Recognition" (09BTJ002).

References

[1] RS Boyer, JS Moore. A fast string searching algorithm. *Communications of the ACM*. 1977; 20(10): 762-772.
[2] AV Aho, MJ Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*. 1975; 18(6):333-340.
[3] Jang-Jong Fan, Keh-Yih Su. An Efficient Algorithm for Matching Multiple Patterns. *IEEE Transactions on Knowledge and Data Engineering*. 1993; 5(2): 339-351.
[4] Alba LM. Evolution of Methods of Money Laundering in Latin America. *Journal of Financial Crime*. 2002.

[5] Alvaor E. Monge. An Adpative and Efficien Algorithm for Detecting Approximately Duplicate Database Record. Califonia State University, Janeg. 2000.
[6] Jun Tang, Huan Liang, The building and perfection of watch list of the financial institution. *Xinan finance*. 2011.
[7] Guogen Wan, Zhiguang Qin, Improved AC-BM Algorithm for Matching Multiple Strings. *Journal of University of Electronic Science and Technology of China*. 2006; 35(4): 531-541.
[8] Siwei Zhou, Yong Cai. Mend of AC- BM A lgorithm and Application in Intrusion D etection Technique. *Microcomputer Applications*. 2007; 28(1): 27-31.
[9] Brooks B. Money Laundering and Gambling: The New Zeal and Experience. *Journal of Money Laundering Control*. 2003.
[10] Baker R. CaPitalism's Achiles Heel: Dirty Money and How to Renew the Free-Market System. John Wiley and Sons, 2005.
[11] Lu Jun, Zhang Zhuo, Mo Juan, LV xingfeng. Multi-pattern Matching Methods Based on Numerical Computation. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(3); 1497-1405.
[12] Ye Peixin, Wei Xiujie. Lebeseguetype Inequality for Orthogonal Matching Pursuit for Microcoherent Dictionaries. *TELKOMINIKA Indonesian Journal of Electrical Engineering*. 2013; 11(1); 213-226.
[13] N.Cereone, A,Tsuehiya eds. Special Issue on Learning and Discovery in Knowledge-BasesDatabase. *IEEE Transactions on Knowlddge and Data Engineering*.1993.
[14] Dwyer T. "Harmful" Tax Competition and the Futrue of offshore Financial Centers. *Journal of Money Laundering Control*. 2000; 15(1); 48-69.
[15] Alexander R. The Role of Whistleblowers in the Fight against Economic Crime. *Journal of Financial Crime*. 2004; 12(2); 131-138.
[16] Doudou La Loudouana, Mambobo Bonouliqui Tarare. Data Set Selection. *Journal of Machine Learning Gossip*. 2003; 1: 11-19.
[17] Haddad E. *Load distribution optimization in heterogeneous multiple Processor systems*. Proceedings of the Workshop On Heterogeneous Processing. 1993; 42-47.
[18] Jiawei Han, Mieheline Kamber. Data Mining Concepts and Techniques. Morgan Kaufinann Publisher. 2000.