*Journal of Information Security*, **2012, 3, 11-17**
http://dx.doi.org/10.4236/jis.2012.31002 Published Online January 2012 (http://www.SciRP.org/journal/jis)

# Two Approaches on Implementation of CBR and CRM Technologies to the Spam Filtering Problem

**Rasim Alguliyev, Saadat Nazirova**
Institute of Information Technology, Azerbaijan National Academy of Sciences, Baku, Azerbaijan
Email: rasim@iit.ab.az, sbunyadova@gmail.com

Received September 26, 2011; revised November 3, 2011; accepted November 17, 2011

## ABSTRACT

Recently the number of undesirable messages coming to e-mail has strongly increased. As spam has changeable character the anti-spam systems should be trainable and dynamical. The machine learning technology is successfully applied in a filtration of e-mail from undesirable messages for a long time. In this paper it is offered to apply Case Based Reasoning technology to a spam filtering problem. The possibility of continuous updating of spam templates base on the bases of which new coming spam messages are compared, will raise efficiency of a filtration. Changing a combination of conditions it is possible to construct flexible filtration system adapted for different users or corporations. Also in this paper it is considered the second approach as implementation of CRM technology to spam filtration which is not applied to this area yet.

**Keywords:** E-Mail Spam; Unsolicited Bulk Message; Theory of Precedent; CBR; CRM

## 1. Introduction

The development of Internet has generated many problems the one of which is spam. Spam is undesirable message appearing in e-mail, search engines, chats, forums, IM (instant messaging). The most known and bothered kind of spam is email spam, as e-mail an effective, fast and cheap kind of communication. Almost each computer user has e-mail, and faces spam problem.

For 2010 year Symantec reports that the total amount of spam in mail traffic was 89.1%, and according to Kaspersky Lab annual report the total amount of spam in mail traffic was 90.8% [1,2]. Such a quantity for spam does electronic communication useless, and sometimes not secured. As spam grows very fast, spammers begin to send harmful software, Trojans, malicious content within it. According Symantec annual report for 2010 there has been registered more than 339,600 various viruses, which are hundreds times more than for 2009 [1]. As seen from above diagram (**Figure 1**) the numbers of registered malicious attacks increased in the summer in 2010, so that they were found in approximately 6% of all emails. According to Ferris Research estimations the worldwide cost of spam email in 2009 was roughly 130 billion dollars [3]. All these facts once again urge us to struggle with spam with most effective new methods. As spam changes too quickly (the body, subject, sender's mail and IP addresses changes) and email filtration should be individual (the message noted as spam by one user for another one may be desirable) the effective anti-spam system should be trainable and personified.

## 2. Related Works

Every day computer users receive in their email boxes hundreds of spam messages from new email accounts. Frequently these messages are come with different subject, body automatically generated by robot software. It is almost impossible to filter them with such traditional methods as black-white lists. Applying artificial intelligence methods to the problem of filtering email accounts from unsolicited messages it is possible to raise efficiency of a filtration of spam. Artificial intelligence methods are [4]:

- Convection—machine learning methods based on a formalism and a statistical analysis;
- Computing—methods of iterative working out and the training based on the empirical data;
- Hybrid—methods using convection and computing methods in common.

One of convection methods is Case Based Reasoning (CBR). In this paper it is considered the possibility of CBR method application to spam filtration problem. CBR is a method of reasoning based on precedents. This is a computing model which uses previous events to understand and solve new problems. In some scientific literature CBR meets as "the theory of precedents". The construction of CBR systems begins in 1982 year from

Copyright © 2012 SciRes. 

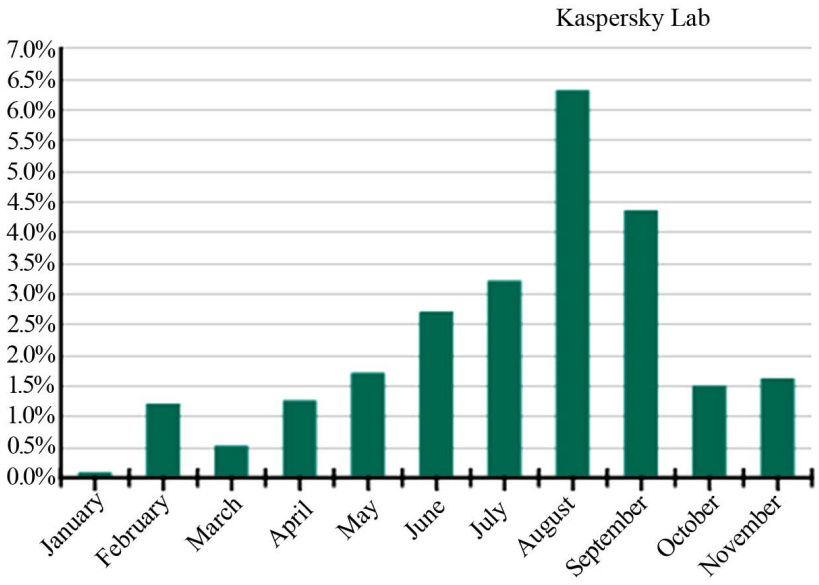


**Figure 1. The percentage of email spam with malicious attachments in 2010, Kaspersky Lab [2].**

Shank's arguments where the notion reminders coordinate the last events with current events to allow generalization and a prediction [5]. Further Kolodner has developed the first CBR system CYRUS expanding Shank's ideas. This CBR system is differing from expert systems. Expert systems store past experience as the generalized rules and objects, whereas CBR systems store past experience as a separate problem, solving episodes [6]. CBR systems try to solve new problem using events from earlier solved problems. So the main princeple of such systems is that one can solve new problems remembering similar events of similar situations.

CBR methods are successfully applied in various areas as classification, diagnostics, forecasting, planning and designing. Independently on a problem for their solving by CBR methods, it is necessary to execute certain sequence of tasks (**Figure 2**).

The basic stages of CBR tasks cycle are considered in such sequence [7]:

1) Choice of the most similar cases of the cases saved up in base.

2) Use of the information and knowledge of this case (set of cases) for the solving new problem.

3) Revision and changes of the solution of the new problem.

4) Preservation of this experience for the solving future problems.

The application of CBR method to spam filtration problem is considered in papers [8-12]. According to these works the classifier based on CBR proves better, than Naive Bayes in spam filtering. Distributed CBR approach can unite in itself spam filtration based on content filtration and collaborative filtration.

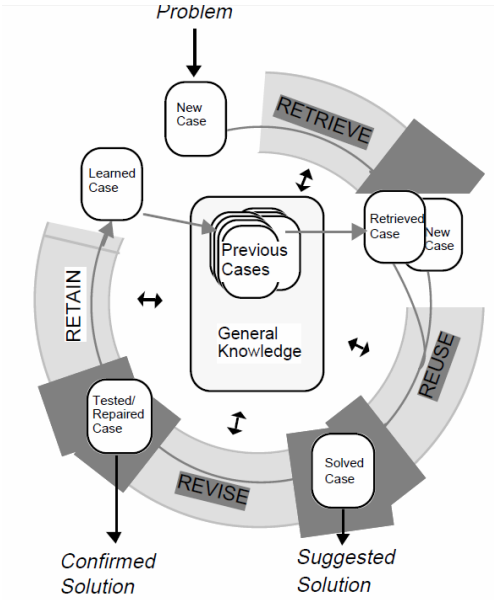


**Figure 2. CBR cycle [6].**

In work [13] there is described the anti-spam system ACABARASE developed on the basis of CBR which after certain training filters spam with less false-positive cases.

The spam filtration model SPAMHUNTING presented in works [14-18] also based on CBR, which applies the disjoint knowledge representation engine. This spam filter able to address the concept drift problem by combining a relevant term identification technique with an evolving sliding window strategy. The idea consists in to identify and remove the obsolete and irrelevant knowledge that has accumulated over to the passage of time.

Copyright © 2012 SciRes. 

Continuous updating technique used in SPAMHUNTING works at two various levels: 1) indexation of the knowledge base; 2) continuous search of its best representation.

Another one machine learning technology is Customer Relation Management (CRM). In spite of the fact that CRM theory has 20 year history, and the expression customer relationship management has been in use since the early 1990s, it did not applied to spam filtering problem yet. But there is great practice of implementation of CRM to different problems [19-25].

## 3. CBR and CRM Implementation Approaches

In this paper it is considered the centralized system of a filtration from unsolicited bulk messages, coordinating all Internet Service Providers (ISP) within country and functioning as collaborative spam filter involving e-mail users of this system and all ISP. This mechanism can be realized at ISP level continuously updating system database with new spam templates, white-black-grey lists. ISP can operatively delegate or delete the data from databases, or transfer them to Network Service Provider (NSP) which provides ISP with Internet traffic (**Figure 3**).

The offered system has the multilayered hierarchical structure consisting of three levels: state, corporate and personal. At each level of multilayered hierarchical system there are server nodes in which there exists database of spam templates. In these databases the spam templates coming from lower level nodes or from the ordinary nodes-user's of the same level are collected.

### 3.1. CBR

For above considered spam filtering problem we define the following cycle of tasks according to CBR theory. At the first step when the user of our multilayered hierarchical spam filtration system reports to the server about new coming spam message system indicates is as *a new case*.

This new case is compared to the previous cases which have been saved up in base of cases—database of spam templates, and the most similar gets out. Combining the chosen case with a new case we get a suggested solution. The combined case is called as *a solved case*.

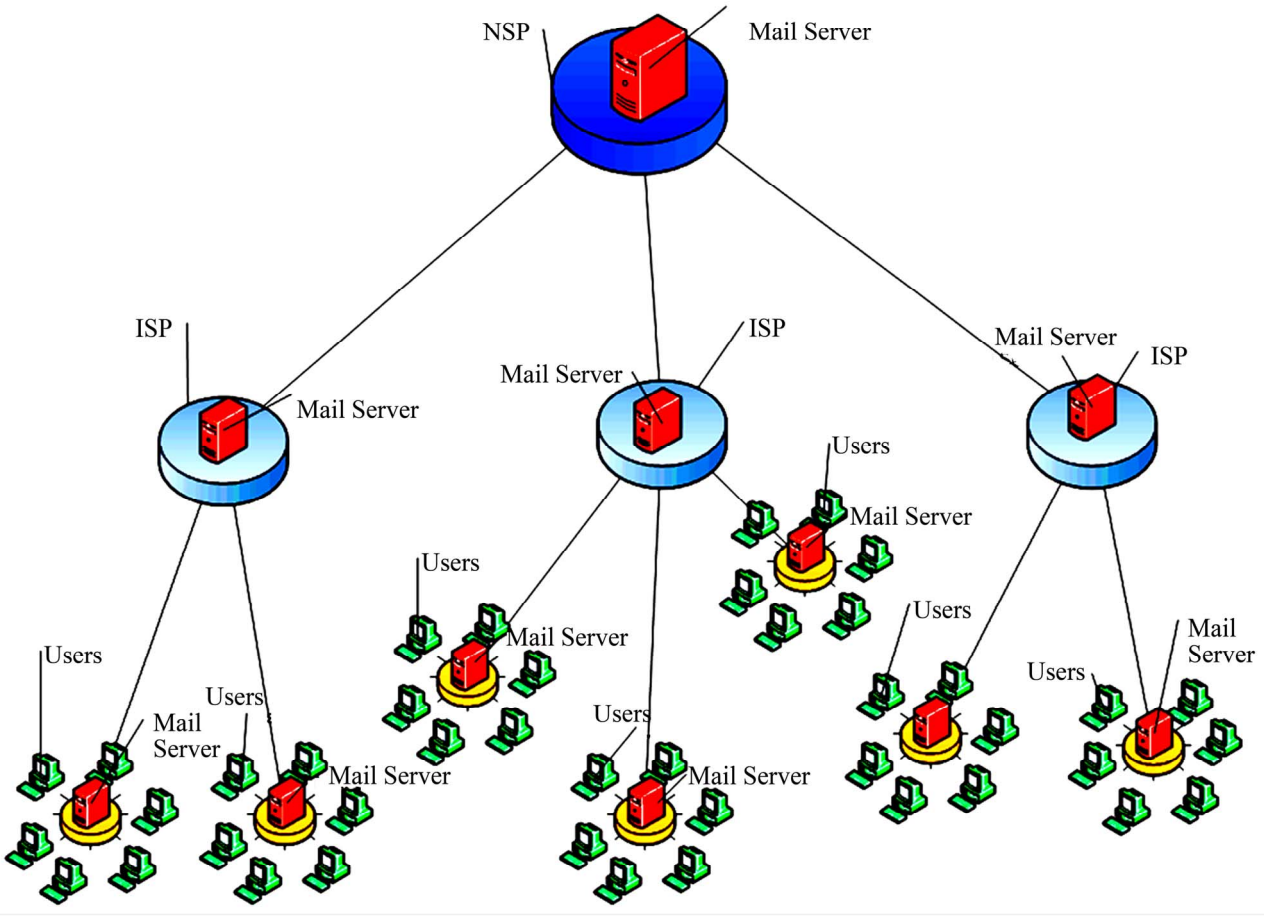


**Figure 3. Architecture of multilayered hierarchical system of spam filtration.**

Copyright © 2012 SciRes. 

Having reconsidered this solution, it is checked on success and applicability to the real world. The solution got at this step is confirmed solution and this case we call as *a tested case*.

In case of a failure the new more suitable case gets out. In a preservation stage the successful case with the corresponding solution registers in base for use in future and is called as *a learned case*.

There should be developed the mathematical methods for solving the tasks belonged to each step. For comparison and extraction of cases one can use the different methods described in works [26,27]. In order to compare new coming message with spam messages collected in database we define the following case parameters—set of characteristics of message:

1) Sender's e-mail address
2) Sender's IP address
3) Subject of message
4) Key words in message body
5) Key phrases in message body
6) Message body

Let's introduce some notations.

$N$ is a number of layers of the offered multilayered system;

$L_i$ is a number of server nodes on $i$ th level, $i = 0, N$;

$K_{l_i}$ is a number of nodes on $i$ th level connected to the server node $l_i$, $i = 0, N$ $l_i = 1, L_i$.

Since the proposed system is assumed dynamic and trainable, and the database of spam templates gradually be updated with new templates, we introduce the parameter of time $t \in T$.

Assume we have $n$ number case parameters, as $x_1, x_2, \cdots, x_n$. In this work $n = 6$.

$s^z_{k_{l_i}}(x_1, x_2, \cdots, x_n, t)$ is $z$th message coming to the node $k_{l_i}$ as spam at a time $t$, with case parameters $x_1, x_2, \cdots, x_n$, where $z \in Z$, $t \in T$, $k_{l_i} = 1, K_{l_i}$. During filtering process each new message, coming to the user $k_{l_i}$ is compared with the spam messages, previously delegated by the same user.

$U_{l_i}(t)$ is a set of spam messages delegated by user $k_{l_i}$ to the server node $l_i$ at a time $t$ until delegation of $z$ th spam message:

$$U_{l_i}(t) = \left\{ s^1_{k_{l_i}}(x_1, x_2, \cdots, x_n, t), \right.$$
$$\left. s^2_{k_{l_i}}(x_1, x_2, \cdots, x_n, t), \cdots, s^{z-1}_{k_{l_i}}(x_1, x_2, \cdots, x_n, t) \right\}$$

where $l_i = 1, L_i$, $i = 0, N$, $t \in T$, $k_{l_i} = 1, K_{l_i}$.

Spam filtration at each level is realized based on the anti-spam policy of that level. Anti-spam policy contains each user's files formed by user's official reports about spam in the received correspondence. On the basis of these official reports-cases spam filtration is realized [28].

The set of legal mails coming to the node $l_i$ is defined by anti-spam policy $P_{l_1}(U_{l_i}(t))$ of the same node:

$$U^*_{l_i}(t) = P_{l_1}(U_{l_i}(t))$$

where $i = 0, N$, $l_i = 1, L_i$, $t \in T$.

Depending on anti-spam policy of each node, comparison can be made by one criterion or by combination of different parameters.

The number of comparisons of two spam messages is

$$N_2 = C^1_n + C^2_n + \cdots + C^n_n = 2^n - 1$$

The number of comparisons of $z$ spam messages is

$$N_z = \frac{z(z-1)}{2}(2^n - 1)$$

In the proposed system it is allowed possibility to withdraw back (restore) the message, previously marked as spam. In this case, the message $s^z_{k_{l_i}}(x_1, x_2, \cdots, x_n, t)$ delegated by the user $k_{l_i}$ as spam at a time $t$ is removed from the set of spam templates $U_{l_i}(t)$. Accordingly, the set of spam templates $U_{l_i}(t+1)$ and the anti-spam policy $P_{l_1}(U_{l_i}(t+1))$ for that level $i = 0, N$ are also changed. The dynamical algorithm of the system will restore the state of a dynamical system in a real time (during the process), using the input information about the system in current discrete time.

In the absence of spam templates no decision is taken for that user. This means that either the user has recently connected to the spam filtration system, or the user is tolerant of spam messages.

## 3.2. CRM

The expression CRM has a variety of meanings. One of them is that CRM is an information industry term for methodologies, software and usually Internet capabilities that help an enterprise manage customer relationships in an organized way [29].

In some papers there have been identified three types of CRM: operational, analytical and collaborative. There are different approaches to these three steps. According to one of them [30]:

- Analytical CRM is responsible for analyzing customers' behavior in terms of sales, marketing or any other service provided. It utilizes data warehouse to extract appropriate data regarding different customers;
- Operational CRM is responsible for automating business processes that are related to customers like marketing and sales etc.;
- Communication/Collaborative CRM as the name implies, is responsible for efficient collaboration/association with the customers through e-mails, fax, phone, SMS or face to face communication.

The graphical interpretation of above steps according

Copyright © 2012 SciRes. 

to Liu & Zhu [31] takes place in **Figure 4**.

Xu & Walton [32] name these steps as main principles of CRM and define them as following:

- Collect information;
- Efficiently usage of collected data;
- Automation of process.

In this paper we consider CRM theory as a management of relation between customers and their choices. By learning relevant information about the customers such as; names, habits, preferences and expectations one-on-one relation can be formed [33]. Learning this information can help to make right decision. Some times during spam filtration process the legal messages indicates as spam and user lost the important mail. Almost in best anti-spam solutions there takes place some percent of false positives. The advantage of using CRM approach is to decrease the number of false positives.

In case of spam filtering problem we consider customer as e-mail user $k_{l_i}$ and choices as messages that indicated by user $k_{l_i}$ as spam $s_{k_{l_i}}^{z}\left(x_1, x_2, \cdots, x_n, t\right)$. Our approach is to use the main idea of CRM theory, that using more information about customer—user, one can increase efficiency of spam filtering. The CRM database containing data, user-profile as preferences, interests, scientific direction, and etc is in the input of our filtration system (**Figure 5**). Processing this profile can automatically manage filtration. Depending on time this profile can be changed by user himself manually or can be organized through automatic analyses of information derived from mails and/or visited Web recourses.

According to the above presented main steps of any CRM system, we can define the following consequence of tasks describing the technology framework of our CRM based spam filtering system (**Figure 6**):

- First one is the construction of analytical CRM system which focuses on data mining tools to gather, analyzes and interprets huge amount of data belonged to users. This data can be derived from e-mail and visited web resources All information belonged to user as his preferences regarding e-mail (which content he like, and which one dislike) and his profile are key points in filtration of his e-mail.

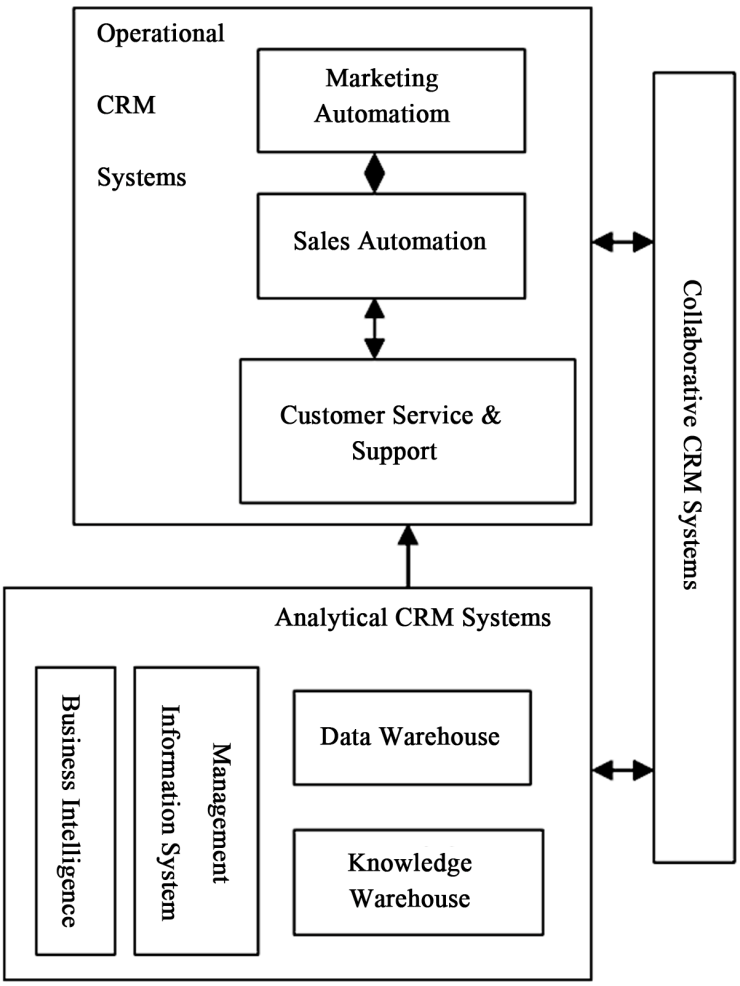


**Figure 4. Technology framework of CRM [31].**

Copyright © 2012 SciRes. 

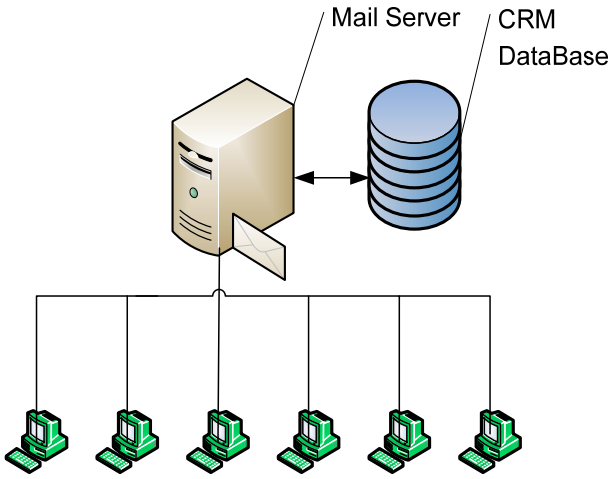


**Figure 5. CRM based spam filtering.**

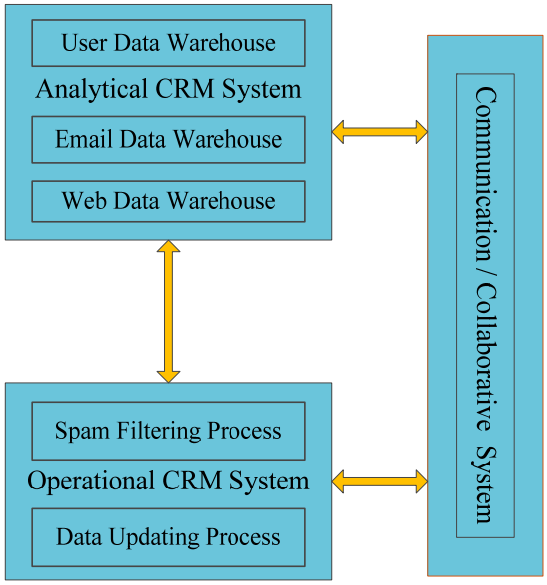


**Figure 6. Technology Framework of CRM based Spam Filtering System.**

- Second step is the construction of operational CRM system. After data collection it should be placed in right place—in CRM based spam filtering system database at the input of the system, also can be assessable to user himself in order to manage this data time by time.
- Third step should be the automated process of filtration. During this process the filtering system can recognize the new coming spam messages, comparing spamness signs of message with corresponding data from spam templates reported by user $k_{l_i}$ and stored in database and also with information from profile.

The efficiency of spam filtration depends on used comparison method and the volume of collected data. So well-trained CRM based spam filtering system will show high efficiency with the less number of false positives.

## 4. Conclusion

In this work it is suggested conception of application of two well-known mathematical apparatus to spam filtering. One of them is CBR technology which is began to apply to spam filtering recently. Another one is CRM technology which is not applied to spam filtering problem yet. These are two machine learning concepts and could be effectively used in spam filtering. As spammers constantly change external signs of spam messages to skip spam filtering systems, there arises a need for adaptive, trainable filtering system. So development of server side personalized e-mail filtering systems that use the learning-based classification algorithms based on CBR and/or CRM technology is a very perspective direction.

## 5. Future Work

Future work will focused on providing methods and experiments to prove the effectiveness of implementation of CBR & CRM technologies onto spam filtration problem.

## REFERENCES

[1] Symantec, “State of Spam and Phishing,” *Annual Report*, 2010 http://www.symantec.com/about/news/release/article.jsp?prid=20101207_01

[2] Kaspersky Security Bulletin, “Spam Evolution 2010,” 2010 http://www.securelist.com/en/analysis/204792163/Kaspersky_Security_Bulletin_Spam_Evolution_2010

[3] Ferris Research, “Cost of Spam is Flattening—Our 2009 Predictions,” 2009. http://www.ferris.com/2009/01/28/cost-of-spam-is-flattening-our-2009-predictions/

[4] E. A. Razdobarina, “Historical Review of Works in Artificial Intelligence,” 2009. http://www.smaut.com/main/public/AiHistoryScool.html

[5] R. Shank, “Dynamic Memory. A Theory of Learning in Computers and People,” Cambridge University Press, New York, 1982.

[6] J. Kolodner, “Case-Based Reasoning,” Magazin Kaufmann, San Mateo, 1993, p. 386.

[7] E. P. Aamodt, “Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches,” *AI Communications*, Vol. 7, No. 1, 1994, pp. 39-59. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.9093&rep=rep1&type=pdf

[8] C. Padraig, N. Niamh, J. D. Sarah, *et al*., “A Case-Based Approach to Spam Filtering that Can Track Concept Drift,” *Proceedings of the ICCBR*03 *Workshop on Long-Lived CBR System*, Trondheim, June 2003.

Copyright © 2012 SciRes. 

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.114.3235&rep=rep1&type=pdf

[9] J. D. Sarah, C. Padraig, T. Alexey, *et al*., "A Case-Based Technique for Tracking Concept Drift in Spam Filtering," *Knowledge Based Systems*, Vol. 18, No. 4-5, 2005, pp. 187-195. doi:10.1016/j.knosys.2004.10.002

[10] J. D. Sarah, C. Padraig and C. Lorcan, "An Assessment of Case-Based Reasoning for Spam Filtering," *Artificial Intelligence Review*, Vol. 24, No. 3-4, 2005, pp. 359-378. doi:10.1007/s10462-005-9006-6

[11] J. D. Sarah, C. Padraig, D. Dónal, *et al*., "Generating Estimates of Classification Confidence for a Case-Based Spam Filter, Case-Based Reasoning Research and Development," *Lecture Notes in Computer Science*, Vol. 3620, 2005, pp. 177-190. doi:10.1007/11536406_16

[12] J. D. Sarah and B. Derek, "Textual Case-Based Reasoning for Spam Filtering: A Comparison of Feature-Based and Feature-Free Approaches", *Artificial intelligence review*, Vol. 26, No. 1-2, 2005, pp.75-87

[13] C. Andres and M. Nunez, "ACABARASE: An Anti-Spam Case-Based Reasoning Systems," *Proceedings of* 3*rd International Conference on IEEE*, *ICONS* 08, New Delhi, 13-18 April 2008, pp. 230-234.

[14] J. R. Mendez, F. Fdez-Riverola, F. Dıaz, *et al*., "Tracking Concept Drift at Feature Selection Stage in SPAM-HUNTING: An Anti-Spam Instance-Based Reasoning System," *Proceedings of the* 8*th European Conference on Case-Based Reasoning*, Fethiye, 4-7 September 2006, pp. 504-518.

[15] J. R. Mendez, C. Gonzalez, D. Glez-Pen, *et al*., "Assessing Classification Accuracy in the Revision Stage of a CBR Spam Filtering System," *Proceedings of the* 7*th International Conference on Case-Based Reasoning System*, Belfast, 13-16 August 2007, pp. 374-288.

[16] F. Fdez-Riverola, E. L. Iglesias, F. Dıaz, *et al.*, "SPAM-HUNTING: An Instance-Based Reasoning System for Spam Labeling and Filtering," *Decision Support Systems*, Vol. 43, No. 3, 2007, pp. 722-736. doi:10.1016/j.dss.2006.11.012

[17] F. Fdez-Riverola, E. L. Iglesias, F. Dıaz, *et al*., "Applying Lazy Learning Algorithms to Tackle Concept Drift in Spam Filtering," *Expert Systems with Applications*, Vol. 33, No. 1, 2007, pp. 36-48. doi:10.1016/j.eswa.2006.04.011

[18] J. R. Mendez, D. Glez-Pena, F. Fdez-Riverola, *et al*., "Managing Irrelevant Knowledge in CBR Models for Unsolicited E-Mail Classification," *Expert Systems with Applications*, Vol. 36, No. 2, 2009, pp. 1601-1614. doi:10.1016/j.eswa.2007.11.037

[19] W. Fang and S. Mao, "Analysis on the Application of CRM in Logistics Enterprises," *Proceedings of International Conference on E-Business and E-Government* (ICEE), Guangzhou, 7-9 May 2010, pp. 3087-3089.

[20] Y. Shen, S. L. Song and S. W. Li, "The Design and Implement of CRM Data Mining System for Medium-Small Enterprises Based on Weka," *Proceedings of International forum on Information Technology and Applications IFITA*'09, Vol. 2, 2009, pp. 596-599

[21] B. Liu, G. Zhao and Y. Su, "Research of University Employment Management System Based on CRM," *International Conference on Intelligent Computation Technology and Automation*, Vol. 2, 2010, pp. 1059-1064. doi:10.1109/ICICTA.2010.48

[22] L. Decai and L. Yue, "Research on Application of CRM in Fields of Network Marketing: Illustrated by the Case of Maibaobao Aveyond," *International Conference on Management Science and Electronic Commerce*, *Artificial Intelligence*, Zhengzhou, 8-10 August 2011, pp. 4713-4716.

[23] K. Xiong, "Study on Application of CRM in E-Government Based on Public Service," *Proceedings of International Conference on Electric Information and Control Engineering*, Wuhan, 15-17 April 2011, pp. 4511-4514. doi:10.1109/ICEICE.2011.5777481

[24] B. Liu, G. Zhao and Y. Su, "Employment Management System Based on CRM," *Proceeding of International Conference on Intelligent Computation Technology and Automation* (*ICICTA*), 11-12 May 2010, pp. 1059-1064. doi:10.1109/ICICTA.2010.48

[25] W. Olof, S. Christer and S. Hakan, "Trends, Topics and Under-Researched Areas in CRM Research," *International Journal of Public Information Systems*, Vol. 3, 2009, pp. 192-208. http://www.ijpis.net/issues/no3_2009/ijpis_no3_2009_p3.pdf

[26] R. M. Alguliev, R. M. Aliguliyev and S. A. Nazirova, "Classification of Textual E-Mail Spam Using Data Mining Techniques," *Applied Computational Intelligence and Soft Computing*, 2011. www.hindawi.com/journals/acisc/aip/416308.pdf

[27] S. A. Nazirova, "Mechanism of Classification of Text Spam Messages Collected in Spam Pattern Bases," *Proceedings of* 3*rd International Conference on Problems of Cybernetics and Informatics*, Vol. 2, 2010, pp. 206-209.

[28] R. M. Alguliev and S. A. Nazirova, "Mechanism of Forming and Realization of Anti-Spam Policy," *Telecommunications*, Vol. 12, 2009, pp. 38-43.

[29] B. Francis, "Customer Relationship Management: Concepts and Technologies," Elsevier Ltd., New York, 2009, p. 500.

[30] Basics of CRM, September 2006. http://www.Advancevoip.com/whitepapers/Basics%20of%20CRM.pdf

[31] C. N. Liu and X. W. Zhu, "A Study on CRM Technology Implementation and Application Practices," *Proceedings of International Conference on Computational Intelligence and Natural Computing*, June 2009, pp. 367-370. doi:10.1109/CINC.2009.120

[32] M. Xu and J. Walton, "Gaining Customer Knowledge through Analytical CRM Industrial Management & Data Systems," *Emerald, MCB Limited*, Vol. 105, No. 7, 2005, pp. 955-971.

[33] J. Berfenfeldt, "Customer Relationship Management," Master's Thesis, 2010, p. 104.

Copyright © 2012 SciRes. 