International Journal of Security and Its Applications
Vol.9, No.10 (2015), pp.341-352
http://dx.doi.org/10.14257/ijsia.2015.9.10.31

Spam Filtering based on Knowledge Transfer Learning

Xing Wang, Bin-Xing Fang, Hui He, Hong-Li Zhang

Department of Computer Science and Technology, Harbin Institute of Technology,
Heilongjiang, P.R.China, 150001
yeahwx@gmail.com; {bxfang, hehui, zhl}@pact518.hit.edu.cn

Abstract

Spam is a serious problem not only the number of floods but also more and more volatile type. It has caused a great impact on people's daily lives. Especially fraud spam, even cause huge losses to companies or individuals possibility. Therefore, it is imminent to filter spam efficiently. Existing spam filtering mechanism is mainly based on the character and content of the spam message. However, once the spam filter uses in other user's mailbox, the existing spam filtering techniques can not be well adapted. In this paper, we propose the adaptive spam filtering method for the above shortcomings. The method uses the unlabeled spam data that from other user or domain to enhance the adaptive and opposability of the anti-spam system. We use the transfer learning model to build the spam filtering system. A transfer learning model can use the untagged data, and migrate knowledge between different filter model, and improve the active collaboration of the filter.

Keywords: *Email spam, Transfer Learning, Markov logic network.*

1. Introduction

In recent years, there is a growing advertising and pornographic mail in the mailbox, which not an initiative subscription or sending from acquaintances. The spam email is overwhelming and has brought manifold grave loss for legal email users.

With the proliferation of spam, the anti-spam technology is rapidly developed. Existing spam filtering system can filter out most of spam by some obvious characteristics. Currently, there are two kinds of anti-spam efficient methods. The first is a heuristic filtering technology, which filter the spam depend on the source. The method classifies the spam by the feature of origin, which is a fixed server or domain name. Because of the method block emails before the submission, the network resources, and bandwidth can be well protected. The second is content-based method. The method first parses the body of the email to obtain content features. Then, the method analysis and matching these content features to determine whether the email is a spam. The other important spam filtering method described as follows.

(1) Keywords filtering technology: the keywords filtering method use the statistical analysis to obtain high-frequency words. The high-frequency words are usable in spam-word library creating. If the mail feature match with the spam-word library, the spam email can be well classified.

(2) Black and white list filtering technology: blacklist and whitelist technology is a very efficient filter. Blacklist contains the IP address that is sent from the spammers. The whitelist contains the IP address that is proved to be a trusted sender.

(3) Rule-based filtering technology: the rule-based spam filtering techniques use some local regularity features to implement the judgment of the message class. The local regularity feature includes the using of the word, phrase, location, size, accessories. The rule-based spam filtering often using a neural network algorithm to generate a set of rules, each rule corresponding scores. If the rules appear in ordinary mail, the rules

ISSN: 1738-9976 IJSIA

get a negative number. If the rules appear in spam email, the rules get a positive number.

(4) Accumulate the score of all rules that match the email. If the total score of all rules is greater than the threshold that set by the system, an email is considered as spam.

(5) Probability and Statistics filter technology: anti-spam algorithm based on Bayesian learning algorithm is most effectively. The Naive Bayesian learning algorithm combined with some of the "noise canceling" technology (for example, using HMM Chinese semantic disambiguation, weight smoothing) has a high self-learning ability and anti-jamming capability. By extracting and analyzing the number of words in the training spam mail, the Bayesian method can classify the category of the email.

Many commercial spam filtering systems have been developed. Why our inbox is still full of spam? There are two main reasons.

- The attacker launched attacks based on the weaknesses of the spam filter. However, the spam filter adaptive capacity is far from sufficient to respond the attacks, the untagged data cannot be fully utilized
- The unlabeled user data has not yet been fully utilized in classification spam.
- For different user's mailbox, the existing spam filtering techniques cannot be well adapted. There also exists a concept drift during the classification spam.

In this paper, we propose the adaptive spam filtering method for the above shortcomings. The method uses the unlabeled spam data that from other user or domain to enhance the adaptive and opposability of the anti-spam system. We use the transfer learning model to build the spam filtering system. A transfer learning model can use the untagged data, and migrate knowledge between different filter model, and improve the active collaboration of the filter.

2. Related Work

The more famous spam dataset is the LingSpam. The data set is collected from a linguist-mail list. An important set of data recently released TREC (Text Retrieval Conference), which is a data set for online assessment, classification of the data set can be iterated sorting mail and receive feedback. TREC data set is applied to each user classifier closer to the real environment.

The spam filtering contest includes TREC 2006 ECML / PKDD 2006 and CEAS2007 meeting, which generated a lot of efficient methods. Semi-Supervised Support Vectors Machines and Application to Spam Filtering [1] and a semi-supervised Spam mail detector [2] both use semi-supervised method to filter spam. A Two-Pass Statistical Approach for Automatic the Personalized Spam Filtering [2] proposed automatic personalized spam filter which does not require user feedback. The algorithm builds a statistical model from the training data and updates on the label two model with personal user data.

Present literatures focus on the theoretical model of spam filtering. For spam filtering cooperation, the general practice is to share knowledge between P2P users [3, 4] or to collect spam report on the mail server. However, the sharing will lead to user privacy issues, [5] suggested a privacy-protected P2P spam filtering system. Guoqing proposed a multi-agent system. Mail in local is classified as spam [6]. Only mail cannot be classified in local will be judged by the other agent work together. Garg suggest exchanging the training the classifier, therefore significantly reduces the amount of data transferred [7].

3. Transfer Learning using Markov Logic Networks

In this section, we first introduce the Markov logic networks, and then we describe a method for transfer learning.

3.1 Markov Logic Networks

A variety of machine learning applications requires the ability to learn from and the reason about noisy, or uncertain, multi-relational data. This requirement has motivated the fields of SRL (statistical relational learning) and multi-relational data mining in which MLN is the most powerful theory we are focusing on.

$$P(X = x) = \frac{1}{Z}\exp(\sum_i \omega_i n_i(x))$$

MLN is a probabilistic extension of first-order logic [8]. In MLN, every logic formula is associated with a non-negative real-valued weight. Let be the set of all propositions describing a word. The probability distribution over possible worlds is given by where Z is normalization constant, is the weight of the formula, and is the number of satisfied groundings. MLN enables us to compact represent complex models in non-i.id domains. Conditional probabilities can be computed using MC-SAT, Lazy inference [9, 10], and lifted inference [11].

The weight and structure of MLN can be learned by using pseudo-likelihood training with L-BFGS, discriminative weight learning [12], Bottom-Up [13], LHL [14], LSM [15], etc. An implementation of learning and inference algorithms is publicly available in the Alchemy Package [16].

3.2 Transfer Learning using MLN

Some researchers have addressed the problem of transfer knowledge from a source model to a target domain using the Markov logic network. The TAMAR algorithm first autonomously maps the predicates in source MLN to the target domain and then revises its structure to improve its performance [12]. Then, the SR2LR algorithm was proposed for a setting in which data about only one entity in the target domain is available [13]. The algorithm evaluates possible of source-to-target predicate correspondences based on short-range clauses in order to transfer the knowledge captured in long-range clauses.

The most similar approach to ours in the literature is DTM algorithm [14]. DTM discovers structural regularities by using second-order Markov logic, and explicit, domain-independent knowledge that included broadly useful properties of predicates, like various forms of homophile.

4. System Framework

In this section, we first introduce the knowledge transfer framework for spam filtering. We then describe the predicate design of MLN. Finally, we describe the structure and transfer learning of MLN.

4.2 Knowledge Transfer Framework for Spam Filtering

We propose the adaptive spam filtering solution as shown in figure 1. The solution can take advantage of the unlabeled data for cross-domain adaptive migration. This scheme is based on Markov logic network, and along with the characteristics of first-order logic. The filtered structure and weights have a very high self-adaptability.

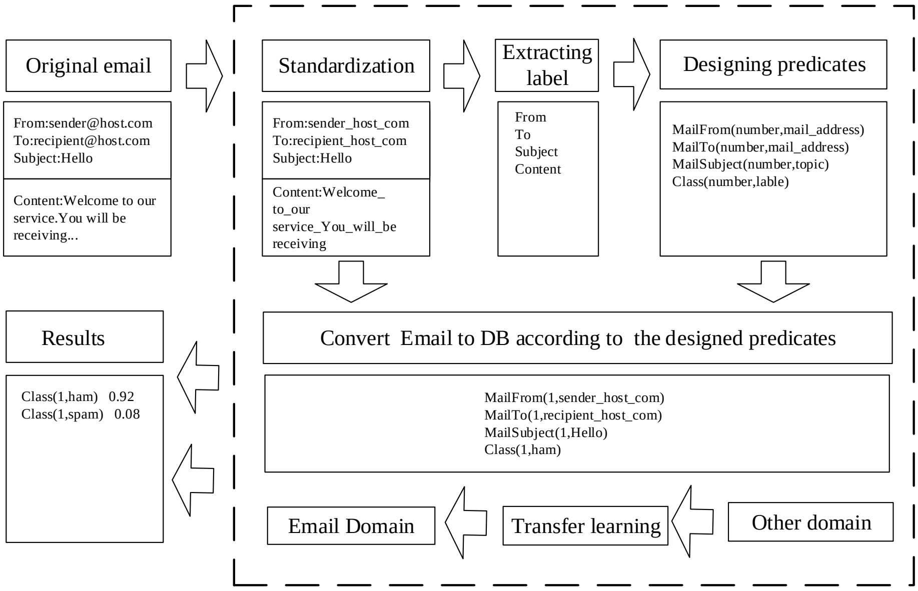


Figure 1. System Framework

4.2 Predicate Design of MLN

According to the characteristics of the spam, we create some suitable predicate. The predicate table is shown in Table 1.

Table 1. Predicates design

Predicates	Meaning
MailDate(number,time)	Date of the Email
Class(number,label)	Label of the Email (ham or spam)
MailFrom(number,mail_address)	Sender of the Email
MailTo(number,mail_address)	Receiver of the Email
MailSubject(number,topic)	Topic of the Email
MailReturnPath(number,mail_address)	Path of the return
MailReceivedFromBy(number, domain_address,domain_address)	Historical transmission
MailMessageID(number,domain_address)	The Message ID includes time to date, the only logo and DNS
FromIP(domain_address,ip)	The sending server's IP address
MailContentType(number,type)	Email type includes text and web forms

Figure 2 is a simple spam MLNs. A represents a spam, and B, C, D, F, representative of mail address. E represents a message sent date. Based on designed predicate, we preprocess the e-mail header information and then create a Markov logic network.

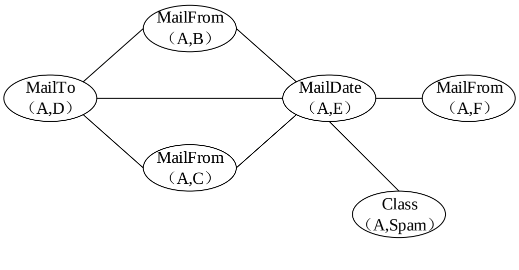


Figure 2. The MLN of spam Email

4.3 Structure Learning of MLN

We use a probability algorithm that Kok and Domingos proposed to learn the structure of MLN. Compared to other methods, Top-down structure learning algorithm (TDSL) can produce more accurate sentence collection. The algorithm learns or corrects a MLN sentence each time. The initial structure is a spatial network or an existing knowledge base. In any case, increase all single clause to the Markov logic network is useful.

(1) Evaluation

As an evaluation measure, the pseudo-likelihood tends to give high-tuple predicate

$$\log P_w^*(X=x)=\sum_{r\in R} c_r \sum_{k=1}^{g_r} \log P_w(X_{r,k}=x_{r,k} \mid MB_x(X_{r,k}))$$

overweight. Therefore, we define the weighted pseudo-likelihood (WPLL).
where R is a first order of a collection of atoms, g_r is a first-order atoms r corresponding to the atom number of closure, $x_{r,k}$ is a true value of k^{th} of atoms r. The selection of the atomic weight C_r is depend on the purpose of the user. By default, we can directly set $C_r = 1/g_r$. The effect of weight on all first-order predicate is same. If the predicate is not important (for example, it always be a part of the evidence predicate); we set its weight to 0. TDSL use $\exp\{-a\sum_{i=1}^{F} d_i\}$ to avoid overfitting, where d_i is the number of words. The current version of the clause with the first clause of the trial. This method is also used in learning Bayesian network.

When evaluates candidate clause by using WPLL, all the optimal weights (maximum WPLL) of clause have to be calculated. The process of numerical optimization may include thousands of times calculation. It will make the algorithms too slow in practical. TDSL uses current weight to initialized L-BFGS and relax the convergence limit to avoid this bottleneck. Second-order quadratic convergence method such as L-BFGS reach the optimal value very fast.

(2) Operating

The errors caused by human experts can be corrected by adding or removing text operation in the clause, which is the underlying operating of TDSL. When adding a text to the clause, TDSL will consider all possible ways. Existing text can share the text variables. To control the size of the search box feel, TDSL restricts the number of distinct variables in a clause. TDSL just removes the text from the original handwritten clause, and we consider removing a text only in the condition that it can leave a shared variable.

(3) Search

There are two alternative search strategies, one faster and another more complete. The first method is to add a clause to MLN each time and using directional search to find the best clause. TDSL adds or deletes each legitimate in each clause text to ensure that the clause is the best, and repeating the process until there is no new clause to improve WPLL. The selected clause has highest WPLL in each search iteration.

The second method is to add k clauses to MLN at the same time. Different from the orientation the search that increases clause of any length, this method try to add all the length of the "good" clause in front of any length clause. The method is also called the first shortest (Shortest-first) search. Shortest-first search is typically higher than the cost of directed search, but it often produces smaller and more precise MLN.

4.4 Transfer Learning of MLN

The first step is to find the best mapping that map the source MLN to the target MLN. The quality of the mapping can be measured by the WPLL score of target MLN. Limited number of predicate may receive an exponential mapping. It will lead to a heavy computational effort. Therefore, TAMAR finds the best local mapping of each independent sentence.

The TAMAR conducts a detailed search for all legitimate mapping space. The legitimate mapping should satisfy the following condition: Each source domain predicate will map to a target domain predicate or an empty predicate. If two predicates have the same dimensions and the type of their predicates is compatible, then the two predicate is compatible. For all legitimate mapping, a type of the source domain of at most consistent to a type of the target domain. TAMAR calculates the WPLL of the MLN that only contains the new generation clause for establishing a proper and legitimate mapping.

Since the predicate mapping algorithm sometimes make an empty map, the source and target domains differ in structure. The mapped structure need to be revised for better fitting the data. The revision algorithm has three steps that are similar to the FORTE algorithm.

(1) Self-Diagnosis: this step force the search only focus on the imprecise part of MLN. The step also checks the source MLN to decide whether the clause should be growth, shorten or delete.
(2) Structure update: this step cut the long clause and extend the short clause. This step use a beam search and WPLL to find a suitable candidate clause.
(3) Finding new clause: this step finds the new clause in the target domain by using the RPF algorithm.

5. Experiments and Results

In this section, we first introduce the experimental environment. We then describe the data preprocessing. Finally, we describe the structure and transfer learning of MLN.

5.1 Experimental Environment

The realization of the proposed model uses Alchemy package. Alchemy is based on Markov logic package, which provides a series of statistical relational learning and probabilistic logic inference algorithm. Alchemy can easily to develop a very wide range of artificial intelligence applications, including classification, link prediction, entity resolution, and social network modeling and information extraction. The module of our design is also added to the Alchemy package for the self-adaptive spam filtering.

5.2 Data Preprocessing

The original spam data set and the organized predicate in the form data set of other domains. Therefore, the pre-processing of the data can be divided into the initial spam handling and the data set processing of outside the field.

In order to deal with the text accurately and efficiently, we use flex and bison to process data sets. Flex and Bison used to design a compiler in the earliest time, and responsible for lexical analysis and parsing. In this experiment, the message format is standard and has a large amount of data, flex and bison could meet our requirements in terms of performance.

After finishing predicate design, we use flex and bison to read the original spam data set, and convert it into a DB form accepted by Markov logic network. Sample results are shown in Table 1.

Table 1. DB sample

MailReturnPath(11,Ke_ccert_edu_cn)
MailReceivedFromBy(11,Fromsea_net_edu_cn,Byhome_ccert_edu_cn)
FromIP(Fromsea_net_edu_cn,IP202_112_5_66)
MailReceivedFromBy(11,Fromtu203027_tsinghua_edu_cn,Bysea_net_edu_cn)
FromIP(Fromtu203021_tsinghua_edu_cn,IP166_111_203_27)
MailDate(11,T2004_Dec_1_14_47_08)
MailMessageID(11,Msea_net_edu_cn)
MailSubject(11,SfEgztLSqsXituDJ2aOxu9zA)
MailTo(11,Ke_ccert_edu_cn)
MailFrom(11,Ccertstaff_ccert_edu_cn)
MailContentType(11,Text_plain)
Class(11,Ham)

For other areas of the data set, we need to learn the corresponding Markov logic network which include the structure and parameter learning. The process may produce the following clause in the form:

Table 2. Markov Logic Network Example Clause

0.127906	MailReturnPath(a1,a2) v MailFrom(a1,a2) v Class(a1,a3)
0.195805	MailFrom(a1,a3) v MailTo(a2,a3) v Class(a1,a4) v a1 = a2

At the beginning of the sentence, the floating point numbers representing the weight of this clause. The greater the weight, the greater impact on the inference results. In Markov logic network, a clause is a clique that may have a positive effect or a negative effect. Therefore, the weight and a simple probability is different. The range of weight is not confined in 0 to 1. A larger absolute value representative a greater impact. The sign of the weights represents the direction of the effect.

5.3 Predicate Mapping

We need to build the predicate mapping for adaptive spam. The purpose of this step is to establish a relationship between spam domain and other domains. We use three datasets from other domains to evaluate the algorithms that are described in this paper. These datasets are publicly available at http://alchemy.cs.washington.edu. The details are shown in Table 3.

Table 3. Data Set

Data Set	Consts	Types	Preds	True Gliterals	Total Gliterals
IMDB	316	4	10	1540	32615
UW-CSE	1323	9	15	2673	678899
WebKB	1700	3	6	2065	688193

IMDB dataset [17] created from the IMDB.com database, describes a movie domain. It contains relationships among movies, actors and directors. For instance, WorkedIn(person,movie), Actor(person), etc. The data is split into five disjoint folds. **UW-CSE dataset** [18] describes the Department of computer Science and Engineering at the University of Washington. Its predicates represent students, faculty, and their relationships. The data is split into five folds. **WebKB dataset** contains webpages from four universities labeled according to the entity they describe. The data from each university is treated as a separate fold.

Table 4. Predicate Mapping Table

IMDB	WebKB	UW-CSE	CDSCE
workedUnder	samePerson	samePerson	MailDate
movie	courseTA	null	MailSubject
director	faculty	professor	null
actor	student	student	Class
null	courseProf	publication	null
null	project	projectMember	FromIP
gender	null	sameQuarter	null
sameMovie	null	sameLevel	null
samePerson	null	taughtBy	MailReceivedFrom By
sameGender	null	yearsInProgram	null
genre	null	samePosition	null
sameGender	null	introCourse	null

As seen in Table 4, the predicate mapping can be built between the predicates or empty. It ensures that no interaction between the predicate mapping. Using predicate mapping results and MLN from other domain, we can transfer knowledge and get spam MLN. The results obtained are as follows:

Table 5. Spam Filter based on Markov Logic Network

0.741011 MailContentType(a1,a2)

2.09276 MailDate(a1,a2)

2.82358 MailReceivedFromBy(a1,a2,a3)

1.00966 Class(a1,a2)

2.52038 MailMessageID(a1,a2)

0.111646 MailReceivedFromBy(a1,a2,a2)

1.92116 MailSubject(a1,a2)

0.0214066 MailFrom(a1,a2)

0.112046 MailTo(a1,a2)

2.52038 FromIP(a1,a2)

0.0438574 MailReturnPath(a1,a2)

0.127906 MailReturnPath(a1,a2) v MailFrom(a1,a2) v Class(a1,a3) v !Class(a1,a4)

0.128654 MailFrom(a1,a2) v Class(a1,a3)

0.195805 MailDate(a1,a5) v MailFrom(a1,a2) v MailTo(a3,a2) v Class(a1,a4) vv !Class(a3,a4)

0.196927 MailReceivedFromBy(a1,a2,a3) v MailFrom(a1,a2) v MailTo(a3,a2) vClass(a1,a4) v Class(a3,a5) v !Class(a1,a5)

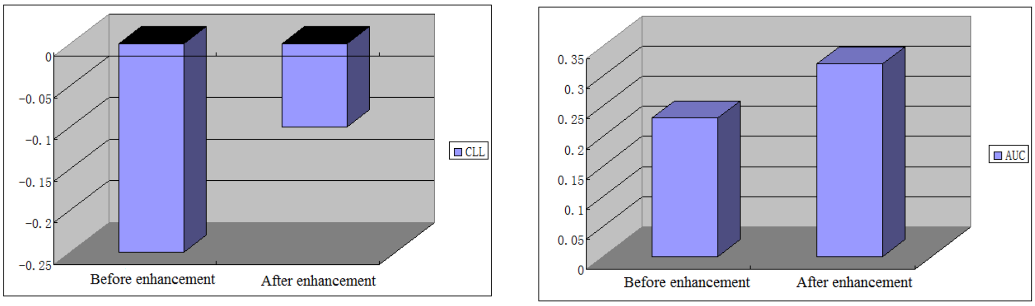
0.195805 MailReturnPath(a1,a3) v MailFrom(a1,a3) v MailTo(a2,a3) v Class(a1,a4)v !Class(a1,a5) v !Class(a2,a4) v a1 = a2

5.4 Spam Detection Results

Using the method mentioned above, we can get the Markov logic network forms of adaptive spam filters. In order to make the contrast results clear and intuitive, we use

Condition Logarithmic Likelihood(CLL) and Area Under the Curve of precision rate and recall rate (AUC) as the measured indicator. The CLL can directly measure the quality of the probability distribution which is estimated by optimizing approximation method. Precision rate and the recall rate are often contradictory. The two values can not be considered separately. Therefore, many works use AUC value.

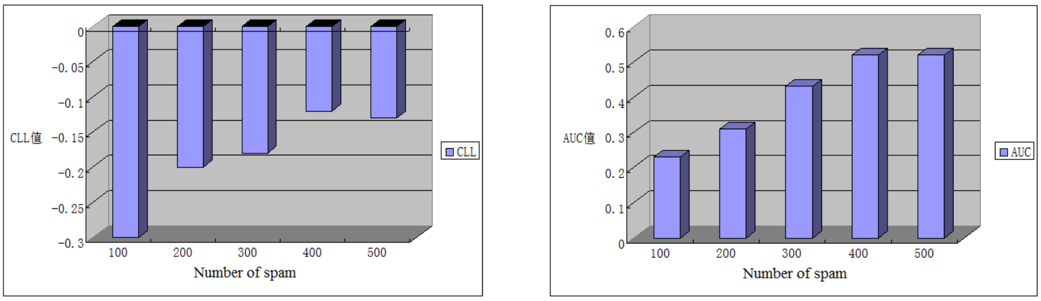
Spam filter based on Markov logic network only have two forms: one is obtained directly from data intensive training. Another is the results of using the existing structure of other domains and a small amount of data of the target domain. On account of these two cases, we complete two experiments are as follows:



(a) CLL

(b) AUC

Figure 3. The CLL and AUC of Expriment Results



(a) Varying the number of spam

(b) Varying the number of spam

Figure 4. The CLL and AUC of Experiment Results when Varying the Number of Spam

Figure 4 reflects the filter obtained by the adaptive migration learning has a better results than the filter obtained by the structure learning no matter which indicator is considered. Figure 5 is comparison results of based on Markov logic network filter and whitelist-based and rule-based filter.

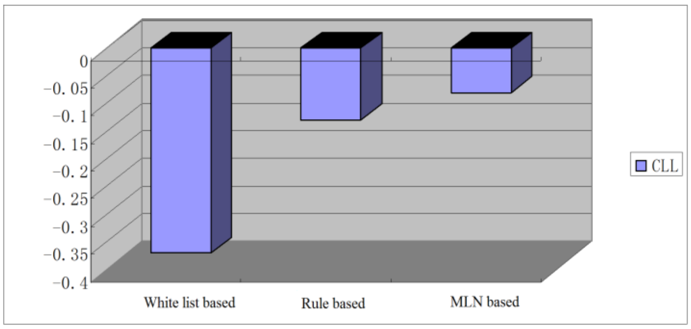


Figure 5. The CLL of Experiment Results

Analysis the above figure, we found that the accuracy rate of the filter and the number of training data of the source domain have a positive correlation. This result is consistent with what we expected. When the spam domain of the training set is infinite, the structure of the source domain hardly has any influence on the filter, the limiting case is the filter trained by using spam datasets. Conversely, if migration spam training set is a small one, the filter structure is the same structure as the source domain.

Table 6. Experiment Result

Predicate	**Single-task**				**Multi-task**			
	10	100	1000	CLL	10	100	1000	CLL
MailReturnPath	0.00654	0.0557	0.0524	-1.71	0.0195	0.0282	0.0284	-2.68
MailReceivedFromBy	0.00004	0.0000	0.0000	-2.91	0.0000	0.0000	0.0000	-1.05
FromIP	0	0	0	0	0	0	0	-2.70
MailDate	0	0	0	-2.17	0	0	0	-2.10
MailFrom	0.01498	0.0741	0.0918	-2.75	0.0272	0.0335	0.0338	-2.71
MailTo	0	0	0	-2.01	0	0	0	-2.66
MailSubject	0	0	0	-2.13	0	0	0	-1.97
MailContentType	0	0.0266	0.0266	-1.13	0.0173	0.0173	0.0423	-0.68
Class	0.19333	0.1933	0.1933	-4.40	0.2944	0.3233	0.3578	-3.29
MailMessageID	0.00420	0.0044	0.0044	0	0.0019	0.0019	0.0019	-2.72
AVG	0.04382	0.0590	0.0614		0.0600	0.0674	0.0774	

Table 6 is the calculation values of AUC and CLL of 100 spam. AUC results of precision are 10,100,1000. The higher the accuracy, the more accurate the results. It can be seen from the table that we have better results of multi-task learning than a single task for the vast majority of the predicate.

6. Conclusion and Future Work

This paper first introduces the research context to give a comprehensive background knowledge. Then describes primary e-mail filtering technology. On this basis, put forward to an enhanced spam filtering. The contributions of this paper are as follows:

1) Create spam filter model based on Markov logic network. Make a detailed presentation about some key technologies such as Data sets transform predicate mapping, migration learning, and inference learning.

2) Build an actual spam filter and test it by using existing data sets. After testing large amounts of data, the spam filter based on Markov logic network structures has a better filtering effect. When the training set and a test set are different greatly, the filter will still show good performance.

Aiming at the present situation of spam filter, we put forward to the Markov logic network-based spam filtering, despite getting some preliminary results; there is much more room for improvement in the future. For example, we can use the reinforce learning in our experiment so that spam filtering system can do self-improvement and self-renewal when a miscarriage occur. So as to achieve better results.

Acknowledgements

This work is partially supported by the High-Tech Research and Development Plan of China (No. 2012AA012506); the National Natural Science Foundation of China (Grant No. 61173145, 61472108). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers.

References

[1] Zien, A., Semi-Supervised Support Vector Machines and Application to Spam Filtering. ECML Discovery Challenge, Berlin, Germany. **(2006)**.

[2] Wei-Lun, T. A Personalized Spam Filtering Approach Utilizing Two Separately Trained Filters. Web Intelligence and Intelligent Agent Technology, **(2008)**.

[3] Lazzari, L., M. Mari, and A. Poggi, CAFE - Collaborative Agents for Filtering E-mails, in Proceedings of the 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise. **(2005)**, IEEE Computer Society. p. 356-361.

[4] Zhou, F., et al., Approximate object location and spam filtering on peer-to-peer systems, in Proceedings of the ACM/IFIP/USENIX 2003 International Conference on Middleware. **(2003)**, Springer-Verlag New York, Inc.: Rio de Janeiro, Brazil. p. 1-20.

[5] Damiani, E., et al. P2P-based collaborative spam detection and filtering. in Peer-to-Peer Computing, 2004. Proceedings. Proceedings. Fourth International Conference on. **(2004)**.

[6] Mo, G., et al., Multi-agent Interaction Based Collaborative P2P System for Fighting Spam, in Proceedings of the IEEE/WIC/ACM international conference on Intelligent Agent Technology. **(2006)**, IEEE Computer Society. p. 428-431.

[7] Garg, A., R. Battiti, and R.G. Cascella, "May I borrow Your Filter?" Exchanging Filters to Combat Spam in a Community, in Proceedings of the 20th International Conference on Advanced Information Networking and Applications - Volume 02. **(2006)**, IEEE Computer Society. p. 489-493.

[8] Richardson, M. and P. Domingos. Markov logic networks. 2006: Kluwer Academic Publishers.

[9] Singla, P. and P. Donaingos. Memory-efficient inference in relational domains. AAAI, **(2006)**. Boston, MA, United states.

[10] Poon, H., P. Domingos, and M. Sumner. A general method for reducing the complexity of relational inference and its application to MCMC. in 23rd AAAI Conference on Artificial Intelligence, **(2008).** Chicago, IL, United states.

[11] Singla, P. and P. Domingos. Lifted first-order belief propagation. in AAAI. **(2008)**. Chicago, IL, United states: American Association for Artificial Intelligence.

[12] Biba, M., S. Ferilli, and F. Esposito. Discriminative structure learning of Markov logic networks. in ILP. **(2008)**. Prague, Czech republic: Springer Verlag.

[13] Mihalkova, L. and R.J. Mooney, Bottom-up learning of Markov logic network structure, in ICML. **(2007)**, ACM: Corvalis, Oregon.

[14] Kok, S. and P. Domingos. Learning Markov logic network structure via hypergraph lifting. in ICML. **(2009)**. Montreal, QC, Canada: Omnipress.

[15] Kok, S. and P. Domingos .Learning Markov logic networks using structural motifs. ICML, 551-558.**(2009)**

[16] S. Kok, P.S., M. Richardson, P. Domingos, M. Sumner, H Poon, and D. Lowd, The Alchemy system for statistical relational AI. Dept. of CSE, Univ. of Washington, **(2007)**.

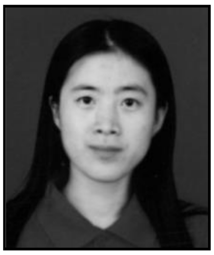
[17] H. T. Mihalkova, Lilyana and R. J. Mooney. Mapping and revising markov logic networks for transfer learning. In Proceedings of the National Conference on Artificial Intelligence, **(2007)**; Vancouver, BC, Canada.

[18] M. Richardson and P. Domingos. Markov logic networks. Machine Learning. 62, 1 **(2006)**

Authors

Xing Wang, Xing Wang received the B.S. and M.S. degree in computer science from Northwest University, XiAn. China. He is now working towards his Ph.D. degree in computer science at Harbin Institute of Technology. His research interests include computer network, machine learning, and public opinion.

Bin-Xing Fang, Bin-Xing Fang received his M.S. and Ph.D. degrees in computer science from the Tsinghua University and Harbin Institute of Technology of China in 1984 and 1989 respectively. He is currently a member of Chinese Academy of Engineering. His current research interests include information security, information retrieval, and distributed systems.



Hui He, Hui He received the B.S., M.S. and Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China. Since September 1999, she has been with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, where she became an Associate Professor in October 2007. Her research interests include network computing, network security.

Hong-Li Zhang, Hong-Li Zhang received her M.S. and Ph.D. in Computer Architecture from the Harbin Institute of Technology on July 1996 and December 1999, respectively. Her research interests are focused in the area of network security, Internet measurement, and network computing. She was awarded 3 Ministry Science and Technology Progress awards and published over 50 papers in journals and international conferences.