ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

A System to Block Online Spam Messages

Shruthi. V, S. Vinodh Kumar

Student, Dept. of Computer Science and Engineering, T. John Institute of Technology, Bangalore, India
Asst. Professor, Dept. of Information Science and Engineering, City Engineering College, Bangalore, India

ABSTRACT: One of the basic problems in Online Social Networks (OSNs) is to provide the dexterity to control the text (messages) posted on their own profile so as to prevent that unwanted content is displayed. In this paper we propose a system that allows OSN users to have a direct hold on the messages posted on their private walls. This can be achieved through a filtering system that allows users to apply their filtering criteria, thereby allowing content-based filtering in support of filtering based upon relationship types.

KEYWORDS: Online Social Networks, Information filtering, Content-based filtering, Filtered wall.

I. INTRODUCTION

Online Social Networks (OSNs) are today one of the most popular interactive medium to communicate, share, and disseminate a considerable amount of human life information. Daily and continuous communications imply the exchange of several types of content, including free text, image, audio, and video data. According to Facebook statistics average user creates 90 pieces of content each month, whereas more than 30 billion pieces of content (web links, news stories, blog posts, notes, photo albums, etc.) are shared each month. The huge and dynamic character of these data creates the premise for the employment of web content mining strategies aimed to automatically discover useful information dormant within the data. They are instrumental to provide an active support in complex and sophisticated tasks involved in OSN management, such as for instance access control or information filtering. Information filtering has been greatly explored for what concerns textual documents and, more recently, web content (e.g., [1], [2], [3]). However, the aim of the majority of these proposals is mainly to provide user a classification mechanism to avoid they are overwhelmed by useless data. In OSNs, information filtering can also be used for a different, more sensitive purpose. This is due to the fact that in OSNs there is a possibility of posting or commenting other posts on particular public/private areas, called in general walls Information filtering can therefore be used to give users the ability to automatically control the messages written on their own walls, by filtering out unwanted messages. We believe that this is a key OSN service that has not been provided so far. Indeed, today OSNs provide very little support to prevent unwanted messages on user walls. For example, Facebook allows users to state who is allowed to insert messages in their walls (i.e., friends, friends of friends, or defined groups of friends). However, no content-based preferences are supported and therefore it is not possible to prevent undesired messages, such as political or vulgar ones, no matter of the user who posts them. Providing this service is not only a matter of using previously defined web content mining techniques for a different application, rather it requires to design ad hoc classification strategies. This is because wall messages are constituted by short text for which traditional classification methods have serious limitations since short texts do not provide sufficient word occurrences.

The aim of the present work is therefore to propose and experimentally evaluate an automated system, called Filtered Wall (FW), able to filter unwanted messages from OSN user walls. We exploit Machine Learning (ML) text categorization techniques [4] to automatically assign with each short text message a set of categories based on its content. The major efforts in building a robust short text classifier (STC) are concentrated in the extraction and selection of a set of characterizing and discriminate features. The solutions investigated in this paper are an extension of those adopted in a previous work by us [5] from which we inherit the learning model and the elicitation procedure for generating pre-classified data. The original set of features, derived from endogenous properties of short texts, is enlarged here including exogenous knowledge related to context from which the messages originate. As far as the learning model is concerned, we confirm in the current paper the use of neural learning which is today recognized as one of the most efficient solutions in text classification [4]. We insert the neural model within a hierarchical two- level

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

classification strategy. In the first level, the RBFN categorizes short messages as Neutral and Non-neutral, in the second stage, Non-neutral messages are classified producing gradual estimates of appropriateness to each of the considered category.

Besides classification facilities, the system provides a Powerful rule layer exploiting a flexible language to specify Filtering Rules (FRs), by which users can state what contents, should not be displayed on their walls. FRs can support a variety of different filtering criteria that can be combined and customized according to the user needs. FRs exploit user profiles, user relationships as well as the output of the ML categorization process to state the filtering criteria to be enforced. In addition, the system provides the support for user-defined Blacklists (BLs), that is, lists of users that are temporarily prevented to post any kind of messages on a user wall.

To the best of our knowledge, this is the first proposal of a system to automatically filter unwanted messages from OSN user walls on the basis of both message content and the message creator relationships and characteristics. The current paper substantially extends [5] for what concerns both the rule layer and the classification module. Major differences include, a different semantics for filtering rules to better fit the considered domain, an online setup assistant (OSA) to help users in FR specification, the extension of the set of features considered in the classification process, a more deep performance evaluation study and an update of the prototype implementation to reflect the changes made to the classification techniques.

The remainder of this paper is organized as follows: Section 2 surveys related work, whereas Section 3 introduces the conceptual architecture of the proposed system. Section 4 describes the text classification method used to categorize text contents, whereas Section 5 illustrates FRs and BLs. Section 6 concludes this paper.

II. LITERATURE SURVEY

The main contribution of this paper is the design of a system providing customizable content-based message filtering for OSNs, based on ML techniques. As we have pointed out in the introduction, to the best of our knowledge, we are the first proposing such kind of application for OSNs. Our work has relationships both with the state of the art in content-based filtering, as well as with the field of policy-based personalization for OSNs and, more in general, web contents. Therefore, in what follows, we survey the literature in both these fields.

Content Based-filtering

Information filtering systems are designed to classify a stream of dynamically generated information dispatched asynchronously by an information producer and present to the user. In content-based filtering each user is assumed to operate independently. As a result, a content-based filtering system selects information items based on the correlation between the content of the items and the user preferences as opposed to a collaborative filtering system that chooses items based on the correlation between people with similar preferences [7], [8]. While electronic mail was the original domain of early work on information-filtering, subsequent papers have addressed diversified domains including newswire articles, Internet "news" articles, and broader network resources [9], [10], [11]. Documents processed in content-based filtering are mostly textual in nature and this makes content-based filtering close to text classification. The activity of filtering can be modelled, in fact, as a case of single label, binary classification, partitioning incoming documents into relevant and non relevant categories [12]. More complex filtering systems include multi label text categorization automatically labelling messages into partial thematic categories.

Content-based filtering is based on ML paradigm according to which a classifier is automatically induced by learning from a set of pre-classified examples. A remarkable variety of related work has recently appeared, which differ for the adopted feature extraction methods, model learning, and collection of samples [13], [1], [14], [3], [15]. The feature extraction procedure maps text into a compact representation of its content and is uniformly applied to training and generalization phases. Several experiments prove that Bag-of-Words (BoW) approaches yield good performance and prevail in general over more sophisticated text representation that may have superior semantics but lower statistical quality [16], [17], [18]. As far as the learning model is concerned, there are a number of major approaches in content-based filtering and text

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

classification in general showing mutual advantages and disadvantages in function of application-dependent issues. In [4], a detailed comparison analysis has been conducted confirming superiority of Boosting-based classifiers [19].However, it is worth to note that most of the work related to text filtering by ML has been applied for long form text and the assessed performance of the text classification methods strictly depends on the nature of textual documents.

The application of content-based filtering on messages posted on OSN user walls poses additional challenges given the short length of these messages other than the wide range of topics that can be discussed. Short text classification has received up to now little attention in the scientific community. Recent work highlights difficulties in defining robust features, essentially due to the fact that the description of the short text is concise, with many misspellings, nonstandard terms, and noise. Zelikovitz and Hirsh attempt to improve the classification of short text strings developing a semi-supervised learning strategy based on a combination of labelled training data plus a secondary corpus of unlabeled but related longer documents. This solution is inapplicable in our domain in which short messages are not summary or part of longer semantically related documents. A different approach is proposed by Bobicev and Sokolova that circumvent the problem of error-prone feature construction by adopting a statistical learning method that can perform reasonably well without feature engineering. However, this method, named Prediction by Partial Mapping, produces a language model that is used in probabilistic text classifiers which are hard classifiers in nature and do not easily integrate soft, multi membership paradigms. In our scenario, we consider gradual membership to classes a key feature for defining flexible policy-based personalization strategies.

Policy based personalization of OSN contents

Recently, there have been some proposals exploiting classification mechanisms for personalizing access in OSNs. For instance, a classification method has been proposed to categorize short text messages in order to avoid overwhelming users of micro blogging services by raw data. The system Twitter is focused and associates a set of categories with each tweet describing its content. The user can then view only certain types of tweets based on his/her interests. In contrast, Golbeck and Kuter propose an application, called Filmtrust that exploits OSN trust relationships and provenance information to personalize access to the website. However, such systems do not provide a filtering policy layer by which the user can exploit the result of the classification process to decide how and to which extent filtering out unwanted information. In contrast, our filtering policy language allows the setting of FRs according to a variety of criteria that do not consider only the results of the classification process but also the relationships of the wall owner with other OSN users as well as information on the user profile. Moreover, our system is complemented by a flexible mechanism for BL management that provides a further opportunity of customization to the filtering procedure.

The only social networking service we are aware of providing filtering abilities to its users is MyWOT, a social networking service which gives its subscribers the ability to: 1) Rate resources with respect to four criteria: trustworthiness, vendor reliability, privacy, and child safety; 2) specify preferences determining whether the browser should block access to a given resource, or should simply return a warning message on the basis of the specified rating. Despite the existence of some similarities, the approach adopted by MyWOT is quite different from ours. In particular, it supports filtering criteria which are far less flexible than the ones of Filtered Wall since they are only based on the four above-mentioned criteria. Moreover, no automatic classification mechanism is provided to the end user.

Our work is also inspired by the many access control models and related policy languages and enforcement mechanisms that have been proposed so far for OSNs, since filtering shares several similarities with access control. Actually, content filtering can be considered as an extension of access control, since it can be used both to protect objects from unauthorized subjects, and subjects from inappropriate objects. In the field of OSNs, the majority of access control models proposed so far enforce topology-based access control, according to which access control requirements are expressed in terms of relationships that the requester should have with the resource owner. We use a similar idea to identify the users to which a FR applies. However, our filtering policy language extends the languages proposed for access control policy specification in OSNs to cope with the extended requirements of the filtering domain. Indeed, since we are dealing with filtering of unwanted

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

contents rather than with access control, one of the key ingredients of our system is the availability of a description for the message contents to be exploited by the filtering mechanism. In contrast, no one of the access control models previously cited exploit the content of the resources to enforce access control.

Finally, our policy language has some relationships with the policy frameworks that have been so far proposed to support the specification and enforcement of policies expressed in terms of constraints on the machine understandable resource descriptions provided by Semantic Web languages. Examples of such frameworks are KAoS and REI, focusing mainly on access control, Protune, which provides support also to trust negotiation and privacy policies, and WIQA, which gives end Users the ability of using filtering policies in order to denote given "quality" requirements that web resources must satisfy to be displayed to the users. However, although such frameworks are very powerful and general enough to be customized and/or extended for different application scenarios they have not been specifically conceived to address information filtering in OSNs and therefore to consider the user social graph in the policy specification process. Therefore, we prefer to define our own abstract and more compact policy language.

III. FILTERED WALL ARCHITECTURE

The architecture in support of OSN services is a three-tier structure (Fig.1). The first layer, called Social Network Manager (SNM), commonly aims to provide the basic OSN functionalities (i.e., profile and relationship management), whereas the second layer provides the support for external Social Network Applications (SNAs). The supported SNAs may in turn require an additional layer for their needed Graphical User Interfaces (GUIs). According to this reference architecture, the proposed system is placed in the second and third layers. In particular, users interact with the system by means of a GUI to set up and manage their FRs/ BLs. Moreover, the GUI provides users with a FW, that is, a wall where only messages that are authorized according to their FRs/BLs are published. The core components of the proposed system are the Content-Based Messages Filtering (CBMF) and the Short Text Classifier modules. The latter component aims to classify messages according to a set of categories. The strategy underlying this module is described in Section 4. In contrast, the first component exploits the message categorization provided by the STC module to enforce the FRs specified by the user. BLs can also be used to enhance the filtering process. As graphically depicted in Fig.1, The path followed by a message, from its writing to the possible final publication can be summarized as follows:

1. After entering the private wall of one of his/her contacts, the user tries to post a message, which is intercepted by FW.
2. ML-based text classifier extracts metadata from the content of the message.
3. FW uses metadata provided by the classifier, together with data extracted from the social graph and users' profiles, to enforce the filtering and BL rules.
4. Depending on the results of the previous step, the message will be published or filtered by FW.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

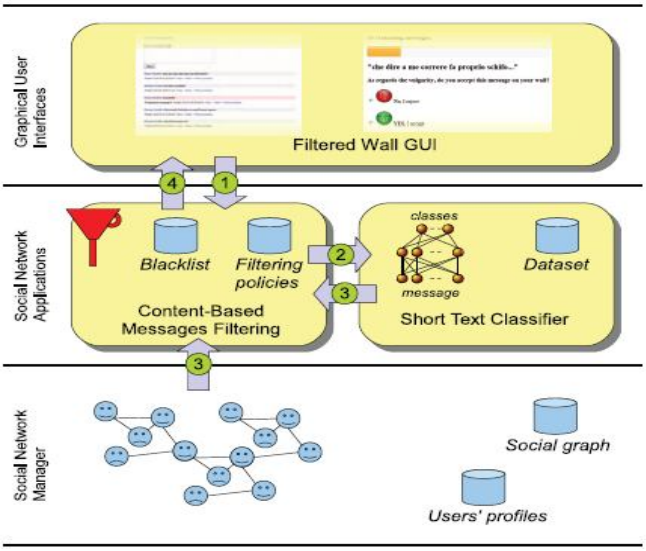

Fig.1 Filtered wall conceptual architecture and flow of messages.

Initially user registers with the provided OSN followed by logging into his corresponding profile by providing the required credentials for authentication. A backend database will be maintained containing such words that are not allowed to be displayed on their private space. Whenever a message occurs, it gets checked against the pre-defined database. If in case a match occurs then the message gets discarded/blocked as it's labelled to be unwanted by the user else it gets displayed on the wall.

IV. TEXT REPRESENTATION

Established techniques used for text classification work well on data sets with large documents such as newswires, suffer when the documents in the corpus are short. In this context, critical aspects are the definition of a set of characterizing and discriminate features allowing the representation of underlying concepts and the collection of a complete and consistent set of supervised examples. Our study is aimed at designing and evaluating various representation techniques in combination with a neural learning strategy to semantically categorize short texts. From a ML point of view, we approach the task by defining a hierarchical two-level strategy assuming that it is better to identify and eliminate "neutral" sentences, then classify "non-neutral" sentences by the class of interest instead of doing everything in one step. This choice is motivated by related work showing advantages in classifying text and/or short texts using a hierarchical strategy [1]. The first-level task is conceived as a hard classification in which short texts are labelled with crisp Neutral and Non-neutral labels. The second-level soft classifier acts on the crisp set of non-neutral short texts and, for each of them, it "simply" produces estimated appropriateness or "gradual membership" for each of the conceived classes, without taking any "hard" decision on any of them. Such a list of grades is then used by the subsequent phases of the filtering process.

Text Representation

The extraction of an appropriate set of features by which representing the text of a given document is a crucial task strongly affecting the performance of the overall classification strategy. Different sets of features for text categorization have been proposed in the literature [4]; however, the most appropriate feature set and feature representation for short text messages have not yet been sufficiently investigated.

Proceeding from these considerations and on the basis of our experience [5], we consider three types of features, BoW, Document properties (Dp) and Contextual Features (CF). The first two types of features, already used in [5], are endogenous, that is, they are entirely derived from the information contained within the text of the message. Text representation using endogenous knowledge has a good general applicability; however,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

in operational settings, it is legitimate to use also exogenous knowledge, i.e., any source of information outside the message body but directly or indirectly related to the message itself. We introduce CF modelling information that characterizes the environment where the user is posting. These features play a key role in deterministically understanding the semantics of the messages [4]. All proposed features have been analysed in the experimental evaluation phase in order to determine the combination that is most appropriate for short text.

The underlying model for text representation is the Vector Space Model (VSM). In the BoW representation, terms are identified with words. In the case of non-binary weighting, the weight w_{kj} of term t_k in document d_j is computed according to the standard term frequency—inverse document frequency (tf-idf) weighting function, defined as

$$tf - idf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#Tr(tk)}$$

Bad words: They are computed similarly to the correct words feature, where the set K is a collection of "dirty words" for the domain language.

Capital words: It expresses the amount of words mostly written with capital letters, calculated as the percentage of words within the message, having more than half of the characters in capital case. The rationale behind this choice lies in the fact that with this definition we intend to characterize the willingness of the author's message to use capital letters excluding accidental use or the use of correct grammar rules. For example, the value of this feature for the document "To be OR NOt to BE" is 0.5 Since the words "OR" "NOt" and "BE" are considered as capitalized ("To" is not uppercase since the number of capital characters should be strictly greater than the characters count).

Punctuations characters: It is calculated as the percentage of the punctuation characters over the total number of characters in the message. For example, the value of the feature for the document "Hello!!! How're u doing?" is 5=24.

Correct Words: It expresses the amount of terms t_k 2T\ K, where t_k is a term of the considered document d_j and K is a set of known words for the domain language.

Exclamation marks: It is calculated as the percentage of exclamation marks over the total number of punctuation characters in the message. Referring to the aforementioned document, the value is 3=5.

Question marks: It is calculated as the percentage of question marks over the total number of punctuation characters in the message. Referring to the aforementioned document, the value is 1=5.

V. FILTERING RULES AND BLACKLIST MANAGEMENT

In this section the user defines a database containing such words that are not allowed to be displayed on their private space. Whenever a message occurs, it gets checked against the pre-defined database. If in case a match occurs then the message gets discarded as it's labelled to be unwanted by the user else it gets displayed on the wall.

A further component of our system is a BL mechanism to avoid messages from undesired creators, independent from their contents. BLs is directly managed by the system, which should be able to determine who are the users to be inserted in the BL and decide when users retention in the BL is finished. To enhance flexibility, such information is given to the system through a set of rules, hereafter called BL rules. Such rules are not defined by the SNMP; therefore, they are not meant as general high-level directives to be applied to the whole community. Rather, we decide to let the users themselves, i.e., the wall's owners to specify BL rules regulating who has to be banned from their walls and for how long. Therefore, a user might be banned from a wall, by, at the same time, being able to post in other walls. Similar to FRs, our BL rules make the wall owner able to identify users to be blocked according to their profiles as well as their relationships in the OSN.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

Therefore, by means of a BL rule, wall owners are, for example, able to ban from their walls users they do not directly know (i.e., with which they have only indirect relationships), or users that are friend of a given person as they may have a bad opinion of this person. This banning can be adopted for an undetermined time period or for a specific time window. Moreover, banning criteria may also take into account users' behaviour in the OSN. More precisely, among possible information denoting users' bad behaviour we have focused on two main measures. The first is related to the principle that if within a given time interval a user has been inserted into a BL for several times, say greater than a given threshold, he/she might deserve to stay in the BL for another while, as his/her behaviour is not improved.

In contrast, to catch new bad behaviours, use the Relative Frequency (RF) that let the system is able to detect those users whose messages continue to fail the FRs. The two measures can be computed either locally, that is, by considering only the messages and/or the BL of the user specifying the BL rule or globally, that is, by considering all OSN users walls and/or BLs.

VI. CONCLUSION

In this paper, we have presented a system to filter undesired messages from OSN walls. Initially user registers with the provided OSN followed by logging into his corresponding profile by providing the required credentials for authentication. A backend database will be maintained containing such words that are not allowed to be displayed on their private space. Whenever a message occurs, it gets checked against the pre-defined database. If in case a match occurs then the message gets discarded/blocked as it's labelled to be unwanted by the user else it gets displayed on the wall.

As a part of future enhancement, a concept called blacklist is introduced. In this case a user will be blacklisted if his messages are discarded continuously for certain number of times thereby not allowing him to interact with the specified opponent. As a result of which filtering time and effort can be significantly reduced.

REFERENCES

1. A. Adomavicius and G. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. knowledge and Data Eng., vol. 17, no. 6, pp. 734-749, 2005.
2. M. Chau and H. Chen, "A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis," Decision Support Systems, vol. 44, no. 2, pp. 482-494,2008.
3. R.J. Mooney and L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization," Proc. Fifth ACM Conf. Digital Libraries, pp. 195-204, 2000.
4. F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.
5. M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-Based Filtering in On-Line Social Networks," Proc.ECML/PKDD Workshop Privacy and Security Issues in Data Mining and Machine Learning (PSDML '10), 2010.
6. N.J. Belkin and W.B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" Comm.ACM, vol. 35, no.12, pp. 29-38, 1992.
7. P.J. Denning, "Electronic Junk," Comm. ACM, vol. 25, no.3, pp.163-165, 1982.
8. P.W. Foltz and S.T. Dumais, "Personalized Information Delivery: An Analysis of Information Filtering Methods," Comm.ACM, vol. 35, no. 12, pp.51-60, 1992.
9. P.S. Jacobs and L.F. Rau, "Scisor: Extracting Information from On- Line News," Comm. ACM, vol. 33, no. 11, pp. 88-97, 1990.
10. S. Pollock, "A Rule-Based Message Filtering System," ACM Trans. Office Information Systems, vol. 6, no. 3, pp. 232-254, 1988.
11. P.E. Baclace, "Competitive Agents for Information Filtering," Comm. ACM, vol. 35, no. 12, p. 50, 1992.
12. P.J. Hayes, P.M. Andersen, I.B. Nirenburg, and L.M. Schmandt, "Tcs: A Shell for Content-Based Text Categorization," Proc. Sixth IEEE Conf. Artificial Intelligence Applications (CAIA '90), pp.320-326,1990.
13. G.Amati and F.Crestani, "Probabilistic Learning for Selective Dissemination of Information," Information Processing and Manage- ment, vol. 35, no. 5, pp.633-654,1999
14. M.J. Pazzani and D. Billsus, "Learning and Revising User Profiles: The Identification of Interesting Web Sites," Machine Learning, vol. 27, no. 3, pp.313-331, 1997.
15. Y. Zhang and J. Callan, "Maximum Likelihood Estimation for Filtering Thresholds," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 294-302, 2001.
16. C. Apte, F. Damerau, S.M. Weiss, D. Sholom, and M.Weiss, "Automated Learning of Decision Rules for Text Categorization," Trans. Information Systems, vol. 12, no. 3, pp. 233-251, 1994.
17. S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive Learning Algorithms and Representations for Text Categorization," Proc. Seventh Intl Conf. Information and Knowledge Manage- ment (CIKM '98), pp. 148-155, 1998.
18. D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. 15th ACM Int'l Conf. Research and Development in Information Retrieval (SIGIR '92), N.J. Belkin, P. Ingwersen, and A.M. Pejtersen, eds., pp. 37-50, 1992.
19. R.E. Schapire and Y. Singer, "Boostexter: A Boosting-Based System for Text Categorization," Machine Learning, vol.39, nos. 2/3, pp. 135-168, 2000
20. H. Schütze, D.A. Hull, and J.O. Pedersen, "A Comparison of Classifiers and Document Representations for the Routing Problem".