International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A System to Filter Unwanted Messages from OSN User Walls using Trustworthiness

Swapnali V. Jadhav[1], Y. B. Gurav[2]

[1]Department of Computer Engineering, PVPIT, Bavdhan. Pune, India

[2]Assistant Professor, Department of Computer Engineering, PVPIT, Bavdhan. Pune, India

Abstract: ***The core problem in today's Online Social Networks (OSNs) is to allocate users the authority to manage the messages posted on their private space to avert that unwanted content. The unwanted data may contains political, vulgar, non neural etc. message filtering systems are designed for unstructured or semi-structured data, as opposed to database applications, which use very structured data. In this paper we proposed a System with the flexible rules to filter the unwanted messages posted on user wall. After crossing threshold value the notification message is send to that user. This allows users to customize the refining criteria to be applied to their walls, and a Machine Learning-based classifier automatically classify the messages and labelling messages in support of content-based filtering.***

Keywords: Flexible rules, message filtering, online social networks, short text classification

1. Introduction

On-line Social Networks (OSNs) are today one of the most popular interactive medium to communicate, share and disseminate a considerable amount of human life information.
An OSN is a web-based service that allows individuals to:

1) Construct a public or semi-public profile within the service,
2) Articulate a list of other users with whom they share a connection,
3) View and traverse their list of connections and those made by others within the service.

Daily and continuous communications imply the exchange of several types of content, including free text, image, and 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 [2], [3]. However, the aim of the majority of these proposals is mainly to provide users 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 the possibility of posting or commenting other posts on particular public/private areas, called in general walls.

Information and communication technology plays a significant role in today's networked society. It has affected the online interaction between users, who are aware of security applications and their implications on personal privacy. There is a need to develop more security mechanisms for different communication technologies, particularly online social networks. 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. 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.

The aim of the system to propose and experimentally evaluate an automated system, called Filtered Wall (FW), able to filter unwanted messages from OSN user walls. The key idea of the proposed system is the support for content based user preferences. This is possible thank to the use of a Machine Learning (ML) text categorization procedure able to automatically assign with each message a set of categories based on its content. We believe that the proposed strategy is a key service for social networks in that in today social networks users have little control on the messages displayed on their walls. In contrast, by means of the proposed mechanism, a user can specify what contents should not be displayed on his/her wall, by specifying a set of filtering rules. Filtering rules are very flexible in terms of the filtering requirements they can support, in that they allow to specify filtering conditions based on user profiles, user relationships as well as the output of the ML categorization process. In addition, the system provides the support for user defined blacklist management, that is, list of users that are temporarily prevented to post messages on a user wall.
This System we design to show the effectiveness of the developed filtering techniques. Finally, we have provided a prototype implementation of our system having Facebook as target OSN, even if our system can be easily applied to other OSNs as well. 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[4].

2. Literature Survey

M. Vanetti[5] proposes a system enforcing content-based message filtering conceived as a key service for On-line Social Networks (OSNs). The system allows OSN users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system, that allows a user to customize the filtering criteria to be applied to their walls, and a Machine Learning based soft classifier automatically producing membership labels in support of content-based filtering. They have presented a system to filter out undesired messages from OSN walls. The system exploits a ML soft classifier to enforce customizable content-depended filtering rules. Moreover, the flexibility of the system in terms of filtering options is enhanced trough the management of BLs. The proposed system may suffer of problems similar to those in the specification of privacy settings in OSN. As future work, They said that to exploit similar techniques to infer BL and filtering rules.

Gediminas Adomavicius[6] gives an overview of the field of recommender systems and describes the current generation of recommendation methods that are usually classified into the following four main categories: content-based, collaborative, Policy-based personalization and hybrid recommendation approaches. This paper also describes various limitations of current recommendation methods and discusses possible extensions that can improve recommendation capabilities and make recommender systems applicable to an even broader range of applications. In this paper, they reviewed various limitations of the current recommendation methods and discussed possible extensions that can provide better recommendation capabilities. These extensions include among others, the improved modeling of users and items, incorporation of the contextual information into the recommendation process, support for multicriteria ratings, and provision of a more flexible and less intrusive recommendation process.

Bharath Sriram[7] states microblogging services such as Twitter, the users may become overwhelmed by the raw data. One solution to this problem is the classification of short text messages. As short texts do not provide sufficient word occurrences, traditional classification methods such as "Bag-Of-Words" have limitations. To address this problem, they propose to use a small set of domain-specific features extracted from the author's profile and text. The proposed approach effectively classifies the text to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages. They have proposed an approach to classify tweets into general but important categories by using the author information and features within the tweets. With such a system, users can subscribe to or view only certain types of tweets based on their interest.

Michael Beye [8] discussed, In recent years, Online Social Networks (OSNs) have become an important part of daily life for many. Users build explicit networks to represent their social relationships, either existing or new. Users also often upload and share a plethora of information related to their personal lives. The potential privacy risks of such behavior are often underestimated or ignored. For example, users often disclose personal information to a larger audience than intended. Users may even post information about others without their consent. A lack of experience and awareness in users, as well as proper tools and design of the OSNs, perpetuate the situation. This paper aims to provide insight into such privacy issues and looks at OSNs, their associated privacy risks, and existing research into solutions.

Josie Maria[9] discussed Effective Web content filtering is a necessity in educational and workplace environments, but current approaches are far from perfect. They discuss a model for text-based intelligent Web content filtering, in which shallow linguistic analysis plays a key role. In order to demon strate how this model can be realized, they have developed a lexical Named Entity Recognition system, and used it to improve the effectiveness of statistical Automated Text Categorization methods. they have performed several experiments that confirm this fact, and encourage the integra tion of other shallow linguistic processing techniques in intelligent Web content filtering. They discussed that shallow linguistic analysis in general, and Named Entity Recognition in particular, can be used to improve the effectiveness of text classification in the framework of intelligent Web content filtering.

3. Implementation Details

3.1 Filtering Types

3.1.1 Content-based Content Based filtering system recommends a document by matching the document profile with the user profile, using traditional information retrieval techniques such Term Frequency and Inverse Document frequency (TF-IDF). User characteristics are gathered over time and profiled automatically based upon a user's prior feedback and choices. The system uses item to item correlation in recommending the document to the user. The system starts with the process of collecting the content details about the item, such as treatments, symptoms etc. for disease related item and author, publisher etc. for the book items. In the next step, the system asks the user to rate the items. Finally, system matches unrated item with the user profile item and assign score to the unrated item and user is presented with items ranked according to the scores assigned.

News Dude, is one of the examples of content based filtering system which uses short term TF-IDF technique and long term Bayesian classifier for learning on an initial set of documents provided by the user. Content based information filtering systems are not affected by the cold start problem and new user problem, as the system focuses on the individual user needs Content based information filtering systems are not suitable for multimedia items, such as images, audio, video. Multimedia documents must be tagged

with a semantic description of the resource which will be a time consuming process. Content-based filtering methods cannot filter documents based on quality and relevance.

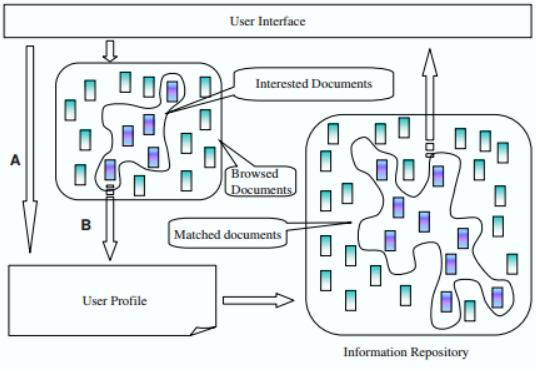

Figure 1: Content based filtering

Limitations: Although content-based filtering has proven to be effective in recommending textual items relevant to a topic, it also has its limitations:

- Content-based filtering more than often provides recommendation in a literal sense, because all the information is selected and recommended based on textual contents. Even though a product was essentially useful, it might be under-valued because of the ambiguous and thus misleading appearance of the textual contents. In addition, it is indistinguishable in quality of the recommended products. This is because the term vector for a product simply captures the frequency of each term in an article, and a poorly worded article can well have an equal or even higher similarity value than a finely written one.
- Content-based filtering generally works well with sufficient textual information. However, other multimedia files such as images, audio and video streams are not applicable if the metadata do not have enough textual information

3.1.2Collaborative filtering: Collaborative filtering systems filters information based on the interests of the user (past history), and the ratings of other users with similar interests. It is widely used in many filtering systems or recommender systems, especially in ecommerce applications. One of the examples of such system are Amazon.com and e-Bay, where a user's past shopping history is used to make recommendations for new products.

3.2 Proposed Work

Despite the efforts in the fields mentioned above, other important issues have been explored include user privacy, trustworthiness and context-aware recommendation. One of user concerns to use recommender systems freely and comfortably is user privacy. Users are usually reluctant to disclose their private information such as purchase, reading, browsing records. However, most current filtering algorithms need to obtain user private information for further analysis and recommendation services. Some work has studied on how to protect user privacy in recommender systems .Current filtering techniques assume that user ratings are trustable and treat all users equally. However, some may argue that the opinions of experts should be more emphasized than that of novices.

The main goal of the system is to design an online message filtering system that is deployed at the OSN service provider side. Once deployed, it inspects every message before rendering the message to the intended recipients and makes immediate decision on whether or not the message under inspection should be dropped. 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 to automatically assign with each short text message a set of categories based on its content. First the message is filtered with filtering rules.

3.2.1 Hybrid filtering systems

The hybrid filtering systems combines features of both the content and collaborative filtering systems. The hybrid system overcomes the problem of cold start and early rater problem by using the content based approach in the initial stage. The first is the simple combination model, which combines results from the collaborative and content-based filters as shown in following figure2.

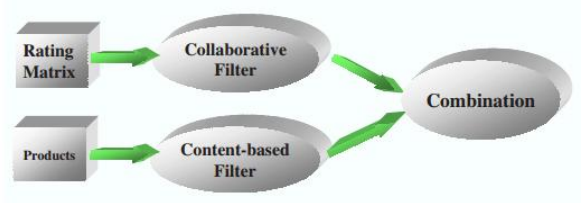

Figure 2: Hybrid filtering system

(a) Filtering Rules: 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.

(b) Online setup assistant for FRs thresholds: OSA presents the user with a set of messages selected from the dataset discussed. For each message, the user tells the system the decision to accept or reject the message. The collection and processing of user decisions on an adequate set of messages distributed over all the classes allows to compute customized thresholds representing the user attitude in accepting or rejecting certain contents. Such messages are selected according to the following process. A certain amount of non neutral messages taken from a fraction of the dataset and not belonging to the training/test sets, are classified by the ML in order to have, for each message, the second level class membership values.

Suppose that Bob is an OSN user and he wants to always block messages having an high degree of vulgar content. Through the session with OSA, the threshold representing the user attitude for the Vulgar class is set to 0.8. Now, suppose that Bob wants to filter only messages coming from indirect friends, whereas for direct friends such messages should be blocked only for those users whose trust value is below 0.5. These filtering criteria can be easily specified through the following FRs5:

- ((Bob, friendOf, 2, 1), (Vulgar, 0.80), block)
- ((Bob, friendOf, 1, 0.5), (Vulgar, 0.80), block)

(c) Blacklists: A further component of our system is a BL mechanism to avoid messages from undesired creators, independent from their contents. BLs are 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 are given to the system through a set of rules, hereafter called BL rules.

BL rule:- A BL rule is a tuple (author,

- creatorSpec, creatorBehavior, T) where author is the OSN user who specifies the rule, i.e., the wall owner;
- creatorSpec is a creator specification;
- creatorBehavior consists of two components RFBlocked and minBanned.

RFBlocked = (RF, mode, window) is defined such that:– RF = #bMessages/#tMessages , where #tMessages is the total number of messages that each OSN user identified by creatorSpec has tried to publish in the author wall (mode = myWall) or in all the OSN walls (mode = SN); whereas #bMessages is the number of messages among those in #tMessages that have been blocked; window is the time interval of creation of those messages that have to be considered for RF computation; minBanned = (min, mode, window), where min is the minimum number of times in the time interval specified in window that OSN users identified by creatorSpec have to be inserted into the BL due to BL rules specified by author wall (mode = myWall) or all OSN users (mode = SN) in order to satisfy the constraint.

T denotes the time period the users identified by creatorSpec and creator Behavior have to be banned from author wall.

3.3 Algorithm

3.3.1 Preprocessing

The primary aim of the pre-processing phase is to remove from the input message all characters and terms that can possibly affect the quality of group descriptions.

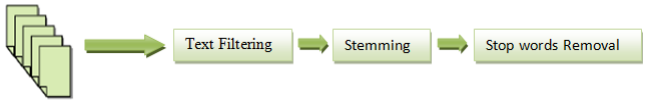

Figure Pre-processing of Message

3.3.2 Pre-processing steps

/** Phase 1: Preprocessing */
for each document
{
do text filtering;
identify the document's language;
apply stemming;
mark stop words;
}

Algorithm :

1: d← input message
{STEP 1: Preprocessing}
2: **for** all d € D do
3: perform text categorization
4: **if** d!=null **then**
Filter text for unwanted symbols
5: apply stemming and mark stop-words in d;
6: **end for**

There are three steps to the preprocessing phase: Text filtering, Stemming and Stop words marking.

(a) Text filtering: In the text filtering step, all terms that are useless or would introduce noise in filtering process are removed from the input message. Among such terms are:

- HTML tags (e.g. <table>) and entities (e.g. &) if any.
- non-letter characters such as "$", "%" or "#" (except white spaces and sentence markers such as '.', '?' or '!') Note that at this stage the stop-words are not removed from the input.

(b) Stemming: Stemming algorithms are used to transform the words in texts into their grammatical root form, and are mainly used to improve the Information Retrieval System's efficiency. To stem a word is to reduce it to a more general form, possibly its root. For example, stemming the term interesting may produce the term interest. Though the stem of a word might not be its root, we want all words that have the same stem to have the same root.

(c)

(d) Elimination of Stop Words: After stemming it is necessary to remove unwanted words. There are 400 to 500 types of stop words such as "of", "and", "the," etc., that provide no useful information about the message. Stop-word removal is the process of removing these words. Stop-words account for about 20% of all words in a typical document. These techniques greatly reduce the size of the searching and matching each word in message. Stemming alone can reduce the size of an index by nearly 40%.

3.4 Mathematical Model

3.4.1 For Filtering Rules:

(a) Input: Filtering Rules are customizable by the user. User can have authority to decide what contents should be blocked or displayed on his wall by using Filtering rules. For specify a Filtering rules user profile as well as user social relationship will be considered.

FR= {Trustier, SOUs, Rule, TuV}

FR is dependent on following factors

- Trustier
- Set of Users (SOUs)
- Rule
- Action

Trustier is a person who defines the rules.
SOUs denote the set of OSN user.
Rule is a Boolean expression defined on content.

(b)Process: FM= {SOUs, Rule==category (Violence, Vulgar, offensive, Hate, Sexual), TuV}

- FM
- SOUs
- Rule
- TuV

Here, FM Block Messages at basic level.
SOUs Denotes set of users
Rule Category of specified contents in message.
TuV is the trust value of sender.

In processing, after giving input message, the system will compare the text with the different categories which are prevented. If message found in that prevented type of category then message will display to the user that "can't send this type of messages", and still the user wants to send the message he/she can continue with sending the message. The Trustier, who gets the message, but the words which are defended in the rule are sent in **** format. After getting the message the Trustier will give the Feedback (FB) to the sender and the sender will gain the TuV accordingly. Process denotes the action to be performed by the system on the messages matching Rule and created by users identified by SOUs.

E.g. FM== {Friends, Rule==category (Vulgar, Sexual), TuV>50}

i.e. Trustier will accept the message from friends but message should not contain vulgar or sexual words. Message containing such words will affect the TuV of sender. Now the question arises, calculation of TuV.

(c) Trust Value Calculations: The trust value of any user in OSN is dependent on the feedback they gain by the user to whom they sent a message. Feedback from the user must also be trust worthy. That's why the FB can be categorized into following:-

1) **Positive with content (PC)** - Good FB, message is acceptable with objectionable content. This will increase the TuV of sender.
2) **Positive without content (PWC)** - Good FB, message is acceptable as this message does not contain objectionable content. This will increase the TuV of sender.
3) **Negative with content (NC)** - Bad FB, such messages must not be sent again, which are against the Rule. This will decrease the TuV of sender.
4) **Negative without content (NWC)** - Bad FB, message doesn't contain any objectionable content but the Trustier is giving negative FB. Such type of FB from Trustier will affect the TuV of its own, and the TuV of sender will remain same.

So, based on above categories the TuV will be calculated as follows:-
FB as 1 and 2 TuV= TuV+ abs [(PC+PWC)/ (NC+NWC)]
FB as 3 TuV= TuV-[1+ (NC+NWC) / (PC+PWC)] for [(NC+NWC) / (PC+PWC)] <1
Otherwise, send system generated message to sender, FB Negative with content exceeds limit of Threshold Value (ThV) and deduct 5 points from TuV, so ThV=TuV-5.
FB as 4 TuV= TuV of sender, but TuV= TuV-[1+ (NC+NWC) / (PC+PWC)] for Trustier.

(d) Output: PFM= {Rule, M||Y}
PFM Percentages of filtered message in a year or month.
In general, more than a filtering rule can apply to the same user. A message is therefore published only if it is not blocked by any of the filtering rules that apply to the message creator.

3.4.2 Blacklists*:* BLs are directly managed by the system. This should be able to determine the users to be inserted in the BL and decide when to retain user back from the BL. To enhance flexibility, such information is given to the system through a set of rules, hereafter called BL rules.
(a) BL rules: INPUT = {Sender, FB, TuV, ThV} Where

- Sender is the OSN user who is sending the message;
- FB is the FeedBack gain by the sender after sending the message
- TuV is the new Trust Value calculated as formulas specified in A.3.
- ThV is the Threshold Value.

BL Rules: ThV= PC+ PWC when, PC+PWC=NC+NWC.
For sender, when 5 points are deducted by system, which means sender cross the ThV put sender into BL for a specific duration.
For Trustier, after giving feedback, check ThV, if true, put Trustier in BL for specific duration.

3.5 Performance Study*:*

As we can see here the graph of accuracy. Our proposed method i.e. Trust value calculation gives more accuracy (93%) than existing RBFN algorithm(85%).

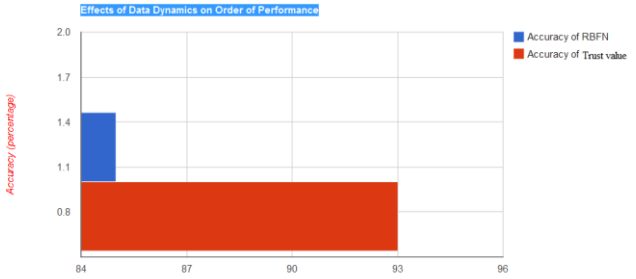

Figure 2: Graph of Accuracy

Results for Message Neutrality:

Table 1: Result for message neutrality

Classification		*Neutral*			*Non-Neutral*		
RBFN	**TV**	**P**	**R**	**F1**	**P**	**R**	**F1**
84%	94%	93%	90%	95%	95%	92%	93%
85%	95%	87%	98%	93%	89%	97%	94%

Here P is Precision, R is Recall and F1 is F-measure. We have calculated these values by using following formula:

Precision=(No. of True Positives)/(No. of true positives +No. of false positives)
Recall=(No. of True Negatives)/(No. of True Negative + No. of false positive)
F1-measure=(2*(Precision*Recall)/(Precision+Recall)).

Results for Non-neutral Classes Identification:

Table 2: Result for non-neural classes identification

Violence			*Vulgar*			*Hate*		
P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
87%	93%	90%	89%	94%	91%	90%	97%	94%
98%	84%	83%	94%	82%	84%	89%	92%	95%

4. Conclusion

In this report, we have discussed the literature survey of the filtering system. We are developing a system to filter undesired messages from OSN walls. The wall that restricts the unwanted message called as the Filtered Wall (FW). In this report we discussed the idea about the system. Additionally, we studied strategies and techniques limiting the inferences that a user can do on the enforced filtering rules with the aim of bypassing the filtering system, such as for instance randomly notifying a message that should instead be blocked.

References

[1] Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo, "A System to Filter Unwanted Messages from OSN User Walls" VOL. 25, NO. 2, FEBRUARY 2013.
[2] A. Adomavicius, G.and Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transaction on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.
[3] 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.
[4] N. J. Belkin and W. B. Croft,"Information filtering and information retrieval: Two sides of the same coin?"Communications of the ACM, vol. 35, no. 12, pp. 29–38, 1992.
[5] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo E. Ferrari" Content-based Filtering in On-line Social Networks".
[6] Gediminas Adomavicius, Member, IEEE, and Alexander Tuzhilin, Member, IEEE," Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, JUNE 2005.
[7] Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu Murat Demirbas " Short Text Classification in Twitter to Improve Information Filtering "
[8] Michael Beye, Arjan Jeckmans, Zekeriya Erkin, Pieter Hartel, Reginald Lagendijk and Qiang Tang," Literature Overview - Privacy in Online Social Networks".
[9] Josie Maria Giomez Hidalgo, Francisco Carrero Garcia, and Enrique Puertas Sanz," Named Entity Recognition for Web Content Filtering".
[10] Hongyu Gao Yan Chen Kathy Lee Diana Palsetia Alok Choudhary, " Towards Online Spam Filtering in Social Networks ".
[11] Antonio da Luz, Eduardo Valle, Arnaldo Araujo," Content-Based Spam Filtering On Video Sharing Social Networks".NPDI-LAB---DCC/UFMG, Belo Horizonte, MG, Brazil. Federal institute of Tehnology of Tocantins-IFTO, Paraiso, TO, Brazil. RECOD Lab—IC/UNICAMP, Campinas,SP, Brazil.
[12] Jennifer Golbeck, "The Twitter Mute Button: A Web Filtering challenge", CHI 2012, May 5-10, 2012, Austin,Texas, USA.
[13] George Forman," An Extensive Empirical Study of Feature Selection Metrics for Text classification",Journal of Machine Learning Research 3(2003)1289-1305,Hewlett-Packard Labs Palo Alto,CA,USA.