A Proposed Supervised Clustering Approach for the Identification of Concerned HIV-Related Messages in Web Forums

1A3-IOS-3a-3

Chaochang Chiu[*1] Nan-Hsing Chiu[*2] Re-Jiau Sung[*1] Chih-Hao Hsiao[*1]

[*1]Dept. of Information Management, Yuan Ze University, Taiwan
[*2]Dept. of Information Management, Chien Hsin University of Sci. and Tech., Taiwan

Web forums become the means of online communication and information sharing sources for the learning about healthcare and related treatment knowledge. By adopting web crawlers and natural language processing techniques, the automatic identification and classification approach of the concerned HIV-related messages is proposed to facilitate the health authorities or social support groups in instant counseling. The proposed supervised GA/k-means for classification approach can help effectively distinguish "concerned" and "unconcerned" HIV-related messages on the web forum.

1. Introduction

The widespread use of the Internet as a source of health information has allowed people to find healthcare decision aids. In usual clinic–patient relationships, many factors motivate patients to seek help from the Internet. Studies have shown that most people living with HIV and their caregivers visited medical websites for HIV-specific information, social support, coping mechanisms, and shared information from the Internet. By adopting web crawlers and natural language processing techniques, the automatic identification and classification approach of the concerned HIV-related messages is proposed to facilitate the health authorities or social support groups in instant counseling. By adopting information retrieval techniques, natural language processing and the proposed classification approach, this research attempts to detect the concerned HIV-related messages. This study used the keywords, AIDS and HIV, to retrieve a total of 2,083 posts from January 1, 2005 to December 31, 2011 from "Yahoo Knowledge" in Taiwan. Among the entire 2,083 posts, 470 (22.6%) posts are labeled as those require concerned.

Using information retrieval and text mining approaches can effectively distinguish "concerned" and "unconcerned" HIV-related messages. The obtained classification performances are compared with Support Vector Machine (SVM), C4.5 decision tree, Naïve Bayes (NB). Further, the correspondence analysis is applied to illustrate the different usages and grouping patterns of HIV-related messages. We proposed supervised GA/k-means for classification approach to help construct an effective identification and classification model with acceptable classification performance accompanied with its full flexibility to develop different fitness functions in accordance with the need of different requirements.

2. The Literature Review

Surfing cyberspace has become one of the most popular activities. The widespread use of the Internet as a source of health information has allowed people to find healthcare decision aids (Morris et al., 2008). Active participant engagement, such as through group interactive sessions, one-on-one counseling, and peer leadership usually results in the significant decreases in risk behavior. With the rapid progress in speed of Internet connections and proliferation of the interactive capabilities of websites, computer-mediated interpersonal communication between interventionists and at-risk individuals is available.

People with human immunodeficiency virus/acquired immune deficiency syndrome (HIV/AIDS) face significant social stigma and isolation. Affected families share their experiences which all point to the stigma that has always been associated with the infection (Bogart et al. 2008; Mahajan et al. 2008). Internet overcomes or reduces many of the barriers by providing information and support that is ubiquitous, anonymous, timely and user-controlled. Studies have shown that most people living with HIV and their caregivers visited medical websites for HIV-specific information, and shared information from the Internet (Horvath et al., 2009; Courtenay-Quirk et al., 2010; Samal et al., 2011).

Internet HIV support groups have been shown to provide information support, emotional support, and network support, are crucial for the people living with HIV (Mo and Coulson, 2008). However, effective and timely identification of the HIV-related messages that should be concerned from the voluminous posts in web forums becomes another emerging issue. Therefore, the automatic monitoring and classification mechanisms for the concerned HIV-related messages could facilitate the health authorities or social support groups to provide instant messages counseling for HIV/AIDS prevention.

3. The Proposed Approach

Fig. 1 shows the proposed system architecture. The process can be subdivided into four main steps: (1) data acquisition, (2) text preprocessing and features selection, (3) classification and evaluation, and (4) content analysis.

Data Acquisition

Yahoo! currently comprises the largest users and enjoys a popular reputation in Taiwan (Alexa.com, 2012). Yahoo Knowledge has about 300 categories of discussion communities. A Web crawler is developed to retrieve Web posts from Yahoo Knowledge in which there exists a major Chinese-based portal

Contact: Dr. C. Chiu, Dept. of Inf. Magt., Yuan Ze Univ., 135 Yuan-Tung Rd., Chungli, Taiwan, 886-34638800 ext 2601, 886-34352077, imchiu@saturn.yzu.edu.tw

for HIV/AIDS related information. All posts including follow-up discussions related to AIDS keywords were captured and manually labeled as "Concerned" or "Unconcerned".

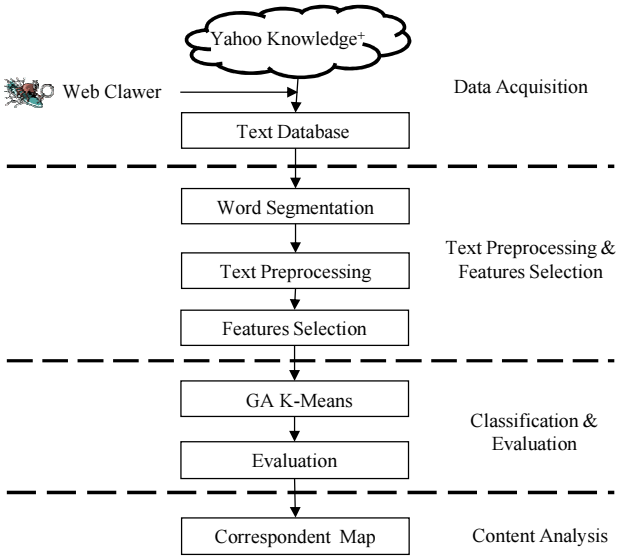

Figure 1: The Research Design of HIV/AIDS Posts Classification

Text Preprocessing and Features Selection

Chinese Knowledge Information Processing (CKIP), a Chinese word segmentation system developed by Academic Sinica, is applied to word segmentation and part-of-speech tagging that can be more feasible for the task of identifying Chinese phrases than the ones used in English. Table 1 shows both the original posts and the contents after words segmentation via online CKIP.

Table 1: The Example of Contents before and after Terms Segmentation

Before Segmentation	我和陌生人共用杯子喝東西, 陌生人用過的杯子沒有經過任何處理, 就直接讓我喝東西, 我會因此感染 HIV 嗎？
After Segmentation	我(Nh)和(Caa)陌生人(Na)共用(VC)杯子(Na)喝(VC)東西(Na), (COMMACATEGORY) 陌生人(Na)用(VC)過(Di)的(DE)杯子(Na)沒有(D) 經過 (VCL) 任何 (Neqa) 處理 (VC), (COMMACATEGORY) 就 (D) 直接 (VH) 讓 (VL) 我 (Nh) 喝 (VC) 東西 (Na), (COMMACATEGORY) 我 (Nh) 會 (D) 因此 (Cbb) 感染 (VJ)HIV(FW) 嗎 (T) ？ (QUESTIONCATEGORY)

The terms obtained after segmentation is tagged with syntactic labels. Synonyms were categorized into a single term. Stop words and common words that bear little or no content information, such as pronoun, preposition, conjunctions, interjection, human name, place name, time words, digitals, article, ask, reply etc. were filtered out. Finally, 140 candidate terms were adopted after text preprocessing and features selection process.

Term frequency (TF) is used to measure how often a word appears in a specific post to evaluate how important a word is in a collection terms. TF is considered as an important factor for features selection. Document frequency (DF) is the number of documents in which a term occurs. DF thresholding, the simplest technique with the lowest cost in computation, has better performance in Chinese text categorization than Information Gain, Mutual Information and $\chi 2$ statistic measure (Shan et al., 2003; Dai et al., 2004). In this study, candidate terms were selected based on TF ≥ 5 and DF ≥ 3.

The vector space model was adopted to represent documents as vectors, i.e. each document d is described by a numerical feature vector w(d). Thus a weight w(d, t) for a term t in document d is computed by term frequency tf (d, t) times inverse document frequency idf (t)－defined as idf (t) = log(N/n_t), where N is the size of the document collection D and n_t is the number of documents that contain term t. Length normalization is used to ensure that all documents have equal chances of being retrieved independent of their lengths.

Classification and Evaluation

In this research, a supervised GA/k-means for classification is proposed to detect the concerned HIV-related messages. The k-means algorithm is commonly used for clustering large data sets due to its relative computational efficiency and ease of implementation. It divides a group of data in multidimensional space into k groups by randomly choosing k centers and assigning data into different clusters through Euclidean distance. However, k-means algorithm which is an iterative and hill climbing clustering algorithms usually converge to a local minimum.

GA is a stochastic search and optimization technique based on the principles of evolution and natural genetics. The GA explores a complex space in an adaptive way, guided by the biological evolution of selection, crossover, and mutation. Since stochastic optimization approaches are good at avoiding convergence to a locally optimal solution, GA can provide near-optimal solutions for fitness function of an optimization problem.

Traditional clustering is used in unsupervised learning method. Some researchers have tried to apply supervised learning mechanism into clustering to generate those clusters that have high data densities and minimal impurity (Zeidat et al., 2005; Li and Ye, 2006). Additionally, Traditional clustering methods regard all data fields to be equally important when determining the partitions for a data set. Since we attempt to classify objects with high probability density in respect to a single class into the same cluster, this can be achieved by assigning different weights to those fields according to the significant relationships with the class labels. Choosing the appropriate set of weights can be regarded as a solution searching problem that can be tackled by GA. GA and k-means algorithms are integrated in attempt to adopt GA global optimization capability to search for maximum evaluation metrics of all clusters with the computed weights that are applied to the input attributes.

The process of supervised GA/k-means for classification is shown in Figure 2. Initial clusters are randomly generated from the GA by assigning the weights to input attributes. After a group of weights are generated, k-means is then activated. Major steps are illustrated as follows:

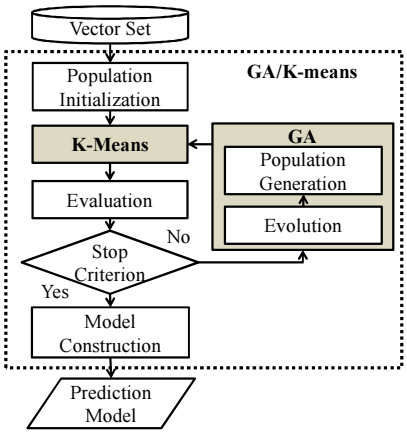

Figure 2: The Process of Supervised GA/k-means for Classification

K-means Module.

K-means algorithm employs an iterative approach to group the data into a pre-determined *k* number of clusters by minimizing the sum of squared Euclidean distance (SSE) of each point from its nearest cluster center. Given a set of n points in d-dimensional space R^d and an integer k, determine a set of *k* cluster centers in R^d. SSE is defined as $\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left\|x_{i,j}-c_j\right\|^2$ where $c_j=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{i,j}$ is the the centroid for cluster *j*, and $x_{i,j}$ is the *i*th data point in cluster *j* for $i = 1, 2, \ldots, n_j$.

Since the aim of supervised GA/k-means for classification is to help the k-means to partition the data into its different class labels. It requires a weighted metric by assigning greater weights to those fields which have a more significant relationship with those class labels.

$$\delta=\sum_{i=1}^{n}w_i\left(x_{i,j}-c_j\right)^2 \qquad (1)$$

δ is called a weighted Euclidean metric. w_i generated by GA is the weight assigned to the *i*th data.

GA Module.

The evolution of the possible solutions is guided by the fitness function that measures the optimality of a solution. Different fitness functions emerge for test data evaluation according to which test goal is pursued. The most frequent and basic measures for information retrieval effectiveness are accuracy and precision. Accuracy is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications. Precision measures the accuracy provided that a specific class has been predicted. In addition to measure the classification performance of model, the social support groups might pay close attention to identify the HIV-related messages that should be concerned. To achieve the above purposes, two different fitness functions are developed:

Fitness Function 1: Max. Accuracy. Let $S = \{s_1, s_2, s_3, \ldots, s_q\}$ represents that there are *q* different outcome classes that are pre-defined in advance. Class purity is used as a criterion to indicate the cluster quality that is measured by the percentage of instances of the majority class, i.e., the class that occurs most frequently in a cluster. The class purity of each cluster is calculated as:

$$\rho_j=\frac{n_{j,p}}{n_j},\ 1\le p\le q \qquad (2)$$

where

$n_{j,p}$: the number of majority instances of the *p*th class in the *j*th cluster.

n_j : the total number of instances in the *j*th cluster.

Fitness value is as the optimal sum-up purity for each cluster using the following formula:

Fitness function 1: $$\sum_{j=1}^{k}\left(\rho_j\times\frac{n_j}{n}\right) \qquad (3)$$

where $\frac{n_j}{n}$ is defined as the relevant weighted count in order to express the corresponding fractional presence of the *j*th cluster.

Fitness Function 2: Max. Precision. Fitness value is as the optimal sum-up purity of the *j*th cluster with majority instances of the *T*th target class using the following formula:

Fitness function 2: $$\frac{\sum_{j=1}^{k^T}n_{j,T}}{\sum_{j=1}^{k^T}n_j} \qquad (4)$$

Where

k^T : the number of clusters with majority instances of the *T*th target class.

$n_{j,T}$: the number of the *T*th target class in the *j*th cluster.

n_j : the total number of instances in the *j*th cluster.

The GA is continuously executed till various evaluation metrics are optimally realized. *K* has been tried with different values to examine its impact on the classification performance. At the end, each cluster is assigned with specific class label corresponding to the majority class of objects in the cluster.

Evaluation. The performance measures for both accuracy and precision are evaluated through an average 5-fold cross-validation process. The original sample data set is randomly partitioned into 5 subsets of nearly equal size. A single subset is used as the validation data set for testing purpose, and the remaining 4 subsets are used for the training purpose. The cross-validation process is repeated 5 times, with each of the 5 subsets used once as the validation data. The 5 results are averaged to obtain an aggregate measure to estimate the classification performance. The obtained GA/K-means classification performances are compared with those from Support Vector Machine (SVM), C5.0, Naïve Bayes (NB).

Content Analysis

To gain a better understanding the terminology used of HIV-related messages, the relationships of terminology used in "Concerned" and "Unconcerned" messages are examined. Furthermore, correspondence analysis is conducted to visually show the relationships between the variables under study, with

the distance on the map representing the correspondence (closeness). The closer proximity means greater perceived similarity (Greenacre, 1984 & 2006). This tool can provide importance information for social support groups to prejudge the messages that should be paid with more attention.

4. The Results

Data Description

This study adopted the keywords, AIDS and HIV, to retrieve a total of 2,083 posts from January 1, 2005 to December 31, 2011 from "Yahoo Knowledge" in Taiwan. Among the entire 2,083 posts, 470 (22.6%) posts are labeled as those require concerned.

Experimental Design

The number of clusters (k) is a question of resource allocation. Determining the decent number of clusters is typically made by a researcher of the application area. A common approach is to generate various values of k and compare them against some evaluation metrics. The selection of k value is a trade-off between computation time and accuracy. That is, a larger k value would increase the larger size of search space that leads to more computation process and vice versa. In general, k is set as more than the number of predicted outcome attributes and is applied on the data sets for different number of clusters. In this study k is set with 3, 5, 10, and 15, respectively.

The weights assigned to each attribute are bounded between 0 and 1 in order to limit the solution search space. The crossover rate, mutation rate, and population size for GA execution are set with 0.5, 0.1 and 100, respectively. The stopping condition is set when the improvement of the fitness value change is less than 0.01% in the last 20 generations.

Classification Performance

The experimental results of performance measures based on the different fitness functions are shown in Table 2. The classification accuracy rates GA/k-means based on maximum accuracy range from 80.2% to 82.2% while SVM with 82.9%, C5.0 with 81.7%, and NB with 77.8%. SVM outperforms other classifiers. GA/k-means at k=3 and k=5 outperform C5.0 but inferior at k=10 and k=15. Secondly, the precision rates of GA/k-means based on maximum precision range from 78.3% to 83.8% while SVM with 77.8%, C5.0 with 75.4%, and NB with 70.6%. GA/k-means outperforms other classifiers. The experimental results demonstrate that the proposed supervised GA/k-means for classification approach can achieve acceptable performance accompanied with its full flexibility to develop different fitness functions in accordance with the need of different requirements.

Contents Analysis

Figure 3 illustrates the comparison of the top 20 most commonly used terminology rate of "Concerned" and "Unconcerned" HIV-related messages respectively. It is shown that there are quite different terminology usage patterns between "Concerned" and "Unconcerned" HIV-related messages. Although there are 15 terms are the same in both types of messages, the terms frequency rate are obviously different. The "Concerned" posts are those potential subjects who have described engaging in sexual behavior. For example, "Oral Sex" (6.7% of the sum of features selected from "Concerned" messages) is commonly used in "Concerned" messages, comparatively, only 1.1% of that in "Unconcerned" messages. Comparatively, the most common used terms in "Unconcerned" messages are those who is at very low or no risk for infection fearing that they have the HIV, such as "needle", "mosquito", and "incubation period".

Table 2: Summary of the Experimental Results

Classifier	Max. Accuracy	Max. Precision
GA/k-means k= 3	82.2%	78.9%
GA/k-means k= 5	82.0%	83.8%
GA/k-means k=10	80.9%	78.7%
GA/k-means k=15	80.2%	78.3%
SVM	82.9%	77.8%
C5.0	81.7%	75.4%
NB	77.8%	70.6%

In summary, those terms used in "Concerned" HIV-related messages focus on the terms about "risky sexual behavior", "sexual partner", "sexual intercourse", and "condom" that are directly associated with sexual behavior. Comparatively, most of the messages with these terms including "body fluid", "symptom", "homosexuality", "AIDS" that belong "Unconcerned" messages are of no immediate need to be paid attention.

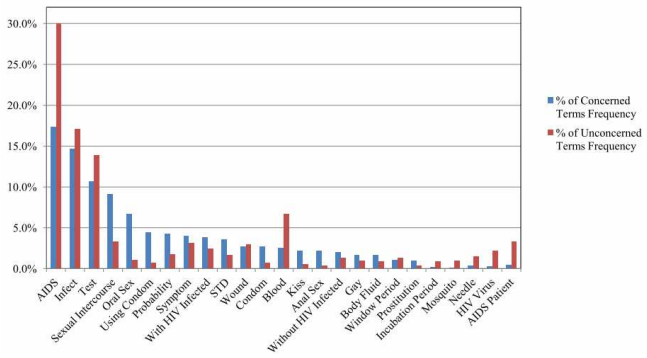

Figure 3: The Comparison of Terminology Used in "Concerned" and "Unconcerned" HIV-related messages

5. The Conclusion

People living with HIV usually surf on the Internet for seeking health information. However misinformation and unfounded claims are frequently encountered online. Although quality information cannot replace quality health services, people do need health information with enough depth to allow them to weight all the options for prevention and treatment.

In light of a borderless network that the Internet presents, t, it has become increasingly common for individuals to use online support groups to help manage their health problems. Web posts cannot be effectively managed due to the lack of adequate personnel and technical support. If the detection of concerned

messages from the target groups can be automatically identified, then appropriate information and the follow-up social support can be provided in more effective ways. This research employs the information retrieval and natural language processing techniques to collect and discover the knowledge from the web posts regarding HIV messages. In the detection of concerned messages, the proposed supervised GA/k-means for classification algorithm demonstrates its effectiveness and flexibility in classification tasks.

However, the supervised GA/k-means for classification algorithm may still bear some constraints. For example, k-means is not able to appropriately handle mixed types of data attributes, and result in longer computation time when many input attributes are involved. Other, automatic assigning appropriate k values can avoid manual trials, and thereby saving more time in model construction.

The labeling the data set is another limitation of our study. There were seven annotators joining this work to engage in annotating and tagging the test set with inter-annotator agreement. To avoid the tagging errors for the concerned HIV-related messages, the test set was re-validated by one of the authors who is a senior officer with over 20 years work experience in Department of Health. Since all annotators are not the professionals in the fields of HIV treatment and prevention, those messages might be annotated differently from professional HIV social support groups. Nevertheless, we propose a viable approach as an initiating such an pilot task that hopefully may draw out more attention from many other research groups and social support agencies to develop more robust and complete solutions to assist those people living with HIV.

With the aid of correspondence analysis, the most frequently used terms in “Concerned” HIV-related messages are directly associated with sexual behavior. Our finding can offer information regarding the distinguish patterns in web forum. Comparing the terminology used in “Concerned” and “Unconcerned” messages. The most common used terms in “Concerned” messages are focus on risky sexual behavior whereas “Unconcerned” messages are those who of worried well. In particular, the terminology used is quite different. Using information retrieval and text mining approaches can effectively distinguish “Concerned” and “Unconcerned” HIV-related messages.

Our research findings is believed to be useful to support health authorities or social support groups in providing immediate, complete, accurate, and personalized healthcare-related activities, encouraging more positive coping mechanisms to those people living with HIV and the potential patients.

References

Alexa [accessed January 15, 2013]. Available at http://www.alexa.com/topsites/coun-tries/TW

Bogart, L.M., C. Burton O, K. David, R. Gery, M. Debra A, E. Jacinta, and S. Mark A. 2008. "HIV-related stigma among people with HIV and their families: A qualitative analysis." AIDS and Behavior 12 (2): 244-54.

Chiu, C., K. Yungchang, L. Ting, and C. Yuchi. 2011. "Internet auction fraud detection using social network analysis and classification tree approaches." International Journal of Electronic Commerce 15 (3): 123-47.

Courtenay-Quirk, C., H. Keith J, D, Helen, F. Holly, M. Mary, K. Rachel, O. Ann, R. B.R.Simon, and H. Eileen. 2010. "Perceptions of HIV-related websites among persons recently diagnosed with HIV." AIDS Patient Care and STDs 24 (2): 105-15.

Dai, L., H. Haiyan, and C. Zong. 2004. "A Comparative Study on Feature Selection in Chinese Text Categorization." Journal of Chinese Information Processing 18(1): 26-32.

Greenacre, M. 1984. Theory and Applications of Correspondence Analysis. London: Academic Press. ISBN 0-12-299050-1.

Greenacre, M. 2006. "Tying Up the Loose Ends in Simple, Multiple and Joint Correspondence Analysis. " Keynote Address, COMPSTAT 2006, In Proceedings in Computational Statistics, (pp.163–186). Springer-Verlag.

Horvath, K.J., C-Q. Cari, H. Eileen, F. Holly, K. Rachel, M. Mary, O. Ann, and R. B.R. Simon. 2009. "Using the internet to provide care for persons living with HIV." AIDS Patient Care and STDs 23 (12): 1033-41.

Li, X., and N. Ye. 2006. "A supervised clustering and classification algorithm for mining data with mixed variables." IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans 36 (2): 396-406.

Mahajan A.P., S. Jennifer N., P. Vishal A., R. Robert H, O. Daniel, S. Greg, and C. Thomas J. 2008. "Stigma in the HIV/AIDS epidemic: A review of the literature and recommendations for the way forward." AIDS (London, England) 22 (2): 67-79.

Mo, P.K.H., and C. Neil S. 2008. "Exploring the communication of social support within virtual communities: A content analysis of messages posted to an online HIV/AIDS support group." Cyberpsychology and Behavior 11 (3): 371-4.

Morris, D., D. Elizabeth, S. Anton, B. Carol, and O. Annette. 2008. "Can people find patient decision aids on the Internet?" Patient Education and Counseling 73 (3): 557-60.

Samal, L., S. Somnath, C. Geetanjali, K. P. Todd, S. Rashmi K., S. Victoria, C. Jonathan, M. Richard D., and B. Mare Catherine, 2011. "Internet Health information seeking behavior and antiretroviral adherence in persons living with HIV/AIDS." AIDS Patient Care and STDs 25 (7): 445-9.

Shan, S.W., F. Shicong, and L. Xiaoming. 2003. "A comparative study on several typical feature selection methods for Chinese Web page categorization." Journal of the Computer Engineering and Application 39 (22): 146-8.

Zeidat, N., E. Christoph F, and Z. Zhenghong. 2005. "Supervised Clustering: Algorithms and Application." Suffolk University Law Review 25: 94-9.