Building adaptive user profiles by a genetic fuzzy classifier with feature selection

María J. Martín-Bautista and María-Amparo Vila
Dpt. of Computer Science and Artificial Intelligence
Granada University
Avda. Andalucía 37
18071 Granada, Spain
{mbautis, vila@decsai.ugr.es}

Henrik L. Larsen
Department of Computer Science
Roskilde University, P.O. Box 260
DK-4000 Roskilde, Denmark
hll@ruc.dk

Abstract-**In the present paper, a genetic algorithm is used to build user profiles from a collection of documents previously retrieved by the user. A fuzzy classification and a genetic term selection process provide a better utilization of valuable knowledge for genetic algorithms in order to get an improvement of the quality of the current and near future information needs in the areas of interest to the user. A gene in the chromosome of the genetic algorithm is defined by a term and a fuzzy number of occurrences of the term in documents belonging to the class of documents that satisfy the user's information need. In this way, the terms that allow the system to discern between good and bad documents are selected and stored as a part of the user's profile to be used in future queries to the system. The fuzzy classifier implements an inductive derivation of the current, experience based, interest profile in terms of an importance weighted conjunction of genes.**

I. INTRODUCTION

In the last years, some techniques mainly coming from Machine Learning have been exported to solve the problems of classification and rules generation in other disciplines, such as Data Mining and Knowledge Discovery.

The settling of the differences between the terms Web Mining and Text Mining comes from the aspects of the information access based on the user needs tasks versus the tasks of organization, classification and categorization of the textual information from the Web [22] or other sources. Therefore, we understand by *Web Mining* the results of the application of traditional mine tasks to the information access and retrieving processes in relation to the web user needs.

When a user retrieves documents from an information system (for instance, from the Internet), systems typically lack means for learning the user's interest profile. Hence, the user cannot use the previous queries to the system for future requests. A learning of the user's needs is becoming a fundamental stage in the process of information retrieval.

These needs may be represented by terms extracted from those documents that the user has evaluated as good ones. One of the main problems studied in this field is the construction of user profiles by means of the discovering of the most relevant and representative terms (features) which information filtering systems can use to determine the most useful information to a given user. We must distinguish between those terms that best represent the information user needs, and those that allow us to discern between relevant and non-relevant, that is, the discriminatory terms for a certain classification.

This problem can be tackled in the classification task by the Feature Selection (FS), one of the first stages of the classification process. By this preprocessing, the complexity of the problem is reduced by the elimination of irrelevant features to consider later in the classification stage. Several approaches with Genetic Algorithms (GAs) to solve the feature selection problem have been proposed in the literature [3, 7, 16, 19, 20, 21]. GAs is an adaptive search technique which improvement over random and local search methods has been demonstrated [6]. The relative insensitivity of the GA to noise, and the requirement of no domain knowledge make GAs be a powerful tool to identify and select the best subset of features to be used by the rule induction module in a classification system.

II. RELATED WORK

Several approaches using GAs related to user profiles construction topic can be found in the literature. In BEAGLE [5], the author builds a population of user profiles, which represent the best subset of keywords for distinguishing relevant documents from non-relevant ones.

There are some others approaches that use other techniques. Reference [2] combines different learning methods such as Rocchio, C4.5 and AQ15, and measures coming from Information Retrieval and Machine Learning, to evaluate the influence of text features on user profiles, partitioning the document set into relevant and non-relevant ones. In [14] a Bayesian classifier is used to define user

0-7803-5877-5/00/$10.00 © 2000 IEEE

profiles, and the expected information gain of the most informative terms is calculated for feature selection.

Collaborative filtering is seen as a classification task in [15]. The dimensionality of document terms is reduced by the selection of the most informative terms based on the singular value decomposition (SVD) of an initial matrix of user evaluations.

Other approach using GAs is presented in [10], where a user profile is built from the user preferences, represented by a population of chromosomes. Each chromosome is a vector of fuzzy genes, with each gene representing by a fuzzy set the number of term occurrences that characterizes preferred documents.

III. Problem Formulation

One of the first stages of the classification process is the Feature Selection (FS), by means of which the complexity of the problem is reduced by the elimination of irrelevant features to consider later in the classification stage.

The problem of FS in the framework of text databases, therefore, can be studied from two points of view:

1) If the documents are not previously classified, the selection of the most relevant features (terms in documents) give us those terms that best describe the documents.

2) On the other hand, if there is a previous classification (categorization) of the documents, a selection of the features would find the most discriminatory terms, that is, those terms that allow us to distinguish the different existent classes in a later stage.

   A study of the importance of the term reduction to improve some significant text categorization methods can be found in [23].

Genetic Algorithms (GAs) [6] has been revealed as a powerful tool to identify and select the best subset of features to be used by the rule induction module in a classification system. In a population of chromosomes $C_1,...,C_p$, each chromosome would represent a potential solution of the problem, that is, a minimal relevant set of features for a certain sample set. The relative insensitivity of the GA to noise, and the requirement of no domain knowledge make GAs a powerful tool for optimizing the process of classification, in particular when the domain knowledge is costly to exploit or unavailable [21].

A. *Fuzzy Set Genes*

Most of the techniques used in text classification are determined by the occurrences of the words (terms) appearing in the documents, combined with the user feedback over the documents retrieved. However, in our model, the most relevant terms will be selected from a previous fuzzy classification given by the genetic algorithm guided by the user feedback, but using techniques from Machine Learning.

In our model, a gene in a chromosome is defined by a term and a fuzzy number η of occurrences of the term in documents belonging to the class of documents that satisfy the user's information need. This fuzzy number is characterized by the membership function $\mu_{\equiv\eta}$.

Let $\Delta=\{D_1, ..., D_m\}$ be the set of documents evaluated by the user, and let $u_i \in [0, 1]$ be the user's evaluation of the document D_i, $i=1, ..., m$, meaning the degree to which the user finds that the document D_i satisfies his needs. We shall assume that an evaluation $u>0.5$ indicates a good document, with $u=1$ representing a highly relevant document, while $u \leq 0.5$ indicates a bad document, with $u=0$ representing a document which is not relevant at all.

We will, without loss of generalization, assume that Δ is ordered decreasingly by u_i in $\Omega = \{D_1,...,D_k,D_l,...,D_m\}$. The subsets $\{D_1,...,D_k\}$ and $\{D_l,...,D_m\}$ contains, respectively, the good and the bad documents.

Let $T = \{t_1,...t_n\}$ be the set of terms extracted from Ω, and x_{ij} the relative frequency (which means it is normalized) of term t_j in document D_i. The estimation of the expected value of x_j in good and bad documents is given by the weighed average of the relative occurrence frequency of a term t_j in the good and bad document by $\bar{x}_j$ and $\bar{x}'_j$, respectively:

$$\bar{x}_j = \frac{\sum_{i=1}^{k}(u_i \cdot x_{ij})}{\sum_{i=1}^{k} u_i} \qquad \bar{x}'_j = \frac{\sum_{i=l}^{m}((1-u_i)\cdot x_{ij})}{\sum_{i=l}^{m}(1-u_i)} \tag{1}$$

Therefore, a gene G is a pair $G(t,\eta)$, where t is a term, and η is a fuzzy number characterized by the membership function $\mu_{\equiv\eta}$ as follows:

$$\mu_{\equiv\eta}(x) = \begin{cases} 0 & x = 0 \\ e^{-\frac{1}{2}\left(\frac{x-\eta}{\sigma}\right)^2} & x > 0 \end{cases} \tag{2}$$

The membership function is a Gaussian one, assuming that the relative term occurrence frequency is normal distributed $N(\eta,\sigma^2)$, where the parameters η, $\sigma \in \Re^+$, with $\eta = \bar{x}_j$ and σ_j^2 defined by:

$$\sigma_j^2 = \frac{\sum_{i=1}^{k}(u_i x_{ij} - \bar{x}_j)^2}{\sum_{i=2}^{k} u_i} \tag{3}$$

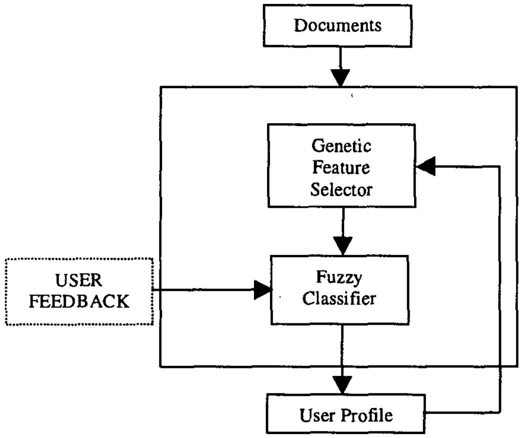

Fig. 1. A general view of the system.

IV. Description of the System

Two main modules can be distinguished in our system, namely the genetic feature selection and the fuzzy classifier (see Figure 1). We describe these modules in the following.

A. The Genetic Selector Module

In the presented model, the Genetic Selector Module allows us, given a previous classification, to select the most discriminatory terms for a certain classification. In this case, we can select the discriminatory terms as derived from the fuzzy classification of the documents previously retrieved by the user. Through identifying and applying these terms, the system learns the user's interests, thus improving the quality of the fuzzy classification process when the user makes a new query.

1) Fitness Function: The fitness function to maximize is based on the discriminatory power of every term through all the population as well as the accumulated discerning in the chromosomes.

$$F(y) = S^{*}(C_y, V_h) + q_y \tag{4}$$

where $S^{*}(C_y, V_h)$ represents a similarity measure between a chromosome and the discriminatory vector. Initially, two different measures of similarities will be considered. The first one is based on the Jaccard's score $S^{J}(C_y, V_h)$, while the second is calculated from the Dice's coefficient $S^{D}(C_y, V_h)$ [18]. Both measures will be weighted by an individual evaluation of the discriminatory power $G(t_j)$ of every term in the resulting vector:

$$S^{J}(C_y,V_h) = \frac{\sum_{j=1}^{|T|} C_y(t_j)\cdot V_h(t_j)\cdot G(t_j)}{\sum_{j=1}^{|T|} C_y(t_j) + \sum_{j=1}^{|T|} V_h(t_j)\cdot G(t_j) - \sum_{j=1}^{|T|} C_y(t_j)\cdot V_h(t_j)\cdot G(t_j)} \tag{5}$$

$$S^{D}(C_y,V_h) = \frac{2\cdot\left[\sum_{j=1}^{|T|} C_y(t_j)\cdot V_h(t_j)\cdot G(t_j)\right]}{\sum_{j=1}^{|T|} C_y(t_j) + \sum_{j=1}^{|T|} V_h(t_j)\cdot G(t_j)} \tag{6}$$

where:

- $C_y(t_j) = \mu_{\equiv}(t_j)$, as defined in (2), and represents the relative frequency of every term t_j appearing in the gene of the chromosome C_y.
- $V_h(t_j) = \mu_{\equiv}(D_i(t_j))\cdot\mu_{\equiv}(D_k(t_j))$ is the discriminatory vector based on the comparison of the term t_j appearing in documents D_i and D_k, where $i,k \in |\Omega|$.
- $G(t_j) = \frac{g(t_j)}{\sum_{i=1}^{|T|} g(t_i)}$, with $g(t_j)$ being the accumulated value of t_j for all the discriminatory vectors, and q_y represents the capability of the chromosome itself, calculated by adding the accumulated discriminatory values of every term presented into the chromosome, as it is shown below.

$$q_y = \frac{\sum_{j=1}^{|T|} C_y(t_j) g(t_j)}{\sum_{j=1}^{|T|} g(t_j)} \tag{7}$$

B. The Fuzzy Classifier Module

The fuzzy classifier implements an inductive derivation of the current, experience based, interest profile in terms of an importance weighted conjunction of genes.

One of the most remarkable models of representation of documents is the vector space model, in which the terms of queries and documents are the components of vectors. These terms may be viewed as features which binary values 0 and 1 indicate the absence or presence of a term in the document, respectively. However, a more accepted representation of a document comes from a weighted vector, where every position indicates the term frequency, that is, the number of times that the term appears in the document, or the term importance indicator calculated by the product of the term frequency and the inverse document frequency, indicating the term frequency of a word in a document relative to the entire collection of documents [17].

A document score in a gene is been given by $S_i(G_j) = \mu_{\equiv\eta}(x)$, where x is the relative occurrence frequency of the term t_j in the document D_i, and $j=1,\ldots,ChrmLength$ (length of the chromosome).

Obviously, the aggregation of the document score in the genes to its overall score for the chromosome should apply an AND-like operator $\otimes$.

$$S_i(C_k) = \overset{chrmlength}{\underset{j=1}{\otimes}} S_i(G_j) \tag{8}$$

where $k=1,\ldots,popSize$ (size of the population).

On the other hand, the aggregation of the overall scores for the chromosomes in the population Z should be an OR-like operator $\oplus$.

$$S_i(Z) = \overset{popsize}{\underset{k=1}{\oplus}} S_i(C_k) \tag{9}$$

(From now, we will call $S_i(Z)$ simply S_i).

The selection of these operators $\otimes$ and $\oplus$ for the production system should be based on their evaluation in an experimental setting.

1) *Document Evaluation*

The evaluation of the documents by the population of the Genetic Algorithm in this stage will be guided by the maximization of a function based on the combination of fuzzy precision and fuzzy recall.

The fuzzy recall-precision measure is applied in experimental situations where documents in the collection queried have all been evaluated by the user. Let u_i and S_i be, respectively, the (expert) user's and the system's evaluation (population score) of the document $D_i \in \Omega$, $i=1,\ldots m$.

We define the fuzzy recall-precision τ by:

$$\tau = \rho^{v_1} \psi^{v_2} \tag{10}$$

where ρ is the fuzzy recall, and ψ is the fuzzy precision defined by:

$$\rho = \frac{\sum_{i=1}^{m} min(u_i, S_i)}{\sum_{i=1}^{m} u_i} \qquad \psi = \frac{\sum_{i=1}^{m} min(u_i, S_i)}{\sum_{i=1}^{m} S_i} \tag{11}$$

and $v_1, v_2 \in [0,1]$ are the importance weights of high recall and high precision, respectively.

For and Internet information retrieval system, we expect that precision is more important than recall, and therefore $v_1 < v_2$.

Notice, that the fuzzy recall-precision τ measures how close the system's evaluation is to the user's evaluation.

V. Experimental Settings

In order to test the system, both the operators of aggregation and the parameters of the algorithm must be determined.

As it was mentioned in the previous section, the operator of aggregation of the document score for all the genes should be an AND-like operator. Initially, we will set this operator to the minimum of the gene's scores of a document. However, for the aggregation of the document score in a chromosome, the average of all the chromosomes scores will be calculated instead of a more OR-like operator, since the maximum was tested in the earlier experiments, and the presence of a very good chromosome in the population gave a good evaluation of a document without taken account the fitness of the rest of the population, which could be not so good.

Regarding the parameters of the GA, they are being fixed as follows:

- *Number of Generations*: 1000
- *Size of the population :* 80
- *Chromosome Length:* 10
- *Probability of Crossover:* 0.6
- *Probability of Mutation:* 0.1

The collection of documents is corresponding to the follow query in the INSPEC database of Jul-Sep 1998: *"Information Retrieval and Classification"*.

The number of documents retrieved was 22, and the number of different terms extracted (after removing stop-list words and stemming) is 616. The number of total terms is 1074.

Let suppose that the user needs are oriented to those documents (in the collection of the 22 documents retrieved previously), regarding topics such as *Web* and *Internet*, to which the user will give the higher evaluation. Since the previous query must be refined specially in terms of precision, the importance weights of the fuzzy recall (v_1) and fuzzy precision (v_2) will be set as follows:

$v_1 = 0.67 \qquad v_2 = 1$

The performance of the system is shown in Table 1.

VI. Concluding Remarks

The presented model has been revealed as a valid tool for the construction of user profiles and classification of documents ranked by relevance. The good performance of the implemented system demonstrates the necessity of a preprocessing stage in the classification process and the Web Mining task.

The use of Genetic Algorithms, a technique relatively insensitive to noise, aids in the searching of the huge space of terms extracted from the documents retrieved.

The user profile is built from the population of the GA, leaded by the best chromosome.

TABLE I
FUZZY RECALL-PRECISION IN QUERY "*INFORMATION RETRIEVAL AND CLASSIFICATION*" WITH *WEB* AND *INTERNET* AS USER PREFERENCES.

Fitness Function	***Fuzzy Recall-Precision***	***Fuzzy Recall***	***Fuzzy Precision***
Fuzzy Value	0.680403	0.762738	0.630512
Jaccard Score	0.724196	0.823568	0.664702
Jaccard+Discrimin.	**0.759182**	**0.889671**	**0.683005**
Dice Coefficient	0.700840	0.717755	***0.689785***
Dice+Discrimin.	0.742266	***0.902801***	0.651436

In this way, a term selection is carried out as a preprocessing stage in the classification of documents, where not only the terms best describing the documents, but also those terms more discriminatory for a certain classification are found.

Finally, the evaluation of the documents by means of the fuzzy recall-precision measure allow to obtain a ranked list of documents ordered by relevance, which can be compared to the relevance of such documents given by the user previously.

Two fitness functions have been proposed, based on the similarity between the chromosome and a discriminatory vector calculated by the Jaccard and Dice coefficients, and weighted by the discriminatory power of such value and the chromosome itself. The Jaccard's score weighted by the mentioned discriminatory factor has been revealed as the best fitness function to get the highest fuzzy recall-precision value.

VII. FUTURE WORK

Some directions for further research may be oriented to the extension of the experimental stage, in order to study the performance of the system when the set of documents, coming from other retrieval systems (also the Web) is larger.

Other aggregation operators among the genes in the chromosome, and among all the chromosomes in the population must be considered.

Finally, a more deep study of the influence of the genetic parameters in the system performance must be also carried out in future works.

REFERENCES

[1] J.E. Baker (1985) "Adaptive Selection Methods for Genetic Algorithms". In *Proc. on the First International Conference on Genetic Algorithms and their applications,* pp.101-111, Grefenstette, J.J. (ed). Hillsdale, New Jersey: Lawrence Earlbaum.

[2] F. Bloedorn, I. Mani, and T.R. MacMillan (1996). "Machine Learning of User Profiles: Representational Issues". In *Proceedings of AAAI'96*, . pp. 433-438. Portland, OR

[3] K.J. Cherkauer and J.W. Shavlik (1996) "Growing Simpler Decision Trees to Facilitate Knowledge Discovery". In *Proc. Of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96),* pp.315-318. Portland, OR: AAAI Press.

[4] M. Dash, and H. Liu (1997). "Feature selection for Classification". In *Intelligent Data Analysis, vol 1, no.3.*

[5] S. Ferguson (1995). "BEAGLE: A genetic algorithm for Information Filter Profile Creation". *Technical Report CS-692.*University of Alabama.

[6] J.H. Holland (1992) *Adaptation in Natural and Artificial Systems.* Massachusetts: MIT Press.

[7] J.D. Kelly and L. Davis (1991) "Hybridizing the Genetic Algorithm and the K Nearest Neighbors Classification Algorithm". In *Proc. of the Fourth International Conference on Genetic Algorithms and their Applications (ICGA),* pp.377-383.

[8] D.H. Kraft, F.E. Petry, B.P. Buckles, and T. Sadasivan (1995). Applying genetic algorithms to information retrieval systems via relevance feedback. In P. Bosc & J. Kacprzyk (Eds.), *Fuzziness in Database Management Systems* (pp. 330-344). Germany: Physica-Verlag.

[9] H.L. Larsen, T. Andreasen, and H. Christiansen (1998) "Knowledge Discovery for Flexible Querying". In Christiansen, H., Andreasen, T. and Larsen, H.L. (Eds.) *Flexible Query Answering System.* Lecture Notes in Artificial Intelligence, vol. 1495, pp. 227-235. Berlin: Springer-Verlag.

[10] M.J. Martín-Bautista, H.L. Larsen and M.A. Vila (1999) A "Fuzzy Genetic Algorithm Approach to an Adaptive Information Retrieval Agent". *Journal of the American Society for Information Science,* vol.50, number 9, pp.760-771.

[11] M.J. Martín-Bautista and M.A. Vila (1998) "Applying Genetic Algorithms to the Feature Selection Problem in Information Retrieval". In *Lecture Notes On Artificial Intelligence (LNAI), 1495.* Springer-Verlag.

[12] M.J. Martín-Bautista and M.A. Vila (1999) "A Survey of Genetic Feature Selection in Mining Issues". In *Proc. of IEEE Conference on Evolutionary Computation. (CEC'99),* vol.2, pp.1314-1321. July 1999, Washington.

[13] M. Mitchell (1996) *An Introduction to Genetic Algorithms.* Cambridge, MA: MIT Press.

[14] M. Pazzani and D. Billsus (1997) "Learning and Revising User Profiles: The identification of interesting web sites". *Machine Learning 27,* pp. 313-331.

[15] M. Pazzani and D. Billsus (1998) "Learning Collaborative Information Filters". In *Proc. of the Fifteenth International Conference on Machine Learning.* San Francisco, CA: Morgan Kauffman Publishers.

[16] W.F. Punch, E.D. Goodman, M. Pei, L. Chia-Sun, P. Hovland and R. Enbody (1993) "Further Research on Feature Selection and Classification Using Genetic Algorithms". In *Proc. of the Fifth International Conference on Genetic Algorithms and their Applications (ICGA),* pp.557-564. San Mateo, CA: Morgan Kauffmann Publishers.

[17] G. Salton (1989) Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading, PA: Addison-Wesley.

[18] G. Salton and M.J. McGill (1983) *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

[19] W. Siedlecki and J. Sklansky (1989) "A Note on Genetic Algorithms for Large-Scale Feature Selection". *Pattern Recognition Letters 10,* pp.335-347, North Holland: Elsevier Science Publishers.

[20] J.E. Smith, T.C. Fogarty and I.R. Johnson (1994) "Genetic Selection of Features for Clustering and Classification". In *IEE Colloquium on Genetic Algorithms in Image Processing & Vision,* IEE Digest 1994/193, London.

[21] H. Vafaie and K. De Jong (1992) "Genetic Algorithms as a Tool for Feature Selection in Machine Learning". In *Proceeding of the 4th International Conference on Tools with Artificial Intelligence,* Arlington, VA, November.

[22] M.R. Wulfekulher and W.F. Punch (1998) "Finding Salient Features for Personal Web Page Categories". In *Hyper Proc. of the Sixth International World Wide Conference.*

[23] Y. Yang and J. Wilbur (1996) "Using Corpus Statistics to Remove Redundant Words in Text Categorization". *Journal of the American Society for Information Science, vol.47(5).*pp.357-369. John Wiley.

[24] L.A. Zadeh (1965). Fuzzy Sets. *Information and Control, 83,* 338-353.