A genetic fuzzy classifier to adaptive user interest profiles with feature selection

María J. Martín-Bautista **María-Amparo Vila** **Henrik L. Larsen**

Dpt. of Computer Science and Artificial Intelligence
Granada University
Avda. Andalucía 37
18071 Granada, Spain
{mbautista, vila@decsai.ugr.es}

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

Abstract

We propose a method for guiding genetic algorithms for information retrieval by fuzzy classification and a genetic feature selection process of terms from documents evaluated by the user. The fuzzy classifier implements an inductive derivation of the current, experience based, interest profile in terms of an importance weighted conjunction of genes. A gene is defined by a symbol and a fuzzy number of occurrences of the symbol in documents belonging to the class of documents that satisfy the user's information need. Once the classification of the documents is made, a genetic selection from the most discriminatory terms is carried out. 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 classification and term selection processes provide a better utilization of valuable knowledge for genetic algorithms in order to get an improvement of the quality of the estimates of the current and near future information needs in the areas of interest to the user.

Keywords: User profiles, Fuzzy Classification, Feature Selection, Genetic Algorithms, World Wide Web.

1 Introduction

The classification of documents in Web servers and document bases is a determinant aspect in Information Retrieval. The efficiency of the different methods utilized in query and matching processes is usually diminished by both a poor classification of the documents and a lack of personalization in the representation of the user's needs.

The classification problem has been widely studied in others disciplines such as Numerical Taxonomy and Machine Learning, and many techniques generated in these fields are being exported to solve the problems of classification and rules generation in Data Mining and Knowledge Discovery in all kinds of information systems, including Web and Textual ones.

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 (Wulkefuhler & Punch, 1998) 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), most of actual systems have a lack of building a user profile. Hence, the user can not 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 no-relevant, that is, the discriminatory terms for a certain classification.

2 Problem Formulation

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.

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 $u=0.5$ is neutral, while $u>0.5$ indicates a good document, $u=1$ representing a highly relevant document, while $u<0.5$ indicates a bad document, $u=0$ representing a document which is not relevant at all.

We will, without loss of generalization, assume that Δ is ordered decreasingly by u_i:

$$\Delta = \{D_1, \ldots, D_k, D_{k+1}, \ldots, D_{l-1}, D_l, \ldots, D_m\}$$

such that the subsets $\{D_1, \ldots, D_k\}$, $\{D_{k+1}, \ldots, D_{l-1}\}$, and $\{D_l, \ldots, D_m\}$ contains, respectively, the good, the neutral and the bad documents.

Let $T = \{t_1, \ldots t_n\}$ be the set of symbols extracted from Δ, and x_{ij} the relative frequency of symbol 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 symbol t_j in the good and bad document by $\bar{x}_j$ and $\bar{x}'_j$, respectively:

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

In our model, a gene in a chromosome is defined by a symbol $\mu_{\tilde{=}\eta}$ and a fuzzy number of occurrences of the symbol in documents belonging to the class of documents that satisfy the user's information need. Then G is a pair $G(t, \tilde{\eta})$, where t is a term, and $\tilde{\eta}$ is a fuzzy number characterized by the membership function $\mu_{\tilde{=}\eta}$ .as follows:

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

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

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

3 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.

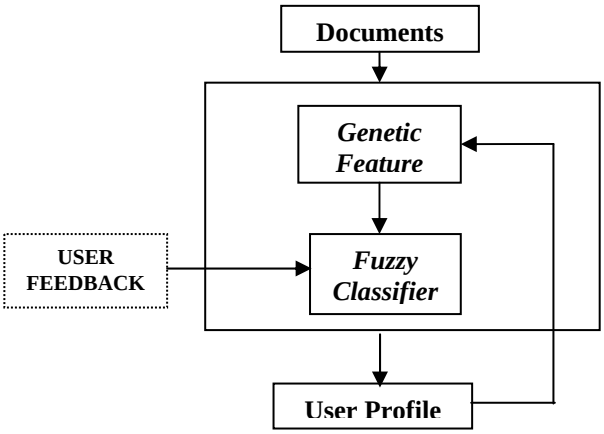

Figure 1. A view of the system.

3.1 The Genetic Selector Module

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 which describe the documents better.
2) On the other hand, if there is a previous classification (categorization) of the documents, a selection of the features would carry out the search of the most discriminatory terms, that is, those terms which allow 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 (Yang and Wilbur, 1996).

Given a previous classification, this module allows us 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.

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.

$$p_y = s_y + q_y \qquad (4)$$

where s_y represents the similarity between a chromosome and the discriminatory vector, and it is based on the

Jaccard's score (Salton & McGill, 1983) weighted by an individual evaluation of every term:

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

where:

- $C_y(t_j)=\mu_{\cong}(t_j)$, as defined in (2), and represents the relative frequency of every term t_j appearing in the gene of the chromosome C_y.
- $V_h(t_j)=\mu_{\cong}(D_i(t_j))\cdot\mu_{\cong}(D_k(t_j))$ is the discriminatory vector based on the comparision of the term t_i appearing in documents D_i and D_k.
- $G(t_j)=\frac{g(t_j)}{\sum_{j=1}^{|T|}g(t_j)}$, with $g(t_i)$ being the accumulated value of t_i 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.

$$q_y=\frac{\sum_{j=1}^{|T|}C_y(t_j)\cdot g(t_i)}{\sum_{j=1}^{|T|}g(t_j)} \quad (6)$$

3.2 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 (Salton, 1989).

A document score in a gene is been given by $\alpha=\mu_{\cong\eta}(x)$, where x is the number of occurrences of the symbol t in the document.

Obviously, the aggregation of a document score in the genes of a chromosome query to its overall score for the chromosome query should apply an AND-like operator. On the other hand, the aggregation of the document-in chromosome scores in the whole population should be an OR-like operator. The selection of these operators for the production system should be based on their evaluation in an experimental setting.

Fitness Function

The evaluation of the Genetic Algorithm in this stage will be guided by the maximization of the fitness function, calculated from 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 $\Omega=\{\omega_1,\ldots,\omega_{n_\Omega}\}$ be the collection queried, and let u_i and s_i be the (expert) user's and the system's evaluation of the document ω_i.

We define the fuzzy recall-precision τ by:

$$\tau=\rho^{v_1}\psi^{v_2} \quad (7)$$

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

$$\rho=\frac{\sum_{i=1}^{n_\Omega}min(u_i,s_i)}{\sum_{i=1}^{n_\Omega}u_i} \qquad \psi=\frac{\sum_{i=1}^{n_\Omega}min(u_i,s_i)}{\sum_{i=1}^{n_\Omega}s_i} \quad (8)$$

and v_1, v_2 are the importances 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 applied to the subset of Ω comprising the documents evaluated by the user in answer retrieved by the system, the fuzzy recall-precision τ measures how close the system's evaluation is to the user's evaluation.

4 Related Work

Several approaches using GAs related to this topic can be found in the literature. In BEAGLE (Ferguson, 1995). The author builds a population of user profiles which represent the best subset of keywords that allow us to discern among all the documents set those to be relevant.

There are some others approaches that use other techniques. Bloedorn et al. (Bloedorn, Mani & MacMillan, 1996), combine 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 (Pazzani and Billsus, 1997), a Bayesian classifier is used to define user profiles, and the expected information gain of the most informative terms is calculated as feature selection stage.

Collaborative filtering is seen as a classification task in (Pazzani and Billsus, 1998). 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 (Martín-Bautista, Larsen and Vila, 1998), where a user profile is built from the user preferences, represented by a population of chromosomes. Each chromosome is a vector of fuzzy genes, where every gene represents by a fuzzy set

the number of occurrences that a term should have in the document required by the user.

References

Baker, J.E. (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.

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

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

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

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

John, G.H., Kohavi, R. and Pfleger, K. (1994) "Irrelevant Features and the Subset Selection Problem". In *Proc. of the Eleventh International Conference on Machine Learning*, pp.121-129. San Francisco, CA: Morgan Kauffmann Publishers.

Kraft, D.H., Petry, F.E., Buckles, B.P., & Sadasivan, T. (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.

Langley, P. (1994) "Selection of Relevant Features in Machine Learning". In *Proc. of the AAAI Fall Symposium on Relevance*. New Orleans, LA: AAAI Press.

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

Martín-Bautista, M.J., Larsen, H.L. and Vila, M.A. (1999) A "Fuzzy Genetic Algorithm Approach to an Adaptive Information Retrieval Agent". *Journal of the American Society for Information Science* (Toappear).

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

Martín-Bautista, M.J. and Vila, M.A. (1998) "A Survey of Genetic Feature Selection in Mining Issues". In *Proc. of Conference on Evolutionary Computation. (CEC'99).* July 1999, Washington. (To appear).

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

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

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

Quinlan, J.R. (1992) *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

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

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

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

Vafaie, H. and De Jong, K., (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.

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

Yang, J., Pai, P., Honavar, V. and Miller, L. (1998) "Mobile Intelligent Agents for Document Classification and Retrieval: A Machine Learning Approach". In *Proc. of the Fourteenth European Meeting on Cybernetics and Systems Research (EMCSR'98).* Vienna, Austria.

Yang, Y. and Wilbur, J. (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.

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