Using Genetic Feature Selection for Optimizing User Profiles

H.L. Larsen*, N. Marín**, M.J. Martín-Bautista*** and M.A. Vila**
P.O. Box 260. DK-4000 Roskilde, Denmark
Avda. Andalucia 37. 18071 Granada, Spain
*** mbautis@decsai.ugr.es

Abstract

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. A feature selection process is carried out through a Genetic Algorithm in order to find the most discriminatory terms to be stored as the user profile.

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

1 Introduction

Classification of documents in Web servers and document bases is essential in Information Retrieval. The efficiency of the different approaches to querying and matching in Information Retrieval Systems (IRS) is typically limited, due to both a poor classification of the documents and a lack of personalization in the representation of the user’s needs.

An important issue in this situation 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. In general, current IRSs do not provide any means for building user profiles based on the experienced results from earlier query-answering sessions with the user. Hence, the user cannot utilize results previous retrieval tasks for future requests. Thus, a learning of the user’s needs is becoming a fundamental issue in the process of information retrieval. These needs may be represented by terms extracted from those documents that the user has evaluated as relevant.
In selecting terms (features) for the user profile, we must distinguish between the terms that best represent the user’s information needs, and those that allow us to discern between relevant and non-relevant, that is, the discriminatory terms for a certain classification.

In our system, a Genetic Algorithm is used to find the most discriminatory terms. A gene in a chromosome is defined by a term identifier and a fuzzy number of occurrences of the term in documents belonging to the class of documents that satisfy the user’s information needs. 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 (relevant) and bad (non-relevant) 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 for improving the quality of the estimates of the current and near future information needs in the areas of interest to the user.

2 Genetic Algorithms and the Feature Selection Problem

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 through the elimination of irrelevant features that may be considered later in the classification stage.

The selection must be carried out without losing information from the original set, once the irrelevant or noise features have been eliminated. Moreover, the selection must not diminish the performance of the classification system.

Let consider a set of examples $E$ described by a set $F$ of variables, each variable referring to a feature of the sample object to be classified. The objects belong to different classes, and the aim is to find a subset $R$ of the most relevant features for the classification task. The relevance of a feature is defined as the ability to distinguish among the examples belonging to different classes. This ability must be higher than a pre-fixed threshold.

The first experiments of feature selection were carried out in the field of Machine Learning, where most of the classification and feature reduction methods have been developed (Liu and Motoda, 1998).

The ability and versatility of the GAs as a tool for solving the feature selection problem have been widely analysed in several works (Punch et al., 1993), (Siedlecki and Sklansky, 1989), (Vafaie and DeJong, 1992). Other optimisation techniques have been utilised to solve the feature selection problem (Liu and Motoda, 1998). However, some studies of comparison among such techniques have shown the advantages of the GAs, especially in problems with a huge amount of features, as typically occurring in the framework of documents with terms. Moreover, the good performance of the GAs compared to the Branch and Bound and the Greedy techniques are shown in (Siedlecki and Sklansky, 1988) and (Vafaie and Imam, 1994).
The solution of the feature selection problem using GAs has been named “Genetic Feature Selection” (Martin-Bautista and Vila, 1999). In this case, supposing we have the set of example objects $E$ and the set of features $F$, we use a population of chromosomes $C_1, ..., C_H$ to represent the solution of the problem, where $H$ is the size of the population. Therefore, each chromosome represents a potential solution of the problem, i.e., a subset of the features comprising the most relevant features for the classification of the examples.

2.1 The Feature Selection Problem in a Information Retrieval framework: Term Selection

In the framework of document classification, we can identify the documents as example objects to be classified, and the terms as features. Thus, the problem can be studied from two points of view:

1. If the documents are not previously classified, the selection of the most relevant features gives us the terms that best describe the documents in the class.

2. On the other hand, if the documents have been classified (categorized) previously, a selection of the features would identify the most discriminatory terms, that is, the 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 (Yang and Wilbur, 1996).

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 representing the best subset of keywords for distinguishing relevant documents from non-relevant ones.

There are some others approaches that use other techniques. Bloedorn et al. (Bloedorn, Mani and 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 for the feature selection process.

(Pazzani and Billsus, 1998) see collaborative filtering as a classification task. 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.

(Martin-Bautista, Larsen and Vila, 1998) presented an approach using GAs, 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 characterizes the documents considered relevant by the user.
3 Problem Formulation

In this paper, we present a scheme for maintaining experience based knowledge on user preferences in user profiles. In querying, the user expresses his preferences through selecting terms related to the required information, joined with logical connectives. However, the user is usually not completely satisfied with the answer, and may indicate the subset of the documents presented that he finds more interesting. This feedback can be transmitted to the system in two different ways.

- On the one hand, the user may query the system again through a new or a modified query. The problem with this method is that it is difficult for the user to give a useful expression of his needs in a query.

- On the other hand, the user feedback may be given by the evaluation of some of the documents in the first set retrieved.

In this last case, we must establish a scheme for defining the representation of the terms based on the ability to describe the relevance of the documents. We introduce such scheme in the following.

Let $\Theta = \{D_1, ..., D_m\}$ 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. The relevance threshold from which we consider a document relevant is $\alpha \in [0, 1]$, initially fixed to 0.5. We shall assume that an evaluation $u_i > 0.5$ indicates a good document, with $u_i = 0$ representing a highly relevant document, while $u_i < 0.5$ indicates a bad document, with $u_i = 0$ representing a document that is not relevant at all.

We will, without loss of generalization, assume that $\Theta$ is ordered decreasingly by $u_i$ in $\Theta = \{D_1, ..., D_k, D_{k+1}, ..., D_m\}$ with the subsets $\Theta_R = \{D_1, ..., D_k\}$ and $\Theta_{NR} = \{D_{k+1}, ..., D_m\}$ containing the relevant documents and the non-relevant documents, respectively.

Let $T = \{t_1, ..., t_n\}$ be the set of terms extracted from document base $\Theta$, and $x_{ij}$ the relative (normalized) frequency of term $t_j$ in document $D_i$. The estimation of the expected value of $x_{ij}$ in good and bad documents is given by the weighted average of the relative occurrence frequency of a symbol $t_j$ in the collection:

$$\bar{x}_j = \frac{\sum_{i=1}^{m} (u_i \cdot x_{ij})}{\sum_{i=1}^{m} u_i}$$  \hspace{1cm} (1)

The knowledge of the system about the user preferences is kept in the population of a Genetic Algorithm. A gene in a chromosome is defined by a term and a fuzzy number $\bar{\eta}$ 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 $\mu_{x \in \bar{\eta}}$ defined by:

$$\mu_{x \in \bar{\eta}}(x) = \begin{cases} 0 & x = 0 \\ \exp\left(-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right) & x > 0 \end{cases}$$  \hspace{1cm} (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 \( \eta \in [0,1] \) and \( \sigma > 0 \) are calculated as follows.

The estimation of the expected value of \( x_j \), represented by \( \eta \), in relevant and non-relevant documents (following the classification given by the user initially) is given by the weighted average of the relative occurrence frequency of a term \( t_j \). The parameter \( \sigma \) will be given by the standard deviation of the expected value \( \eta \).

\[
\sigma_j^2 = \frac{\sum_{i=1}^{m} u_i (x_{ij} - \bar{x}_j)^2}{\sum_{i=1}^{m} u_i}
\]  

(3)

We can now represent the feature characterizing the occurrence of \( t_j \) in a document by (3) with \( \eta - \bar{x}_j \) and \( \sigma = \sqrt{\sigma_j^2} \). The information on the size of \( \bar{x}_j \) is applied in selections of the most relevant terms, according to the user feedback on the classifications derived by the GA.

4 Description of the System

4.1 Chromosome Description

Let us consider a population \( C_1, \ldots, C_H \) of chromosomes, where each chromosome \( C_h \) is a fixed number of genes. A gene \( G \) is a pair \( (t, \tilde{\eta}) \), where \( t \) is a term, and \( \tilde{\eta} \) is a fuzzy number characterized by the membership function \( \mu_{\tilde{\eta}} \) expressed in (2). Therefore, the population of our Genetic Algorithm represents a collection of documents. Each chromosome can be interpreted as a query of the user or simply a set of terms representing the user preferences.

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

4.2 The Genetic Selector Module

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 maintains an increasingly efficient representation of the user’s interests, thus improving the quality of the fuzzy classification process when the user makes a new query.

The fitness function \( P(h) \) to be maximized measures the discriminatory power of every term through all the population as well as the accumulated discerning in the chromosomes:

\[
P(h) = S(C_h, V_g) + q_h
\]  

(4)
where \( S(C_h, V_y) \) represents the similarity between a chromosome \( C_h \) and the discriminatory vector \( V_y \), and it is based on the Jaccard’s score (Salton and McGill, 1983) weighted by an individual evaluation of every term:

\[
S(C_h, V_y) = \frac{\sum_{j=1}^{P} C_h(t_j) \cdot V_y(t_j) \cdot G(t_j)}{\sum_{j=1}^{P} C_h(t_j) + \sum_{j=1}^{P} V_y(t_j) \cdot G(t_j) - \sum_{j=1}^{P} C_h(t_j) \cdot V_y(t_j) \cdot G(t_j)}
\]  

(5)

where:

- \( C_h(t_j) = \mu_{z_j}(x_j) \), as defined in (2), with \( x_j \) being the relative frequency of the term \( t_j \) appearing in the gene of the chromosome \( C_h \), and \( \eta_j = \bar{x}_j \).
- \( G(t_j) = \frac{g(t_j)}{\sum_{j=1}^{P} g(t_j)} \), with \( g(t_j) \) being the accumulated value of discrimination for the term \( t_j \) for all the discriminatory vectors, and
- \( q_h \) 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.

\[
q_h = \frac{\sum_{j=1}^{P} C_h(t_j) \cdot g(t_j)}{\sum_{j=1}^{P} g(t_j)}
\]

(6)

- \( V_y(t_j) \) is the discriminatory vector based on the comparison of the term \( t_j \) appearing in documents \( D_i \) and \( D_k \) defined as follows:

\[
V_y(t_j) = \begin{cases} 
\mu_{z_x}(D_i(t_j)) \cdot \mu_{z_x}(D_h(t_j)) & \text{if } \mu_{z_x}(D_i(t_j)) \neq 0 \text{ and } \mu_{z_x}(D_h(t_j)) \neq 0 \\
\mu_{z_x}(D_i(t_j)) & \text{if } \mu_{z_x}(D_i(t_j)) \neq 0 \text{ and } \mu_{z_x}(D_h(t_j)) = 0 \\
0 & \text{if } \mu_{z_x}(D_i(t_j)) = 0 \text{ and } \mu_{z_x}(D_h(t_j)) \neq 0
\end{cases}
\]

(7)

### 4.3 The Document Fuzzy Classifier Module

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

A document score in a gene is then given by \( S_i(G_h) = \mu_{z_i}(x_i, k=1,...,K) \), where \( x_i \) is the relative occurrence frequency of the term \( t_j \) in the document \( D_i \), and \( K \) is the chromosome length (i.e., the number of genes in a chromosome). The presence of a term in the gene is then given by the scheme given in (2). Obviously, the aggregation of the document score in the genes to its overall score for the chromosome should apply an AND-like operator \( \odot \).

\[
S_i(C_h) = \bigodot_{k=1...K} S_i(G_h)
\]

(8)
where \( h=1,\ldots,H \), being \( H \) the size of the population (i.e., the number of chromosomes in the population).

The evaluation of a document \( D_i \) by a chromosome could be given by the expressions (9) and (10) corresponding to the minimum and average operators, respectively:

\[
S_i(C_h) = \min_{k \in \{1\ldots K\}} S_i(G_k) \quad (9)
\]

\[
S_i(C_h) = \frac{\sum_{k \in \{1\ldots K\}} S_i(G_k)}{K} \quad (10)
\]

Other aggregation operators could be applied for the operator \( \odot \) in (8), for instance the min operator, or an OWA operator representing the linguistic quantifier ‘most’ (Yager and Filev, 1994).

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

\[
S(Z) = \bigoplus_{h=1\ldots H} S_i(C_h) \quad (11)
\]

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

The selection of these operators \( \odot \) and \( \oplus \) for the production system should be based on their evaluation in an experimental setting.
**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 been all 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 $D_i$ for $i=1, ..., m$.

We define the fuzzy recall-precision $\tau$ by:

$$\tau = \rho^v_1 \psi^v_2$$  \hspace{1cm} (12)

where $\rho$ is the fuzzy recall, and $\psi$ is the fuzzy precision defined by:

$$\rho = \frac{\sum_{i=1}^{m} \min(u_i, S_i)}{\sum_{i=1}^{m} u_i} \hspace{1cm} \psi = \frac{\sum_{i=1}^{m} \min(u_i, S_i)}{\sum_{i=1}^{m} S_i}$$  \hspace{1cm} (13)

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

For an 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 $\tau$ measures how close the system’s evaluation is to the user’s evaluation.

**5 Experimental Stage**

The main objective of the system is the construction of a user profile. This profile is usually generated in filtering process more than in retrieving process where no information about the user is considered. Therefore, our main task is to filter the document collection using the information retrieved in the first query. This kind of situation appears when a user queries an information system, and the system retrieves a huge amount of documents as the answer. It would be desirable that the user could feedback the system by expressing the preferences through the evaluation of some documents. This new information can be considered as filter of the first answer.

**5.1 Genetic Components**

The genetic components have been determined by preliminary tests. The selection scheme for probabilities is the inverted linear ordering (Bäck, 1992). The selection mechanism is the universal stochastic sample (Goldberg, 1989), with an elitism model. To carry out the crossover, we have considered the one-point crossover operator, with a probability of 0.6. As for the mutation, we choose by a standard mechanism a gene to mutate. The term in the selected gene will be substituted by another random term in the set of indexed terms $T$. The probability of mutation
is set to 0.2 because is the only way to introduce new terms and keep the diversity in the population.

Finally, the size of the population is 80 chromosomes and the chromosome length is 10. We have considered 1000 generations for each run of the algorithm, calculating the average of three runs with different random seeds to get the final results.

5.2 Results

We have considered two different examples to carry out the experimental stage.

The first example is a collection of documents 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's information 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. Initially, the documents are evaluated by an expert, where there are 8 relevant documents and 14 non-relevant ones, with the relevance threshold set to 0.5.

Since the previous query must be refined in particular in terms of precision, the importance weights of the fuzzy recall ($v_1$) and fuzzy precision ($v_2$) will be set as follows:

$$v_1 = 0.67, v_2 = 1$$

The second example is a collection of 100 documents, corresponding to the query "Information and Retrieval" to the INSPEC database of 1999-2000. The number of terms extracted for these documents is 11210, where 1641 terms are different.

The preferences of the user in the framework of this query are located in the field of Genetic Algorithms. As in the first example, the user will give the highest relevance to the most preferred documents. In this case, the number of relevant documents is 13 and there are 87 non-relevant documents.

In Table I can be observed that the fuzzy recall-precision values with and without a genetic feature selector module. When a pre-processing stage of feature selection is carried out using the Jaccard coefficient, the values of the retrieval measures remain the same, though we notice a slight, but insignificant increase in the efficiency, in particular in the precision. However, let us notice that the number of features has been reduced as a maximum to the size of the population.

In the first example, where the system starts from 616 terms, the reduction is not significant. However, in the second example, the 1641 starting terms are reduced to the number of terms different in the population (800 as a maximum). Moreover, these terms are the most discriminatory terms for the classification of the retrieved documents. The user profile will be built with these terms without any worsening in the performance of the system.
6 Concluding Remarks and Future Work

An optimization of the user profile in filtering processes has been presented in this paper. The system is able to keep only the most discriminatory terms in the population of a Genetic Algorithm. This selection is based in the Jaccard’s coefficient combined with a discriminatory vector. Though the tests showed a small increase in the performance when feature selection is used, in particular for the precision, the increase is not significant. The results reveal the good performance of the system with this feature selection, as the value of the fuzzy recall-precision measure is maintained, while the number of terms needed to obtain this performance is reduced significantly. Thus, the main advantage in applying an initial feature selection stage seems to be a considerably reduction of the search space of the GAs.

Since the number of terms in large document bases, like the web, the inclusion of such an initial stage allows for an efficient application of GAs for discovering users’ information needs based on the success of their previous retrieval tasks.

In a future work, we will study the effect of others coefficients for measuring the quality of the selected terms. Also the general performance of the system when the document base increases must be analyzed in a future work.

References


