Genetic-based Fuzzy Clustering for Automatic Web Document Categorization

Vincenzo Loia Paolo Luongo
Dipartimento di Matematica ed Informatica
Soft-Computing Lab. - Università di Salerno
84081 Baronissi (Salerno) Italy
loia@unisa.it

ABSTRACT

The work herein described presents a genetic-based fuzzy clustering methodology able to cluster web documents into thematic categories. Our approach, exploiting evolutionary computation, applies the clustering on the context of the document, as opposite to content-based clustering that works on the complete document information. The final results reveal interesting performances if compared with a large categories directory implemented manually.

1. INTRODUCTION

Popular engines (Altavista, Netscape and Lycos) changed themselves from crawler-based into a Yahoo!-like directories of web sites composed of million of terms handled by thousands of human editors.
This drawback has an immediate effect in a lost of precision: the most popular Web search engines, in recent experiments, return only a fraction of the URLs of interest to user [3], have a small coverage of available data [2], suffer of instability in output of same queries submissions [4].
This work presents a clustering-based Web document categorization that exploits genetic computation to better cope with the complexity of the domain. Our clustering technique is performed only on the information related to the context of the document rather than on its content.

2. CONTEXT IN A WEB DOCUMENT

Let us consider a link in a Web page: in general we note the existence of sufficient information spent to describe the referenced page. Thus this information may be used to categorize a document. The process starts with an initial list of URLs, and, for each URL, retrieves the web document, analyzing the structure of the document expressed in terms of its HTML tags. For each meaningful tag found are extracted contextual data. For example, when the <A> tag is found containing an URL, an URL Context Path (URL: C_1: C_2:...: C_n) is defined, containing the list of the context strings C_i so far associated to the URL.
Any URL is analyzed through a breadth-first visiting: first the complete page is analyzed, then for each external link a new visiting is triggerred on the corresponding host. Next step regards the clustering process that exploits the Context Paths database and the categories-based catalogue in order to evaluate the membership value of each URL to a category.

SAC 2001, Las Vegas, NV

3. CLUSTERING METHODOLOGY

Let **T** be the set of the noun phrases. $\forall x \in T$ we define $\tilde{x}$ as the *fuzzy set* "noun phrases *similar to* x", formally:

$$\tilde{\mathbf{x}} = \{(t, \mu_x(t)) \mid \forall t \in T\}$$

with $\mu_x : T \rightarrow [0,1]$ as membership function.
The function is defined in order to give higher values for the noun phrase that generalize the original term of the category. Its calculus takes into account the synonyms for each simple term contained into the noun phrase of the category, rejecting the terms that are not synonyms or related terms. Any synonym of the simple term has a weight: the weights are higher for hypernym synonyms (generalization terms) and lower for hyponym synonyms (specialization terms), hence the clustering method brings up generalization with respect to each document matched. The membership value of a noun phrase, derived from a combination of simple terms, is given as an average of the synonyms weights.
Given **P(T)** as the power set of T, let us define the following similarity measure:

$$\begin{aligned} &Let\ x = (t_1, \ldots, t_n) \in P(T)\ and\ t_i \in T\ \forall i = 1..n \\ &\qquad y = (h_1, \ldots, h_p) \in P(T)\ and\ h_j \in T\ \forall j = 1..p \end{aligned}$$

$$\mathbf{S_K(x,y)} = \sum_{j=1}^{p} \sum_{i=1}^{n} (\mu_{t_i}(h_j))^K \ (shortly\ x \oplus_k y) \qquad (1)$$

where **K** is the *similarity factor* of the measure.
Given a couple $(x,y) \in P(T)^2$ we define **G**: P(T)xP(T)→ [0, 1] as the *coverage* of y on x:

$$\mathbf{G(x,y)} = \frac{|\{h_j |\ h_j \in y\ and\ \exists\ t_i \in x\ \ \exists'\ \mu_{t_i}(h_j) > 0\}|}{|x|} \qquad (2)$$

$$(shortly\ x \sqcap y)$$

Each category (or sub-category), defined by its noun phrases,

is viewed as a cluster $\mathbf{C_j} \in \mathbf{P(T)}$. Objects of the cluster are URLs extracted from the Web documents: each URL has an associated **Context Path** as *feature vector*, represented by $\mathbf{CP_i} \in P(T)$ (for the i^{th} context path).

In order to evaluate the membership grade μ_{ij} of the $\mathbf{CP_i}$ on cluster $\mathbf{C_j}$, a *familiarity grade* $\mathbf{A_{ij}}$ is defined; this parameter is the weight returned by the matching between context path and category, computed as the similarity measure on $\mathbf{P(T)}$ between $\mathbf{C_j}$ and $\mathbf{CP_i}$.
Up now the clusters are statically defined (their noun phrases are fixed).The dynamical behavior is provided by the genetic exploration (as defined in the next paragraph) and by a *specialization grade* s for each cluster, that allows us to vary the cluster dimension. The specialization grade exploits the *similarity factor* K that enables to modify the incidence of each similarity grade for the single terms. The next formula defines the familiarity grade using the specialization grade s_j for cluster C_j.

Familiarity Grade:

$$A_{ij} = \frac{C_j \oplus_{s_j} CP_i}{\text{noun phrases matched by } CP_i \text{ on } C_j} \qquad (3)$$

$A_{ij} \in [0,\ 1]$

Membership Grade:

$$\mu_{ij} = A_{ij} \cdot (C_j \sqcap CP_i)^{'} \quad \mu_{ij} \in [0,\ 1] \qquad (4)$$

Our clustering method exploits the concept of the *overlapping* flexibility; it allows objects to belong to all clusters.

Overlapping Property:

$$\sum_{j=1}^{|C|} \mu_{ij} \geq 0 \qquad (5)$$

Finally, the clustering method maximizes the following *Index of Quality* $\boldsymbol{J(C)}$, for which an *Influence Grade* $\mathbf{m}$ is introduced in order to reduce the impact of lower μ_{ij} values. At the increasing of $\mathbf{m}$ more relevant will be the weight of the clusters characterized by a higher specialization (membership grade).

Index of Quality:

$$J(C) = \sum_{j=1}^{C} (J_j) \qquad (6)$$

$$J_j = \begin{cases} (\sum_{i=1}^{N} \mu_{ij})^m & \text{no subcategory in } C_j \\ ((\sum_{i=1}^{N} \mu_{ij} + 1) \cdot \sum_{c}^{\text{subcategs } C_j} J_c)^m & \text{otherwise} \end{cases} \qquad (7)$$

with $m \in [1,\infty)$ and J_j as Index of Quality for the j^{th} category.

Index of Quality is skilled to specialize the categories, in order to contrast the generalization spur arising from the computation of matching weights.

4. GENETIC FRAMEWORK

1. **Representation of genomes** – the genome is constituted by a representation of a structured hierarchy of thematic categories, named *Category Forest*, composed by a set of tree structures, our *Category Tree*. Each Category Tree is viewed as a Root Category (it identifies a thematic category). Starting from a Root Category we find the subcategory nodes (specialization of category) which may be parents of other subcategories of lower level.

   Each root node is supported by three threshold values useful for the specialization grade of the thematic category (each subcategory is accompanied by a specialization grade). The subcategories can be defined **fixed** in the parent category, by means of a marker; this is useful to do not move the subcategory into other parent categories as effect of the mutation operator.

2. **Definition of the fitness function** - It is composed of two different evaluations. The first, named *Clustering Fitness*, is computed by the clustering methodology in terms of Index of Quality. The second factor is the *Quality of Distribution* (QoD), measuring the quality of distribution of the documents into thematic categories. This value is computed by averaging the membership grades of the documents, for each category or subcategory.

3. **Definition of the Crossover operator** – The crossover point is chosen randomly taking into account that root categories that can not be broken by crossover.

4. **Definition of mutation operators** – Different mutation operators are defined. They operate on root category and subcategory entities in order to better customize the system in generalizing or specializing its behavior.

5. CONCLUSIONS

Our clustering algorithm is based on a fuzzy clustering method that searches the best categories catalogue for web document categorization. The categorization is performed by context, this means that the clustering is guided by the context surrounding a link in an HTML document in order to extract useful information for categorizing the document it refer to. [1] is the first concrete effort of a context-based categorization even though their approach does not support fuzzy partitioning and the search of the better partitioning could suffer of the usual drawbacks of traditional clustering algorithms.

6. REFERENCES

[1] Attardi, G., Di Marco S., and Salvi, D. (1998). Categorisation by Context. *Journal of Universal Computer Science*, 4:719-736.

[2] Lawrence, S. and Giles, C. L. (1999). *Nature*, 400:107-109.

[3] Selberg, E. (1999) *Towards Comprehensive Web Search*.PhD thesis, University of Washington.

[4] Selberg, E and Etzioni, O. (2000). On the Instability of Web Search Engine. *RIAO 2000.*