Proximity Fuzzy Clustering for Web Context Analysis

Vincenzo LOIA
Department of Mathematics & Informatics
University of Salerno
84081 Baronissi (SA), Italy
loia@unisa

Witold PEDRYCZ
Department of Electrical & Computer Engineering
University of Alberta Edmonton
T6R 2G7 Canada
pedrycz@ee.ualberta.ca

Sabrina SENATORE
Department of Mathematics & Informatics
University of Salerno
84081 Baronissi (SA), Italy
ssenatore@unisa.it

Abstract

This study extends the web classification approach through a proximity-based fuzzy clustering sensible to the influence of the page. The proximity-based fuzzy clustering works in an unsupervised manner, augmented by a certain auxiliary supervision mechanism. The supervision scheme is realized via a number of proximity "hints" (constraints) that specify an extent to which some pairs of patterns are regarded similar or different. The hints are provided externally to the clustering algorithm and improve the searching activity by customizing the user's navigation. In this paper we focus on the feature spaces corresponding to the Web data characterizing the context analysis and we discuss how the knowledge extraction process identifies the right context.

Keywords: fuzzy clustering, proximity measure, Web mining, Proximity Fuzzy C-Means (P-FCM), proximity hints (constraints), fuzzy multiset.

1 Introduction

Many techniques for web page classification are based on text analysis, neglecting hypertext context-based investigation. Clustering algorithms provide promising results when are based on combined term-similarity and hyperlink-similarity measures. However all typologies of clustering techniques require improvements in term or word vector representation of web pages, especially for Web collections dealing with one or few specific topics.

The huge quantity of digital collections of data disseminated on Internet can be regarded as a great information space. It is common idea considering Web pages or hypertext documents as points or vectors in that space, where each vector dimension represents a chosen term. Each dimension has an associated weight that indicates the frequency of a corresponding word [6]. In this approach we consider an extension of weight vector dimensions applied to P-FCM clustering [4].

2 Our Approach

Search engines have many remarkable capabilities but nevertheless important progresses done in these last years in improving their design and overall behavior many open problems remains unsolved.

In general search engines suffer from limited coverage, outdated database, and imprecision in treating the query. Of course, this is due to a still ongoing growth of the Web, but part of this responsibility stays in a weak treatment of the query and in ranking strategies still too simple [2]. In order to reinforce the relevance and quality of the searching, it is necessary to improve the engine capability for a deeper understanding of the document. Automatic classification/categorization is attracting major efforts, considering the impressive costs deriving from human-powered directories.

It is a general tendency to realize suitable classifiers that index web pages space, through a relevance judgment according with the context analyzed. Automated classifications define, through partitioning process, different categories and typologies of web pages, grouping them on the basis of static/dynamic correlations found during context

analysis. Our approach to Web context analysis employs the following techniques:

1) Automatic indexing
2) HTML structure analysis and vector space representation with fuzzy multi-set
3) Applying P-FCM algorithm

Next sections explain each step.

3 Automatic indexing

The purpose of automatic indexing is to automatically identify the content of each web document in term of associated features, i.e. words or phrases. Traditional automated indexing techniques filter all words and phrases in the document, removing stop-word, stemming terms. Our indexing extracts nouns and adjective and counts the frequency of each word. Then it calculates an evaluation of each word using the TF-IDF method for each Web page. Finally the system stores the evaluation into a lexicon, and, for each page, it builds the data set for the next clustering-based phase. Terms with highest occurrence are candidate to become part of feature space.

4 HTML structure analysis and vector space representation with fuzzy multi-set

This task starts analyzing the structure of web pages in terms of HTML tags; at this stage, we consider the following tags:

`<TITLE>,<H*>,<META name=keywords…>,<A>`

Whenever one of these tags is found, a context phrase is analyzed. As in previous phase, we consider these terms eventually stemmed, discarding stop words, conjunctions and articles. Tags related to layout or emphases are discarded. As result, we obtained a sequence of words for each selected tag.

Then we define the vector space, through fuzzy multiset. Each web page is treated as an n-dimensional vector, where each component is a fuzzy multiset representing the membership of each term extracted from that HTML tag.

The choice of terms is an essential step to define a correct features space. In order to guarantee an adequate context identification we use, in the indexing phase, TF-IDF algorithm to count the frequency by which a term appears in a web page.

Besides, each term is evaluated in the area of web page where it is found: we build a fuzzy multiset to characterize the membership of terms that appear in different tags.

The rules defined to correlate a membership to a term take into due consideration the area where the term appears:

- *If the term appears in the tag section, its membership value is 1.*
- *If a synonym of the term appears, its membership value is 0.8*
- *If an antonym of the term appears its membership value is 0.1*
- *…*
- *If the term does not appear in the tag section, its membership value is 0*

The adoption of a thesaurus allows handling the correlated words (synonyms, antonyms, related noun, etc.) of given matrix terms in order to improve the quality and the approximation of the results. Now let us deepen the formal aspect.

4.1 Extended Multiset-valued term-document matrix in P-FCM algorithm

We consider an ordered fuzzy multiset, where each value corresponds to a membership for a specific tag in the web page. Let be:

T ={TITLE, H1, H2, META, A} the HTML tags (in case of multiple occurrences, a numbering process avoids ambiguity).

$W=\{w_1, w_2, \ldots w_h\}$ a set of keywords (in features space).

S: $W \rightarrow \mathcal{FM}(\mathcal{P})$ a function where $\mathcal{FM}(\mathcal{P})$ is a collection of all fuzzy multisets of web pages $\mathcal{P}=\{p_1, p_2, \ldots p_k\}$, such that $S(w_j)=\{(\mu_{\tau 1}(w_j), \mu_{\tau 2}(w_j),\ldots,\mu_{\tau m}(w_j))/p_1, (\mu_{\tau 1}(w_j), \mu_{\tau 2}(w_j),\ldots,\mu_{\tau m}(w_j))/p_2, \ldots, (\mu_{\tau 1}(w_j), \mu_{\tau 2}(w_j), \ldots, \mu_{\tau m}(w_j))/p_k\}$ (1)
where τ_h is a tag in T, with h=1,…, m.

Given a web page p, its context Cp is given by a set of terms $\{cp_1, cp_2, \ldots cp_n\}$ extracted from p, through automatic indexing technique. We define:

$$\mu_{\tau h}(w_j) = \max_{c_{pi} \in Cp} \mu_{\tau h}(c_{pi}) \quad \forall \tau h \in T, \forall c_{pi} \in Cp \qquad (2)$$

4.2 A sketched example

Just to give some descriptive information, let us consider some HTML sections of the following web page.

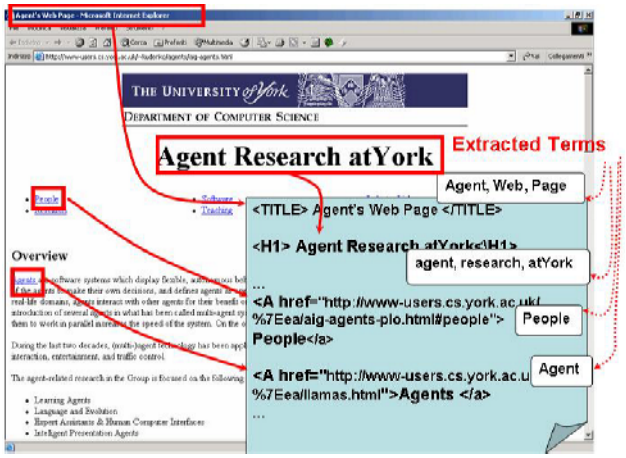

Figure 1: Extraction of significative terms from selected HTML tags

As example of feature space let us consider the following set:

W= {SPIDER, FUZZY, CRAWLER, etc. }

Considering the HTML sections (see Figure 1), and focusing on the word SPIDER, we calculate the membership value related to the section TITLE, +according to (2) :

μ_{TITLE} ("SPIDER")=max {0.8, 0.3, 0}= 0.8

where {0.8, 0.3, 0} represents the membership of terms *Agent, Web, Page* in TITLE, according to rules defined in Section 4.

Applying the function (1), we obtain:

S("SPIDER")= {(0.8, 0.8, 0, 0.8,…)/p }

The value 0.8 is the (max) membership in TITLE and H1 tags: in both tags we find the term *agent* that can be consider as a synonym of term "SPIDER". In the same way, the value 0 is the membership of first A tag, 0.8 is the membership in second A tag and so on (in this case we are considering a single web page p).

In this way, each entry in the data matrix is a fuzzy multiset that represents the set of membership value of a term in each selected HTML tags, for a given web page.

5 Applying P-FCM algorithm

Proximity-based FCM, or shortly P-FCM [4], is an extension of the well-known fuzzy c-means clustering (FCM) algorithm [1] particularly useful for Web exploration and data organization on the Web. In fact, many factors may play an important role in a human judgment concerning the "proximity" of Web pages (layouts, backgrounds, links, texts, …). Many of these factors are difficult to quantify and to translate into computationally meaningful features. The textual information is the most evident and it is almost the exclusive contributor to the feature space when determining structures in a collection of Web pages. The use of the proximity hints can compensate for the consideration of a subset of the feature space: we cluster Web pages in the subspace of textual information and the proximity values provided by the user are useful to enrich this subspace by capturing hypermedia or cognitive information.

5.1 The concept of proximity and its relationship to partition matrices

The concept of proximity between two objects (patterns) is one of the fundamental notions of high practical relevance. Formally, given two patterns "a" and "b" , their proximity, prox(a, b), is a mapping to the unit interval such that it satisfies the following two conditions

prox (a, b) = prox (b, a) symmetry

prox (a, a) =1 reflexivity

The notion of proximity is the most generic that constitutes a minimal set of requirements; what we impose is straightforward: "a" exhibits the highest proximity to itself and the proximity relation is symmetric. In this sense, we can envision that in any experimental setting, these two properties can be easily realized. Given a collection of patterns, the proximity results obtained for all possible pairs of patterns are usually arranged in a matrix form known as a proximity relation P.

5.2 An overview on P-FCM algorithm

P-FCM computing scheme comprises of two nested phases, as given in Table 1. The upper level deals with the standard FCM computing (iterations) and follows the well known scheme encountered in the

literature, while the one nested is aimed at the accommodation of the proximity requirements and optimizes the partition matrix on this basis.

Given: number of clusters, fuzzification coefficient, distance function, partition matrix and a termination condition (small positive constant ε).

Table 1. A general flow of optimization of the P-FCM algorithm

Repeat

main external loop

Compute prototypes and partition matrix U using standard expressions encountered in the FCM method

Repeat

internal optimization loop

Minimize some performance index V guided by the collection of the proximity constraints

Until no significant changes in its values over successive iterations have been reported (this is quantified by another threshold δ)

Until a termination condition has been met (namely, a distance between two successive partition matrices does not exceed ε).

The accommodation of the proximity requirements (constraints or hints) is realized in the form of a certain performance index whose minimization leads us to the optimal partition matrix. As stated in the problem formulation, we are provided with pairs of patterns and their associated level of proximity. The partition matrix U (more specifically the induced values of the proximity) should adhere to the given levels of proximity. Bearing this in mind, the performance is formulated as the following sum:

$$V = \sum_{k_1=1}^{N} \sum_{k_2=1}^{N} (\hat{p}[k_1, k_2] - p[k_1, k_2])^2 \, b[k_1, k_2] d[k_1, k_2] \quad (3)$$

The notation $\hat{p}[k_1, k_2]$ is used to describe the proximity level induced by the partition matrix. It becomes apparent that using directly the values of the membership (corresponding entries of the partition matrix) is not suitable. Simply, if two patterns k_1 and k_2 have the same distribution of membership grades across the clusters, these membership grades are usually not equal to 1 as the proximity value could be close or equal to 1. The value d[k1, k2] represents the distance between the patterns; p[k1, k2] characterizes the partition matrix (where each coordinate is a membership defined as average of elements of the correspondent fuzzy multiset); b[k1, k2] assumes binary value (it returns 1 if there exists a proximity value between k1 and k2, otherwise zero). With the partition-proximity defined in this way, (3) reads as follows

$$V = \sum_{k_1=1}^{N} \sum_{k_2=1}^{N} (\sum_{i=1}^{c} (u_{ik_1} \wedge u_{ik_2}) - p[k_1, k_2])^2 \, b[k_1, k_2] d[k_1, k_2] \quad (4)$$

The optimization of V with respect to the partition matrix does not lend itself to a closed-form expression and requires some iterative optimization.

6 Conclusions

The work herein described extends previous results reported in using P-FCM as a useful technique for Web mining. As in [3], the extension is based on the role of fuzzy multiset as a more robust approach to derive keywords. A possible extension of our model converges toward a better characterization of web pages capturing implicit contextual and conceptual information. It always more clears the importance to evidence the main topics of web pages, delineating the right contest. Our future development t aims to substitute the TF-IDF method towards conceptual fuzzy sets (CFS) approach [5] that works realizing a conceptual matching between input keywords and Web pages.

References

[1] J.C. Bezdek, *Pattern Recognition and Fuzzy Objective Function Algorithms*, Plenum Press, N. York, 1981

[2] Lawrence S., Giles L. , Context and Page Analysis for Improved Web Search, *Internet Computing, IEEE* , Volume: 2 Issue: 4 , Jul/Aug 1998, pages 38 -46

[3] Miyamoto S., Information clustering based on fuzzy multisets *Information Processing and Management* 39 (2003) 2: 195-213

[4] Pedrycz W., Loia V., Senatore S., P-FCM: A Proximity – Based Fuzzy Clustering, accepted on Special Issue of *Fuzzy Sets and Systems* on "Web Mining using Soft Computing", 2003

[5] Takagi T., Tajima M., Query Expansion Using Conceptual Fuzzy Sets For Search Engine, Proc. FUZZIEEE 2001, 1303-1308.

[6] Vlajic N., Card H.C., Categorizing Web pages on the subject of neural networks, *Journal of Network and Computer Applications,* **21**, 91-105, 1998.