USING WEB SITES EXTERNAL VIEWS FOR FUZZY CLASSIFICATION

Georgi Furnadzhiev

Abstract: In the paper a fuzzy sets implementation into web sites classification is considered. Web sites external features are addressed and the possibility to use them for the classification is proved. An example with five different categories classification is given.

Keywords: web mining, fuzzy sets, classification

Introduction

There are more than 8.10^9 Google indexed web pages in the World Wide Web. Finding relevant information is very difficult. Searching information is a main problem. When we find many results, it is a good idea to classify them.

Using web search engines we can choose: result file type, language, domain, etc. Often we receive a message "This web site is added to directory X in category (ies)..." in the result list. This directory contains qualitative, but very small subset of all web sites in the world, and for most results, we do not have any information about their types. This makes a big part of our result uncategorized. We can group them by region or language, for example, but not regarding their content. It will be good if we can, using a web crawler or metasearch engine, to specify a web site type from a given set at least. Other useful opportunity will be to classify uncategorized part of the result list of our search query. It is not the same using Google to find word "accommodation" in science conferences' web sites or travel agencies' web sites.

Automatically web sites categorization provides two main advantages for the end user. He or she can search information in specific group of web sites at first. The result will be always classified at second. The user will receive grouped results and it makes easier finding relevant information. A search engine supporting automatically web sites classification will allow more flexible queries or well arranged results.

The main part of realised web sites classification is based on the internal document representation and structure. The authors never forget the web page is a text document, and the web site is a text documents structure. The main efforts are addressed to find relevant text and structural features to the web site.

But in the other hand users can classify the web sites without knowledge in web development. The web directories editor can categorize a new site without reading meta-tags information or finding other web sites linked to the regarded one. Here we use external web sites views for their classification.

Web Sites Classification

Unknown objects classification is a main part of machine learning and data mining research. When we classify a set of objects we need

- formal object and classes descriptions
- classification model
- training set and training mechanism
- rules adding unknown objects into a class

There are created many automatical classification approaches, based on artificial neural networks, decision trees, genetic algorithms, etc. The classification process follows the steps:

- Model choice.
- Training. We use a relatively small and labelled subset, called training set. The labels mean belonging to a class. Based on this training set, we construct the classifier.
- Unknown objects classification.

Web resources classification is an application of traditional data mining techniques in respect to the specific area. ([8], [9]) The datasets contain web sites. All web sites classification could be possible, if we have a good ontology describing the current state of the art. However, it is a very difficult activity. We have to know at least the current situation in the entire web. A rational idea is to have a specific sub-ontology and use it to decide on the particular problems.

Talking about web sites classification, we have to keep in mind two main arguments.

- The hyperlinks between web sites do not reflect on their types. The authors are not obligated to relate their web sites to any other ones.
- The most adequate web sites description approach is using quality data. We can detect features or count them only.

There are many realised approaches to determine web site type or automatical construction of web directories. In general, we can find two very popular directions – adapted for web documents text mining techniques ([2]) and web structure mining techniques ([5], [6], [7]). In first case, authors prefer to weight different parts of the web sites or the web pages, and in the second – to use web structure in general. There are examples for domain specific classification ([3], [4]).

Here we try to prove how the type of web sites affects their external features. We try to find how the content influences the external view.

Web Sites External Features

Firstly we have to define what *an external feature* is. Every web site can be considered from two points of view:

- Internal – this is the site structure, meta tags, technologies, formal languages used in site creation, etc
- External – this is the visible part of the site

For example, when the user clicks “Sign in”, it could be a button, or (GIF or JPEG) image or text hyperlink in the different cases. It can start a script, written in some formal language, providing one and the same semantics.

When talking about links here, we mean external views of the same web site’s links.

Web Sites Features and Fuzzy Sets

The fuzzy sets [1] are good mechanism for describing the features of the web sites classes. There are not any formal models for the web sites creation and the authors are not obligated to include anything. Moreover, main purpose in the web is to be distinctive. However, content and specific area has an effect on the language, structure, representation of the data, etc. We can expect similar content to be presented in similar ways. From this point of view we cannot say a given feature is specific for a web sites class or not, but we can define a relative belonging into a set of features describing the class. That makes the fuzzy web sites description very relevant. We can regard the web site like a fuzzy set of features.

In other hand a given web site can belong into different categories. Sometimes the site category is not exactly defined. In these cases it might be as well to use fuzzy belonging into a web sites category. It is possible to regard the web sites categories as fuzzy set of web sites.

Semi structured nature of the web make the fuzzy models very useful for its explanation.

How to Prove

To define a fuzzy set describing a class, we need to discover a relatively small training set of web sites and their descriptions. For every member of this set we have to find the features contained in them at first, and compare the given results at second. With a simple comparison and counting, we find relatively belonging into a set of features for this class (and this small set). This makes our results as accurate as our training set is representative.

We need to prove whether our fuzzy sets are relevant or not. Of course, the initial fuzzy set is not enough for the classes’ description. It is possible to find one or more elements for all classes, but we have to find the specific

ones. In a formal model if we have the classes C_1 … C_n, and T_i i= 1 … n are the fuzzy sets given from the first step, we actually are interested in sets

$$T_i \setminus \bigcup_{j \neq i} T_j (*)$$

for every i= 1, …, n. Here we can use the equation

$$A \setminus B = A \bigcap \overline{B}$$

where A and B are arbitrary sets. This representation will help us to apply definitions for the section and the union of fuzzy sets. If for every i = 1, …, n all of the sets (*) are not empty and are not the universal set, we can say we have found lists of features describing given classes.

This model is temporary because of the temporary nature of the web. It is exact for the training set only, not for all web sites in the world, belonging into the classes. Moreover, it provides correlations among the given classes, but not among all classes, which could exist in the world around. To improve the model correctness and accuracy we have three ways:

- Using carefully selected and relatively big training sets
- Frequently testing the training set for changes and actualise the features database and sets (*)
- Using model for classification web sites with “expected” types

How to Classify

The next task is to find a rule for unknown web site evaluation. A natural approach is to consider every web site description like a fuzzy set too and find all of the distances between this description and the fuzzy sets, associated with the classes. An uncategorized web site belongs to a class, if and only if, the distance between the site and the class is the smallest. The distance can be defined in many different ways. Actually, this is clustering web with preliminary defined cluster centres. In our works, we compare the Hamming and the Euclidean metrics. The metrics choice can be automated. It is necessary to have program applying two or more metrics or similarity functions. In the second case, the system must prove how similarity is bigger. The system can simultaneously follow two criteria:

1. Better total correctness, and in case they are equal -
2. The web sites distribution after the test. The statistical dispersion is good measure there.

For metrics choice, the same training set can be used.

In other hand we can use similarity function for fuzzy classification. It makes our model fuzzy in general. It is not difficult to see that

$$n(x, y) = \frac{1}{1 - d(x, y)}$$

when d is a given metric is a similarity function. If the objects x and y are equal, $n(x,y)$=1.

Objects and Classes Descriptions

Web sites’ descriptions in this model are simple. For every one we define a vector $V_i(v_{i1}, v_{i2}, \ldots, v_{in})$ where v_{ij}=1 if the feature j is found in the site, and v_{ij}=0 if the feature is not found in the site.

If we have m web sites belonging into a given class, we define the vector T with components:

$$t_j = \frac{1}{m} \sum_{k=1}^{m} v_{kj}$$

It is not difficult to see that

- vector *T* defines a fuzzy set
- if we have two or more classes and mark them with T_i their vectors, then $D_i = T_i \setminus \bigcup_{k \neq i} T_k = T_i \bigcap \overline{(\bigcup_{k \neq i} T_k)}$ is a fuzzy class descriptor for every *i*.

Experiment

We made experiments with 100 web sites from five following types

T1. University web sites
T2. Newspaper web sites
T3. International unions web sites
T4. Governmental web sites
T5. Parliament web sites

We used 20 web sites by class. Their first nontrivial pages have been considered. Here by *nontrivial page* we mean the first page after simple Enter page. We used Yahoo! Directory for finding representative for all world training sets, from different languages, countries and continents with respect of their relative distribution. When we described these web sites, we obtain 127 different features. We count the features found into the classes. We compare the classes by (*) and obtain classes descriptors. Here we give elements *x* with $\mu(x) \geq 0.5$ for every class. Here $\mu(x)$ is the fuzzy set characteristic function.

T1: University web sites (30 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "Alumni"	0.70
Link "About university"	0.70
Link "Structure"	0.65
Link "Events"	0.65
Link "Library"	0.60
Link "Researches"	0.60
Link "Students"	0.60
One colour background	0.55

T2: Newspaper web sites (60 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "News"	0.60
Link "Sport news"	0.60
Link "Archives"	0.50
Link "Advertising"	0.50

T3: International unions web sites (52 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "About us"	0.55

T4: Governmental web sites (57 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "Searching"	0.50

T5: Parliament web sites – 45 elements with nonzero value of $\mu(x)$ but $\mu(x) \leq 0.45$ for all. The first five are

Feature	Belonging
Language choice	0.45
Link “Contacts”	0.35
Links to institution’s documents	0.35
Link "News"	0.35
One colour background	0.35

In our tests, we compare the Hamming and the Euclidean metrics and test them with 100 random selected web sites – by twenty for class. The results are given in the following tables

Euclidean metrics						
	1	2	3	4	5	Correctness
T1	17		1		2	85 %
T2		18	1		1	90 %
T3			17		3	85 %
T4			4	16		80 %
T5			2	2	16	80 %
Total	17	18	25	18	22	84 %

Hamming metrics						
	1	2	3	4	5	Correctness
T1	18				2	90 %
T2		14	1		5	70 %
T3			12		8	60 %
T4			1	8	11	40 %
T5			1		19	95 %
Total	18	14	15	8	45	71 %

Here “Correctness” is the percent of web sites correctly added into their class’ sets. Based on the results we can say the Euclidean metrics is better and can recommend it.

Less correctness for some types we can explain with classes’ similarity. Distance matrix between classes is as follow

T1	0,00				
T2	2,48	0,00			
T3	1,99	1,71	0,00		
T4	2,04	1,82	1,04	0,00	
T5	1,94	1,80	0,94	1,03	0,00
	T1	T2	T3	T4	T5

when the Euclidean distances between classes descriptors are given. These descriptors are obtained in the training phase and depend on training set only. As we can see, the best results in metrics tests we obtain for most “isolated” classes.

In our example we used the smallest distance between web site and web site class. In fuzzy web sites classification we can talk about biggest similarity.

Conclusions

Based on the experiment results we can say this approach have acceptable correctness for further studies and applications. The best results are observed for less similar classes. The main weak points are similar classes' areas. The approach is applicable to most general web sites categories.

It is a good idea to test the approach in a similar web sites classification. There are many huge categories in the web directories. We can apply the approach for subcategories creation. We can expect similar classes, but the web sites are similar too.

Other result is fuzzy sets are suitable mechanism for web sites classes description and study.

The results manifest how important the web site structure is. The most of the described features are external representation of this structure. It is prove in practise the proposition web sites are independent objects for classification.

References

1. Zadeh L., Fuzzy sets, Information and control, Vol. 8, 1965 (338 – 353)
2. Pierre J., On the Automated Classification of Web Sites. Linköping Electronic Articles in Computer and Information Science, Vol. 6(2001): nr 0. http://www.ep.liu.se/ea/cis/2001/000/. February 4, 2001
3. Ardo A., T. Koch, and L. Nooden. The construction of a robot generated subject index. EU Project DESIRE II D3.6a, Working Paper 1 1999. http://www.lub.lu.se/desire/DESIRE36a-WP1.html
4. Kock T., A. Ardo. Automatic classification of full-text HTML documents from one specific subject area. EU Project DESIRE II D3.6a, Working Paper 2 2000. http://www.lub.lu.se/desire/DESIRE36a-WP2.html
5. Attardi G., A. Gulli, and F. Sebastiani. Automatic Web Page Categorization by Link and Context Analysis. In Chris Hutchison and Gaetano Lanzarone (eds.), Proceedings of THAI'99, European Symposium on Telematics, Hypermedia and Artificial Intelligence, 105-119, 1999.
6. Cho J., H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. In Computer Networks and ISDN Systems (WWW7), Vol. 30, 1998.
7. Rennie J., A. McCallum. Using Reinforcement Learning to Spider the Web Efficiently. Proceedings of the Sixteenth International Conference on Machine Learning, 1999.
8. Han, J. and Chang, K. C.-C. Data Mining for Web Intelligence, IEEE Computer, Nov. 2002
9. M. N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim, Data Mining and the Web: Past, Present and Future, Proceedings of WIDM99, Kansas City, U.S.A., 1999.

Author Information

Georgi Furnadzhiev - Institute of Mathematics and Informatics, BAS, Information Research Department; Acad. Georgi Bonchev St., Block 8, Sofia 1113, Bulgaria; e-mail: furnadjieff@math.bas.bg