Journal of Computer Sciences

Original Research Paper

Fuzzy Modeling for Multi-Label Text Classification Supported by Classification Algorithms

[1]Beatriz Wilges, [2]Gustavo Mateus, [2]Silvia Nassar, [2]Renato Cislaghi and [3]Rogério Cid Bastos

[1]*Department of Information Systems, Federal University of Santa Catarina, Florianópolis, Brazil*
[2]*Department of Informatics and Statistic, Federal University of Santa Catarina, Florianópolis, Brazil*
[3]*Department of Engineering and Knowledge Management, Federal University of Santa Catarina, Florianópolis, Brazil*

Article history
Received: 05-07-2016
Revised: 31-07-2016
Accepted: 08-08-2016

Corresponding Author:
Beatriz Wilges
Department of Information Systems, Federal University of Santa Catarina, Florianópolis, Brazil
Email: beatriz.wilges@ufsc.br

Abstract: The ever-increasing amount of information on the Web is organized in structured, semi-structured and unstructured data. Text classification systems, capable of handling such different structures, may facilitate the work of important tasks such as indexation and information retrieval in search engines. The objective of this research is to develop a method for the classification of documents into multiple categories with fuzzy logic. This method was built from a process of pattern recognition and, also, two variables called similarity and accuracy were used. The proposed fuzzy classification method uses variables that express the ability to analyze the similarity and accuracy of a document through a database of terms. The database of terms is generated by a collection of pre-classified documents in categories of interest. The documents processed according to the similarity and accuracy in the database of terms composes a training set also called knowledge base. From this database, it is possible to identify a pattern that specifies a set of rules through a knowledge discovery process. This process involves the data mining of the knowledge base. Thus, it was possible to define a general model that is used in the creation of rules and membership functions of the fuzzy model for the classification of documents into multiple categories. The general model of the rules identified in the data mining process and implemented in fuzzy model considers the most significant variables and also contributes to the specification of the membership functions, such as the definition of linguistic terms of fuzzy sets. Thus, it was possible to implement a more deterministic approach regarding the input, membership functions and inference rules of the fuzzy model. The results of the proposed method for classification of documents are relevant because they have a satisfactory accuracy rate.

Keywords: Text Categorization, Decision Tree, Fuzzy Modeling

Introduction

The volume of information available is increasing significantly over the years, which means that people have more access to knowledge from any electronic device over the Internet. Thus, automatic categorization systems play an important role in the text classification process, by assisting in information retrieval processes. According to Yang and Pedersen (1997) the process of retrieving documents in properly classified databases is more efficient and the search scope is reduced even if a large volume of information is available.

Li *et al*. (2011) further state that the goal of text classification is to label textual documents with thematic classes from a predefined set. Also according to these authors, many different methods have been applied to text classification tasks including the K-Nearest Neighbor (KNN) approach (Bang *et al*., 2006; Tan, 2006), Naïve Bayesian approaches (Baker and McCallum, 1998; Lewis, 1998; Yang and Liu, 1999), support vector machine (Dumais and Chen, 2000) and decision trees (Lewis and Ringuette, 1994; Quinlan, 1993).

However, most machine learning algorithms were designed for single-label classification in which a document can only belong to one category (Jiang *et al*., 2012). In multi-label text classification, a document can belong to more than one category. Therefore, the focus of this research is on multi-label text classification associated with fuzzy models.

The goal of this research is to analyze a multi-label text classification process, which generated a fuzzy model. In the research (Peters and Koster, 2003) was used a new criterion for Term Selection, which is based on the uncertainty in Term Frequency across categories.

This proposal involved constructing a database of terms organized by category, where classification algorithms were applied to generate Decision Trees (DT) from several text classification tests. With these DT, it was possible to generate a lined up fuzzy model capable of displaying the relevance degrees of the analyzed text for each category.

Thus, a database was built for each category using unstructured text extracted from the Web, mostly online magazines and newspapers. This database was organized into four categories involving Education, Technology, Sports and Economy. In this research, the most frequent terms in each category were considered.

This article is organized as follows: Initially, the construction of the base of terms from different sources will be presented, with an estimate of approximately 3000 words have been used for each category. In sequence, the methodology of the classification process was detailed as well as its variables and indicators. Following, the classification process and their results using Decision Tree algorithms (DT) are presented. The fuzzy model built from the classification process is presented next and finally, considerations and further analysis on fuzzy modeling are suggested.

General Organization Model

The proposed model will allow a collection of documents from the OM to be classified according to categories of interest to the organization. For this, a method of document classification that considers multiple categories is applied.

To perform the implementation of this organizational model, we developed a method of document classification. This method considers the high dimensionality involved in the classification of documents, i.e., as a document contains lots of information that can be represented in several classes, it is categorized in multiple categories. Figure 1 shows the steps of the development of the document organization model based on this method.

Also at this stage the variables are defined and the data to a knowledge base is generated with these variables. This database is built to allow the knowledge discovery process find associations between variables in order to design rules for the fuzzy modeling.

So part of these activities involve a process of Knowledge Discovery in Databases (KDD), specifically in a learning set divided into training and testing. The processing on this training set is done by ranking algorithms. The evaluation of the model generated in this phase is done by calculating the accuracy and error. Both accuracy and error of the model consider the results of the matches between the real class and the expected class.

The proposed method considers a process of Knowledge Discovery (KDD) on the rules that are implemented in a fuzzy modeling. The use of fuzzy logic deals with overlapping categories in a document handling the vagueness of the input variables. The structure of the document organization model is shown in Fig. 2.

The application of the model is made from a pre-processed document, whose similarity and confidence are calculated according to the definition of a database of terms. These two variables are the inputs of the fuzzy modeling. Thus, a collection of documents of an organization can be categorized by assigning its relevance to each category identified as of interest to the organization.

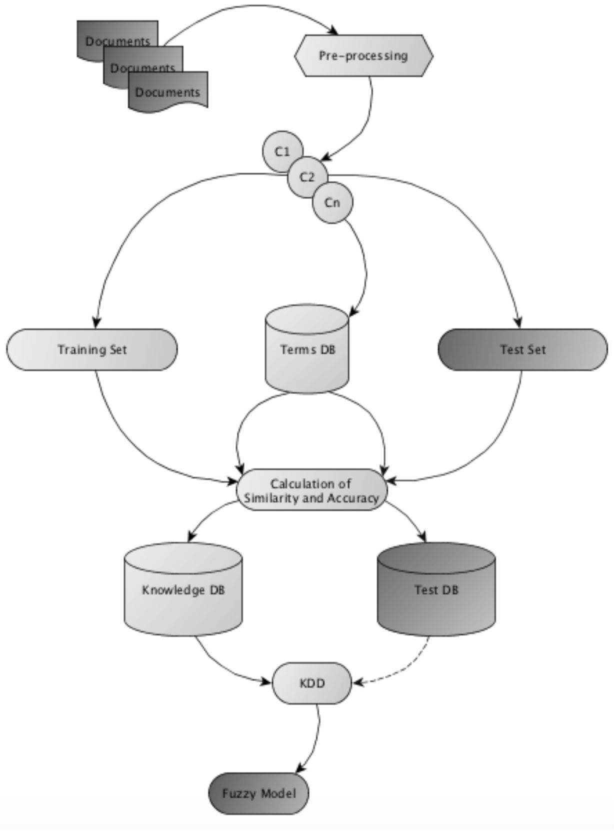


Fig. 1. Model development

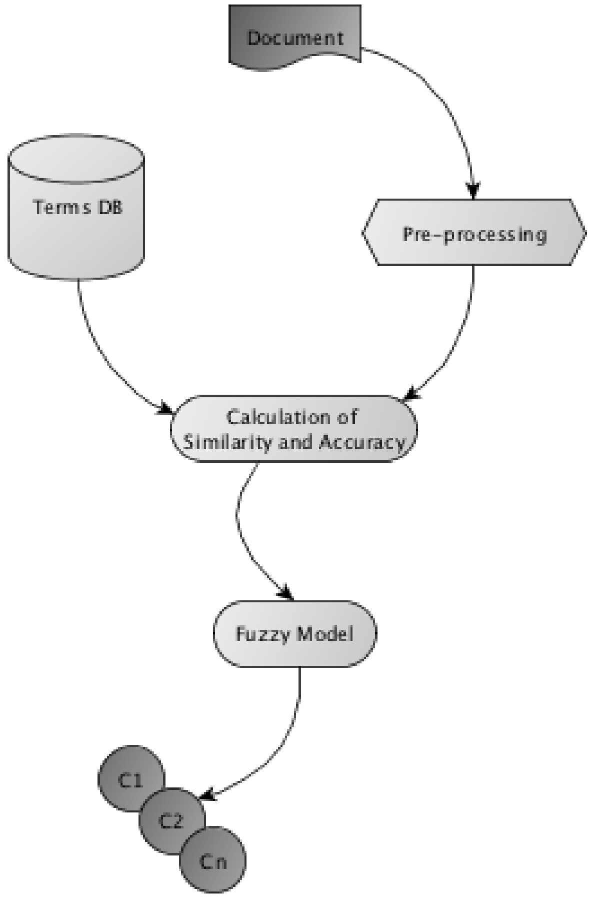


Fig. 2. Model application

Building the Term Database

After selecting a collection of documents belonging to the same category, the stage of text preparation commenced. This step involves certain techniques that facilitate the process of selecting textual features. The proposal in order to build the database (DB) was to identify all words that best express the characteristics of a given category. This DB is constructed based on the most common terms in all documents of the same category. In this case, the document database should be large enough to cover most terms that define the category. From that DB it is possible to define if any text belongs to a category in the DB.

Pre-processing is a very important stage in the text classification, including several steps to transform the set of documents, in natural language, into a list of useful terms and in a format which is compatible for knowledge extraction. In this study, a tool called RapidMiner (2016) was used to perform both pre-processing and DT generation.

Document processing firstly involved the "Extract Content" operator, which extracts the textual content of an HTML document and yields extracted text blocks from the document. Later, the "tokenize" operator divides the text of a document in a sequence of tokens. Further, the "Transform Cases" operator transforms all characters of a document in lowercase or uppercase. Finally, the "Filter Stop words (Dictionary)" operator removes all the words that belong to a list of stop-words, which is loaded from a file on the operator itself. The stop-words list contains words that are not relevant to counting the most common terms in the text document; therefore, the list contains grammatical words that do not affect the definition of the category of the text. Figure 3 shows a part of the DB constructed from document processing.

In Fig. 3 it is observed that, for every text document, all terms in respect to their total occurrence in the text (total occurences) and which document the term was extracted from (document occurences) were analyzed.

At this stage of the text document processing, studies and checks have been conducted to identify the best way to implement this process. It was considered a case-comparison study utilizing both serial and parallel paradigms in the text processing (Wilges *et al*., 2014). In the results, it was observed the higher efficiency of serialized processing. This is because the text files used in processing did not have a size on the order of petabytes.

In this research it is clear that being able to define whether a text belongs to a category is not enough. It is important to know the accuracy of the analysis during the decision making process. Hence, a fuzzy model that could consider the membership degree of a text document to a particular category was developed. Therefore, all terms of each category from the DB were ordered and normalized to the highest and lowest frequency for each category. Thus, an index called relevance degree $\left(\mathrm{RD}_{w_i}^{c_j}\right)$ was obtained for each word. The relevance degree $\left(\mathrm{RD}_{w_i}^{c_j}\right)$ of each word (w) must be calculated for all categories (c), which are composed by many text documents. Thus, the weight of the terms in (DB) for each category, known as the relevance degree of the term $\left(\mathrm{RD}_{w_i}^{c_j}\right)$, is calculated as follows:

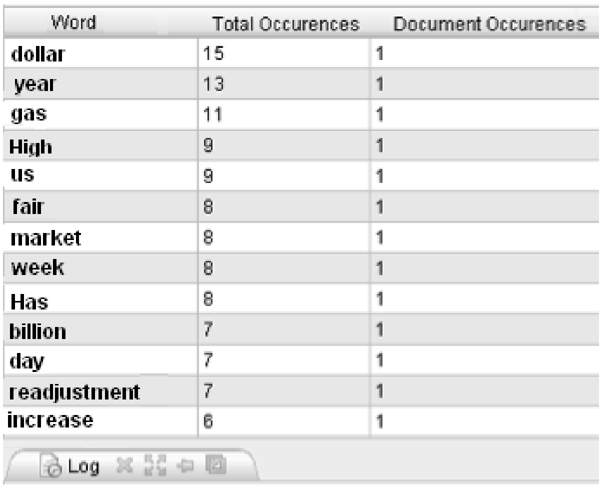
$$\mathrm{RD}_{w_i}^{c_j} = \frac{f_{(w_i,c_j)}}{\max\left(f_{(w_z,c_j)}\right)}\Bigg) \qquad (1)$$

Where:

$\mathrm{RD}_{w_i}^{c_j}$ = Relevance degree of the term w_i in the category c_j

$f_{(w_i,c_j)}$ = Term frequency w_i in the category c_j

$\max\left(f_{(w_z,c_j)}\right)$ = Highest frequency represented by the term w_z in the category c_j



Word	Total Occurences	Document Occurences
dollar	15	1
year	13	1
gas	11	1
High	9	1
us	9	1
fair	8	1
market	8	1
week	8	1
Has	8	1
billion	7	1
day	7	1
readjustment	7	1
increase	6	1

Fig. 3. DB with terms of a category

For each of the analyzed categories in the document organization model, the database should store keys to identify the document, to identify the category, the representation of the term, the frequency of the term in the category and the relevance degree in the category.

The purpose of the terms database is to identify all the words that best express the characteristics of a given category. This database is built according to of the most common terms between the various documents of the same category. Thus, Definition of variables if a particular category has at least 100 thousand words, the designed database should have a minimum sample size of 380 words per category, considering a sampling error margin of 5% with a 95% confidence level.

Definition of Variables

Many studies related to Information Retrieval (IR) work with models of document indexing, document representation and similarity measures from retrieved documents. In this research, the concepts of similarity and accuracy were used. That is, similarity S^c_{PT} between the parties involved: The Parsed Text (PT) and the texts from the (DB). In turn, accuracy A^C_{PT} measured the reliability of results presented by similarity.

The Parsed Text (PT) went through the same text document preparation process presented previously: Extraction, cleanup, stop-words removal and character case transformation. From the frequency of terms in PT_{w_i} , multiplication by the relevance degree of the term in each category from DB_{w_i} was performed. Equation 2 expresses the relationship between the terms of (PT) and the terms of (DB), where f_{PTw_i} is frequency and $RD^c_{w_i}$ is the relevance degree of the term for a specific category of the DB. The $V^c_{PTw_i}$ stores the frequency values of PT in relation to their relevance degree for each category in DB. For each evaluated word in the analyzed text the respective $RD^c_{w_i}$ has been observed:

$$V^c_{PTw_i} = f_{PTw_i} * RD^c_{w_i} \tag{2}$$

where, $V^c_{PTw_i}$ represents a value for each word calculated as a function of frequency f_{PTw_i} and the relevance degree $RD^c_{w_i}$ f the word. From the $V^c_{PTw_i}$ of each word the average of $V^c_{PTw_i}$ from the analyzed text has been extracted:

$$AV^c_{PT} = \frac{1}{nw_{PT}} \sum_{i=1}^{nw_{PT}} V^c_{PTw_i} \tag{3}$$

Also, the average has been extracted from all relevance degrees $\left(RD^c_{w_i}\right)$ of each word of a given category from the database:

$$ARD^c_{DB} = \frac{1}{nw_c} \sum_{i=1}^{nw_c} RD^c_{w_i} \tag{4}$$

The similarity S^c_{PT} of the Parsed Text (PT) regarding to the DB Texts is calculated according to the average values $\left(AV_P T^c\right)$ of PT, calculated by Equation 3. The AV^c_{PT} is divided by the Average Degree $\left(ARD^c_{DB}\right)$ of terms for each category of the DB. Thus, S^c_{PT} is obtained through Equation 5:

$$S^c_{PT} = \frac{AV^c_{PT}}{ARD^c_{DB}} \tag{5}$$

Confidence is the number of terms of the analyzed text (PT) corresponding to the database (DB) for each category. Thus the calculation is represented by Equation 6:

$$A^C_{PT} = 1 - \frac{(nw^c_{DB} - nw^c_{DB\cap PT})}{nw_{PT}} \tag{6}$$

Where:

nw^c_{DB} = Total terms for each category of DB

$nw^c_{DB\cap PT}$ = Total terms which are common between the DB and the analyzed text (PT)

nw_{PT} = Total terms from the analyzed text (PT)

Classifying Training Documents

A set of test documents, within the categories specified in the terms database, is selected to build a knowledge base on the results of the calculation of similarity variables and accuracy. Thus, the whole set of initial test documents passes through the preprocessing step, just as the documents that generated the database of terms. After this step, the vector representing the test document is processed and values of similarity and accuracy for each category under consideration are obtained. Furthermore, the prior index of the category in each test document is considered, i.e., the category from which the document has been extracted, is initially identified.

Therefore, each analyzed text has the similarity $\left(S_{PT}^{c}\right)$ and accuracy $\left(A_{PT}^{C}\right)$ taken regarding to all categories from the DB. Table 1 presents the results of this process of extraction from S_{PT}^{c} and A_{PT}^{C} conducted on two texts. The attribute "belongs" used in the table corresponds to the actual text classification based on their category. For example, Text 1 belongs to the category "Economy", thus similarity and accuracy were evaluated and assigned to each of the four categories stored in the database, according to the aforementioned equations.

In all results, the Similarity equation pointed exactly which category the text belonged to, without error. Accuracy, as shown by the results, has supported Similarity when it was not the highest value, being at least the second highest resulting value. These results were carried for the classification of 352 texts and were considered satisfactory to design a fuzzy model. To construct the fuzzy model all results of similarity and accuracy were used from the analyzed texts, according to the data presented in Table 1.

Data Classification Process

From the results shown in Table 1, there was a process of classification for the 352 texts. The process of data classification uses Decision Tree (DT) algorithms and aims to provide a more deterministic approach to the construction of the fuzzy model. The DT is built considering the concept of entropy, which measures the information level of an attribute. The smaller the entropy value, the lower the uncertainty and the most useful the attribute is for classification. Figure 4 shows the tree generated under the tested dataset by applying the ID3 classification algorithm (Quinlan, 1986) in the RapidMiner tool. On this DT the values of the results from accuracy and similarity were normalized to percentages.

Figure 4 shows the first analysis of the mining process, with lowest entropy in the similarity attribute already categorized. The results of variables similarity and accuracy were clustered into three groups: High, medium and low. Due to this construction, the linguistic variables include high, medium and low for both the similarity and accuracy attributes. That is two thresholds were established for each variable (similarity and accuracy). Thus, the limits that classify the results into high, medium and low were obtained. In Fig. 5 we present the results of the processing from the second DT algorithm, C4.5 (Quinlan, 1993), to build the tree.

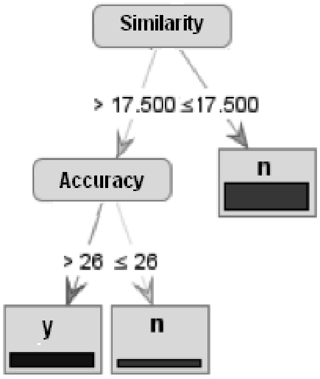


Fig. 4. Result of the ID3 process

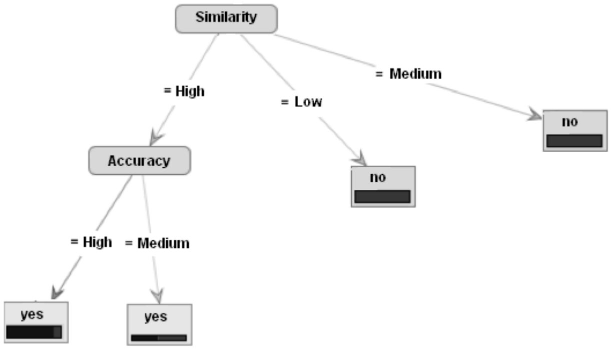


Fig. 5. Result of the C4.5 algorithm

Table 1. Results of text analysis

Text	Category	Similarity (S)	Accuracy (A)	Belong (B)
2	Economy	1	1	yes
2	Education	0,04	0,09	no
2	Technology	0	0	no
2	Sport	0,1	0,09	no
3	Economy	1	1	yes
3	Education	0,08	0,8	no
3	Technology	0	0	no
3	Sport	0,34	0,85	no

The proposal of classifying data was performed to adjust, in the best possible way, the membership functions and this was only possible through the DT in Fig. 5. A fuzzy model was generated according to this tree. It is observed that, in this tree, when the Similarity is High (SH) and the Accuracy is High (AH) the text Belongs (BY) to a category with, at least, 85% of certainty. In cases where the Similarity is High (SH) and the Accuracy is Medium (AM) the text belongs (BY) with about 50% of certainty to the category and considering the medium (SM) and low (SL) similarities, the text does not belong (BN) to the category with certainty close to 100%. Thus, the membership functions and fuzzy model rules have been adjusted to meet this purpose.

Table 2. Performance evaluation matrix of the DT model

		Real class	
Correspondence matrix		Yes	No
Expected class	Yes	26	6
	No	8	88

The model verification was conducted with a set of 32 test documents (Fig. 5). This model is represented by the DT that was generated by processing the training set. The results presented by the matrix in Table 2 evaluate the performance of the DT model.

The accuracy and the error of this model consider the results of the matches between the real class and the expected class observed in Table 2 and followed the definitions of the functions as presented in the methodology of this study. Thus, the calculation of the accuracy from the model resulted in 89%. This result was considered adequate for the implementation of the rules in the fuzzy model based on the DT in Fig. 5.

Building the Fuzzy Model

The proposal of the fuzzy model for text classification in multiple categories is relevant because it can handle the imprecision of knowledge coming from gaps and blanks present in the dataset. Moreover, fuzzy logic is able to summarize the results of a classification considering two inputs (similarity and accuracy) into an output with level of relevance.

In this study, MATLAB (2016) was used to build and simulate the Fuzzy Model. The MATLAB platform is optimized for solving engineering and scientific problems, moreover it has a Fuzzy Logic Toolbox that provides functions, apps and a Simulink block for analyzing, designing and simulating systems based on fuzzy logic.

The fuzzy modeling is composed of a fuzzification process, a knowledge base and a defuzzification process. The knowledge base is represented by the rules, inference engines and aggregation functions. The structure of the described fuzzy modeling is shown in Fig. 6.

The fuzzy model was built based on the results of the DT that allowed the creation of a more precise approach for both the input variables as in the rule base. From that, the fuzzy model was able to treat the uncertainty in the classification process of texts.

The input function called similarity was developed with the linguistic variables high, medium and low and the input function accuracy was built with only the high and medium linguistic variables as presented in the classification process of the DT. The output function called belong was developed with the linguistic variables yes, maybe and no. All membership functions have been implemented with Gaussian distribution functions, which have a simple curve. In particular *gaussmf* function was used, which has a bell curve with maximum of 1 and minimum at 0. Figure 7 shows the overview of the fuzzy model.

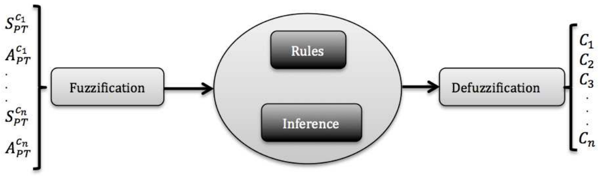


Fig. 6. The structure of the fuzzy modeling

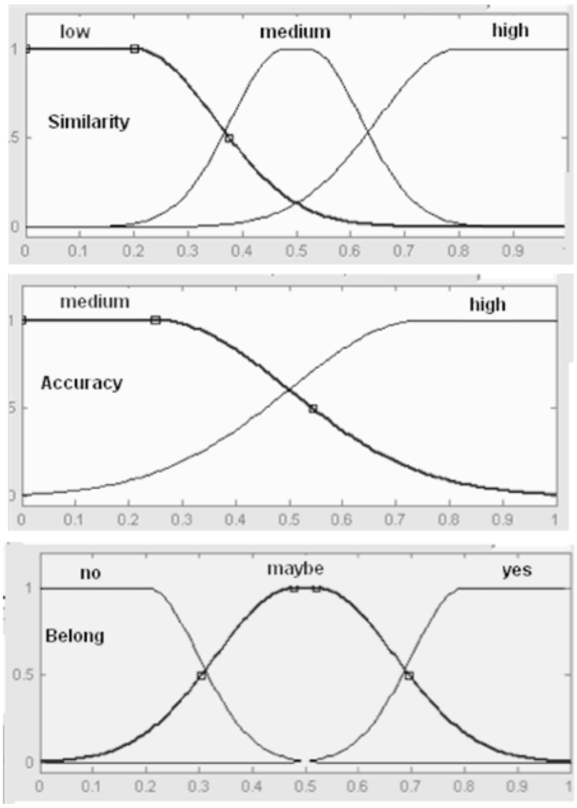


Fig. 7. Fuzzy modeling in text classification

The controller of the inference engine used was Mamdani, implemented with the minimum operator and the defuzzification method was the smallest of maximum.

According to Wang (1997), defuzzification is the reverse procedure of fuzzification. In this phase, we design a mapping from a fuzzy set to a crisp value. There are several defuzzifiers, such as Center of Gravity (Centroid), Center Average, Smallest of Maxima (SoM), Mean of Maxima (MoM) and Largest of Maxima (LoM). In this study we have used the Smallest of Maxima (SoM) as the defuzzification method.

Acording (Yaguinuma *et al*., 2013), when using fuzzy DL with concept definitions, the defuzzification queries available are SoM, MoM, LoM, which consider the extremes of maximum degree. Depending on the situation, they may lose information compared with other defuzzification methods that are based on the shape of the fuzzy set. In the experiments of this study the results with the SoM method were more accurate.

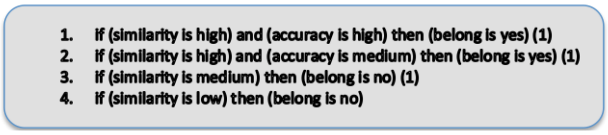
The Smallest of Maxima (SoM) represents the choice of smallest point of the universe with the highest degree of relevance. We calculate the So Musing the following equation:

$$\mu_{SoM} = \min(x_i), such\ that \varphi_A(x_i) = \varphi_{\max,}$$

where, φ_A is the membership function of set A.

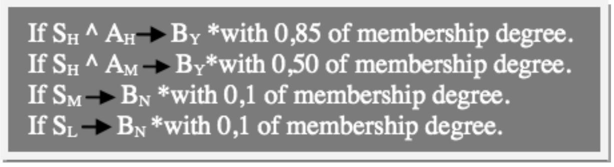
Four inference rules that cover all linguistic variables were defined, each rule consists of the operator and associated with the method minimum. The aggregation of the rules is made by the method maximum. Figure 8 presents the rule base.

The rules shown in Fig. 8 were built according to the definitions presented in the C4.5 decision tree algorithm (Fig. 5). Thus, the membership functions of the fuzzy model were adjusted so that the output could correspond to the values presented by the DT, according to Fig. 9.



1. If (similarity is high) and (accuracy is high) then (belong is yes) (1)
2. If (similarity is high) and (accuracy is medium) then (belong is yes) (1)
3. If (similarity is medium) then (belong is no) (1)
4. If (similarity is low) then (belong is no)

Fig. 8. Fuzzy modeling rules in text classification



If S_H ^ A_H → B_Y *with 0,85 of membership degree.
If S_H ^ A_M → B_Y*with 0,50 of membership degree.
If S_M → B_N *with 0,1 of membership degree.
If S_L → B_N *with 0,1 of membership degree.

Fig. 9. Fuzzy model rules adjusted by the decision tree

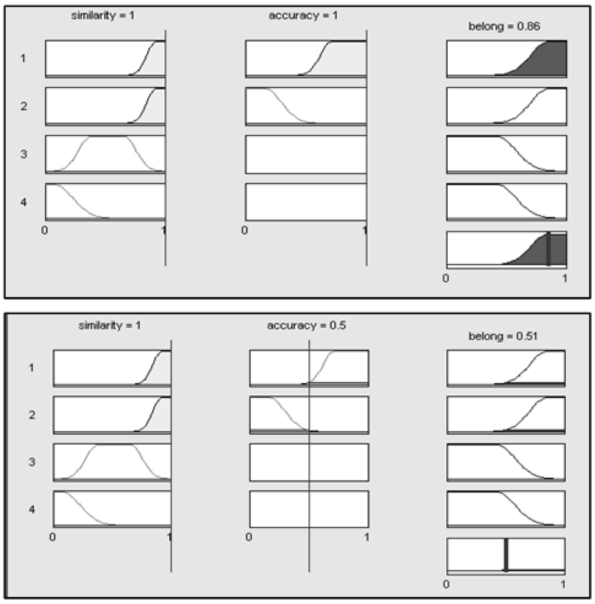


Fig. 10. Results after applying the first and second rules in the fuzzy model

The results in fuzzy model are relevant because they show the activation of the membership function in at least two points for each of the input functions. Figure 10 provides the fuzzy categorization when the similarity is high and accuracy is high, showing similar results to that of the DT. Also, when the similarity input is high and accuracy is medium.

Results

It is known that there is, usually, a predominant category when performing the classification process of the document. Yet, there are other relevant values associated with the document for each of the categories considered in the domain. These values vary in a range of [0,1].

The results of the model, through the fuzzy classifier, with their respective inputs for accuracy and similarity to each category are shown in Table 3. These results show the synthesis of the defuzzified output from the fuzzy modeling for each entry.

The overall assessment was performed in relation to the allocation of the defuzzified output value for the category in the original document indexing. In some cases, it is clear that even if the similarity is relatively high, if the accuracy does not have a significant value to the category, the result of fuzzy output is also not favorable. The fuzzy model prioritizes the combination of the values of both of the most significant indicators (similarity and accuracy) in the result set.

Table 4 shows the summary of the overall performance, considering a set of 97 documents in relation to the prior indexing of the document to a particular category and the fuzzy output generated by the model.

According to Table 4 the hit rate was about 78%, that is, from the 97 documents of the training set, 75 had the defuzzified output value corresponding to the original indexing.

Table 3. Results of text analysis with fuzzy output

Text	Category	Similarity (S)	Accuracy (A)	Belong (B)	Fuzzy out
2	Economy	1	1	yes	0,877
2	Education	0,04	0,09	no	0,121
2	Technology	0	0	no	0,265
2	Sport	0,1	0,09	no	0,123
3	Economy	1	1	yes	0,877
3	Education	0,08	0,8	no	0,467
3	Technology	0	0	no	0,117
3	Sport	0,34	0,85	no	0,490

Table 4. General results

Indexed category	Score
Yes	76
No	21
Total	97

Conclusion

In this research, a database of terms with four different categories of texts was constructed. Each term was normalized within their category in order to evaluate their degree of importance within it. A multi-label text classification process was built based on two concepts, similarity and accuracy. In the preliminary results, the model has worked consistently as expected while keeping correspondence between the text and its category.

The most experiments in text classification have been made from statistical techniques. Considering the advantages of text classification, the ideal would be that more experiments should be designed to treat other languages, while taking into account the peculiarities of each, especially regarding the semantic analysis of texts. According to (Manning *et al*., 2008) the statistical text classification require a number of good example documents for each class. Furthermore, the need for manual classification isn't eliminated because the training documents come from a person who has labeled them.

In order to increase the effectiveness of the presented results by the statistical analysis of this research, a way to implement a fuzzy model was studied, enabling it to inform the relevance degree of the text to each category. Decision tree algorithms were used to build a fuzzy model that could match the settings of results in a database. This approach has served its purpose, because the constructed model represents closely the results achieved with the analysis. Moreover, it was possible to generate output indicating the relevance in which the analyzed text belonged to a given category. Text classification problems are well represented when solutions involve fuzzy models, since classification meets the required expectations and also indicate what is the.

Acknowledgement

We acknowledge the Federal University of Santa Catarina and the Engineering and Knowledge Management Program for providing the the opportunity to develop this research.

Funding Information

Authors would like to thank the Ministry of Education for providing the financial support directly to the Engineering and Knowledge Management Program to which the research is linked.

Author's Contributions

Beatriz Wilges: Main researcher in the project. Designed the research and developed the proposed solutions. Responsible for the writing of the majority of the paper.

Gustavo Mateus: Prepared the workflow for the experiments, drew illustrations and also organized data of the tables.

Silvia Nassar: Organized the writing and structure of the manuscript.

Renato Cislaghi: Responsible for editing some parts of the paper and commenting research ideas.

Rogério Cid Bastos: Research advisor. Supervision and monitoring of the research.

Ethics

This article is original and contains unpublished materials. The corresponding author confirms that all other authors have read and approved the manuscript and there are no ethical issues involved.

References

Baker, L.D. and A.K. McCallum, 1998. Distributional clustering of words for text classification. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 24-28, Melbourne, Australia, pp: 96-103. DOI: 10.1145/290941.290970

Bang, S.L., J.D. Yang and H.J. Yang, 2006. Hierarchical document categorization with k-NN and concept-based thesauri. Inform. Process. Manage., 42: 387-406. DOI: 10.1016/j.ipm.2005.04.003

Dumais, S. and H. Chen, 2000. Hierarchical classification of Web content. Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval, Jul. 24-28, Athens, Greece, pp: 256-263. DOI: 10.1145/345508.345593

Lewis, D.D. and M. Ringuette, 1994. Comparison of two learning algorithms for text categorization. Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, (AIR' 94), pp: 81-93.

Lewis, D. D, 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. European conference on machine learning. Springer Berlin Heidelberg.

Jiang, J.Y., S.C. Tsai and S.J. Lee, 2012. FSKNN: Multi-label text categorization based on fuzzy similarity and k nearest neighbors. Expert Syst. Applic., 39: 2813-2821. DOI: 10.1016/j.eswa.2011.08.141

Wang, L.X., 1997. A Course in Fuzzy Systems and Control. 1st Edn., Prentice Hall PTR, Upper Saddle River, ISBN-10: 0135408822, pp: 424.

MATLAB, 2016. Fuzzy logic toolbox design and simulate fuzzy logic systems. The MathWorks, Inc, MATLAB.

Manning, C.D., P. Raghavan and H. Schütze, 2008. Introduction to Information Retrieval. 1st Edn., Cambridge University Press. ISBN-10: 1139472100.

Li, M., L. Liu and C.B. Li, 2011. An approach to expert recommendation based on fuzzy linguistic method and fuzzy text classification in knowledge management systems. Expert Syst. Applic. Int. J., 38: 8586-8596. DOI: 10.1016/j.eswa.2011.01.062

Peters, C.M.E.E. and C.H.A. Koster, 2003. Uncertainty and term selection in text categorization. Int. J. Unc. Fuzz. Knowl. Based Syst., 11: 115-115. DOI: 10.1142/S0218488503001977

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. 1st Edn., Morgan Kaufmann, San Mateo, ISBN-10: 1558602380, pp: 302.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learn., 1: 81-106. DOI: 10.1023/A:1022643204877

RapidMiner, 2016. http://rapid-i.com/

Tan, S., 2006. An effective refinement strategy for KNN text classifier. Expert Syst. Applic. Int. J., 30: 290-298. DOI: 10.1016/j.eswa.2005.07.019

Wilges, B., G. Mateus, R. Bastos and M. Dantas, 2014. A case-comparison study of automatic document classification utilizing both serial and parallel approaches. J. Phys. Conf. Series.

Yaguinuma, C.A., W.C. Magalhães Jr, M.T. Santos, H.A. Camargo and M. Reformat, 2013. Combining fuzzy ontology reasoning and mamdani fuzzy inference system with HyFOM reasoner. Proceedings of the International Conference on Enterprise Information Systems, Jul. 4-7, Angers, France, pp: 174-189. DOI: 10.1007/978-3-319-09492-2_11

Yang, Y. and J. Pedersen, 1997. A comparative study on feature selection in text categorization. Proceedings of 14th International Conference on Machine Learning, (CML' 97), Morgan Kaufmann Publishers, San Francisco, US pp: 412-420.

Yang, Y. and X. Liu, 1999. A re-examination of text categorization methods. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, Berkeley, CA, USA, pp: 42-49. DOI: 10.1145/312624.312647