International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:7, 2007

A Comparative Study of Web-pages Classification Methods using Fuzzy Operators Applied to Arabic Web-pages

Ahmad T. Al-Taani, and Noor Aldeen K. Al-Awad

***Abstract*—** In this study, a fuzzy similarity approach for Arabic web pages classification is presented. The approach uses a fuzzy term-category relation by manipulating membership degree for the training data and the degree value for a test web page. Six measures are used and compared in this study. These measures include: Einstein, Algebraic, Hamacher, MinMax, Special case fuzzy and Bounded Difference approaches. These measures are applied and compared using 50 different Arabic web-pages. Einstein measure was gave best performance among the other measures. An analysis of these measures and concluding remarks are drawn in this study.

***Keywords*—** Text classification, HTML, Web pages, Machine learning, Fuzzy logic, Arabic Web pages.

I. Introduction

WITH the rapid growth of the internet, there is an increasing need to provide automated assistance to Web users for Web page classification. Such assistance is helpful in organizing the vast amount of information returned by search engines, or in constructing catalogues that organize Web documents into hierarchical collections [1]. Classification is expected to play an important role in future search services. For example, Chen *et al.* [2] showed that users prefer navigating through catalogues of pre-classified content. In order to meet such a strong need, we need automated Web-page classification techniques.

Web-page classification is harder than free text classification because of the noisy information founded in them such as advertisement represented through images, media sounds, navigation bars, and page formatting. So we need to summarize and benefit from these data and make them useful for end user who needs to manage and plan their work depending on a more accurate classification process. It is an essential matter to focus on the main subjects and significant content. As a result the critical task to deal with ambiguous web pages and their embedded structure through studying HTML language to remedy the process and then using some classification method such as machine learning, or fuzzy set theory [3].

Recently much work has been done on Web-page classification [1][4 -15]. In these approaches different methods are proposed. These methods includes: Web summarization-based classification, fuzzy similarity, natural language parsing web page classification and clustering to find reliable list answers, text classification approach using supervised neural networks, machine learning methods, *k*NN model-based classifier, and fuzzy classifiers.

In this study, an analysis and comparison of six fuzzy similarity approaches applied to Arabic web pages classification is presented. The clustering scheme is built and known for each category from training documents and the similarity between a test document and a category is measured using a fuzzy relation.

This relation is called fuzzy term-category relation; where the set of membership degree of words to a particular category represents the cluster prototype of the learned model. Based on this relation, the similarity between a document and a category's cluster center is calculated using fuzzy conjunction and disjunction operators [4].

After that the calculated similarity represents the membership degree of document to the category, and each membership functions of fuzzy sets take values in [0,1] that is used for testing different test documents in order to come out with the weakness points as well as the strength points. It may be observed then that representing the document as a Boolean features vector [4] simplifies greatly the fuzzy similarity formula and reduces it to a major factor.

II. Methods

A. Overview of the Proposed Approach

A fuzzy similarity approach is used for Arabic web-pages classification. The proposed system is composed of five stages: Training, Noise Elimination, learning, classification, and testing stages. The clustering scheme is built and already known for each category from training documents and the similarity between a test document and a category is measured using a fuzzy relation.

This relation is called fuzzy term-category relation, where the set of membership degree of words to a particular category represents the cluster prototype of the learned model. Based on this relation, the similarity between a

Ahmad T. Al-Taani is with the Department of Computer Sciences, Yarmouk University, Irbid, Jordan (phone: +962-7-77438520; fax: +962-2-7211128; e-mail: ahmadta@yu.edu.jo).

Noor Aldeen K. Al-Awad is a graduate student at the Department of Computer Sciences, Yarmouk University, Irbid, Jordan (e-mail: noor_kamel@yahoo.com).

document and a category's cluster center is calculated using fuzzy conjunction and disjunction operators. After that the calculated similarity represents the membership degree of the test document to the category, and each membership function takes a value between 0 and 1. This value is used for testing different test documents in order to come out with the weak points as well as the strength points. The document is then represented as a Boolean features vector which greatly simplifies the fuzzy similarity formula and reduces it to a major factor [4].

Six well known measures using fuzzy operators are used and applied to different Arabic web pages. These measures include: Einstein, Hamacher, bounded difference, Algebraic, MinMax, and Scfuzzy approaches. Then comparison between these measures is presented in this study. Finally, concluding remarks are drawn at the end of this study.

B. Categories with their Related Terms

Text documents are represented as a set of categories: C = $\{c_1, c_2, \ldots, c_n\}$. The category of a document D: $c(D) \in C$, where c(D) is a categorization function whose domain is D and whose range is C [16].

To compute the similarity model of each text document, each document is passed to an HTML stripper, to a stop word eliminator, and to a stemmer. Then weights are computed for each term. These weights represent statistical similarity between documents.

Background knowledge is used in the classification process. Such background knowledge provided us with a corpus of text that contains information both about the importance of words (in terms of their membership values in a large corpus), and the probability of words (what percentage that a test document will participate in the sameness process with the document). This gave us a large context in which to test the similarity of a training example with a new test example. Then this context is used in conjunction with the training examples to label a new web page.

C. Fuzzy Conjunction/Disjunction based Algorithm

Each document d has one category c; and as a result each category has one or more document. A set of n documents and their related categories are represented as an ordered pair where D = $\{\langle d_1, c(d_1)\rangle, \langle d_2, c(d_2)\rangle, ..., \langle d_n, c(d_n)\rangle\}$. The resulted documents that have many terms are stored with their relevant categories, as each row represents (term, document no., category no.). Then each term is counted for each document and is represented as the term and it's frequency as follows: p = $\{\langle t_1,f(t)\rangle, \langle t_2,f(t_2)\rangle, ..., \langle t_n,f(t_n)\rangle\}$. Where $f(t_i)$ is the frequency of the term t_i in the document or web page p.

Now the frequencies for each term are summed up for all the documents of a category to give the repetition of terms among their relevant categories. The membership value for each term is obtained by using (1)

$$M(t_i, c_j) = \frac{\sum f(t_i)\ \{(c(p_k) = c_j), (f(t_i) \in, (p_\kappa \in D\}}{\sum f(t_i)\ \{(f(ti) \in, (p_\kappa \in D\}} \qquad (1)$$

The membership value in fuzzy set theory denotes the degree of relevance of term t_i to category c_j. There are multiple categories with their membership degree values related to each term [4].

There are some consequences about the documents that are classified in many categories; and this may result implicitly into moderately convergence between each degree of voting or distribution for the term being examined.

D. Fuzzy-based Similarity Approach

After computing the membership values of the individual terms in each category, then we need to measure the likelihood for a given test web page to be classified into the existing categories of the training datasets.

The test web page can be classified correctly if each of its terms is participated in the process of comparison or similarity. Let a test web page w = $\{\langle t_1, Deg(t_1)\rangle, \langle t_2,Deg(t_2)\rangle, ... , \langle t_n,Deg(t_n)\rangle\}$, where $Deg(t_i)$ represents the membership degree for a (t_i) to be associated with a test web page and computed by (2)

$$Deg(t_i) = \frac{f(t_i)}{Max(f(t_i))} = \frac{f(t_i)}{f(t_m)} \qquad (2)$$

Then the similarity between w and a category c_j is given by (3)

$$sim(w,c_j) = \frac{\sum_{t \in w} M(t,c_j) \otimes Deg(t)}{\sum_{t \in w} M(t,c_j) \oplus Deg(t)} \qquad (3)$$

where $\otimes$ and $\oplus$ denote the fuzzy conjunction and disjunction operators, respectively. For more details about these operators refer to [4][17].

E. The Classification Task

Depending on the value of $sim(w,c_j)$, we need to repeat the same calculation for the next category with the same document and the next category till the last one. Then the category of the test web page p is the one that represents its contents by selecting the largest value of the equation outputs, i.e. Cat(p) = MAX($sim(w,c_1)$, $sim(w,c_2)$, . . . $sim(w,c_n)$).

III. Experimental Results

We have first collected the training data from different sources, and then applied the different stages of the fuzzy similarity approach to these data. Fig. 1 below shows the results after applying the six measures on 50 Arabic web pages, 5 web pages for each category. These categories are: *1: Autobiography (Auto), 2: Children's Stories (Child), 3: Economics (Eco), 4: Health and Medicine (Hlth), 5: Interviews (Intrv), 6: Religion* (Rlg), *7: Science(Scnc), 8: Short Stories(Short), 9: Sociology (Socio), 10: Tourist and Travel (Trst).*

From Fig. 2, we can conclude that the algorithms perform differently for all categories depending on their precision, the category itself, or the whole test data set.

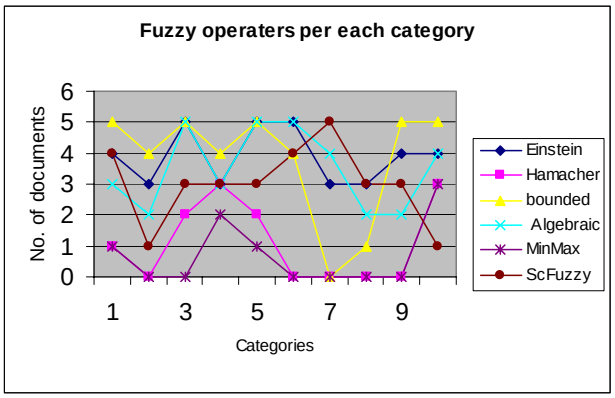

Fig. 1 Measures Performance

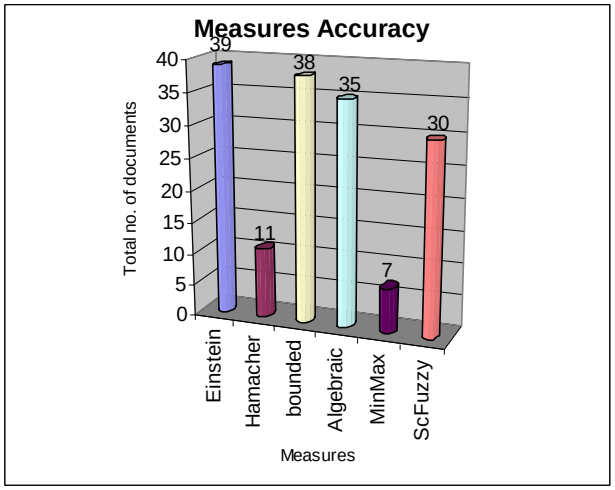

Fig. 2 Measures Accuracy

Einstein measure was gave best performance among the other measures and then the Bounded measure followed by Algebraic measure.

IV. Conclusions and Future Work

We have presented in this paper a fuzzy similarity approach for Arabic web-page classification. The approach used fuzzy term-category relation by manipulating membership degree for the training data and the degree value for a test web page. We used and compared six measures in this study. These measures are: Einstein, Hamacher, bounded difference, Algebraic, MinMax, and Special case fuzzy (Scfuzzy). The best performance is achieved by the Einstein measure then the Bounded measure followed by Algebraic measure.

The training data is first collected from different sources, and then normalized by passing it through the noise elimination module. The approach also includes the HTML stripping, stop word removing, and stemming. The learning process began by representing terms as numbers to reduce their representation. The final step in the process was to apply the six measures to the web pages.

Future work will consider the use of hyperlinks embedded in each web page to some depth and find out the synonyms of their text terms, i.e. classifying pages depending on their hyperlinks, in which each web page is categorized based on the group of web pages that it refers to, and recursively get the category label with the most proposed one.

References

[1] Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, Wei-Ying Ma. Web-page Classification through Summarization. Proc. of the ACM SIGIR 04, , July 25–29, 2004. Sheffield, South York Shire, UK.

[2] H. Chen and S. T. Dumais. Bringing order to the Web: Automatically categorizing search results. Proc. of CHI2000, 2000, 145-152.

[3] D. Michie, D.J. Spiegelhalter, C.C. Taylor. February 17, 1994. Machine Learning, Neural and Statistical Classification, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge, U.K.

[4] Dwi H. Widyantoro and John Yen, Department of Computer Science Texas A&M University, 1999. A Fuzzy Similarity Approach in Text Classification Task, Texas, USA.

[5] Hui Yang, Tat-Seng Chua. Effectiveness of Web Page Classification on Finding List Answers. Proc. of the ACM SIGIR 04, July 25–29, 2004, Sheffield, South York Shire, UK.

[6] Stephanie W. Haas, and Erika S. Grams. Page and Link Classifications: Connecting Diverse Resources. Proc. of the ACM, 1998. 99-107. Digital Libraries 98 Pittsburgh PA USA.

[7] Michelangelo Ceci and Donato Malerba. Hierarchical Classification of HTML Documents with WebClassII. F. Sebastiani (Ed.): ECIR 2003, LNCS 2633, pp. 57-72, 2003.

[8] Rongbo Du, Reihaneh Safavi-Naini and Willy. Web Filtering Using Text Classification, 2002, supported by Smart Internet Technology Cooperative Research Centre, Australia.

[9] Lawrence Kai Shih and David R. Karger. Using URLs and Table Layout for Web Classification Tasks. WWW2004, May 17–22, 2004, pages 193-202, supported by ACM, New York, USA.

[10] Eric J. Glover1, Kostas Tsioutsiouliklis, Steve Lawrence, David M. Pennock , Gary W. Flake. Using Web Structure for Classifying and Describing Web Pages. WWW2002, May 7–11, 2002, , pages 562-569, supported by ACM, Honolulu, Hawaii, USA.

[11] Gongde gue, Hue Wang, David bell, Yaxin bi, and Kairan Greer. A KNN model-based approach and its application in text categorization, 2002, supported by European Commission project ICONS, project no. IST-2001-32429.

[12] Anders Ardö, DTV, Lyngby, Denmark Traugott Koch, NetLab, Lund, Sweden. Automatic classification applied to the full-text Internet documents in a robot-generated subject index, 1999. Manuscript of a forthcoming publication in proceedings of the Online Information 99 Conference, London.

[13] Aijun An, Yanhui Huang, Xiangji Huang, and Nick Cercone. Feature Selection with Rough Sets for Web Page Classification, 2002. Supported by natural Sciences and Engineering Research Council (NSERC) of Ontario, Canada and the Institute for Robotics and Intelligent Systems (IRIS).

[14] Hans Roubos, Magne Setnes, and Janos Abonyi, 2000. Learning Fuzzy Classification Rules from Data.

[15] Heiner Stuckenschmidt, Jens Hartmann and Frank van Harmelen, 2002, American Association for Artificial Intelligence (www.aaai.org). Learning Structural Classification Rules for Web page Categorization. Bremen, Germany.

[16] Sarah Zelikovitz, Haym Hirsh, 1999. Improving Short-Text Classification Using Unlabeled Background Knowledge to Assess Document Similarity. Computer Science Department, Rutgers University, USA.

[17] Włodzisław Duch. Similarity-based methods: a general framework for classification, approximation and association, Control and Cybernetics vol.29 (2000), Grudzia¸dzka, Toru´n, Poland.