I.J. Modern Education and Computer Science, **2016, 11, 20-27**
Published Online November 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2016.11.03

A Supervised Approach for Automatic Web Documents Topic Extraction Using Well-Known Web Design Features

Kazem Taghandiki
Department of Computer Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
Email: taghandiky@gmail.com

Ahmad Zaeri
Department of Computer Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
Email: zaeri@eng.ui.ac.ir

Amirreza Shirani
Department of Computer Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
Email: shiraniamirreza@gmail.com

***Abstract*—**The aim of this paper is to propose an efficient method for identification of web document topics which is often considered as one of the debatable challenges in many information retrieval systems. Most of the previous works have focused on analyzing the entire text using time-consuming methods and also many of them have used unsupervised approaches to identify the main topic of documents. However, in this paper, it is attempted to exploit the most widely-used Hyper-Text Markup Language (HTML) features to extract topics from web documents using a supervised approach.

Hiring an interactive crawler, we firstly try to analyze HTML structures of 5000 webpages in order to identify the most widely-used HTML features. In the next step, the selected features of 1500 webpages are extracted using the same crawler.

Suitable topics are given to each web document by users in a supervised learning process. A topic modeling technique is used over extracted features to build four classifiers- C4.5, Decision Tree, Naïve Bayes and Maximum Entropy- which are separately adopted to train and test our data. The results of classifiers are compared and the high accurate classifier is selected. In order to examine our approach in a larger scale, a new set of 3500 web documents is evaluated using the selected classifier. Results show that the proposed system provides remarkable performance which is able to obtain 71.8% recognition rate.

***Index Terms*—**Topic extraction, web document, supervised active learning, Topic modeling

I. Introduction

The world wide web has become the most important source of diverse information. From personal webpages to real-time news around the world. It encompasses heterogeneous webpages which are considered as one of the most challenging domains for information retrieval. Not only is the content of web documents constantly changing and new information is being added in every second, but they are written in different languages in different forms. In a 2015 research report [1], about 49 billion web pages have been indexed by Google and Bing crawlers. This shows an unprecedented growth in the volume of unstructured data in the web which has caused major problems in detection and access to relevant information.

Recent trends in creating massive and unstructured web documents demonstrate the urgent need to manage the large amount of data stream or, at least, extract the most important features in order to meet today's demands [2]. By unstructured, we mean that there are no computer-readable annotations that tell the computer the semantic meaning of the words in the text.

The topic or subject category of a web document is any word or sequence of words that indicates the main concept of the web document. It is to help readers finding the right type of content. Moreover, the topic is considered as one of the useful features in web mining or many other applications like the automatic construction of ontologies, document summarization and classification [3].

Most existing approaches for topic identification rely on the analysis of the entire document to calculate measures to analyze the distribution of terms like Term Frequency and etc. They mostly use unsupervised learning approaches to cluster documents.

In this paper, however, a topic selection method which considers just the most widely-used HTML features is presented. And unlike other works, classification algorithms is used to give a right label to each document in a supervised learning process. The procedure is explained as follows:

Using java libraries, an interactive crawler was created, and with the help of this crawler, HTML structure of 5000 webpages is examined to identify and extract the most widely-used webpage features. These 5000 webpages are all in English and coming from 114 internet domains. In our first experiment, only 1500 webpages are analyzed and three users are employed to assign one relevant subject to each webpage. Then MALLET topic modeling toolkit [4] is used over extracted features to build four classifiers- C4.5, Decision Tree, Naïve Bayes and Maximum Entropy- which are separately adopted to identify the topics of web documents. The results of different classifiers are compared and it is found that Maximum Entropy with the accuracy of 88% yields the highest possible result. In order to evaluate our system in a larger scale, in our next experiment, Maximum Entropy is used to extract the topics of the other 3500 web documents. Results show that the proposed system is able to obtain remarkable performance with the recognition rate of 71.8%.

The rest of the paper is organized as follows. Section II discusses the previous works about topic extraction methods. Section III introducing the proposed model also explains the implementation of data set and classifiers used in this paper. Section IV presents experimental results of the proposed system and Section V concludes the paper.

II. Related Work

Numerous studies on the topic extraction of web documents have been performed. Most studies rely on analyzing the distribution of terms and applying methods like TF-IDF, co-occurrence or n-gram over the entire documents. In an early work [5], title tag and Term Frequency measure are used to identify keywords of documents obtaining 32% precision. In many works like [6] an ontology structure is utilized in order to determine topics in web documents. They used Term frequency measure as preprocessing task and tested their approach on 100 web documents obtaining 76.62% and 59% Precision and Recall respectively.

In [7] an approximate version of the TF–IDF measure suitable to work on the continuous data stream is proposed. And in order to meet fast response requirements and overcome storage constraints when processing a continuous flow of data stream, the parallel GPU architecture is implemented. Similar to this work, the approach described in [8] uses a modified version of the TF–IDF measure called the TF-PDF (Term Frequency–Proportional Document Frequency) that recognizes the terms trying to explain the main topics in the news archive in a weekly basis. TF-PDF is designed in a way that it would assign heavy term weight to these kinds of terms and thus reveals the main topics.

In [9] they proposed a technique to extract interesting topics from online reviews using bi-grams, single-nouns and assign sentiment labels to these topics which are considered suitable for recommendation systems. In [3] TF-IDF, Phi-square, mutual information and variance measures are used to extract both single-words and multi-words as key-terms (topic) in a language independent method; Portuguese, English and Czech were the languages experimented.

Using supervised learning approach, [10] proposed a keyword extraction technique that uses lexical chains. In order to build lexical chains, WordNet was used to define word senses and semantic relations between words. In feature extraction phase, seven features used in the system are: the first and the last occurrence of the word in text randomly the number of word occurrences in text, lexical chain score, direct lexical chain, lexical chain span score and direct lexical chain span score of a word. C4.5 decision tree algorithm is used and the system reached 45% precision for full texts and 20% for abstracts when 5 keywords are extracted.

In many works like [11, 12] methods such as stop-words removing, lemmatization, part-of-speech tagging and syntactic pattern recognition are used to extract key words and noun phrases. In many other studies, LDA has been used to extract the topic of a document. For example, in [13] LDA model is applied for topic identification. Their method can consider the density of each topic and compute the most unstable topic structure through an iteration process.

In [14], authors proposed a three-phase approach for topic-based k-means++ document clustering. Firstly, the method can determine the best topic model and the best number of topics by the concept of significance degree of topics and the topic selection criterion. They show that their clustering approach based on the best topic model improves the performance of topic based document clustering.

To the best of our knowledge, topic extraction of web documents using extracted designing features of webpages has not been fully investigated. Most previous works have focused more on the computation of the measures like TF–IDF over the entire content of a document. Computing the whole content requires more computing time and resources. Unlike most previous works using unsupervised learning approach, in this paper, 4 classifiers are adapted to train and test our model in a supervised learning process.

III. Experiment Procedure

The implementation process of the proposed approach is described in four steps shown in Fig.1.

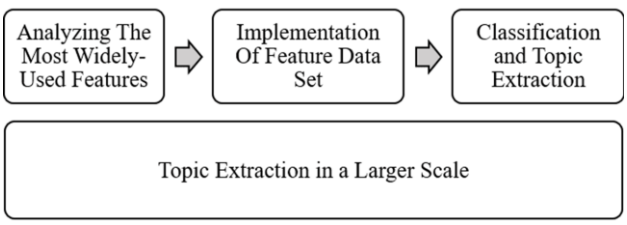

Fig.1. Implementation steps

In the first step, we try to analyze 5000 web documents in order to understand the most widely-used HTML features. In the second step, selected features of 1500 web documents are extracted and the feature data set is constructed. In the third step, 4 different classifiers are employed to classify topics of documents then performance evaluation of proposed model is performed. We dedicate the last part of this paper to evaluate the proposed model for the topic extraction of new 3500 web documents.

In this paper, the MALLET topic modeling toolkit is used. MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. Also it includes sophisticated tools for document classification. We assist this toolkit to classify topics from features we have extracted before. Also, in our work, an interactive crawler is constructed using libraries in Java programming language. This crawler pursues the following goals:

1- To find the most widely-used HTML features of webpages

2- To extract selected features from webpages.

Following sections present more details about the construction of features data set.

A. Analyzing The Most Widely-Used Features In Web Documents

As mentioned before, unlike previous works which mostly consider and analyze the entire document, in this paper, we tried to use important HTML features to identify the main topic of the pages. In order to find the most important features, five features which are relevant to the content of the web document are specified: "Title tag", "Heading tag", "Paragraph tag", "Description metadata" and "Keyword metadata". Analyzing the percent usage of these 5 features in our 5000 webpages, we conclude that "Title Tag", "Keywords Metadata" and "Heading tag" are three important and the most widely-used features to distinguish the topics of the webpages. All 5000 webpages are in English coming from 114 web domains. Fig.2 shows the percent usage of HTML features over 5000 web documents.

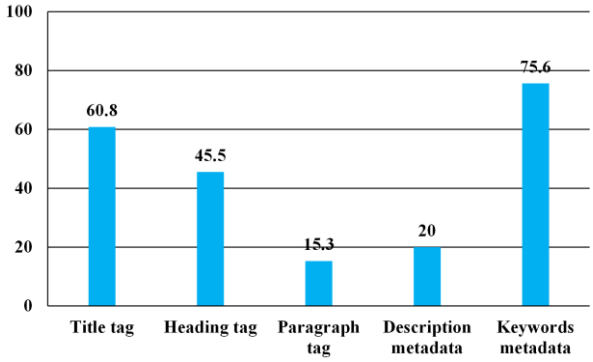

Fig.2. The percent usage of HTML important features.

As it is shown in fig. 2, "Keyword metadata" with 75.6 percent is the most popular HTML feature, then "Title tag" and "Heading tag" with 60.8 and 45.5 percent are three important features which we consider in this paper. "Description metadata" and "Paragraph tag" with under 20 percent, have more blank values and we decided to remove them from feature set.

To have a better understanding of three mentioned features Table 1 presents features vector of FileHippo website before applying any pre-processing tasks and assigning labels. This website is a download website that offers mostly windows software.

Table 1. Features vector from FileHippo webpage

FileHippo.com - Download Free Software	Title tag
download software freeware shareware program filehippo file hippo	Keyword metadata
The Latest Versions of the Best Software	Heading tag

B. Feature Extraction And Implementation Of Feature Data set

In this phase, with the help of interactive crawler, three popular features -"Keyword metadata", "Title tag" and "Heading tag"- from 1500 web documents are extracted. Three users are employed to assign a suitable topic to each web document. This process is done in an interactive action with users. 500 pages are randomly shown to each user who is asked to pick relevant topics for pages. Users assist existing topics from Open directory [15] project to assign a relevant topic. This raw open-source directory is used by google, Netscape search and AOL. In general, the directory is similar to the table of contents in a book. At the end, all 1500 web documents are assigned relevant topics. (Shown in Fig 3)

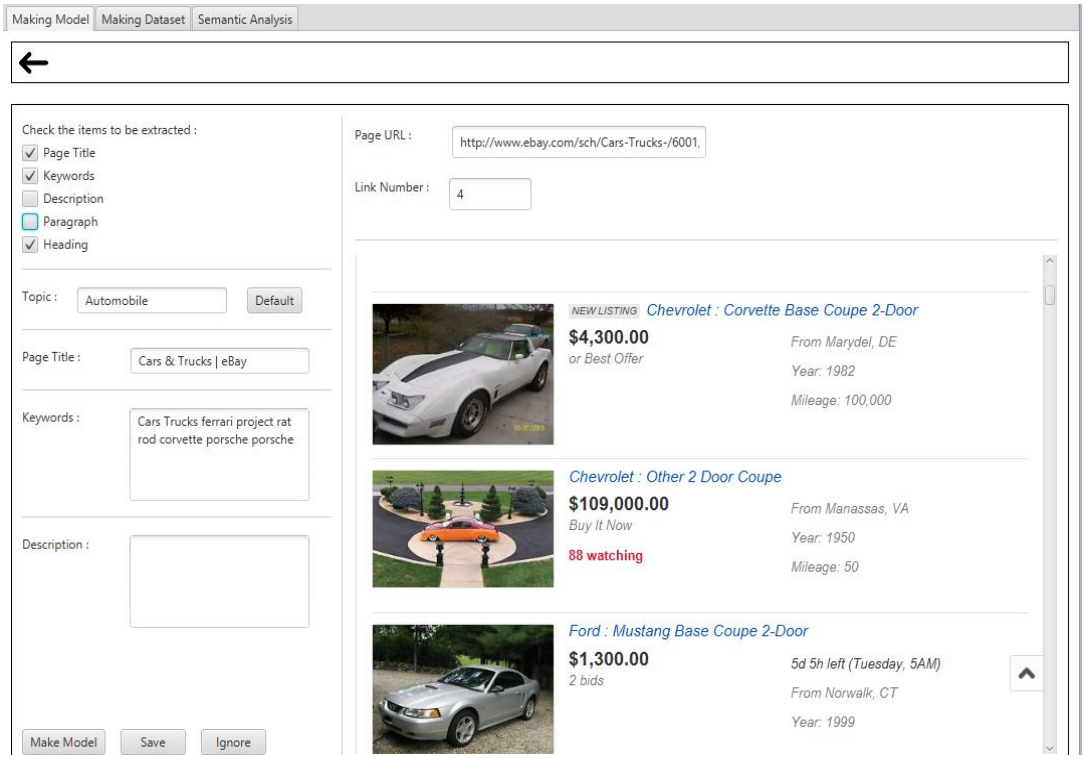

Fig.3. Features extracted from a web page by interactive crawler

Fig. 3 shows an example of the interactive process of indication a relevant topic for a car selling web page. In this process, important HTML features and the content of web page are shown and the user is asked to pick a relevant topic. The given topic fills the Label field in feature data set. So our feature data set consists of three features and a label field.

Using Mallett Toolkit, pre-processing tasks like "tokenization", "case folding", and "stop words elimination" along with "stemming" process using Lucene libraries are done on all features and labels in our data set. At the end, we have feature data set for 1500 webpages with correct given topics. Table 2 presents features vector of FileHippo website after applying pre-processing tasks and assigning a relevant label.

Table 2. Features vector of FileHippo webpage

download free software	Title tag
download software freeware shareware program	Keyword metadata
best software	Heading tag
software	Label

C. Classification and Topic Extraction

After computation of diverse features and selection the optimal set, they are fed into classifiers. In order to build our topic model, 4 classifiers - Naïve Bayes, C4.5, Decision Tree and Maximum Entropy are separately employed to train and test our data. In the following, a brief description of the mentioned classifiers is presented:

Naïve Bayes: Bayesian classifiers are statistical classifiers which are based on Bayes' theorem. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. They also yield high accuracy and speed when applied to large databases. One of the simplest implementation of Bayesian classifiers is known as the naïve Bayes classifier. Naïve Bayes is built on this assumption that the effect of an attribute value on a given class is independent of the values of the other attributes [16].

Decision Tree: Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values [17].

C4.5: C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [18] which is a successor of ID3. C4.5 can be used for classification, and for this reason, it is often referred to as a statistical classifier [17]. At each node, it chooses the best features to effectively split samples to two separate subsets.

Maximum Entropy: Unlike Naive Bayes, the Max Entropy does not assume that the features are conditionally independent of each other. The Maximum Entropy is based on the Principle of Maximum Entropy which selects the model with the largest entropy. The main idea behind Maximum Entropy is that the most uniform models are more preferable, and also they should satisfy given constraints [19] .

10-fold cross-validation is well-known for its relatively low bias and variance [16]. To validate the results, 10 Fold cross validation is used over the data set, in 10

iterations; 80% of data is considered as trained set and the remaining 20% as a test set. In order to have a better comparison, Accuracy and F-measure in percent and computing time in second of all 4 classifiers are computed and are shown in Fig 4.

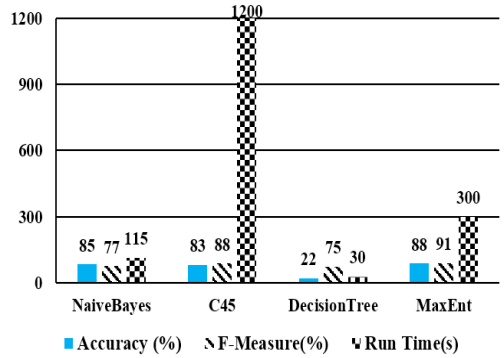

Fig.4. The Accuracy, F-Measure and Run Time of four classifiers -C4.5, Decision Tree, Naïve Bayes and Maximum Entropy

As it is shown in Fig. 4, Maximum Entropy yields better results compared to other classifiers; it is able to achieve 88% accuracy and 91% F-measure in a reasonable execution time. Naïve Bayes has better execution time, but considerable lower F-measure and Accuracy are achieved. Maximum Entropy seems the best option, so we choose Maximum Entropy as the main classifier. The problem with C4.5 is its long execution time. And the accuracy of Decision tree is not satisfying (22%).

The scripts of importing the feature data set as well as pre-processing tasks are shown in Fig. 5. We ignore a standard list of very common stop words in English like adverbs, conjunctions, pronouns and prepositions. We also convert all words to lowercase. Also, the scripts of creating classifiers together with evaluating the results in Mallet Toolkit is presented in Fig. 6.

```
bin\mallet import-file ^
--input Data-Crowled\feauterdatasetClassification.csv ^
--output train-input\classification-import.mallet ^
--remove-stopwords ^
--stoplist-file Data-Crowled\stoplists(en).txt ^
--preserve-case
--label 1 ^
--name 0 ^
--data 3 ^
```

Fig.5. Importing file and pre-processing tasks

```
bin\mallet train-classifier ^
--input train-input\classification-import.mallet ^
--trainer NaiveBayes ^
--trainer DecisionTree ^
--trainer C45 ^
--trainer MaxEnt
--training-portion 0.8 ^
--cross-validation 10 ^
--output-classifier classifier\classifier.mallet ^
--report train:accuracy test:f1:Actor
```

Fig.6. Creating classifiers and performance evaluation

As it is shown in Fig.6, Four different classifiers are separately employed and 10-fold cross validation is used to have a better evaluation. In each iteration, 80% of data is specified as a training portion and the remaining 20% as the test portion. The output is "classifier.Mallet" file which is a trained topic model.

D. Topic Extraction in a larger Scale

In the previous section, trained topic model is constructed. In this section, in order to test our method in a larger scale, we create a topic inference tool based on the current, trained model to identify topics of new 3500 webpages. Table 3, presents some examples of given topics by our model which indicates that the proposed system is able to extract relevant topic from each web document.

Table 3. Extracted topics from 10 sample web documents

URLS OF WEBPAGES	TOPIC
www.theguardian.com/environment/food	FOOD
http://filehippo.com/	SOFTWARE
www.divxcrawler.tv/latest.htm	MOVIE
https://actor.im/	ACTOR
https://mp3skull.wtf	SONG
www.foodnetwork.com	FOOD
http://www.miniclip.com	GAME
http://www.artwork.fm/	ART
www.theguardian.com/world/animals	ANIMAL
http://www.harvard.edu	HARVARD
http://www.bbc.com/sport	SPORT
http://www.npr.org/	MUSIC
vimeo.com	VIDEO
https://en.wikipedia.org/wiki/Automobile	Automobile
http://www.dell.com/	Dell
www.webopedia.com	Dictionary
http://www.animaljam.com/welcome	Animal
http://mobile.softpedia.com/	Mobile
http://www.softpedia.com/	Program

The proposed system is able to extract the topics of 81.8% of all new 3500 webpages (shown in Fig. 7). The main reason behind failure to extract topics of the remaining documents (18%), is that these documents do not full-fill the value of specified features properly or appropriate labels do not assign by users to cover the topics of web documents. Perhaps, by extracting more features from web documents we could enhance the ability of our system to extract topics of more web documents. However, by using more features the computation time would be increased dramatically. In the next section performance evaluation is presented.

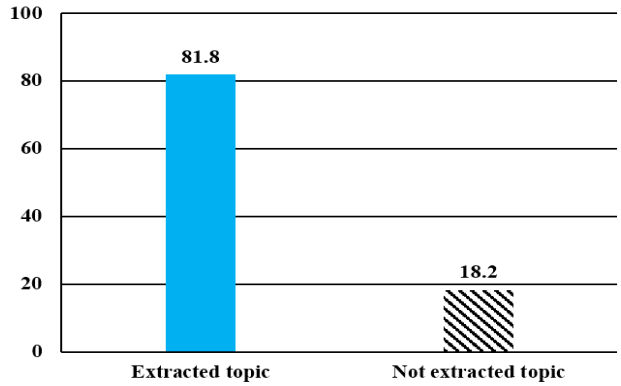

Fig.7. The Percentage of web documents whose topics have been extracted

IV. Performance Evaluation

As explained before, we employ Maximum entropy to train and test our model. The set of 1500 web documents is considered as the train set and new 3500 as the test set. In order to evaluate the given results, three users are employed to examine extracted topics of 3500 pages in a supervised learning process. Fig. 8 shows the statistical results of proposed system on 3500 web documents.

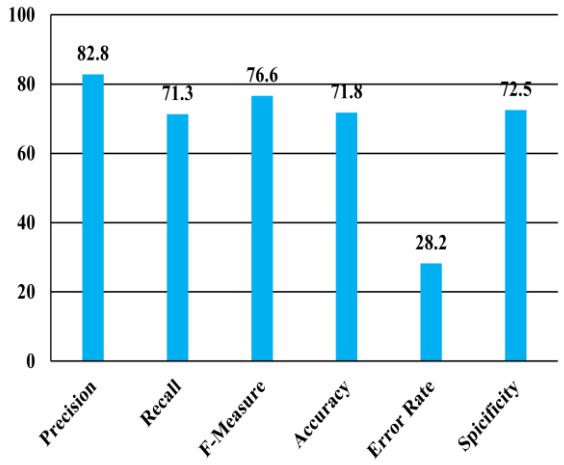

Fig.8. The statistical results of proposed system on 3500 web documents

Fig. 8 shows statistical measures like precision, recall, F-measure, sensitivity, specificity and error-rate. These measures are widely used in many classification programs. Precision can be thought of as a measure of exactness, whereas recall is a measure of completeness. The F-measure is the harmonic mean of precision and recall. It gives equal weight to precision and recall. specificity is the proportion of negative tuples that are correctly identified. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. where the error rate is the percentage of tuples misclassified by the classifier [16].

Experimental results obtained in this study demonstrate that by the precision of 82.8% and the Recall of 73.7%, all extracted topics are correct and the error rate of just 28.8% shows the strength and efficiency of the system.

As mentioned before, the huge advantage of the proposed model is its ability to reduce computation time. Unfortunately, we do not have computation time of previous works to make a comparison. However, in order to measure the speed and address the scalability of the proposed system, another experiment is performed to assess the required time to process 875, 1750, 2625 and 3500 webpages. Fig.9 shows linear regression line, which indicates the relationship between execution time and the number of webpages. Y indicates a number of hyper-links and R indicates the accuracy of Y.

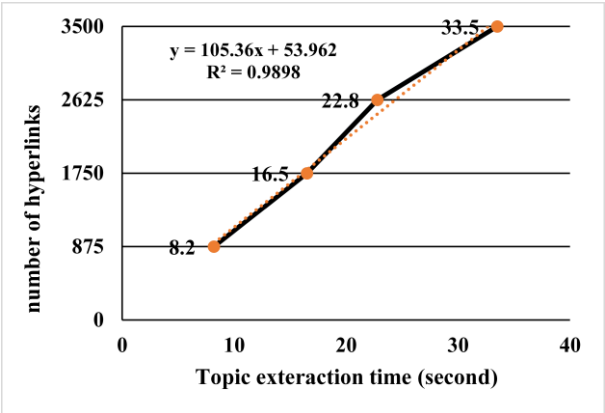

Fig.9. Scalability Diagram

As it is shown in Fig.9, the topic of all 3500 web documents are identified in 33.5 seconds, which indicates remarkable ability of proposed system. Table 4 presents hardware specification of performing system.

Table 4. Hardware Specification

Processor	Intel Core i5 2450M (2.50 GHz)
Main memory	4.00 GB
System type	64-bit Operating System

V. Conclusion and Future Work

In this paper, we proposed an efficient method for topic identification. Unlike previous works which emphasis more on analysis of the entire content of documents, in this paper, the proposed approach examines important HTML features of webpages.

First, the structures of a group of 5000 web documents are analyzed by an interactive crawler and the most important HTML features like “Keyword metadata”, “Title tag” and “Heading tag” are identified. Then selected features are extracted from web documents and stored in a feature dataset. In our first experiment, in order to select the high-performance classifier - C4.5, Decision Tree, Naïve Bayes and Maximum Entropy- are selected as four classifiers to train and test 1500 web documents. Results show that Maximum Entropy with the accuracy of (88%) yields better results compared to three other classifiers. In order to evaluate our model in a

larger scale, Maximum Entropy as our main classifier is used to test new 3500 web pages and three users are employed to examine the correctness of extracted topic; this process is performed in a supervised learning process. Through the testing results, we realized that our approach obtained high accuracy rate (71.8%), compared to previous works for topic selection.

However, for future work, our topic selection system intends to use hierarchical concepts, Taxonomy, T-BOX and A-BOX level concepts in DBpedia ontology to assist the topic recognition of web documents. In other words, particular topics like “Gulf” will be better recognized in general form by, for example, a label like “Sport”.

Also the information in features like meta tags, comments, menus and the URLs of web pages can be useful to understand more specific detail about each web documents and as a result, our system would be able to recognize the topic of more web documents with higher accuracy. Although adding more features to the system would increase the computation time, this problem will be solved with the help of more advanced hardware of performing systems.

ACKNOWLEDGMENT

The authors wish to appreciate Arash Safavi for his contribution to developing the used interactive crawler.

REFERENCES

[1] M. d. Kunder. (2015, 26 Oct). *The size of the World Wide Web (The Internet)*. Available: http://www.worldwidewebsize.com/

[2] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications,* vol. 19, pp. 171-209, 2014.

[3] L. Teixeira, G. Lopes, and R. A. Ribeiro, "Automatic extraction of document topics," in *Technological Innovation for Sustainability*, ed: Springer, 2011, pp. 101-108.

[4] A. K. McCallum. (2002). *MALLET: A Machine Learning for Language Toolkit*. Available: http://www.cs.umass.edu/~mccallum/mallet

[5] C.-Y. Lin, "Knowledge-based automatic topic identification," in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, 1995, pp. 308-310.

[6] H. Kong, M. Hwang, G. Hwang, J. Shim, and P. Kim, "Topic selection of web documents using specific domain ontology," in *MICAI 2006: Advances in Artificial Intelligence*, ed: Springer, 2006, pp. 1047-1056.

[7] U. Erra, S. Senatore, F. Minnella, and G. Caggianese, "Approximate TF–IDF based on topic extraction from massive message stream using the GPU," *Information Sciences,* vol. 292, pp. 143-161, 2015.

[8] K. K. Bun and M. Ishizuka, "Topic extraction from news archive using TF-PDF algorithm," in *null*, 2002, p. 73.

[9] R. Dong, M. Schaal, M. P. O'Mahony, and B. Smyth, "Topic extraction from online reviews for classification and recommendation," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2013, pp. 1310-1316.

[10] G. Ercan and I. Cicekli, "Using lexical chains for keyword extraction," *Information Processing & Management,* vol. 43, pp. 1705-1714, 2007.

[11] F. Liu, D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts," in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, 2009, pp. 620-628.

[12] J. M. Cigarrán, A. Peñas, J. Gonzalo, and F. Verdejo, "Automatic selection of noun phrases as document descriptors in an FCA-based information retrieval system," in *Formal concept analysis*, ed: Springer, 2005, pp. 49-63.

[13] A. T. Misirli, H. Erdogmus, N. Juristo, and O. Dieste, "Topic selection in industry experiments," in *Proceedings of the 2nd International Workshop on Conducting Empirical Studies in Industry*, 2014, pp. 25-30.

[14] Y. Ma, Y. Wang, and B. Jin, "A three-phase approach to document clustering based on topic significance degree," *Expert Systems with Applications,* vol. 41, pp. 8203-8210, 2014.

[15] (2016 Jun 19). *DMOZ - The Directory of the Web*. Available: https://www.dmoz.org/

[16] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*: Elsevier, 2011.

[17] T. N. Phyu, "Survey of classification techniques in data mining," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2009, pp. 18-20.

[18] S. L. Salzberg, "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993," *Machine Learning,* vol. 16, pp. 235-240, 1994.

[19] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *IJCAI-99 workshop on machine learning for information filtering*, 1999, pp. 61-67.

Authors’ Profiles

Kazem Taghandiki is a graduate student at the University of Isfahan in Software Engineering. He received his B.S in computer engineering from Birjand University in 2013. His area of interest includes Data mining, Information Retrieval and Semantic Web. He is currently working on his thesis in the area of Noisy Hyperlinks Removing.

Ahmad Zaeri is an assistant professor of software engineering in University of Isfahan, Iran. He received his B.S from Shahid-Beheshti University in 1998 and his Master and Ph.D. from the University of Isfahan in 2001 and 2012 respectively. His area of interest includes Semantic Web, Knowledge Mining and Software Development.

Amirreza Shirani is a graduate student at the University of Isfahan who majors in software engineering. He earned his B.S in computer engineering from Shahid-Beheshti University (National University of Iran). His area of interest includes Machine Learning, Pattern Recognition, Information Retrieval and Data Mining.

How to cite this paper: Kazem Taghandiki, Ahmad Zaeri, Amirreza Shirani,"A Supervised Approach for Automatic Web Documents Topic Extraction Using Well-Known Web Design Features", International Journal of Modern Education and Computer Science(IJMECS), Vol.8, No.11, pp.20-27, 2016.DOI: 10.5815/ijmecs.2016.11.03