Social Network Analysis and Mining (2018) 8:53
https://doi.org/10.1007/s13278-018-0532-z

ORIGINAL ARTICLE

An optimal method for URL design of webpage journals

Zahra Mousavilou[1] · **Rozita Jamili Oskouei**[2]

Received: 11 March 2017 / Revised: 8 August 2018 / Accepted: 9 August 2018

Abstract
In spite of a large number of accessible conference or online research papers conducted by various researchers, their URL structures lack content-related information. This has made the search of studies in a certain field difficult. Note that a webpage URL is not merely an address to reach the content in Internet. The importance of information in a URL structure becomes more apparent when the optimal URL structure design and the use of keywords in URL are emphasized in search engine optimization. In this study, a mode is offered to optimize the URL structure of webpage journals. The related URL becomes more readable by determining the journal titles and inserting this title in the URL related to the pages containing abstracts. Therefore, researchers can easily find the articles. The proposed model was tested on 1300 articles in 160 journals including Elsevier, Springer, IEEE, and Willie's Online Library using artificial intelligence. The results are acceptable using F-measure evaluation criteria.

Keywords Webpage classification · Scientific article title determination · Journal webpage URL · SEO

1 Introduction

To create a proper structure and classify a large number of web documents, we have to identify their titles. Identifying titles can help to better understand their content (Rathore and Roy 2014). In many cases, individuals seek a certain topic. However, reviewing and understanding all accessible documents seems impossible to find the related documents. Even if studying all documents is possible, it is time consuming. On the other hand, since the abstracts are not more than one or more paragraphs, it is difficult for individuals and sometimes machines to determine the scope of the subject using the abstracts. Therefore, recognizing the title of a document can reduce the time it takes to search the document and optimize search results (Baghdadi and Ranaivo-Malançon 2011). Title webpage recognition is an important application of title recognition. Webpage categorization is a predefined class (Baykan et al. 2011). Webpage classification has numerous applications in detecting user's behavior, offering the best option to users. At the moment, there are numerous applications for the webpage classification. Each of these methods uses various information such as text, webpage links, and webpage location characterized by URL.

Page content is generally used to classify webpages. However, using URL for classification has numerous advantages over other methods including accelerated classification and the filter of improper webpages prior to download (Kan 2005; Priyatam et al. 2013). Although there are many advantages for URL-based classification, it faces certain shortcomings. For instance, the evaluation of offered methods shows that the URL classification efficient depends heavily on the URL design type. Studies in this regard show that when URL is mentioned in the webpage title such as Open Directory Project (ODP, http://www.dmoz.org/Computers/), the designed system is precisely capable of classifying. In contrast, when the webpage title is not mentioned, not much information can be practically obtained (Bayakan et al. 2009, 2011; Kan 2005; Priyatam ey al. 2013). This is also true for search engines. It is known that a webpage is not merely an address to reach the webpage in the internet. The importance of information in a URL structure is more apparent when the optimal URL structure design and the use of keywords in URL are emphasized in Search Engine Optimization (SEO).

✉ Rozita Jamili Oskouei
rozita2010r@gmail.com

Zahra Mousavilou
z.mousaviloo@gmail.com

[1] Zanjan University of Medical Sciences, Zanjan, Iran

[2] Department of Computer and Software Engineering, Mahdishahr Islamic Azad University, Mahdishahr, Iran

Published online: 23 August 2018

The webpage address is advised to be a description of the page content or site. The webpage address, as much as possible, needs to be a summary of the webpage content. In SEO, a URL is optimal which simply and clearly introduces the page content for the search engines (SEO Best Practice, https://moz.com/learn/seo/url, Accessed: 18-Jul-2016). A URL is believed to be well designed when it is easily read by users and they can realize the content accordingly. Not only must URL be readable for search engines, but also the user needs to be aware of the content prior to downloading (A guide to the perfect SEO-friendly URL structure. http://seositeheckup.com/articles/a-guide-to-the-perfect-seo-friendly-url-structure, Accessed: 21-Jul-2016). At the moment, there is not much knowledge available about the content of journal in webpages' URL addresses. These websites provide certain information about the author and publisher in the digital object identifier (DOI) form in their website URLs. DOI is a numerical code, acting like a finger print and it is unique for each article. In other words, DOI is uniquely assigned to every article. It is considered the main link key between the online product and others (here other articles) in cyberspace. Most prestigious publishers in the field of scientific articles and books have used this unique code for their products on the internet. Since the document code is selected by the publishers, they use their own standards for every journal. The code consists of two parts: publisher and document separated by/symbol (DOI handbook. https://www.doi.org/hb.html, Accessed: 26-Jun-2016). Using DOI is essential in journal URL; however, the gap is felt in journal webpage URL. URL needs to be able to provide clear information about the webpage content for users. The purpose of this study is to meet this shortcoming. Our method can automatically convert the page title to a sequence of keywords. Using this series in the journal webpage address can help to use it for different webpage classification purposes. A hierarchy of algorithms is mapped in this method for different scientific fields. According to the instruction, the best title is selected for the webpage. Finally, the method is applied for 1300 papers. The results show that our proposed method can be very helpful for designing the modern URL for webpages.

This study is split into seven sections. The second part belongs to the keyword extraction. The third section is literature review. The fourth part is the model proposal. Fifth and sixth sections are the model implementation and discussion and future studies, respectively. Finally, the conclusion is provided.

2 Basic concepts

Some of the concepts which are related to this article is described in the below.

2.1 Keyword extraction

Keyword extraction is widely used for text retrieval. Keyword extraction is the selection of a small subsection of words in the text. These words help the understanding of the background and title. There are two general methods for keyword extraction: finding predefined words and extracting keywords. Keywords are mainly used by search engines to classify the webpage and text database (Beliga 2014; Lott 2012). In this section, we have listed some of the most-widely used methods for the keyword extraction. Our proposed method for keyword extraction is a semi-automatic method based on the human intelligence. The human intelligence precision is far greater than machine-based methods.

2.1.1 TF-IDF

Statistical methods are based on the repetition of the words in the text and known as term frequency–inverse document frequency (TF-IDF). In this technique, the weight of words equals with the number of repetition in a text (Li et al. 2011). If t is a word, d is the text, n is the number of repetition, and N is total number of words, the following formula is used to determine the term weight in d text using TF-IDF:

$$W(t,d) = \frac{tf(t,d) * \log\left(\frac{N}{n1} + 0.01\right)}{\sqrt{\Sigma(t \int d)\left[tf(t,d) * \log\left(\frac{N}{n1} + 0.01\right)\right]2}}.$$

Keyword extraction is a common method offered by many researchers to improve the algorithm. A weakness of this method is that it needs too many texts with similar subjects in learning stage in order to precisely determine the keywords.

2.1.2 Word co-occurrence relationships

Although most keyword extraction algorithms are based on the number of a term repetition, there are some methods using others than counting keywords. For example, word co-occurrence algorithm calculates the word co-occurrence. The algorithm output is a matrix in which the rows and columns show the words and the entry shows the number of co-occurrence of two words. This is a symmetric matrix. Lexical chain-based is also used to extract the keywords, using the semantic relationships among the words (Saeedeh Momtazi et al. 2010).

Word co-occurrence analysis is used in varies research areas such as text mining, content mining, ontology, etc.

Generally, the main goal of word co-occurrence is to find similarities of meaning between difference word pairs or among word patterns.

2.1.3 Key graph

It is a common keywords extraction method. In this method, the whole text is shown through a graph. Its nodes are the words in the text and edges usually display the relationships among the words, which appeared simultaneously in the text (Fukiko Kobayashi; Yumiko Nara 2014). When the nodes and their relationships are drawn, the graph is divided into sub-graphs according to the density-based clustering. Among sub-graphs, the nodes which connect two clusters are selected as keywords.

2.2 Machine learning methods

There are two types of machine learning: Supervised and Unsupervised. Supervised methods are generally used for the keyword extraction. Supervised machine learning methods extract the keywords based on the learnt model. These methods require a manually-set trainer. Decision tree, Naïve Bayes, and support vector machine (SVM) are the most common ones in this regard. Accordingly, this type of keyword extraction method requires an educational set and field definition. In this paper, we introduced a method where keywords are mapped through a graph. The greatest weight is considered "Embeddable webpage URL" using the weights of nodes which is equal to the number of repetitions added to the number of children repetition of the same node.

3 Literature review

Titles can be determined automatically or manually (Cusick 2015). The automatic method is known as various terms depending on the type of keyword determination, such as ontology, graph-based, and n-gram, etc. (Gencosman et al. 2014; Xuan et al. 2015; Rathore and Roy 2014). S. Poulimenou (Cusick 2015) proposed a keyword extraction algorithm which identifies the very important words in scientific articles. The algorithm is based on the Vector Space Model. In this method designed for showing the article titles, each word is considered a vector. Every vector has three characteristics: the number of word characters, word importance and power, and category of the word location. A score is attributed to each of above features in the title based on the weight calculated by this algorithm. Compared to the score of other related words, the score can determine the appropriateness of the word for the selection or non-selection of the word for article retrieval. Texts and articles can be extracted in a digital library with great precision using this algorithm. Azcarraga et al. (2012) proposed a neural network, known as Back Propagation Neural Network, to generalize the title and archive article contents to follow certain features, such as the location of the word in the sentence, paragraph, or input text, formats such as the subject, and other predefined features. The network function presented in this paper is such that a rule extraction method is used to extract the symbolic data extraction taken from the re-propagated network. Then the extracted rules are mapped on a decision tree. The author believes that the decision tree acts more precisely than the neural network. It also has a more understandable and easier format. Rathore and Roy (2014) offered an automatic method to detect the webpage title. In this method, webpage titles are determined ontology so that the keywords are extracted from various HTML tags and then keywords are extracted by word co-occurrence. After that, the extracted keywords are mapped on ontology. Finally, the webpage title is extracted. Baghdadi Ranaivo-Malançon (2011) used the divide and conquer method so that the whole text is divided into sentences. Then the topic of each sentence is determined. Finally, the weights of titles are calculated. Titles with greater weights are selected as text titles.

Beliga (2014) offered a detailed description of keywords extraction from different perspectives. In this study, the authors divided keyword highlighting techniques into two categories: detection and extraction. In terms of keyword detection, a dictionary with hierarchically sorted words is used. Here, the text is mapped into this dictionary and the class is identified. On the other hand, there exits the keyword extraction, directly extracted from the text. In this study, keyword extraction is examined from the supervised and unsupervised perspective. Finally, a graph-based method, which is an unsupervised method, is offered. The core of the model, proposed by Haggag (2013), is the semantic similarity and relationships among words so that the semantic relationships examine various aspects of relationships among words such as day and night or light and heavy. Using these relationships helps discover such relationships. However, semantic similarity only considers the semantic similarity dimension, such as running and walking or cat and dog. Here, the author made an initial list of the words with certain keywords using common statistical methods. Then all possible pairs are listed. Here, the power between words is evaluated using word–word semantic similarity. The text words are mapped based on the meaning similarity with a weight criterion. Finally, the strongly linked words will have the greatest ranking in semantic similarity for the keyword selection. Link algorithm is used to examine the similarity of paired words. The results are normalized and keywords are extracted using the semantic relationships among paired words.

4 Proposed method to determine scientific paper title

The purpose of the study is to determine the title and to use it in the webpage URL. For this purpose, a hierarchy of algorithms and keywords related to each topic is collected. A hierarchy of algorithms is "the name of algorithms along with the subsections". For example, Kernel algorithms and neural networks are subclasses. The structure is defined in a hierarchy. In this study, the hierarchy of algorithms and abbreviations of AI algorithms is used. This hierarchy is implemented in a tree structure. Each node of this tree has five fields.

- In The Keyword field, the name of algorithm and the synonyms and abbreviations are recorded.
- A phrase which can be a better synonym of the algorithm (Scientific branch) is placed in ID field.
- Parent Node field shows the node's father's ID.
- Width field shows the number of repetitions in the Keyword field. The tree is manually created once for every scientific journal. Thereafter, it is a mapped level-by-level from the root to the leaves and the number of repetition and synonyms are calculated. That we claim from the root to the leaf means that if we assume that all of the existing branches are implemented for a single journal, over 1000 titles might have been defined. Level-by-level examination from the root to the leaf makes the withdrawal of unrelated words possible, thereby making the search optimized. Finally, the branch with the maximum Total value is identified and their nodes IDs are used as URL code.

In the proposed algorithm in this study, the words and key phrases of each node of the tree are searched and the number of word repetition of each node is calculated using n-gram method. Here, n shows the number of alphabets of the word. For example, $n=5$ for "*Clas*s". The words of each node can be a phrase, such as a few words in succession with a line of distance or an abbreviation of the same statement. Following this stage, the minimum number of repetition of all related words is taken into account to calculate the number of repetition. For example, for "Support Vector Machine", if the *Machine* repetition is 17, Vector is 15, and Support is 12, the number of repetition is considered 12. Synonym words and phrases are separated using comma. After calculating the number of repetition, the numbers are summed. In the following stage, after calculating the number of algorithm synonym repetitions and inserting in node repetition field, father's repetition number is added to the number of children's repetition, and then placed in Total Width field. Finally, the branch with the greatest total width weight from the root to the leaf is considered the main branch and the highlighted words of this branch are retrieved in a series. Inserting this series in the URL can help users find the algorithms used in the article.

Table 1 shows a sample of the designed data structure to find the title of the article (Murty and Raghava 2016) using the aforementioned algorithms. The table lists the algorithm ID, keywords, algorithm synonyms, parent node, algorithm's father's keyword, node weight, and total node weight which is equal to the total of node's and children's weights.

When the table output is added to the Springer URL, URL is as follows: (for better understanding, the series is shown in red).

The URL that can represent for the source (Murty and Raghava 2016) is as follows:

http://link.springer.com/chapter/10.1007/978-3-319-41063-0_5/Artificial-Intelligence/Data-Mining/Classification/Kernel/SVM.

Above URL shows that the page is related to an AI article on the subsection of data mining, which used Kernel and SVM algorithm for the classification. As it can be seen, this is a far better URL compared to the existing one which is as below:

https://link.springer.com/chapter/10.1007/978-3-319-41063-0_5.

Table 1 Sample of solution

ID	Parent node	The keywords	Width	Total width
Data mining	Artificial intelligence	Data mining, DM	1	133
Classification	Data mining	Class	49	132
Kernel	Classification	Kernel	43	83
SVM	Kernel	Support vector Machine, SVM	40	40
Neural network	Classification	Neural network	0	0
Back propagation	Neural network	Back propagation	0	0
Preprocess	Data mining	Process	0	0

Table 2 Name of journals which are Used for evaluating proposed method

	Journal name	Number of tested papers	Number of correct classified papers
1	Elsevier Journals	275	267
2	Springer Journals	270	251
3	The ACM Digital Library	210	192
4	IEEE Xplore Digital Library	190	181
5	Journal of Machine Learning Research	92	82
6	Morgan and Claypool Publishers	70	59
7	Wiley-online library	55	48
8	Other	150	132

5 Proposed method evaluation

After implementing the proposed idea in Sect. 4, we must evaluate the performance of the proposed method by online and offline examination. For this purpose, the proposed method is tested on 160 journals which are declared in Table 2 and ranked as valid journals on http://www.scimagojr.com on AI.

As we can see in Table 1, for evaluating the accuracy of the proposed method, we tested almost 1300 AI articles using sensitivity and specificity factors. First of all, the method was tested for detecting and classifying AI articles. After classification, sensitivity and specificity factors were calculated using Eq. 1 and Eq. 2. Their averages were taken into account at this stage. After that, the same procedure was carried out on the classes to separate the subclasses. The values are then recorded. The same procedure is carried out for five stages. The results were controlled by 100 computer students. The values are recorded in Tables 3, 4. The values can be seen in Fig. 1.

Table 3 Sensitivity opposed to specificity

Computer	Sensitivity	Specificity
1	99.5	85
2	99.7	87.2
3	99.5	88
4	98.6	97.2
5	99.1	98.3

Table 4 Evaluation results of the proposed method

	The evaluated items	How to evaluate	The result based on the standard F-measure
1	The full-text articles	Offline	94%
2	The abstract of the articles	Online	73%

The results were compared with manual results, done by 100 computer students. Tables 2 and 3 and Fig. 2 show the results.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) = (\text{Number of true positive assessment})/(\text{Number of all positive assessment}), \quad (1)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) = (\text{Number of true negative assessment})/(\text{Number of all negative assessment}). \quad (2)$$

As the results show, the method is capable of detecting the title using the proposed algorithm.

The analysis shows that the greater the specificity is which we get close to the sub-branches, the higher the sensitivity is seen. Therefore, the accuracy rises; however, specificity does not change much. Another test was run on 1300 complete offline articles (downloaded) and the system was examined on the abstracts of the same online article. In the online method, titles, abstracts, and keywords were used to detect the page title.

The purpose is to offer a systematic work which helps website designers. It is therefore claimed that the proposed method is capable of detecting the page title using the aforementioned algorithm. Experimental studies show that the method can optimally detect the webpage title from the abstract on the user's side. However, since the abstract is limited, the series taken from the abstract is shorter compared to the whole text. On the other hand, the difference between the two methods indicates that the abstract page is not alone capable of describing the whole text. When the whole text is accessible, the method can better determine the title. Therefore, if journals

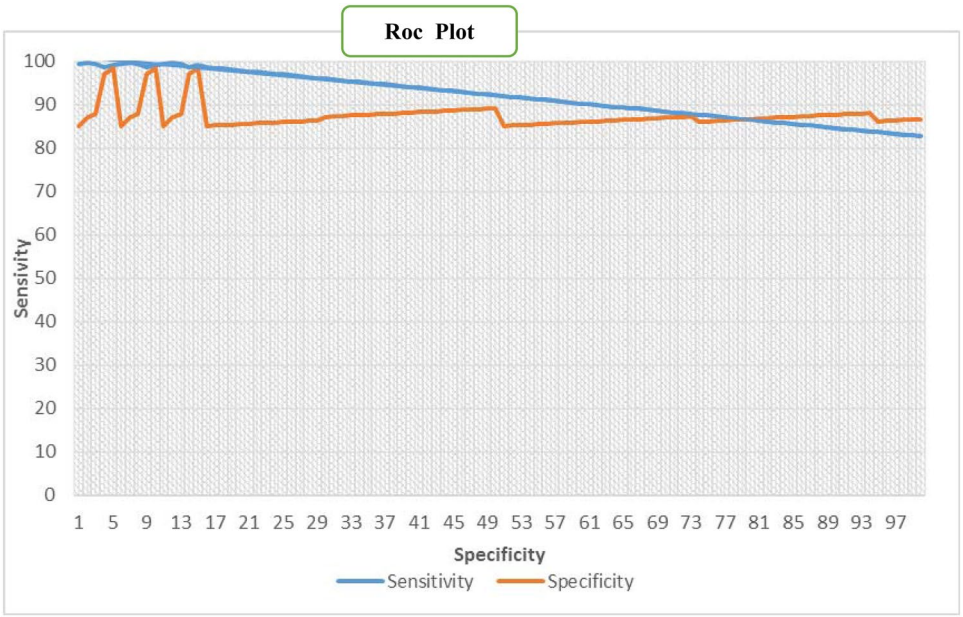

Fig. 1 Diagram of Sensitivity Opposed to Specificity

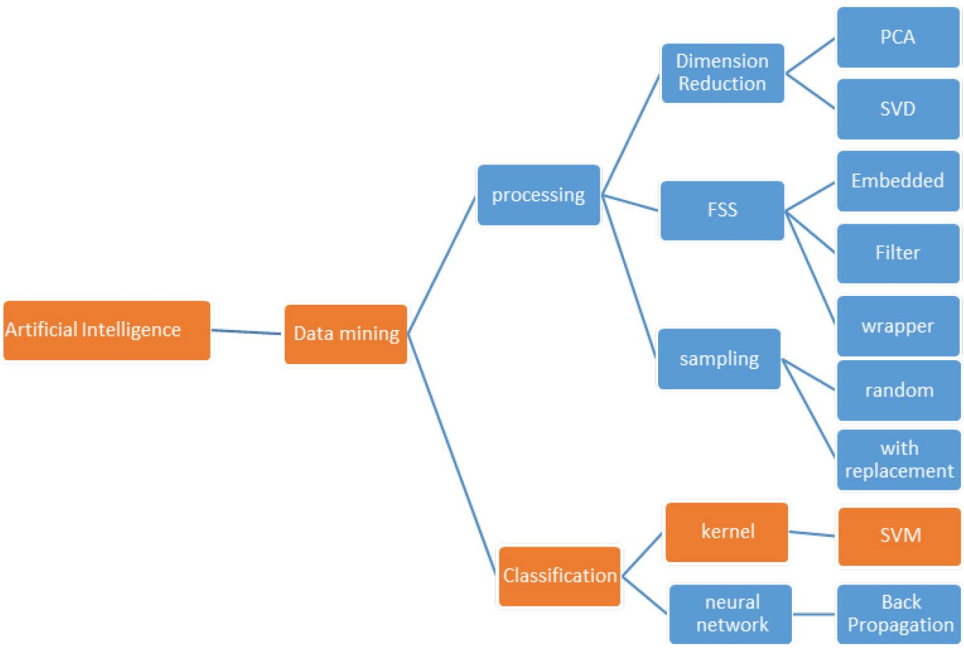

Fig. 2 A part of designed hierarchy for system

mention the article subject in URL, the researcher is able to decide whether or not the article is related to the field of study.

6 Discussion

Webpage classification is an important subject related to webpages. However, the content-based classification is common. Certain advantages, such as high speed make the URL classification more effective. This method can display better performance in terms of classification when sufficient information exists in URL. Due to the SEO importance, many websites consider the page title in address. However, scientific journals have not taken it into account. This makes the search more difficult for researchers. If the title is available in address, finding related articles becomes far easier.

In this study, we proposed a method to determine the title based on the proposed algorithms. The output of the article is a series of algorithm hierarchy. The system was evaluated

by online (examination of abstract) and offline (examination of the whole text) methods. The results show that the review of the whole text gives similar series of algorithms, such as SVM, classification, or data mining. However, the review of abstracts sometimes leads to dissatisfactory results. Therefore, online evaluation cannot be as good as offline evaluation. It is, therefore, concluded that if the page title (Hierarchy algorithms used in this article) is extracted from the whole text by the researcher, it can give better information to search engines and users. Finally, this series is used in the relevant URL. As a result, the search is optimized for the users and search engines to a great extent. This URL design of journal webpages can be compared by inserting DOI in their URLs.

Our method is very useful and the URL content is more readable. This method was tested on AI articles and can be generalized on other scientific articles. Although this method can detect the algorithms used in an article, if the article contains more than one algorithm, then the more repeated one is selected as the main algorithm. This makes sure that the article covers only one algorithm. This leads to an inaccurate inference from the text. For example, the article (20) was detected as AI and the URL is as follows:

http://ieeexplore.ieee.org/document/6252618/.

This is a wrong inference. Article (Rajola 2013) was detected as neural network article and the URL is as follows:

https://link.springer.com/book/10.1007/978-3-642-35554-7.

This means that the article is related to AI and data mining and talks about the classification and neural network using Back Propagation. However, we know that the article is about different data-mining methods and neural network is one of them. Since it was repeated more, it was selected as the core of the article. Our proposed method can be useful in this matter and help users for selecting more related papers to their favorite topics.

7 Conclusion

In this study, we proposed a method for more readable URL designing, related to journal webpage content. This method can specify the scientific article's title using the proposed method and apply it in abstract page URL. Therefore, researchers can easily find the articles related to their interested subject of study. The results of our study on 1,300 selected AI journals show that the algorithm is capable of correctly detecting the title and inserting in abstract page URL using F-measure criterion with precision of 94% and average sensitivity rate equal to 99.2% and specificity factors rate approximately 90.3%. Therefore, the final goal of this paper, to detect the article content by URL, is met. The future challenge can be related to finding URL of pages based on the requirement and knowledge level of users. For example, when the user is expert may be requiring very special and advanced journal papers, but when he/she is a beginner, it will be very applicable if the URL of tutorial or survey paper can be found by search engine. Therefore, labeling URL of each paper based on the level of paper published in one journal can be very useful. Our next research will attempt to cover this topic.

References

Azcarraga A, Liu MD, Setiono R (2012) Keyword extraction using backpropagation neural networks and rule extraction. In: The 2012 international joint conference on neural networks (IJCNN), pp 1–7

Baghdadi HS, Ranaivo-Malançon B (2011) An automatic topic identification algorithm. J Comput Sci 7(9):(1363–1367)

Baykan E, Henzinger M, Marian L, Weber I (2009) Purely URL-based topic classification categories and subject descriptors. In: Proceedings of the 18th international conference on World wide web. (1109–1110)

Baykan E, Henzinger M, Marian L, Weber I (2011) A comprehensive study of features and algorithms for URL-based topic classification. ACM Trans Web 5(3):15

Beliga S (2014) Keyword extraction: a review of methods and approaches. University of Rijeka, Department of Informatics, Rijeka

Cusick MB (2015) Human generated topics: a gold standard for automated topic evaluation (Order No. 10110614). Available from ProQuest Dissertations and Theses A&I; ProQuest Dissertations & Theses Global. (1798418928). https://search.proquest.com/docview/1798418928?accountid=41306. Accessed Dec 2016

Fukiko Kobayashi; Yumiko Nara (2014) A narrative analysis by text mining technique using key graph. In: 2014 IEEE international conference on data mining workshop (ICDMW)

Gencosman BC, Ozmutlu HC, Ozmutlu S (2014) Character n-gram application for automatic new topic identification. Inf Process Manage 50(6):821–856

Haggag MH (2013) Keyword extraction using semantic analysis. Int J Comput Appl. 61(1), 1–6

Kan M (2005) Fast webpage classification using URL features. In: Proceedings of the 14th ACM international conference on Information and knowledge management, pp 325–326

Li JR, Mao YF, Yang K (2011) Improvement and application of TF * IDF Algorithm. In: Liu B, Chai C (eds) Information computing and applications. ICICA 2011, vol 7030. Springer, Berlin **(lecture notes in computer science)**

Lott B (2012) Survey of keyword extraction techniques. https://pdfs.semanticscholar.org/f9f6/8c217aef0f3f873eb602a03748ceb5806c88.pdf. Accessed Feb 2017

Momtazi S, Khudanpur S, Klakow D (2010) A comparative study of word co-occurrence for term clustering in language model-based sentence retrieval. In: Human language technologies: the 2010 annual conference of the North American Chapter of the ACL, pp 325–328

Murty MN, Raghava R (2016) Kernel-based SVM. In: Support vector machines and perceptrons. Springerbriefs in computer science. Springer, Cham

Poulimenou S et al (2014) Keywords extraction from articles' title for ontological purposes. In: Proceedings of the 2014 international conference on pure mathematics, applied mathematics, computational methods (PMAMCM 2014), pp 120–125

Priyatam N, Iyengar S, Perumal K, Varma V (2013) Don't use a lot when little will do: genre identification using URLs. In: Proceedings of the CICLing 2013

Rajola F (2013) Data mining techniques. In: Customer relationship management in the financial industry. Management for professionals. Springer, Berlin

Rathore AS, Roy. D (2014) Ontology based web page topic identification. Int J Comput Appl 85(6):35–40

Xuan JY, Jie L, Zhang GQ, Luo XF (2015) Topic model for graph mining. IEEE Trans Cybernet 45(12):2792–2803