Indian Journal of Science and Technology, Vol 9(33), DOI: 10.17485/ijst/2016/v9i33/97032, September 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Characteristic Selection with Rough Sets for Web Page Ranking

G. Anuradha and N. Deepak Kumar

Department of CSE, GMRIT, Rajam - 532127, Andhra Pradesh, India; anuradhagovada@gmail.com, Deepakkumar210591@gmail.com

Abstract

Objective: The objective is to classify web pages and assign ranking to web pages using feature selection with rough sets and TF_IDF methodology. **Proposed Method:** Web page ranking is a process to assign position at a particular site appears in the result of web page. A site is said to have a high page ranking when it appear at or near the top of the list of web result. A challenge in web page ranking is to provide relevant information to the user according to query. To finding relevant information from the result set is a tedious process. To obtain a refined result set that contains the URL's more relevant to the user's query, so it is essential to rank. For classification purpose, we are using feature reduction method based Rough Set Theory (RST). **Application:** Feature selection is most essential technique in rough sets as well as the data mining. Attribute selection is a main challenge for expanding the theory and making use of rough set. **Findings**: The proposed method emphases on the removal of the unnecessary attributes as a way to sort the effective reduct set and framing the core of the attribute set. After successful classification procedure, we have to applying TF_IDF methodology for assign the ranking to the documents.

Keywords: Core, Data Preprocessing, Data Mining, Feature Selection, Rough Sets Theory (RST), Reduct, Tf-IDF, Text Mining

1. Introduction

In present days huge number of web pages is uploading in to the www, it's very complicated to searching and getting the absolute result from the data center. Classification is major part in www. In search engines, text classification is an important process is helping in organizing the huge amount of data.

Case in point, most Internet web indexes, for example, Yahoo and ask, separate the indexed web documents[1] into a various categories[2-3] for users. The content based and Latent Semantic Analysis[4] will aggravate effect for easy browsing. Human classification is not possibly to keep rate of development of the web. To overcome this problem automatic web page classification and Machine learning is used. Automatic classification is much inexpensive and quicker than human cataloging. To classify the webpages pre-labeled[5] groups are used. The problems to be overcome by first build the webpage pre cataloguing and generate Rough Set Theory (RST). Once categorizing another page, these different sets of instructions should be used together as a part of some approach to direct the classification or classifications of the new page. This whole procedure is appeared in Figure 1.

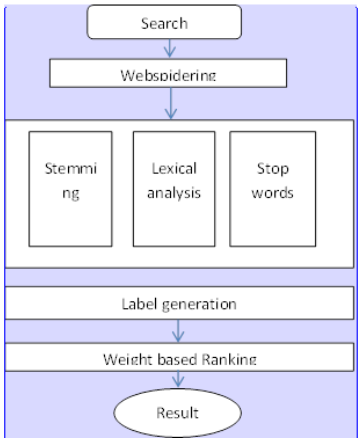

Figure 1. Process flowchart.

*Author for correspondence

1.1 Roughset Theory

In this section, all the important theories related to Rough sets theory have been well-defined. The beginning stage of rough set theory is to divide the universe based on the existing knowledge about the given questions at present, then to determine the support degree to some concepts of each divided composition: positive support, positive non-support or possible support, and three approximate sets are used to indicate positive region, negative region and boundary individually[6].

Give us a chance to say that the universe of talk is *U* furthermore accept that *R* is equivalence relation based on *U*. So the approximation space which is a pair <U,R>, where U={x1,x2,…xi,….., xn}, where x is an element such that x∈U.

T = (U, A, C, D)

U = non empty universal set

A = Attribute non empty finite set

C = condition

D = decision

a ∈A

$f_{a:}$ a:U→V_a

V_a value attribute

Lower and upper approximations can be characterized as below:

Lower Approximation of a Subset: $L_R(x) = \{x \in U: [x]_R \subseteq x\}$

Upper Approximation of a Subset: $U_R(x) = \{ x \in U :[x]_R \cap x \neq$

Boundary Region of a Subset:

$$BN_B(X) == \overline{B(X)} - \underline{B(X)}$$

Boundary region is a non-empty set that is $\overline{B(X)} \neq \underline{B(X)}$ $\overline{B(X)} \neq \underline{B(X)}$ then the set is called a Rough Set with respect to the B.

A set is called Crisp set when its boundary region is empty that is $\overline{B(X)} = \underline{B(X)} \overline{B(X)} = \underline{B(X)}$

1.2 Reduct and Core

In rough set theory reduct and core are two most critical concepts. Reduct used in attribute collection process to reduce unnecessary characteristics and summarize the attributes which holds the precision of the original set for making applications.

- Reduct of a decision table is a set of state attributes that is enough to define the decision attribute.
- Any reduct empowers us to reduce condition attribute.

It reduces calculation cost for rule generation.

Rough Set Theory (RST) is a mathematical tool for demonstrating inadequate or loose data. In this work, top frequent words are figured from the training data set web pages. To remove unnecessary features from training data set we applied feature selection[7] method in RST[8] and then apply the tf * idf concept[9-12] for finding the ranking to documents. The Proposed work is concentrate on eliminating noise from Web page and applies summarization to Web-page classification[13-15].

2. Architecture

2.1 Implementation of Proposed Approach

We arbitrarily selected over 1450 web pages from the Yahoo category. Then each category collected nearly 100 example web pages[16-17] for pre-labeled classification purpose. In this paper we have categorized 10 pre-labeled classes. We just downloaded around thousand depictions of Web pages that are physically made by human editors. Since it is a period devouring assignment to run investigates this expansive information set, we arbitrarily evaluated these pages with depictions for our analysis reason. The extricated subset incorporates 1000 pages, which are disseminated among 10 classifications (we just consider the main two level classes)[18]. Example pre-labeled classes are education, sports, entertainment...etc shown in Table 1.

Table 1. Distribution of the training data.

Category	Number of web pages
Business	120
Computers	90
Education	220
Entertainment	130
Government	85
Health	251
News	141
Sports	213
Science	85

Economics	115
Total	1450

3.1 Information Table

Table 2 comprises of 11 uid summaries. uid summaries consists many attributes. In this table, rows and columns are attributes and objects. Here {a, b, c, d} are the attributes.

Table 2. A collection of url id summaries

Objects	**Attributes, condition**			
	a	b	c	d
U				
uid1	Education	games	school	science
uid2	School	education	science	college
uid3	Education	university	research	health
uid4	Education	school	depart - ment	development
uid5	Education	college	develo - pment	university
uid6	Education	school	news	science
uid7	develo - pment	education	research	news
uid8	School	college	research	university
uid9	School	college	news	university
uid10	School	media	student	teach
uid11	School	education	science	college

In this table, rows are Objects and columns are Attributes. Every row has web page summary.

This summary is collected from various urls (uid). Each column contains attributes. Here we have utilized just a couple bits of attributes[19].

3.2 Decision Table

Here we are considering each attribute contained a specific number. We are pre-labeled some attributes in education sites data. Based on the pre-labeled class numbers we are working out conditional attribute. This example some frequently terms related to education it should labeled as {education-1, school-2, university-3, college-4, research-5, science-6, department-7} other than these label's we have taken {games, health, development, news, media} these are the unrelated terms to the education site .so we are eliminating the attributes from the core and reduct concept. Table 3 demonstrates the Decision table, same or ambiguous articles might be symbolized ordinarily and a portion of the qualities might be pointless (repetitive). That is, their evacuation can't influence the characterization.

Table 3. Decision table

	a	**b**	**c**	**d**	**D**
U					
uid1	1	0	2	6	N
uid2	2	1	6	4	Y
uid3	1	3	5	0	N
uid4	1	2	7	0	N
uid5	1	4	0	3	N
uid6	1	2	0	6	N
uid7	2	4	5	3	Y
uid8	2	4	5	3	Y
uid9	1	3	5	0	N
uid10	2	0	0	0	N
uid11	2	1	6	4	Y

3.3 Decision Algorithem

If (education||school||university||college||research||science||department) then (decision=yes)
Else
(Decision=no)

Indiscernibility relation attributes are essential in set approximation. Minimal subsets of attributes are classified as "reduct", it is negligible subset of all attributes that empowers the same gathering of components of the universe as the entire arrangement of attributes that don't have a place with a reduct are repetitive with respect to classification of elements of the universe.

In Table 4 uid3 and uid9, uid7 and uid8, uid2 and uid11 are similar attributes so we apply core-reduct method to reduce the complexity. Hear we only considering the unique attributes row's only so eliminate the uid3, uid7, uid11.after elimination reduced table is called final kkd table.

Table 4. After core and reduct

	a	b	c	d	D
U					
uid1	1	0	2	6	N
uid2	2	1	6	4	Y
uid4	1	2	7	0	N
uid5	1	4	0	3	N
uid6	1	2	0	6	N
uid8	2	4	5	3	Y
uid9	1	3	5	0	N
uid10	2	1	6	4	Y

4. Ranking Methodology

We apply ranking methodology after successful completion of the classification using Rough Set Theory (RST). Hear we using the TF_IDF (term frequency and invers document frequency)[20] for ranking the documents in successive ordering. This weight is a numerical measure used to assess how many time a word appeared in a document. Tf_idf can be effectively used for filtering stopwords and discovering root words is appeared in Figure 2 in different subject fields including text summarization and classification.

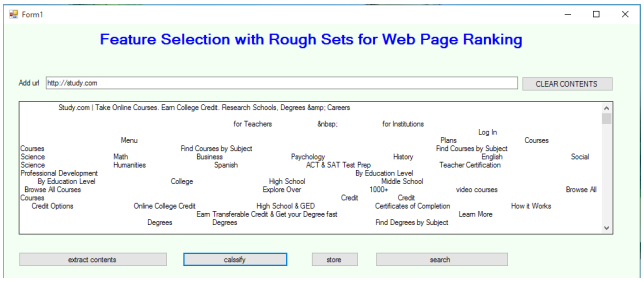

Figure 2. Extracting the root words from url.

4.1 TF (Term Frequency)

Term recurrence in the given report is basically the quantity of number of times a given term come into view in that record. TF used to gauge the significance of thing in a record, the quantity of events of every word in the archive. Each report is characterized as a vector comprising of words, for example,

D={word1,word2,word3… … .word n}

Where D implies the Document and word appears on that report and n speaks to the quantity of words in the record.

$$F\ (q,d) = \sum_{t \in q} f_{t,d}$$

Hear document d, query q, Calculate W(q,d)

$$TF(q,d) = \sum_{t \in q} tf_{t,d}$$

Term frequency is shown in Figure 3 for education is18 in study.com. According to TF, ranking is assigned to particular URL is shown in Figure 4.

4.2 IDF (Inverse Document Frequency)

The Inverse Document Frequency calculated by following equations:

$$idf_t = \log_{10} \frac{N}{df_t}$$

$$W_{i,j} = TF * IDF$$

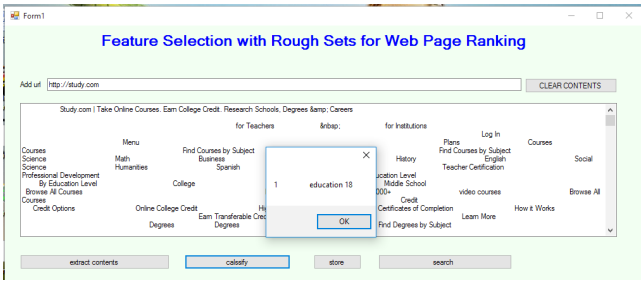

Figure 3. Classifying the url.

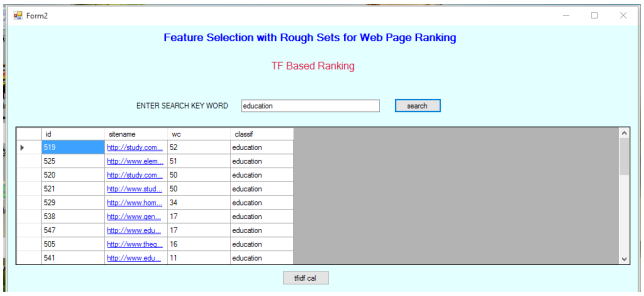

Figure 4. Ranking based on tf.

At the end ranking is assigned based on TF*IDF is shown in Figure 5.

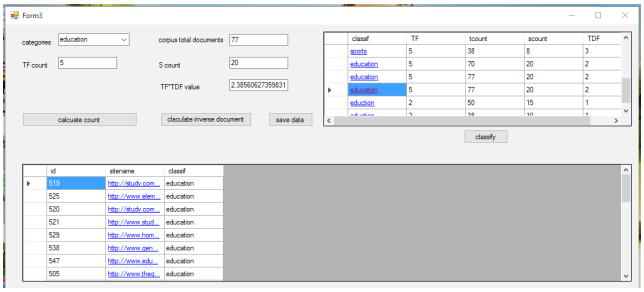

Figure 5. Ranking the document based on TF*IDF.

5. Conclusion

In this study, Rough sets are used for dimension reduction technique applied for effective result[21-24]. It is fit of decreasing the repetition among the attributes. RST strategy registers the best component for minimized redundancy in contrast to Information Gain (IG).This proposed method is approached for feature selection for text classification.Pre-labeled indexing will helped for effective classification[25] and high dimensionality reduction. Stemming and lexical analysis used for better web text extraction purpose. After successful classification applying TF*IDF method[26] for ranking the web documents.

6. References

1. Lewis DD, Ringuette M. Comparison of two learning algorithms for text categorization. Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94); Las Vegas, NV. 1994 Apr 11-13. p. 81-93.
2. Debole S, Debole F, Sebastiani F. Supervised term weighting for automated text categorization. Proceedings of SAC-03, 18th ACM Symposium on Applied Computing; Melbourne, US. 2003. p. 784-8.
3. Yang L, Yang Y, Liu X. A re-examination of text categorization methods. SIGIR-99; 1999. p. 42-9.
4. Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990; 41(6):391-407.
5. Wiener E, Pedersen JO, Weigend AS. A neural network approach to topic spotting. Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95); Las Vegas, US. 1995. p. 317-32.
6. Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer; 1991.
7. Mladenic D. Feature subset selection in text learning. Proceedings of the10th European Conference on Machine Learning (ECML98); 1998.
8. Chouchoulas A, Shen Q. Rough set-aided keyword reduction for text categorization. Applied Artificial Intelligence. 2001; 15(9):843-73.
9. Tf-idf: A Single-Page Tutorial - Information Retrieval and Text Mining. Available from: http://www.tfidf.com.
10. Mohod SW, Dhote CA. Feature Selection technique for text document classification: An alternative approach. IJRITCC. 2014; 2(9):2914-7.
11. Salton G, Buckley C. Term-weighing approach sin automatic text retrieval. In Information Processing and Management. 1988; 24(5):513-23.
12. Nir O. Reexamining tf.idf based information retrieval with genetic programming. Proceedings of SAICSIT South African Institute for Computer Scientists and Information Technologists; Republic of South Africa. 2002. p. 1-10.
13. Aizawa A. An information-theoretic perspective of tf–idf measures. 2003; 39(1):45-65.
14. Wang Y, Wang XJ. A new approach to feature selection in text classification. Proceedings of 4th International Conference on Machine Learning and Cybernetics; Guangzhou, China. 2005 Aug 18-21. p. 3814-9.

15. Lee LW, Chen SM. New methods for text categorization based on a new feature selection method and new similarity measure between documents. IEA/AEI; France. 2006. p. 1280-9.
16. Gunasundari R, Karthikeyan S. A study of content extraction from web pages based on links. 2012 May; 2(3):23-30.
17. Kumar V, Singhal N, Dixit A, Sharma AK. A novel architecture of perception oriented web search engine based on decision theory. Indian Journal of Science and Technology. 2015 Apr; 8(7):635-41.
18. Porter MF. An algorithm for stripping. Program. 1980; 14(3):130-7.
19. Scott S, Matwin S. Text classification using word net hypernyms. Proceedings of the Conference on the Use of WordNet in Natural Language Processing Systems; 1998. p. 45-51.
20. Jing LP, Huang HK, Shi HB. Improved feature selection approach TFIDF in text mining. 2002; 2:944-6.
21. An1 A, Huang Y, Huang X, Cercone N. Feature Selection with Rough Set for Web Page. 2005. p.1-15.
22. Buyukkokten O, Garcia-Molina H, Paepcke A. Seeing the whole in parts: Text summarization for Web browsing on hand held devices. Proceedings of WWW10; Hong Kong, China. 2001 May. p. 652-62.
23. Chakrabarti S, Dom B, Indyk P. Enhanced hypertext categorization using hyperlinks. Proceedings of the ACMSIGMOD; 1998. p. 307-18.
24. Chen H, Dumais ST. Bringing order to the Web: Automatically categorizing search results. Proceedings of CHI2000, School of Information Management and Systems, University of California; Berkeley, CA. 2000. p. 145-52.
25. Huang Y. Web-based classification using machine learning approaches [Master's thesis]. Regina: Department of Computer Science, University of Regina; 2002.
26. Agarwal B, Mittal N. Sentiment classification using rough set based hybrid feature selection. 2013. p.115-9.