ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.7

An Ideology for Text Scooping Method

Govindraj Chittapur[1], Chetan Konnur[2]
Assistant Professor[1], Student[2]
Department of MCA
Basaveshwar Engineering College Bagalkot, India

Abstract:
Identifying the similarity between articles is the complex task, here some of the systems like Stanford CS and Crossref are used to compare similarity status between the articles and the used has to pay for it. The system has to perform very quick similarity comparison between the articles using minimum amount of time and the user can able to submit many articles with dif-ferent names at a time and user can get the similarity status of submitted articles. And the system should generate unique group id for every cluster of the articles. This system should be used to identify the similari-ty of the articles which are submitted by the users in two ways; first one is to submit the article by pasting the descrip-tion directly with different article name and second one is by pasting the URL of the article. After the successful submis-sion of articles those will be cleaned by removing the stop words then system should provide unique token id for the each word exist in the cleaned article and then tokenized clean article will be used in the frequency computation, log likelihood computation, RV coefficient computation and clustering process and these each modules will be executed one by one sequentially and finally the processed articles will be grouped with the similarity status and unique group id.

I. INTRODUCTION

System Architecture for Proposed System

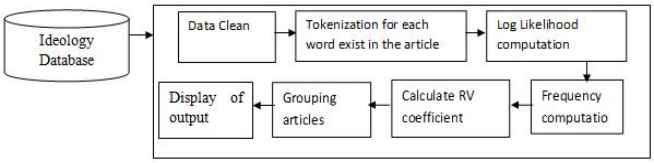

Figure.1. System Architecture for "An Ideology for Text Scooping Method"

In this proposed system the articles will be submit-ted in two ways as shown in the above architecture diagram, in the initial stage the articles will be submitted in the format of offline submission means here the user will submit the article by pasting the article description in the given text are and the user can also submit the article by pasting the URL of the article, in this the article description will be extracted by the URL by converting the URL in to DOM tree. As shown in the above architecture diagram the articles will be stored in the format of online submission or offline submis-sion.

Here in the above diagram the stop words database will be used to store the preserved words which are used in the data cleaning process and these stored stop words will be removed from the article description. And in the data store section the other data like login information, Likelihood computation, clustered ar-ticles, cleaned articles, RV coefficient results and other re-lated data will be stored in this block. In the data cleaning section the data will be cleaned using the stop words which are stored in the stop words da-tabase, and the articles which are submitted by the users will be cleaned in the process, here all the stop words are erased and the submitted articles will be cleaned and doesn't have any meaning for the description. In the tokenization section the cleaned articles will be used to tokenize the words, here the cleaned article de-scriptions which contain core words will be used to tokeni-zation, and here each and every word will be tokenized and unique token id will be given for the each and every word of the article. Here in log likelihood block will be used to calcu-late the similarity between two articles by using the below equations

In this we need to compute
Expected freq= freq of word * freqof others/totalfreq
Log Likelihood (principle)= 2* summation(i-1 to N) gi* log(gi/egi)
G1*log (g1/eg1) +g2*log (g2/eg)+……+gnlog (gn/eg)
G1 is freq of first word in the article
G2 is freq of second word in the article Gn is freq of nth word in the article In the frequency computation module the submitted articles will be used for calculating the frequency of the re-peated words in the comparing articles. RV coefficient is the module which is used to cal- culate the coefficient of the articles and here the words exist in the cleaned escription are tend to intersection of words means here the same words exist in the comparing articles will be calculated using
Summation (x-x*) (y-y*/x**y**)
This is used to get the intersection or common words be-tween two articles. If RV > 0.02 then compared articles are similar or else they are not similar In the final module themes the compared articles will be grouped with the unique group id and shows the sta-tus on which are same and which articles are not same among the given threshold value.

II. LITERATURE SURVEY

Identifying similarity between the articles is a im-portant research topic in research and development project management here some of the authors mentioned with their algorithms, D. Saravana Priya and Dr. M Karthikeyan [1] has proposed Clustering analysis, Fuzzy SOM, knowledge based agent, NGRA

Algorithm, Ontology, R&D and Text mining, Kuwar Aditya et al. [2] Has proposed Self Organiz-ing Method (SOM), Pravin Shinde et al. [3] has proposed Clustering, Classification, Self Organizing Map, Text Min-ing, Optimization, Latent Semantic Indexing, Mr Bhushan Medage et al. [4] proposed Data mining ,Text mining, ab-stract selection, Project Evaluation, Guide Allocation, John Butler et al. [5] proposed Simulation; Ranking and Selec-tion; Multiple Attribute Utility Theory.

III. DESIGN ISSUES FOR "AN IDEOLOGY FOR TEXT SCOOPING METHOD"

Here the articles should be submitted in two ways and each way consist its own interface, the user can submit articles using URL of the article or by directly pasting the article description. And here the URL of the article should be of html type then only the DOM tree procedure will work on div tags and the description should be limited in size.

IV. IMPLEMENTATION OF "AN IDEOLO-GY FOR TEXT SCOOPING METHOD"

An Ideology for Text Scooping Method is a pro-posed solution where the similar types of articles to be find out from the several set of articles, in the earlier solution the manual process is used to find the duplicate and similar types of articles and the manual process is very time con-suming and the result is not accurate and it uses lot of man power with other resources, this process does not consist any of the algorithms and the articles should be analyzed to process the complete task. And now in the proposed solution the several algorithms are used to differentiate the article, here text scooping means the text mining and this is done by the text mining algorithm, this proposed solution consists several stages in it like submit articles, data cleaning, tokeni-zation, frequency computation, log likely hood, RV coeffi-cient, clustering these are the sequential stages to perform the appropriate solution to find the similarity in the article. First thing is there will be a registration and login, perform registration as admin or user, user can submit article offline or online, the admin can view all articles perform data cleaning, tokenization, freq computation, log likelihood computation, RV coefficient algorithm and grouping. Then there will be all users, they can view themes for articles submitted by n users and view grouping of all articles, Now coming to registration in this we should give first name last name user id password email id and then step number 2 is perform basic validation like first name cannot be empty then the last name cannot be empty then the user id should not be empty then the password cannot be empty email cannot be empty all these basic validation should per-form after the successful validation go to the next step oth-erwise show validation error for the user, if validations are successful retrieve list of user id's from the application like u1,u2,u3 etc till Un we check whether the given user id is in the list if yes validation error otherwise user is created and registered successfully, this is the registration process. Then after that next basic module is login for this username and password for admin and user, basic valida-tion's like username cannot be empty and password cannot be empty after this if basic validations are successful then retrieve list of existing user names and check whether given username is in usernames list after that if it is there then pro-ceed otherwise validation error, then in the next step retrieve the actual password for the user password actual then com-pare actual password whether it is equal to given password from the user if both are same means continue otherwise invalid user/password error, otherwise what we do is if password validation is successful then obtain the login type if login type is double equals to 1 then provide admin func-tionality otherwise provide user functionality. In the initial stage of article submission the articles are submitted in two ways 1. Directly copy and paste the content of article which is exist in the local machine 2. Past-ing the URL link, in these two ways we can submit the ar-ticles by giving the appropriate article name. After the successful submission of the articles the articles will be cleaned on the data cleaning method by using pre reserved stop words, and in this proposed solution con-sist an option for inserting and removing stop words, accord-ing to these saved stop words the data cleaning process will be done and in this process the unwanted data which is simi-lar to the stop words only are removed from the articles and after this data cleaning the article descriptions will be mea-ningless and the cleaned article will be stored in the format of cleaned, article name and clean article.

V. RESULT AND DISCUSSION:

The below screenshots are usefull to understand the interface and the procedure of the process

Figure.2. Article submission using URL

The system will fetch the article description from div tags using DOM tree technique

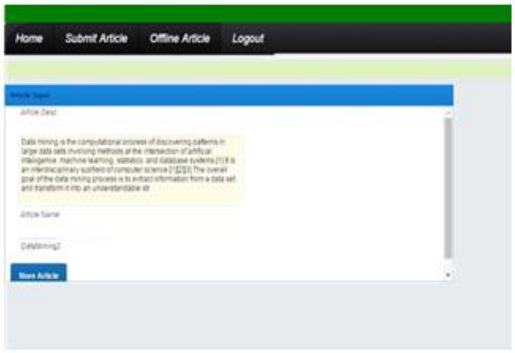

Figure.3. Article submission by pasting description of the article And here the description size should be limited

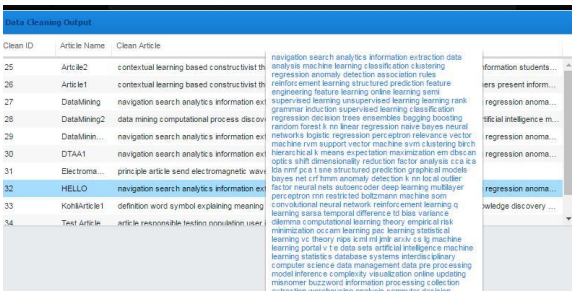

Figure.4. List of all submitted articles

Each of the cleaned articles consist its own clean ID

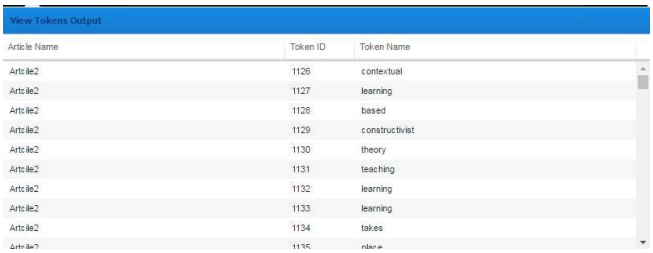

Figure.5. Tokenized each word of the article and the ID will be unique

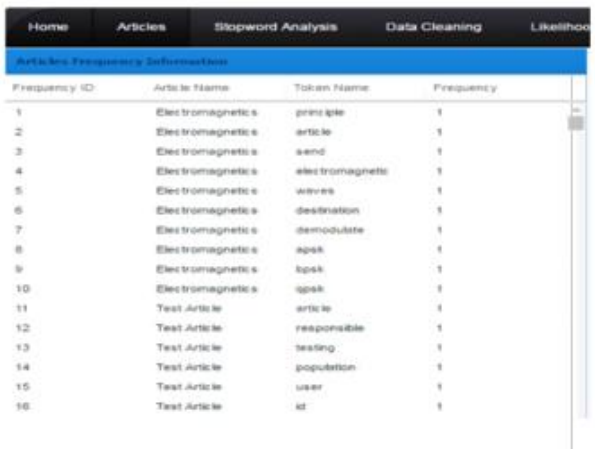

Figure.6. Frequency calculation over the articles according to the number of repeated token names and generates unique fre-quency ID

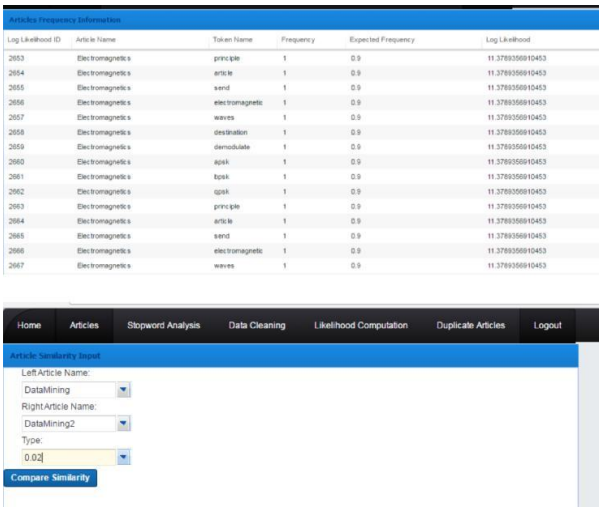

Figure.7. Comparing articles by selecting both left and right different articles with threshold value

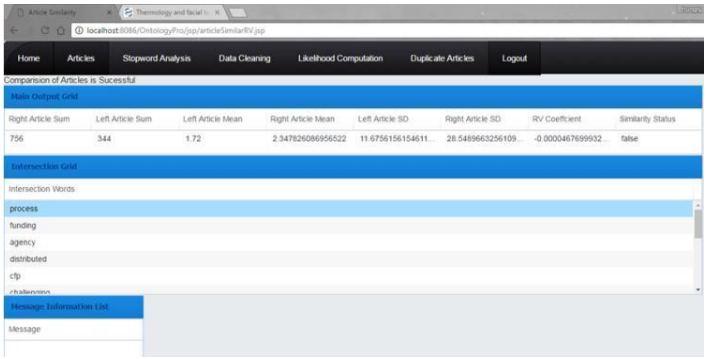

Figure.8. Similarity status of the compared articles

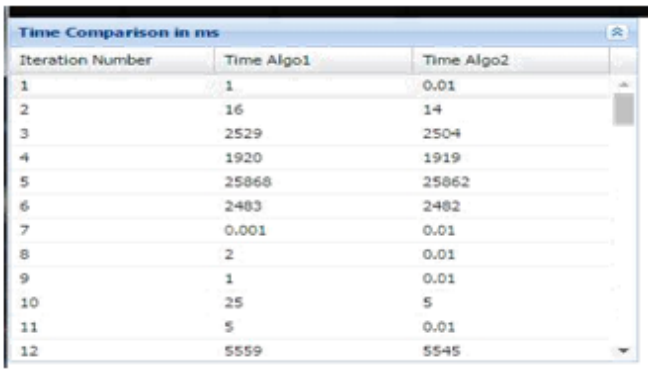

Figure.9. Time consumption comparison with proposed and previous system

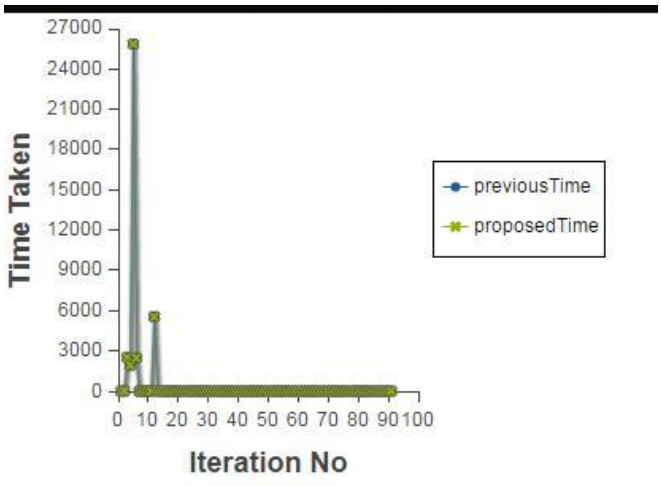

Figure.10. Graphical view of the previous and proposed system time

VI. ACKNOWLEDGMENTS

I would like to acknowledge my sincere thanks to Basa-veshwara Engineering College Department of MCA for giv-ing support, resources for doing this research paper.

VII. REFERENCES

[1].Q. Tian, J. Ma, and O. Liu, “A hybrid knowledge and model system for R&D project selection,” Expert Syst. Appl., vol. 23, no. 3, pp. 265–271, Oct. 2002.

[2].K. Chen and N. Gorla, “Information system project se-lection using fuzzy logic,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 28, no. 6, pp. 849–855, Nov. 1998.

[3].A. D. Henriksen and A. J. Traynor, “A practical R&D project-selectionscoring tool,” IEEE Trans. Eng. Manag., vol. 46, no. 2, pp. 158–170, May 1999.

[4].F. Ghasemzadeh and N. P. Archer, “Project portfolio selection through decision support,” Decis. Support Syst., vol. 29, no. 1, pp. 73–88, Jul. 2000.

[5].J. Butler, D. J. Morrice, and P. W. Mullarkey, “A multi-ple attribute utility theory approach to ranking and selec- tion,” Manage. Sci., vol. 47, no. 6, pp. 800–816, Jun. 2001.

VII. BIOGRAPHIES

Prof. GOVINDRAJ CHITTAPUR received the B.Sc. degree in Computer Science from the University of KUD, Dharwad, Karnataka, in 2002, the M.C.A degree from VTU, Belgaum, Karnataka, in 2005 and awarded M.sc Technology by Research from Mysore University in 2011. Has published 17 international journal and national journal has also served as Reviewer and Editorial Board Member of various interna-tional journal and conferences. His major research area in-cludes Image and Video forensics, Machine learning and Data Mining.

Mr. CHETAN M. KONNUR received the B.Sc. degree in computer science from the Rani Channamma University, Belgaum, Karnataka, in 2014, and pursuing MCA degree under VTU, Belgaum, Karnataka.