Impact Factor Value: 4.029 ISSN: 2349-7084
International Journal of Computer Engineering In Research Trends
Volume 4, Issue 11, November - 2017, pp. 524-528 www.ijcert.org

Relevance Feature Search for Text Mining: A Survey

Rekha R. Kamble*[1], Dattatraya V. Kodavade[2]

[1] *PG Scholar, Dept of Computer Science and Engineering,* DKTE Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute), *416115, India*
[2] *Professor, Dept of Computer Science and Engineering,* DKTE Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute), *416115, India*

rekhakamble2604@gmail.com[1], dvkodavade@gmail.com[2]

Abstract: - To determine the quality of user searched documents is a huge challenge in discovering relevance feature. To search the text, document, image, etc. approximately user want relevant features. The techniques earlier used where term based and pattern based. These days clustering methods like partition based, density based and hierarchical is used along with different feature selection method. Extracting terms from the training set for describing relevant features is known as the term-based approach. Low-level support problem is solved by partition based text mining, but it suffers from a large number of noise patterns. Information content in documents is identified by frequent sequential patterns and sequential patterns in the text documents and the useful features for text mining are extracted from this. Extracted terms are classified into three type's positive terms, general terms and negative terms. To deploy high-level features over low level features positive and negative patterns in text documents are discovered in the present paper.

Keywords: Text mining, text feature extraction, text classification

1. Introduction

To find useful information from huge digital text documents, text mining is used. The quality of text representation is improved by text classification method it also develops high-quality classifier. The information that meets users' needs is retrieved by using text mining model this is to be done within an efficient time interval. To retrieve relevant documents as many as possible is the goal of Traditional Information Retrieval. Non-relevant ones are filtered out at the same time.

In Data Mining process information is extracted from data set and that data is transformed into an understandable structure. Then patterns in data are found which are then characterized as extracting hidden information also previously unknown and useful information is also found from the data.

Feature in text documents that describe the user preferences is a significant matter if the quality of discovered relevance is to be ensured. Because of large-scale terms and data patterns, this is a very challenging task. Term-based approaches were adopted in existing text mining algorithm, in which problems of polysemy and synonymy were there. Synonymy word means having same meaning same as another word. Polysemy word means having two or more meanings. In term based method each term in the document is associated with a value known as weight. Most existing text mining and classification methods use term-based approaches that used in many previous classification methods and text mining.

Closed sequential patterns in text documents are being discovered by Pattern taxonomy mining (PTM). The pattern is a set of terms that are frequently appeared in a paragraph. Information filtering is done by few pattern mining approaches. To remove redundant and noisy patterns, some Data Mining techniques are being developed. As compared to the term based method, the pattern based method performs better in describing user preferences.

The Relevance Feature Discovery introduce a model in which terms are classified into different categories after this term weights are updated then these terms are distributed in patterns efficiently so that performance of text mining can be improved. The objective of Relevance Feature Discovery is to extract high-quality features that can represent what user needs. As compared to Term based Methods and Pattern based methods this system is better.

The paper is organized as follows, Section 2 contain the literature survey of feature selection techniques, Section 3contain architecture of proposed work, Section 4 contain implementation part of preprocessing, in section 5 we discussed results. , Section 6 we conclude about work done, Section 7 contain future work.

2. Literature Review

Y. Li et al. [1] proposed a method to select negative documents that are closed to the extracted features in positive documents. It proposed the mining and revision algorithms which use twice for positive and negative documents. Features in positive documents are found by revision process in the training set which contains higher level positive patterns and low-level terms. After this top-K negative samples are selected from the training set in compliance with standard rules of the positive features. The feature discovery in the positive document is done using pattern mining technique and which is also used to discover negative patterns and terms from selected negative documents. The Revised weight function is obtained by the process of revised initial features. The negative patterns could be useful for revising positive features in training set. But negative patterns cannot largely improve accuracy is the drawback of the system.

N. Zhong et al. [2] mentioned a technique known as effective pattern discovery that overcomes two problems such as the low-frequency and misinterpretation. To refine the discovered patterns in text documents above mentioned technique uses pattern deploying and pattern evolving. It proposed D^- pattern mining algorithm. The training process is described to find the set of D^- patterns. Deploying process is focused that consists of term support evolution and D^- pattern discovery. D^- Pattern is composed of discovering all the positive documents. But how to effectively integrate patterns in both relevant and irrelevant documents is the drawback of the system.

Z. Zhao et al. [3] devised Similarity Preserving Feature Selection-Nesterov's method. Existing algorithms are improved by SPFS framework to overcome the drawbacks like handling feature redundancy. To choose a subset of the original features, Feature selection this is used. For this selection, criterion is taken into account that select a small set of original features. By considering the original features and feature selection, the learning models are improved. It is a learning process. To guide searching for relevant features this is the main objective of learning models. The drawback of the system is the identification of the optimal feature set without any redundancy.

Li et al.[4] proposed FClustering and WFeature algorithm. FClustering algorithm describes the process of feature clustering and after that terms are classified using the FClustering algorithm and WFeature algorithm that are used to calculate term weights. A model for relevance feature discovery is presented which discovers both positive and negative patterns in text documents, these features are described as higher level features and deploys them over low-level features.

Pattern mining technique has raised two issues. The first one is how to deal with low-frequency patterns and second is how to effectively use negative feedback to revise extracted features for information filtering. Another challenging issue in text mining is a long scale pattern. The proposed system will present an innovative technique to overcome all above limitations and problems for finding and classifying low-level terms based on their appearance in the higher-level features and their specificity in a training set.

3. Proposed Work

In the web search Relevance is a big research problem, which discusses a document relevance to a user or a query. To select the relevant features in a text document is its main focus. Relevance Feature Discovery model uses two algorithms that describe user preferences to ensure the quality of discovered relevance features. The first one is used to cluster the terms into three classifications as positive, negative and general. In text documents positive and negative found as higher level features and they are deployed on low-level features. The second algorithm is used to calculate the feature weight. Relevant and irrelevant feedback is used to find useful features. Integration of term and pattern features is done instead of using them in two separated stages.

The proposed system architecture of Text mining process is as follows. (Figure 1)

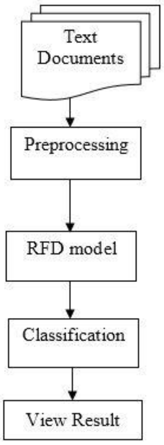


Figure 1. System architecture of Text mining process.

4. Implementation

The first module is preprocessing. In this module, the preprocessing is applied to text documents. The stop words are removed from the documents and stemming process is used to reduce a word to its root form.

The Porter Stemmer Algorithm [9]:

It is a simple rule-based algorithm for stemming. The algorithm consists of seven sets of rules, applied in order.

Consonant is a letter other than A, E, I, O, U, and Y preceded by consonant.

Vowel is any other letter.

With this definition, all words are of the form:

$(C)(VC)^m (V)$

Where, C-string of one or more consonants (con+)

V-string of one or more vowels

The rules are of the form:

(condition) S1 -> S2

Where S1 and S2 are suffixes

m-The measure of the stem

*S-The stem ends with S

v-The stem contains a vowel

*d-The stem ends with a double consonant

*o-The stem ends in CVC (second C not W, X, or Y)

Step 1: removes final -es.

SSES -> SS

IES -> I

SS -> SS

S -> ϵ

Step 2a: gets rid of plurals and -ed or -ing.

(m>1) EED -> EE

(*V*) ED -> ϵ

(*V*) ING -> ϵ

Step 2b:

(These rules are ran if second or third rule in 2a apply)

AT-> ATE

BL -> BLE

(*d & ! (*L or *S or *Z)) -> single letter

(m=1 & *o) -> E

Steps 3 and 4: turns terminal y to i when there is another vowel in the stem.

Step 3: Y Elimination (*V*) Y -> I

Step 4: maps double suffices to single ones. so -ization (= -ize plus -ation) maps to -ize etc. note that the string before the suffix must give m() > 0.

(m>0) ATIONAL -> ATE

(m>0) IZATION -> IZE

(m>0) BILITI -> BLE

Steps 5 and 6:

Step 5: deals with -ic-, -full, -ness etc. similar strategy to step 4.

(m>0) ICATE -> IC

(m>0) FUL -> ϵ

(m>0) NESS -> ϵ

Step 6: takes off -ant, -ence etc., in context <c>vcvc<v>.

(m>0) ANCE -> ϵ

(m>0) ENT -> ϵ

(m>0) IVE -> ϵ

Step 7: removes a final -e if m() > 1.

Step 7a:

(m>1) E -> ϵ

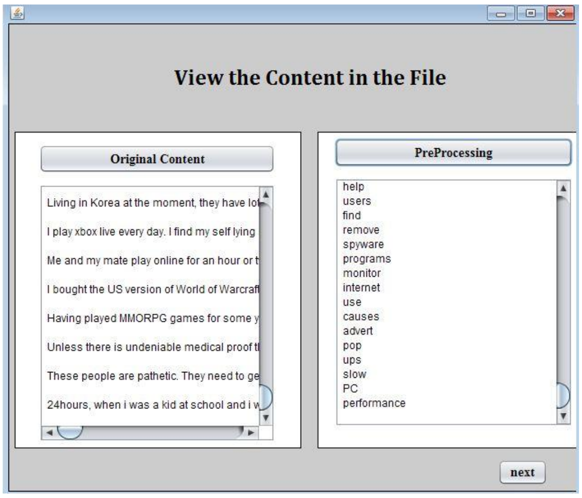
(m=1 & !*o) NESS -> ϵ

Step 7b:

(m>1 & *d & *L) -> single letter

5. Result Obtained

Figure2. Shows the removal of stop-words. The stop-words are removed according to given stop-words list.



Living in Korea at the moment, they have lot

I play xbox live every day. I find my self lying

Me and my mate play online for an hour or t

I bought the US version of World of Warcraf

Having played MMORPG games for some y

Unless there is undeniable medical proof t

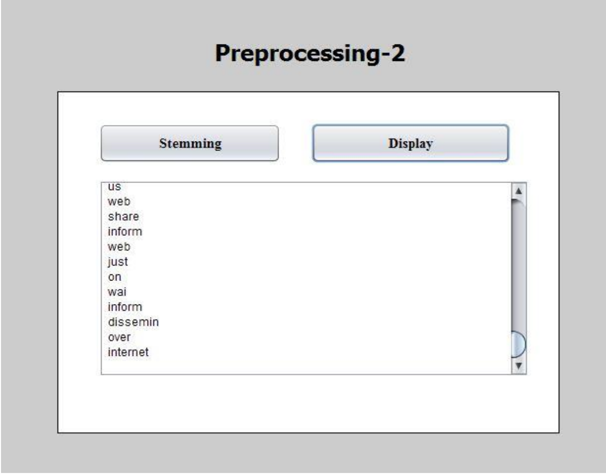
These people are pathetic. They need to ge

24hours, when i was a kid at school and i w

help
users
find
remove
spyware
programs
monitor
internet
use
causes
advert
pop
ups
slow
PC
performance

Figure 2. Removal of stop-words.

Figure.3. shows were stemming process. Stemming is the process of reducing words to their root form.



us
web
share
inform
web
just
on
wai
inform
dissemin
over
internet

Figure 3. Stemming terms

.

6. Conclusion

The first module preprocessing is implemented. The stop-words are removed according to a given stop-words list from the given document. The Porter Stemming algorithm is used for stemming terms. The terms are reduced to their root form.

7. Future work

In future, we develop the Relevance Feature Discovery (RFD) model. The RFD model describes the relevant features about three groups: positive specific terms, general terms and negative specific terms based on their appearances in a training set.

8. References

[1] Y. Li, A. Algarni, and N. Zhong, “Mining positive and negative patterns for relevance feature discovery,” in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2010, pp. 753–762.

[2] N. Zhong, Y. Li, and S.-T. Wu, “Effective pattern discovery for text mining,” in IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 30–44, Jan. 2012.

[3] Z. Zhao, L. Wang, H. Liu, and J. Ye, “On similarity preserving feature selection,” in IEEE Trans. Knowl. Data Eng., vol. 25, no. 3, pp. 619–632, Mar. 2013.

[4] YueLi,, Arif ”Relevance feature discovery for text mining” IEEE transaction on knowledge and data engineering,vol.27,no.6, pp.1656-1669, june2015.

[5] N. Azam and J. Yao, “Comparison of term frequency and document frequency based feature selection metrics in text categorization,”Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768,2012.

[6] X. Li and B. Liu, “Learning to classify texts using positive andunlabeled data,” in Proc. 18th Int. Joint Conf. Artif. Intell., 2003,pp. 587–592.

[7] Y. Li, A. Algarni, S.-T. Wu, and Y. Xue, “Mining negative relevancefeedback for information filtering,” in Proc. Web Intell. Intell.Agent Technol., 2009, pp. 606–613.

[8] G. Salton and C. Buckley, “Term-weighting approaches in automatictext retrieval,” in Inf. Process. Manage., vol. 24, no. 5,pp. 513–523, Aug. 1988.

[9] The Porter Stemmer home page (with the original paper and code): http://www.tartarus.org/~martin/PorterStemmer/ 988.

[10] K.Arun .SrinageshandM.Ramesh,”Twitter Sentiment Analysis on Demonetization tweets in India Using R language.”International Journal of Computer Engineering in Research Trends., vol.4, no.6, pp. 252- 258, 2017.

[11] TekurVijetha, M.SriLakshmi andDr.S.PremKumar,”Survey on Collaborative Filtering and content-Based Recommending.”International Journal of Computer Engineering in Research Trends., vol.2, no.9, pp. 594- 599, 2015.

[12] N.Satish Kumar, SujanBabuVadde,”Typicality Based Content-BoostedCollaborative Filtering RecommendationFramework.”International

Journal of Computer Engineering in Research Trends., vol.2, no.11, pp. 809-813, 2015

[13] B.Kundan,N.Poorna Chandra Rao and DrS.PremKumar,"Investigation on Privacy and Secure content of location based Queries."International Journal of Computer Engineering in Research Trends., vol.2, no.9, pp. 543-546, 2015.

Author Profile

Rekha Kamble completed B.E. in Computer Science & Engineering from DKTE Society's Textile &Engineering Institute, Ichalkaranji, India in 2016. She is pursuing Master of Technology in Computer Science & Engineering from DKTE Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute), India.

Dr .D. V. Kodavade, the Head of Department of Computer Science & Engineering, at DKTE Society's Textile & Engineering Institute, Ichalkaranji (An Autonomous Institute),India. He is a member of the ACM, CSI, IEEE Computer Society. His current research interest includes Artificial Intelligence & Knowledge Based Systems, IoT, Neural Networks, Hybrid Intelligence.