ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 8, August 2016

Pattern Based Approach to Discover Relevant Text Files Using Text Mining Agent

S.Sasirekha,M.Phil Final Year,

Department of Computer Science,

Mother Teresa University,

Chennai- 15

Dr.Mrs.R.Janaki,MCA.,M.Phil.,Ph.D.,

Assistant Professor

Department of Computer Science,

Queen Mary's College (A),

Chennai-600 004

Abstract - In this paper we are proposing a method to discover the text that is already available in the database by using Text Mining Agent. The function of Text Mining Agent is to fetch all the text that are uploading by a user and it will check with each dataset of the text pattern that are available in the database. Text Mining Agent will produce two types of result that are DP+ (Data Pattern Positive) and DP- (Data Pattern Negative). If the Result of Text Mining Agent DP+ means it indicates that the text file a user a uploading already available in the database dataset so if the user upload the file that may create duplication of files and the memory occupied by the duplicate text file becomes waste of memory. So our approach make decision at the time when Text Mining Agent produce result i.e., DP+ means duplicate of text files so the file will not be uploaded to the Database and if the result is DP- means the file uploaded by the user does not match with the any Database so it is not a duplicate file it will be uploaded to the Database.

Key Words: Text mining, text feature extraction, text classification

I. INTRODUCTION

Data Mining is a method of fetching the files from the database. In our earliest approach we are using Features algorithm that is purely Term based approach. Term based approach means it will not compare with the text that are available in both the text files. So we approach for F-Clustering algorithm which means clustering of two data to produce certain results. The drawbacks of existing algorithm are they suffer from problem of polysemy and synonymy i.e., same text files may have different term name or file

All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 8, August 2016

name which result into occur of duplication of files . Similarly a different text files may contain same term name or file name so there is elimination of pure non-duplicate files. So in our propose approach we use F-Clustering algorithm that overcome the drawback of existing approach by using the Pattern Based Approach.

II. MATERIAL AND METHODS

In the Previous approach we use W-Features algorithm that is purely Term based Approach i.e., the algorithm checks all the relevant term from database and match it with file that are uploading. The Term Based Approach produce two type of result T+ and T- where T+(Term Positive) and T-(Term negative). The existing approach does not accept the test conditions of polysemy and synonymy. So two different text may contain same term name i.e., synonymy which may produce duplication of files and polysemy means same text may contain different content which eliminate non-duplicate files.

In the propose concept we use F-Clustering Algorithm which produces two type of results DP+(Data Pattern Positive) and DP-(Data Pattern Negative).If the Text Mining Agent produces DP+ means the text pattern user uploading already exist in the Database so it will not upload the file which overcome the problem of polysemy and synonymy and if it returns DP- the pattern of text the user uploading does not matches any text pattern of datasets available in the Database so it can be uploaded to the Database so the elimination of non-duplicate files will not occur anymore is the system.

To connect with Library user must give their username and password then only they can able to connect the server. If the user already exits directly can login into the server else user must register their details such as username, password and Email id, into the server. Server will create the account for the entire user to maintain upload and download rate. Name will be set as user id. . Logging in is usually used to enter a specific page.

There is “Received” table which shows admin whether Student had got the book. Student can able to update his/her database when he/she received the book. Admin can able to delete the student details from her page when the student had returned the library book. Admin can also upload a file to Library.

Inside College Library you can able to view the categories of books available in the library i.e., Java book, Engineering books etc. When he/she want to pick a book from college Library he/she can able to see the content of Book by pressing Details button after viewing the content if the book is needful to him then you can able to pick it by pressing pick it but

All Rights Reserved © 2015 IJARMATE

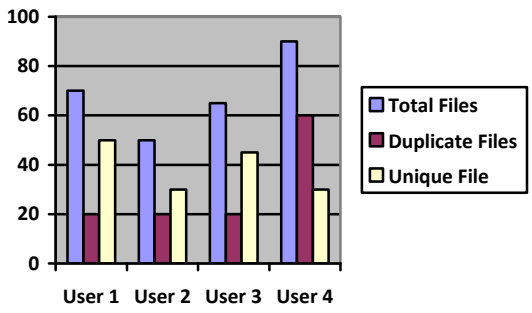
ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 8, August 2016

before picking books you need to give all your details to the admin regarding the book numbers Student name, apartment, id number, semester, E-mail, mobile number, Date etc and you want to type "WAITING" in the Received box i.e., The book is not received to you when the book is received You can update your profile by "Received".

Here Student can able to upload a file to Library and The Data mining Agent able to point out the content of the file is already available in the Database, if the file is already available in the library it will return that the file is already available in the Library otherwise the file will Uploaded Successfully.

Here Student can able to view and download the file directly from the Library. Here Student can able to view the content of uploaded files that are present in library and able to view and Download file from Library.

Student need to update her profile from "WAITING" to "RECEIVED" when he/she received book from Librarian. By giving her name and Book Number he/she received and should enter in Receipt "Received". So that Library admin may be able to know that the student had received the book. Then when the book is returned by the student then the Librarian can able to delete her profile from the Database. Christo Ananth et al. [13] discussed about a method, Wireless sensor networks utilize large numbers of wireless sensor nodes to collect information from their sensing terrain. Wireless sensor nodes are battery-powered devices. Energy saving is always crucial to the lifetime of a wireless sensor network. Recently, many algorithms are proposed to tackle the energy saving problem in wireless sensor networks. There are strong needs to develop wireless sensor networks algorithms with optimization priorities biased to aspects besides energy saving. In this project, a delay-aware data collection network structure for wireless sensor networks is proposed based on Multi hop Cluster Network. The objective of the proposed network structure is to determine delays in the data collection processes. The path with minimized delay through which the data can be transmitted from source to destination is also determined. AODV protocol is used to route the data packets from the source to destination.

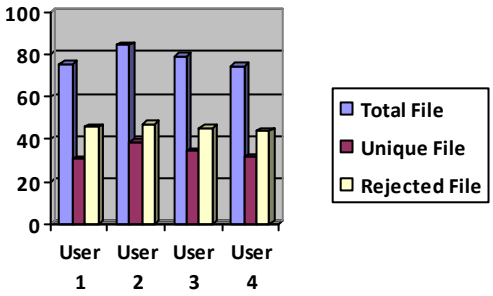


All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 8, August 2016

A. RESULTS

Given a topic, long patterns are usually more specific for the topic, but they usually appear in documents with low support or frequency. If the minimum support is decreased, a lot of noisy patterns can be discovered. Thus the drawbacks of existing algorithm is listed above and the drawbacks of algorithm is overcome by using Feature Clustering algorithm



The Above Table shows the duplicate files and unique files stored in the database.

FClustering ()

Input: Discovered features $< T, DP^+, DP^- >$ and function *spe*.
Output: Three categories of terms T^+, G and T^-.

Method:

1: $G = \emptyset, T^+ = \emptyset, T^- = \emptyset$;
2: foreach $t_i \in T$ do
3: if $t_i \notin \{t | t \in P, P \in DP^+\}$
4: then $T^- = T^- \cup \{t_i\}$;
5: foreach $t_i \in T - T^-$ do {
6: $c_i = \{t_i\}$;
7: $max_{spe}(c_i) = min_{spe}(c_i) = spe(t_i)$; }
8: let $m = |T - T^-|$;
9: let $C = \{c_1, c_2, \cdots, c_m\}$ and $min_{spe}(c_1) \geq \cdots \geq min_{spe}(c_m)$;
10: while ($|C| > 3$){ //start merging process
11: let $k = 1$ and $mind = dif(c_1, c_2)$;
12: for $i = 2$ to $m - 1$ do
13: if $dif(c_i, c_{i+1}) < mind$
14: then $\{k = i; mind = dif(c_i, c_{i+1})\}$;
15: let $c_k = c_k \cup c_{k+1}$;
16: if $min_{spe}(c_{k+1}) < min_{spe}(c_k)$
17: then $min_{spe}(c_k) = min_{spe}(c_{k+1})$;
18: if $max_{spe}(c_{k+1}) > max_{spe}(c_k)$
19: then $max_{spe}(c_k) = max_{spe}(c_{k+1})$;
20: for $i = k + 1$ to $m - 1$ do // delete c_{k+1} from C
21: let $c_i = c_{i+1}$;}
22: if $|C| = 1$ then $T^+ = c_1$
23: else if $|C| = 2$ then $\{T^+ = c_1; G = c_2\}$
24: else $\{T^+ = c_1; G = c_2; T^- = T^- \cup c_3\}$;

All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 8, August 2016

ALGORITHM

FCLUSTERING:

Calculation Feature Clustering depicts the procedure of highlight grouping, where DP+ is the set of found examples of D+ and DP-is the set of found examples of D-.

Preferences:-
It finds both positive and negative examples in content records as more elevated amount highlights and sends them over low-level components (terms).

It additionally orders terms into classes and upgrades term weights in view of their specificity and their dispersions in example.

III DISCUSSION

We believe that the positive feedback is more constructive than the negative feedback since the objective of relevance feature discovery is to find relevant knowledge. However, we believe that negative feedback contains some useful information that can help to identify the boundary between relevant and irrelevant information for improving the effectiveness of relevance feature discovery. The obvious problem for using irrelevant documents is that most of the irrelevant documents are not closed to the given topic because of the very large amount of negative information. Therefore, it is required to choose some useful irrelevant documents (offenders) to decide the groups of terms for the three categories i.e., synonymy and polysemy.

IV CONCLUSION

To protect user privacy, various privacy-preserving classification techniques have been proposed over the past decade. The existing techniques are not applicable to outsourced database environments where the data resides in encrypted form on a third-party server. This paper proposed a novel privacy-preserving k-NN classification protocol over encrypted data in the cloud. Our protocol protects the confidentiality of the data, user's input query, and hides the data access patterns. We also evaluated the performance of our protocol under different parameter settings.

REFERENCES

[1] M. Aghdam, N. Ghasem-Aghaee, and M. Basiri, "Text featureselection using ant colony optimization," in Expert Syst. Appl.,vol. 36, pp. 6843–6853, 2009.

[2] A. Algarni and Y. Li, "Mining specific features for acquiring userinformation needs," in Proc. Pacific Asia Knowl. Discovery DataMining, 2013, pp. 532–543.

All Rights Reserved © 2015 IJARMATE

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 8, August 2016

[3] A. Algarni, Y. Li, and Y. Xu, "Selected new training documents toupdate user profile," in Proc. Int. Conf. Inf. Knowl. Manage., 2010,pp. 799–808.

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

[5] R. Bekkerman and M. Gavish, "High-precision phrase-based documentclassification on a modern scale," in Proc. 11th ACMSIGKDD Knowl. Discovery Data Mining, 2011, pp. 231–239.

[6] A. Blum and P. Langley, "Selection of relevant features and examplesin machine learning," Artif. Intell., vol. 97, nos. 1/2, pp. 245–271, 1997.

[7] C. Buckley, G. Salton, and J. Allan, "The effect of adding relevanceinformation in a relevance feedback environment," in Proc. Annu.

[8] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansionterms for pseudo-relevance feedback," in Proc. Annu. Int.

ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 243–250.

[9] G. Chandrashekar and F. Sahin, "Asurvey on feature selectionmethods," in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.

[10] B. Croft, D. Metzler, and T. Strohman, Search Engines: InformationRetrieval in Practice. Reading, MA, USA: Addison-Wesley, 2009.

[11] F. Debole and F. Sebastiani, "An analysis of the relative hardnessof Reuters-

[12] J. Eisenstein, A. Ahmed, and E. P. Xing, "Sparse additive generativemodels of text," in Proc. Annu. Int. Conf. Mach. Learn., 2011,pp. 274–281.

[13] Christo Ananth, T.Rashmi Anns, R.K.Shunmuga Priya, K.Mala, "Delay-Aware Data Collection Network Structure For WSN", International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST), Volume 1,Special Issue 2 - November 2015, pp.17-21

[14] Y. Gao, Y. Xu, and Y. Li, "Topical pattern based document modellingand relevance ranking," in Proc. 15th Int. Conf. Web Inf. Syst.

Eng., 2014, pp. 186–201.

[15] X. Geng, T.-Y. Liu, T. Qin, A. Arnold, H. Li, and H.-Y. Shum,"Query dependent ranking using k-nearest neighbor," in Proc.Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, 21578 subsets," J. Amer. Soc. Inf. Sci. Technol., vol. 56,no. 6, pp. 584–596, 2005.

All Rights Reserved © 2015 IJARMATE