RESEARCHERID
THOMSON REUTERS

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

RECOMMENDER SYSTEM FOR A SUITABLE QUESTION AND ANSWER FORUM

P.L.K.Priyadarsini*1, Tunikuntla Lakshmi Sai Varshitha[2] & Vadakattu Pratyusha[3]
*1,2&3School of Computing, SASTRA Deemed to be University, Thanjavur, India

ABSTRACT

In current scenario, online question answer platforms play a major role for different purposes. There always exists an ambiguity for upgrading ones knowledge in different subject areas to the users to select the best platform to post their queries in the social media. In this paper, we recommend the most appropriate platform to the user for queries to get a fast, relevant answer. Classification based on Bag of words strategy and hierarchical clustering are used to determine the appropriate platform for the query. Several attributes like number of relevant answers, likes, comments and temporal attribute are considered. Implementation is carried out using R language on the datasets from stack-overflow, twitter and yahoo answers. The results are analyzed and it is found that the proposed algorithm recommends correct question-answer forum based on the question's domains.

Keywords: *Hierarchical clustering, Classification, Bag of Words, Question and Answer Platform, Natural Language Processing.*

I. INTRODUCTION

Social media is the most popular platform to learn, share, and express the knowledge, opinions and experiences of people. A part of it contains only knowledge providing sites where one searches for the answer to their query while some others are discussion forums where one answers to the query and others express their views and opinions to that answer or to that query. These discussion platforms connect people across the world .With the increased usage of these forums, individual's knowledge sources got expanded and hence one saves his time to find a relevant answer. Some of these discussion platforms are Stack-overflow, Twitter, Yahoo etc.

With the increasing technology, the number of such platforms has drastically increased. This led to an ambiguity as users are now confused to select the most appropriate platform for their query. For example if a person chooses to post a query on technological aspects and posts it in a twitter platform where more of general topics are discussed and less of technology are discussed ,his probability of getting a relevant answer is very low. Hence the problem of finding an appropriate platform to post a query.

In Conventional method, to solve this issue, the user needs to post his query in as many platforms as possible and need to compare the results for the most relevant one. Some platforms may not have immediate reply but has a relevant answer whereas some may have immediate reply but no relevant answer, some may not have either or some may have both .This comparison is thus a tedious process and wastes user's time and energy.

In our proposed system, we take the input query from the user and suggest him the appropriate platform where reasonable discussions are held on that topic of query, thus saving time and energy of the user .The data is first categorized into appropriate domains using bag of words strategy and then for each domain, community detection is done by hierarchical clustering. .Later the obtained results are compared with the results of other platforms, thereby recommending the most appropriate platform for his query.

II. RELATED WORK

Several works related to our work are described in this section. The first one is a Bag of words strategy, which is the new technique used when there is a need to compare data with some predefined words. Some of its applications include assessing activities of daily living using wrist watch accelerometer data where the obtained values through accelerometer[1] is compared with existing values in bag of words . This strategy is also used in human emotion classification[2] where the ECG signal is taken for several emotions, classified in which the bag of words contain the code words for the signal.

In [3], problem of spam detection, where false mails are identified from the bulk of mails, is addressed. The semantic text classification method is used to analyze the contents of mails. In [4], Sentiment analysis is used to estimate the opinions of the user, which is very useful for business development applications. There were also some works which use the concepts of semantics [5], lemmatization [6], N-grams [7]. These methods increase the effectiveness of the analysis process of data in natural language.

Some of the works use Agglomerative Hierarchical clustering of data [8]. Next, we have social media platform analysis. In [9], they proposed a method to retrieve the appropriate data from multiple social media platforms. Many works includes the temporal attribute while analyzing the data [10].

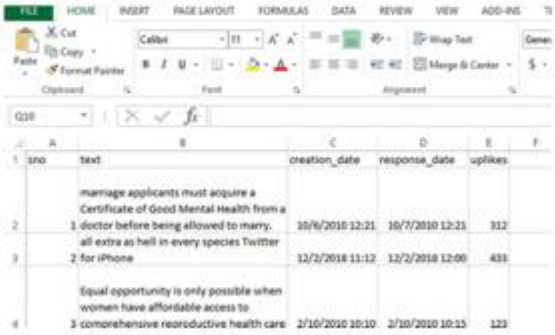
III. PROPOSED WORK

In this work, a recommender system for question answer platform is proposed. Datasets of two platforms namely Stack-overflow and twitter are considered for the recommendation. The method is broadly divided into steps as follows:

STEP 1. Classification of the available data into several domains using Bag of Words strategy STEP 2. Perform hierarchical clustering For each domain
STEP 3. Find the most relevant cluster within each domain
STEP 4. Comparison of the results obtained from different platforms
STEP 5. Creation of Interface and working

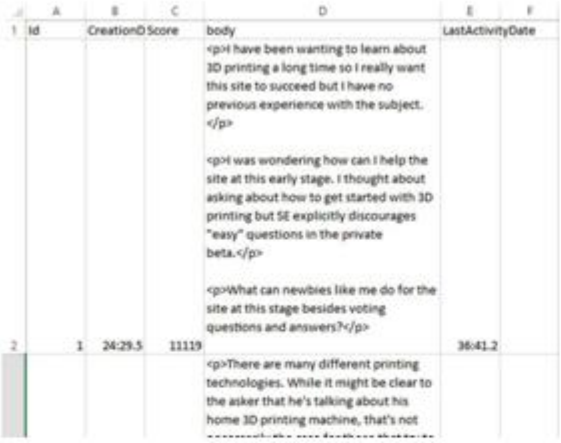
STEP 1:
In *Bag-of-Words* model, a bag represents a text document of group of words where the grammar and word ordering are excluded but the multiplicity is maintained. [11]

Dataset description
Each of the datasets used includes User-ID, Creation-Date, Posts, Response-Date, Up-likes, Down-likes, Comments and Number-Of-Responses. User-ID is the unique ID given to a particular user. Creation-Date is the date and time for a particular post of the user. Post is the actual query given by the user. Response-Date is the date and time of the first response for a particular query .Up-likes are the positive reactions to a particular query. Down-likes are the negative reactions to a particular query. Comments are the discussions for a posted query. Number-Of-Responses are the total number of discussions held for a particular post. The Fig1 & Fig2 shows the snapshot of both the twitter dataset and stack-overflow dataset.



	A	B	C	D	E
1	sno	text	creation_date	response_date	uplikes
2	1	marriage applicants must acquire a Certificate of Good Mental Health from a doctor before being allowed to marry.	10/6/2010 12:21	10/7/2010 12:21	312
3	2	all extra as hell in every species Twitter for iPhone	12/2/2018 11:12	12/2/2018 12:00	433
4	3	Equal opportunity is only possible when women have affordable access to comprehensive reproductive health care	2/10/2010 10:10	2/10/2010 10:15	123

Fig 1 Sample Twitter Data



	A	B	C	D	E	F
1	Id	CreationD	Score	body	LastActivityDate	
2	1	24:29.5	11119	<p>I have been wanting to learn about 3D printing a long time so I really want this site to succeed but I have no previous experience with the subject. </p> <p>I was wondering how can I help the site at this early stage. I thought about asking about how to get started with 3D printing but SE explicitly discourages "easy" questions in the private beta.</p> <p>What can newbies like me do for the site at this stage besides voting questions and answers?</p>	36:41.2	
				<p>There are many different printing technologies. While it might be clear to the asker that he's talking about his home 3D printing machine, that's not		

Fig 2 Sample Stack overflow Data

Classification of the dataset to several domains gives the necessity to have a bag of words for each domain. Bag of words generally mean the set of words used in particular domain. For example, health bag consists of words like health, patient, operation, nutrition etc. These bag of words are taken from the internet. The following steps are followed

- The posts/queries/comments of the user are in natural language. So, before analyzing the text, cleaning should be done. Cleaning of text includes removing punctuation, numbers, stop-words (generally those used for formation of sentences like if, or, for and etc. which are of no use during analysis).
- After cleaning the text, it is split into separate words for easier analysis.
- Now, each post of the user in the dataset is considered and compared with the bag of words of different domains and a count is taken.
- The maximum count defines the domain to which the query is classified.

This is repeated for all the data in the dataset.

STEP2:

Our next step is to find the relevant cluster in each domain. Generally, the relevancy of an answer is identified by considering more Up-likes, less Down-likes, more Number-of-Responses (indicating vast discussion), quick reply (minimum difference between Creation-date and Response-date) etc. So, for our convenience we introduce a new attribute, score which combines all the above mentioned attributes.

Considering Up-likes as U, Down-likes as D, Number-of-Responses as N, Creation-date as C, Response- date as R, Posts as P, Score can be

calculated as

$$\forall P,\ SCORE(P) = {}^{(U-D+N)}{}_{(R-C)} \;\; /$$

For each domain identified in Classification, hierarchical clustering is done. Hierarchical clustering consists of two steps: (i) to find a distance metric and to (ii) hclust the resultant. Distance metric can be found out using methods like Euclidean, Man-Hatten and hclust consists of methods like ‘average’, ’single’, ’complete’, ’WARD.D’. Any of these combinations can be used for clustering. According to our observations WARD.D gives best clustering results.

STEP 3:

After the clustering, the clusters obtained may contain an uneven number of posts. To find the relevant cluster, first we need to normalize all the clusters available.

By considering Normalization as n, Score as S, Total-Score as T, Number of as N

$\forall cluster, n = {}^{S}/_{N}$

The maximum of n indicates the most relevant cluster which is calculated for each domain. This can be further extended to find the most relevantly discussed subtopic in each platform. For example, the subtopic of health includes Ortho, Neuro, Pediatric, etc. To find the relevant topic discussed in a platform, it is enough to consider or examine only the most relevant cluster in each domain. To identify to which sub cluster each post belongs to, the same classification procedure discussed in step-1 is done, with the posts being only the posts of relevant cluster and bag of words topics being those subtopics for further classification.

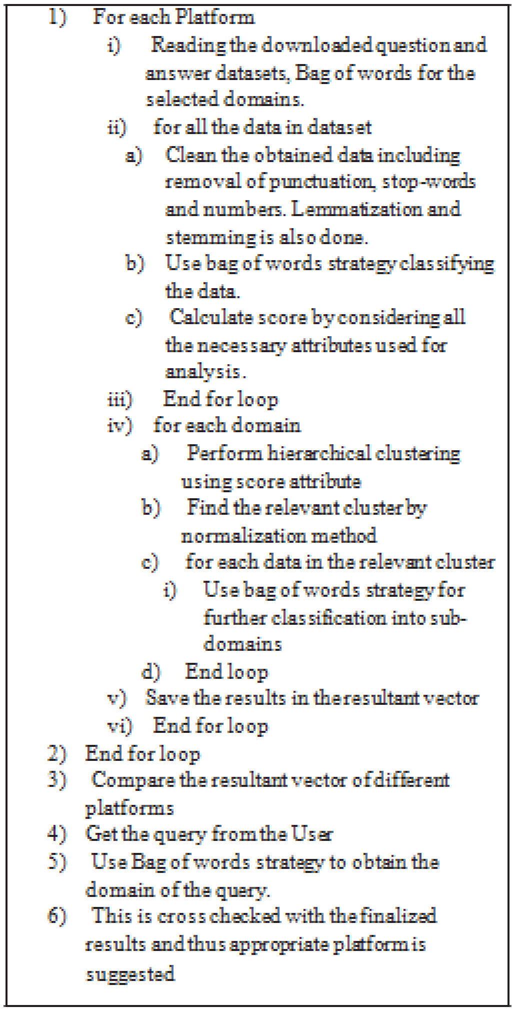
STEP4:

The obtained results for each cluster in each domain is stored in a finalized vector. The above procedure is repeated for all the available platforms and a finalized vector for each platform is obtained. Now, in-order to find where the relevant discussions are held, the corresponding finalized vectors are compared and the results are displayed.

STEP 5:

The final step of the process involves the interaction with the beneficiaries. For this purpose, we utilized a package named 'shiny' in R language. Our interface consists of an input space, submit button and an output space. Input space accepts input query of the user. As soon as the submit button is clicked, the input query is accepted by our algorithm and is first split into individual words and classified to a domain using bag of words technique. Now the domain is checked with the finalized vectors obtained in step 4. This computed result (the relevant platform) is displayed in the output space.

Algorithm



```
1) For each Platform
    i) Reading the downloaded question and
       answer datasets, Bag of words for the
       selected domains.
    ii) for all the data in dataset
        a) Clean the obtained data including
           removal of punctuation, stop-words
           and numbers. Lemmatization and
           stemming is also done.
        b) Use bag of words strategy classifying
           the data.
        c) Calculate score by considering all
           the necessary attributes used for
           analysis.
    iii) End for loop
    iv) for each domain
          a) Perform hierarchical clustering
             using score attribute
          b) Find the relevant cluster by
             normalization method
          c) for each data in the relevant cluster
              i) Use bag of words strategy for
                 further classification into sub-
                 domains
          d) End loop
    v) Save the results in the resultant vector
    vi) End for loop
2) End for loop
3) Compare the resultant vector of different
   platforms
4) Get the query from the User
5) Use Bag of words strategy to obtain the
   domain of the query.
6) This is cross checked with the finalized
   results and thus appropriate platform is
   suggested
```

IV. EXPERIMENTAL RESULTS

We have collected the data from stack overflow and twitter datasets .The input to our algorithm is the query of the user and output is a platform recommendation i.e., name of the platform to the user. For example, the user wants to post a query “what is dynamic programming? ”. Our algorithm takes this as an input, and recommends the relevant platform to post that query such that user could get the best answer in less time. Here example:” stack-overflow” is displayed in the output section.
The analysis of the algorithm is done in the following way:

Initially the complete dataset available is classified according to domain using bag-of words technique. Bag of words can be best explained with the following description. For example, we consider three domains namely politics, technical and health. Every entry data in the considered dataset is compared to each bag and number of matching of the words to each bag is noted. The higher the matching, that entry belongs to that domain.

ISSN 2348 – 8034
Impact Factor- 5.070

Table 1. Table showing the process of classification

Entry	Health bag	Politics bag	Technology bag	Domain
What are the precautions taken for cardiac disease	3	1	0	health
Example of dynamic programming	0	0	2	Technology
Politics in India is becoming worse day by day	0	2	0	politics
How to reduce belly-fat?	1	0	0	health
What food habits are to be followed to have a healthy life?	3	1	0	health

In example 1 "what are the precautions taken for cardiac disease". This after pre-processing the remaining words are precautions, cardiac & disease. Precautions comes in both politics bag with score 1 and in health bag with score 3.Technology bag has no match making it as 0.The maximum match lies within the health bag making it fall under health domain. If suppose there is more than one maximum then we can make it fall into general category.

The Fig 3 shows the classification of sample of 150 records falling in different categories using bag of words strategy. The x-axis indicates the marking of the records according to domain and y axis indicates the marking of the data according to the score of the data.

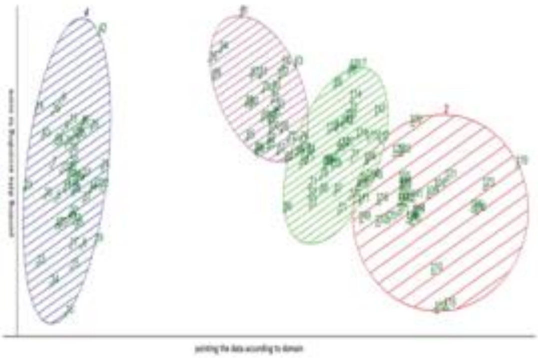


Fig 3 classification to different domains

Next, clustering of data in each domain is done. Agglomerative Hierarchical clustering is carried out to find the most relevant cluster where score (step 2) is taken as primary criteria. The Fig 4 explains the hierarchical clustering of a single domain.

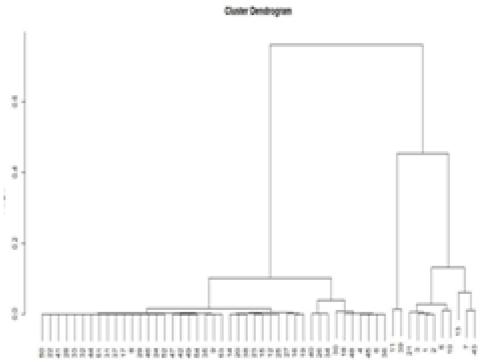


Fig 4 hierarchical clustering

Relevant cluster among the available clusters is calculated as explained in step 3.A vector stores the finalized and normalized scores in each domain. This method is done to all the platforms and finalized scores are stored respectively.

When the user gives his query input, the algorithm first identifies the domain of query. For example, if the user has identified the health domain, then the health attribute of the finalized vectors of all platforms is compared. The maximum of the comparison gives the relevant platform. This platform indicates most relevant discussions held on that domain. Hence it is suggested accordingly.

The Fig 5 shows the statistical analysis of the sample of data for correct classification of queries to separate domains. The y-axis indicates the number of queries. The x-axis indicates the domains of classification. When 50 queries of health, 50 queries of politics and 50 records of technology are considered the Fig 5 shows the classification by our algorithm

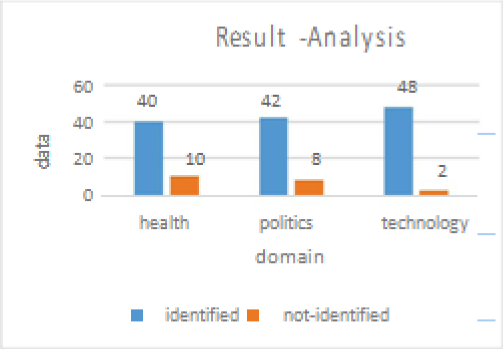


Fig 5 graph showing average performance

ISSN 2348 – 8034
Impact Factor- 5.070

Table 2 Table showing Average Performance

Domain related queries	Correctly classified	Incorrectly classified	Average performance
Health related queries	40	10	0.8
Politics related queries	42	8	0.84
Technology related queries	48	2	0.96
Total	130	20	0.866

The above table indicates the average performance of identification of the queries according to their respective domains. The high probability indicates that the algorithm is effective.

The sample input output is explained in Table 3

V. CONCLUSION

In this paper, we designed an algorithm to know the most relevant platform that could efficiently answer the query of the user. This efficiency includes least response time for the query posted, the maximum valid discussions regarding that domain and relevancy of the answer posted. This paper also utilizes the bag of words strategy where each bag contains all the words related to a particular domain. So, when a new post is posted this bag of words can easily identify the domain in spite of the word which is not frequently used earlier. We preferred agglomerative hierarchical clustering over k-means clustering technique because in k-means, the data first fixes the centers (number given by user) and then assigns other data elements whereas in agglomerative hierarchical method each data element is initially treated itself as a cluster and these clusters are combined at each level till a fixed number of clusters are reached (number given by the user). Altogether it can be concluded that, by using our algorithm, the user can get the best answer in less time. It can be easily observed that this algorithm can be extended to work with more question-answer forums given.

Table 3 Table showing final input and output.

Input	Output
What are precautions taken for cardiac disease	Twitter
Example of dynamic programming	Stack-overflow
Politics in India is becoming worse day by day	Twitter
How to reduce belly-fat?	Twitter
What are the food habits to be followed to have a healthy life?	Twitter

VI. ACKNOWLEDGEMENTS

We would like to thank SASTRA Deemed to be university for providing facilities to carry out our work and for motivating us in every way possible. Also, we are thankful to twitter as well as Stack-overflow communities for providing the necessary data.

REFERENCES

1. *Matin Kheirkhahan, Shikha Mehta, Madhurima Nath,Amal A. Wanigatunga, Duane B. Corbett, Todd M. Manini and Sanjay Ranka :"A bag-of-words approach for assessing activities of daily living wrist accelerometer data", Bioinformatics and Biomedicine (BIBM),2017 IEEE International Conference, December 2017*
2. *Ljiljana Seric and Pero Bogunovic: "Human emotions classification using bag-of-words method on single electrode brain interface", Software, Telecommunications and Computer Networks (Soft-COM), 2017 25th International Conference, November 2017*
3. *Wei Hu, Jinglong Du and Yongkang Xing: "Spam filtering by semantics-based text classification", Advanced Computational Intelligence (ICACI), 2016 Eighth International Conference, April 2016*
4. *Pushpak Bhattacharyya, "Sentiment Analysis", Emerging Trends and Applications in Computer Science (ICETACS), 2013 1st International Conference, December 2013.*
5. *Taimoor Hassan, Shoaib Hassan and Muhammad Asfand Yar:" Semantic analysis of natural language software requirement", Innovative Computing Technology (INTECH), 2016 Sixth International Conference, February 2017*
6. *Daniel Yue Zhang, Dong Wang, Hao Zheng, Xin Mu, Qi Li and Yang Zhang: "Large-scale point=-of-interest category prediction using natural language processing models", Big Data(Big Data), 2017 IEEE International Conference, January 2018*
7. *Dijana Kosmajac, Vlado Keselj: "Language identification in multilingual, short and noisy texts using common N-grams", Big Data (Big Data), 2017 IEEE International Conference, January 2018*
8. *Ryo Ozaki, Yukihiro Hamasuna and Yasunori Endo: "Agglomerative Hierarchical Clustering Based on Local Optimization for Cluster Validity Measures", 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), October 2017*
9. *Daichi Takehara, Ryosuke Harakawa, Takahiro Ogawa, Miki Haseyama: "Extracting hierarchical structure of content groups from different social media platforms using multiple social metadata", May 2017*
10. *Fan Xia, Chengcheng Yu, Linhao Xu, Weining Qian and Aoying Zhou: "Top-k temporal keyword search over social media data", January 2017*
11. *Jiawei Han and Micheline Kamber: "Data Mining: Concepts and Techniques", Second Edition, Elsevier, pp.285-418, 2016.*