Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/13S134

Concept Aware Related Forum Post Retrieval Framework with Increased User Satisfaction Level

L. Rasikannan[1,*] *and P. Alli*[2]

[1] Alagappa Chettiar Government College of Enginering & Technology, Karaikudi, Tamil Nadu, India.
[2] Department of Computer Science and Engineering, Velammal College of Engineering & Technology, Madurai, Tamil Nadu, India.

Received: 2 Nov. 2018, Revised: 1 Dec. 2018, Accepted: 05 Dec. 2018
Published online: 1 Aug. 2019

Abstract: Nowadays, Forums and blogs are considered as ones of the most valuable resources that give the beneficial information to the online users. However, identifying the related forum as well as blog posts from the masses of existing online contents is a very hard task. With the aim of taking out the most alike forum posts from the online resources, several authors introduced numerous research approaches. In the previous work, content Similarity over Intention-based Segmentation (CS-IBS) algorithm is presented. It splits the documents into several text segments from which the content-associated posts are found. But, in these past studies, preprocessing is not carried out, where the contents taken online might comprises changing noisy words that reduce the accurateness of the retrieval and by reason of not taking the sense of the multiple text segments, the accurateness of the clustering is also decreased. As a result, by presenting the new method, Concept aware related Forum Post Retrieval Framework (CR-FPRF), this issue is considered and solved in the presented research technique. According to this research, primarily forum post segmentation is carried out.

Keywords: Similar text retrieval, Concept aware clustering, Semantic analysis, Preprocessing, Feature analysis

1 Introduction

Information retrieval systems are distinguished by the degree at which they work. It is considered to be beneficial for differentiating three noticeable scales [1,2]. The system should offer look for over billions of documents kept on millions of computers in case of web search [3]. Unique problems want to collect documents for indexing, being capable of implement systems, which function proficiently at this massive scale, in addition dealing with specific features of the web, for instance, for increasing their search engine rankings, the manipulation of hypertext in addition to not being misled by site providers manipulating page content, provided the commercial significance of the web [4]. Next is personal information retrieval [5] . Before some years, consumer operating systems have incorporated information retrieval. Mailing applications offer search as well as text classification: they however offer a spam (junk mail) filter, and usually offer manual or automatic ways for categorizing mail for providing the facility of placing directly into specific folders. And some other unique problems here are, dealing with the extensive type of document categories on a pc, and creating the search system maintenance free as well as adequately lightweight in regard to startup, processing, and disk space usage by which it could execute on single machine deprived of aggravating its owner [6]. Concept aware Related Forum Post Retrieval Framework (CR-FPRF) is presented in this research. According to this research, primarily forum post segmentation is carried out. Subsequently, in order to eliminate the recurrent words, start and stop words from the form associated posts, preprocessing is performed. Later preprocessing, clustering of alike forum posts-based concept similarity is performed by means of Latent semantic analysis technique which is presented for meaning extraction of the terms existing in the forum posts. Similar content retrieval is assured, depending upon meaning taken out from the forum posts. The complete structure of the proposed methodology is provided in the following. Diverse literature studies are considered, which deals with retrieval of personalized information in an efficient manner is provided in the part 2. Comprehensive discussion of the presented system with appropriate examples and description is provided in part 3. Complete assessment of the presented system is examined in the

* Corresponding author e-mail: rassikannanmail2018@gmail.com

java simulation environment and evaluated by means of the analysis graph in part 4. Lastly in part 5, depending upon the analysis outcomes, conclusion of the presented system are provided.

2 Related Work

In this segment, various literature survey are analyzed for carrying out effective personalized information retrieval. Information filtering was presented by the authors in [7] ,for mobile augmented reality for the reason that display info in AR is messed with more info. Information filtering signifies to discard the info that could possibly be exhibited by determining and ordering the info, which is appropriate to a user at a specified stage. Accordingly, info is categorized depending upon the customer's physical context and on their present perceptions and goals. Efficiently handling huge volume of info is one among the hardest issues for AR in [8]. The cluttering junk info makes it hard to identify helpful info and even turn out to be disrupting to interface with real-world processes. Augmented reality is completely hopeless that is deprived of a means to filter out garbage info. So, there exist plentiful approaches for letting certain level of filtration. The author stated that information retrieval is enabled by assistances from human-computer interaction studies as well as agent technology with the aim of identifying how and when to provide the info to the user or how finest to play on behalf of the user in [9]. With the objective of enhancing the retrieval efficiency in the upcoming computing environments in regard to the accurateness of the retrieval and user fulfillment, the main problem is the incorporation of techniques from context-awareness as well as information retrieval. Adaptive Hypermedia (AH) is a key to handle composite and deeply structured info in [10] . The AH system keeps a user profile for every people communicating with it, at the start depending upon the preferences of the user. This method alters the content they see and also the pathways they follow via the content consequently bring up-to-date files as the users go via the data space. The filtering is known as a logical extension as well as fine-tuning of the aura utilized with the object distribution for not displaying the entire unrelated info to users in[11]. For assisting the decision process and to consider the various state of affairs, which is met, the user could choose a filtering mode as per his present mission. The authors provided samples from prototype augmented reality systems, which express certain means that is shown in info, highlighting the automated draught of overlay geometrical animations in [12]. In this technique, they concentrated on sight organization according to the context of the user, particularly user's place indoor and outdoor. On the other hand, they presumed that information filtering is carried out and that the whole thing showed must be presented. Distinctive user-directed relevance feedback methods include user marking documents, which are identified appropriately. [13]. from these documents, keywords are taken out and included to the user's request or utilized to re-weight previous request terms. Sometimes, this process is apparent to the user, and some other cases, it is impervious. The request that progresses is supposed of as a depiction or prototypical of the user's interests within a specific search session [14, 15]. The authors assessed an overabundance of particular search approaches that handle a complete unsolved forum post as a query, taking out forum threads conversing alike issues to solve it. As forum posts encompass numerous in appropriate background info, the task is very difficult compared to conventional document retrieval issue. The discussion threads to be taken out are also somewhat dissimilar from conventional formless text documents. The researchers presented and analyzed two diverse approaches for flattening the language model of a blog data depending upon the thread encompassing the post in[18] .With the aim of exploiting thread structures in diverse means, it explores numerous variants of the two techniques. In addition, it produces an individual tagged test data set for forum post retrieval as well it assesses the presented smoothing approaches by means of this data set. The authors presented two modest means of re-ranking the top n of a primary run in [17], In the primary method known as Credibility-inspired re-ranking, depending upon the credibility-inspired score, it re-ranks the top n of a DNA baseline. In the next method known as combined re-ranking, based on their retrieval score, it measures product of the credibility-inspired score of the top n outcomes, and also depending upon this score, it re-ranks. The authors presented a generative model in [19] .for intensifying queries by means of exterior collections wherein dependencies amid requests, documents, and growth credentials are openly modeled. Diverse samples of our archetypal are conversed and create diverse (in) reliance suppositions. The researchers analyze the issue of identifying relevant blog data to a post at hand in[20]. In contradiction of conventional methods for identifying relevant documents, which carry out content evaluations crosswise the content of the posts all together, it takes every data as a group of divisions, a piece printed with a diverse objective. It supports that the likeness amid two posts must be depending upon the resemblance of their corresponding segments, which are anticipated for the identical goal, which is to say that they express similar goal. The authors implemented and positively trialed a fresh Patient-Reported Information Multidimensional Exploration (PRIME), an automated communal of deep learning as well as machine learning techniques depending upon clustering, classification, association rules and Natural Language Processing (NLP) methods to enforce organization on to this huge body of amorphous text produced by free-flowing considerations in OCSG. Depending upon this organization, personalized patient info is taken out into a multidimensional database for future analysis, reporting and picturing. The authors Introduced an active and

interpretable OHF post classification structure in [20,21] . Precisely, it categorizes sentences into three classes which are as follows: indication, medicine, and surroundings. Every ruling is anticipated into a predictable attribute space encompassing UMLS semantic types, labeled sequential patterns and other heuristic features. For classifying OHF posts, a forest-based model is implemented. An interpretation technique is also designed in which the decision rules are openly taken out to obtain awareness of valuable info in texts.

3 Concept Aware Similar Forum Post Retrieval

Forum post retrieval is considered as one among the most focused research area with augmented online information hunters. However, identifying the related forum as well as blog posts from the masses of existing online contents is a very hard task. Concept aware Related Forum Post Retrieval Framework (CR-FPRF) is introduced in the presented system. According to this research, primarily forum post segmentation is carried out. Subsequently, in order to eliminate the recurrent words, by the start and stop words from the form associated posts, preprocessing is performed. Later preprocessing, clustering of alike forum posts-based concept similarity is performed by means of Latent semantic analysis technique is presented for meaning extraction of the terms exist in the forum posts. Similar content retrieval is assured, depending upon meaning taken out from the forum posts. The complete framework of the presented system is provided in figure 1.

In figure 1, the whole organization of the presented system is provided. The comprehensive description of the presented system is provided as follows.

3.1 Problem definition

The problem to be resolved is along these lines: Consider number of documents D, reference document dq. Here the goal is to find the k relevant documents which is similar to the reference document dq which would most probably of interest to a user, which previously takes dq being of interest. The particular process is denoted as document matching.

3.2 Segmentation of documents

There are $2^{|}d| - 1$ possible segmentations exist for a document d. amongst them, we are fascinated in the one that is more precisely bring into line with the diverse goals of the text. Identifying the accurate segmentation is a difficult task. In a better segmentation, each segment is
(i) Clear and well preprocessed
(ii) Greatly detached from its nearby segments

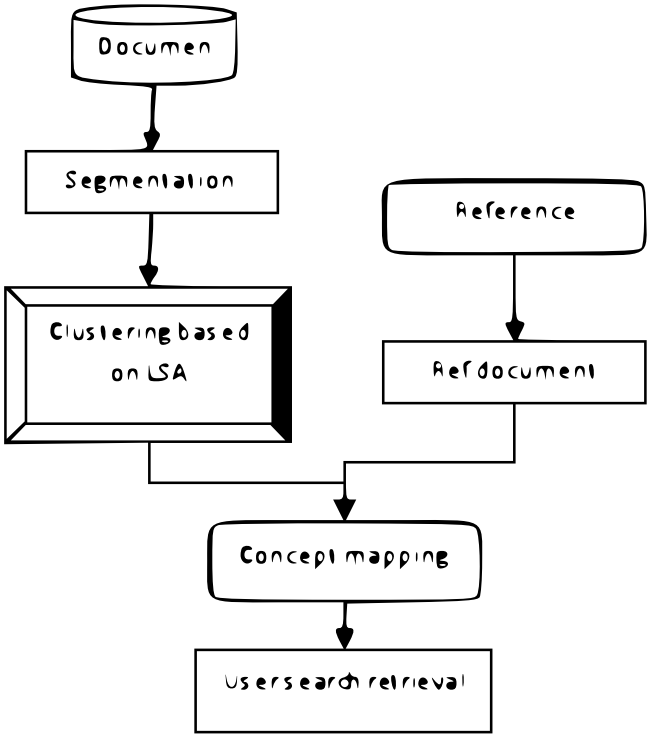

Fig. 1: Overall Flow of Proposed Research Method

As the condition for segmentation is goal-based, the above mentioned two characteristics transform to segmentation in which each segment: (i) express a solitary apparent goal; and (ii) this goal is extremely diverse from those expressed by the nearby segments. Equally, the fore said conditions call for segmentation with in-depth borders. There are three difficulties in identifying a better goal-based segmentation: find out the features to utilize for finding out the goals, gauge the coherence within a segment beside the deepness of the borders of candidate segmentation, and, amongst the candidates, choose the finest segmentation. It is likely that above one segment from the identical document wind up in the identical group, when they contain the identical goal on the other hand are not successive in the document, or because of local optimal values of segment diversity as well as border depth, the border selection technique maintained a border amid them. We create an additional permit over the groups and when those instances are identified, the entire segments, which be in the identical document in a group are integrated into one. Conversely, supposing the grouping C of the segments of a set of documents D, for each group I$\in$ C a novel collection of segments is taken as an alternative, which is built-in this manner: $\{s|\exists d \in D : U_{s'} \in I \wedge s^r \in s^d S'\}$ here the symbol on segments denotes concatenation. Consequently, every document might have as a maximum one segment in every group.

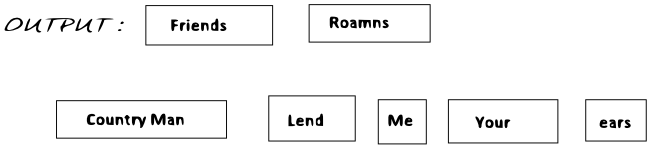

Fig. 2: Output

3.3 Processing documents

The preprocessing of documents procedure is followed as defined in [22]. The entire process is as follows. Subsequently reading the input text documents, texts preprocessing step is carried out, it splits the text document to features that are known as (attributes, tokenization, words, terms), it denotes that text document as a vector space whose elements are that extracted features and their calculated weights that are got by the frequency of every feature identified in the document, subsequently it eliminates the non-informative features for instance (stop words, numbers, and special characters). By means of minimizing them to their root by utilizing the process of stemming, the residual features are subsequently standardized. The dimensionality of the feature space might still be very high, despite the non-informative features elimination as well as the stemming process. So; this research employs particular thresholds for decreasing the feature space size for every input text document dependent upon the frequency of every feature in that text document. The objective of this stage is that it enhances the features' quality and simultaneously decreases the trouble of mining process.
Tokenization: Tokenization doesn't merely isolate strings into fundamental processing units. For producing higher level tokens, it as well interprets and clusters separated tokens. Raw texts are preprocessed as well as segmented into textual units. There are three operations involved in data processing: the primary operation is to transform document to word counts that is equivalent to bag of word (BOW). The next operation is eliminating blank sequence that is to say that this step encompasses cleansing and filtering (for instance stripping extraneous control characters, whitespace collapsing,). Lastly, every input text document is segmented into a collection of features (words,tokens, terms or attributes).
Specified a character series in addition to a defined document unit, tokenization is splitting it up into pieces, known as tokens, possibly all at once discarding some characters, for instance punctuation. Consider the below sample of tokenization:

Input: Friends, Romans, Countrymen, lend me your ears; Output in figure output.
Here tokens can also termed as words or terms, however it is significant to create a type/token distinction. An occurrence of a series of characters in a document, which are assembled as a valuable semantic element for processing is known as a token. A type is known as the class of whole tokens encompassing the similar character series. A term is nothing but a (possibly normalized) kind, which is incorporated in the IR system's dictionary. The collection of index terms is completely discrete from the tokens, for example, they are semantic identifiers in taxonomy, and on the other hand actually in current IR systems, they are powerfully associated with the tokens in the document. Conversely, more willingly than being accurately the tokens, which seem in the document, they are resultant from them by numerous normalization processes.
Stop Words Elimination: A stop words list is known as a list of generally recurrent features that arise in each text document. The general features for instance conjunctions for instance and, or, but and pronouns he, she, it and so on. Should be eliminated because of it doesn't have consequence and these words include an extremely little or no value on the categorization process (that is to say every feature must be eliminated when it matches any feature in the stop words list). So, if the feature is a number or a special character at that time that feature must be eliminated. With the intention of identifying the stop words, we could assemble our list of terms by frequency and take the high recurrent ones as stated by their shortage of semantics value. They must be eliminated from words, could as well eliminate extremely infrequent words, for instance, words that merely happen in m or a smaller amount document, for instance m=6.
Stop words are a "single set of words". It could mean diverse things to diverse applications. E.g., in certain applications eliminating all stop words right from determiners (for instance the, a, an) to prepositions (for instance above, across, before) to certain adjectives (for instance good, nice) could be a suitable stop word list. To certain applications on the other hand, this could be disadvantageous. For example, in case of sentiment analysis eliminating adjective terms for instance 'good' and 'nice' and negations for instance 'not' could throw methods of their tracks. In those cases, one couldselect to utilize a minimal stop list encompassing simply determiners or determiners with prepositions or merely coordinating conjunctions based upon the requirements of the application.
Some instances of minimal stop word lists that you couldutilize:
• Determiners – It is inclined to mark nouns where a determiner typically would be subsequent to a noun examples: a, an, the, another
• Coordinating conjunctions – It link phrases, words, and clauses examples: an, norfor,, but, yet, or, so

• Prepositions – It's how temporal or spatial relations examples: under, in, towards, before
Stemming: It is described as the process of eliminating affixes (prefixes andsuffixes) from features that is to say that the process derived for decreasing modulated words

to their stem. The stem need not to be found to the real morphological root of the word and it is typically adequately associated through words map to the identical stem. This process is utilized to decrease the number of features in the feature space as well as enhance the clustering performance while the diverse forms of features are stemmed into a solitary feature.E.g.: (connect, connects, connected, and connecting) from the stated above instance, the collection of features is conflated into a solitary feature by elimination of the diverse suffixes -s, -ed, -ing to obtain the single feature connect. For identifying the root words in the document, this research employed standard Porter Stemming Algorithm. There are numerous kinds of stemming techniques that vary in terms of accuracy and performance how some stemming hitches are overwhelmed.

A modest stemmer searches the modified form in a lookup table. The benefits of this method are as follows: it is modest, rapid, and simply deals with exceptions. The drawbacks of this method are that each and every inflected forms should be openly listed in the table: novel or unacquainted words are not handled, albeit they are seamlessly regular (for example iPads iPad), and the table might be large. For languages with modest morphology, such as English, table sizes are simple; on the other hand, extremely modulated languages such as Turkish might have hundreds of probable modulated forms for every root. A lookup technique might utilize primary part-of-speech tagging to evade over stemming.

The production technique: The lookup table utilized by a stemmer is usually created semi-automatically. E.g., when the word is ”run”, at that time the inverted technique may automatically produce the forms ”runs”, ”running”, ”runned”, and ”runly”. The final two forms are valid creations, on the other hand they are improbable.

Suffix-stripping algorithms: it doesn't depend upon a lookup table, which encompasses modulated forms and root form associations. As an alternative, a naturally minor list of ”rules” is kept that gives a path for the algorithm, specified an input word form, to identify its root form. Certain samples of the rules comprise:

- When the term ends in 'ed', eliminate the 'ed'
- Whenthe term ends in 'ing', eliminate the 'ing'
- Whenthe term ends in 'ly', eliminate the 'ly'

Suffix stripping methods delight in the advantage of being much humbler to keep than brute force algorithms, supposing the maintainer is adequately familiar in the difficulties of semantics and morphology and encrypting suffix stripping rules. These techniques are considered as crude provided the weak performance while handling exceptional associations (such as 'ran' and 'run'). The solutions created by suffix stripping techniques are restricted to those lexical types that contain renowned suffixes with a small number of exceptions. This, on the other hand, is an issue, as not all parts of speech contain such a well-articulated collection of rules. Lemmatisation tries to enhance upon this trouble.

3.4 Segment weightings

To define the text feature in text categorization, Vector Space Model (VSM) is a distinctive technique. In order to calculate the term weighting in every aspect of the text feature, it accepts TF-IDF weights. On the other hand, it just takes the association among the term and the entire text on the other hand ignore the association amid diverse terms. Focusing on this issue, an enhanced TF-IDF weights function is presented that utilizes the allotment info amongst classes as well as within a class

3.5 Improved TF-IDF Weights

A huge amount of practical as well as hypothetical examination displays that the typical TF-IDF function clearly is insufficiency. TF-IDF takes the text set as a complete, particularly the IDF part of TF-IDF function, it just takes the association amid the IDF features as well as the text count wherein it seems however abandoned distribution of the feature item in a class and in diverse types. In order to resolve the issues of the typical TFIDF Function, with the appropriate knowledge, we study the info distribution in line, At that time this research present the purpose of distribution amongst classes DIDA and the purpose of distribution in a class DIDC, that denotes the dissimilarities of features amid diverse types and in the identical type correspondingly.

(1) Distribution degree amongst classes

$$DI_{DA}(t_k) = \frac{\sqrt{\sum_{i=1}^{m}(tf_i(t_k)) - \overline{(tf(t_k))}^2/(m-1)}}{\overline{(tf_i(t_k))}} \tag{1}$$

Here, the frequency of feature tk in category I is denoted by $tf_i(t_k)$ the average frequency of feature t_k in all groups is denoted by ,$\overline{(tf_i(t_k))}$the number of types is denoted by m and $\overline{(tf_i(t_k))}$=$1/m\sum_{i=1}^{m}(tf_i(t_k))$

(2) Distribution degree in a class

$$DI_{DC}(t_k) = \frac{\sqrt{\sum_{j=1}^{n}(tf_j(t_k)) - \overline{(tf'(t_k))}^2/(n-1)}}{\overline{(tf'(t_k))}} \tag{2}$$

Here$tf_j(t_k)$is known as the frequency of feature t_k in text j, $\overline{(tf'(t_k))}$ is called the average frequency of feature t_k in all texts, the number of texts in a class is denoted by n and $\overline{(tf'(t_k))}$=$1/n\sum_{i=1}^{m}(tf_j(t_k))$

In keeping with the purpose of allocation degree amongst classes, we understood that while some features merely look in a type, it contains its allocation degree DI_{DA} = 1 no classification capability, when the word frequency of some feature contains the similar value in every group. According to the above examination, we could prove that the features giving out degree amongst

classes DI_{DA} is relative to the categorization capability. According to the function of allocation degree in a class, when a characteristic merely looks in one text of some type its allocation degree DI_{DC} = 1 , in keeping with that we could predict that this text is probably a distinct instance of this type and the feature contains the poor classification capability. It contains its distribution degree DI_{DC} = 0 as well as solidest classification capability, when the word occurrence of some feature contains the similar value in every kind. So we could acquire that the features allocation degree in a class DI_{DA} is contrariwise relative to the categorization capability. According to our research, uniting the allocation degree amongst classes and in a class, the novel improved IF - IDF weights function is stated in this manner:

$$W(t_{i,k}) = f * min(0,1) \tag{3}$$

Here f is known as the task value calculated by traditional TF-IDF weights function, the purpose of allocation amongst classes is denoted by DI_{DA} and the purpose of allocation in a class is defined by DI_{DC}

3.6 Clustering based on latent semantic analysis

The clustering technique utilized in this research is dependent upon the usage of LSA and it encompasses 3 stages. A term-document matrix is built in the primary stage, and utilizing LSA decayed to a concept space. Subsequently, the dimensionality of the perception space is condensed, then in the third phase, hierarchical clustering is carried out.

3.7 Decomposition of term document matrix

The decomposition of term document matrix of the document is adopted as defined in [23]. In the beginning, the whole input credentials are preprocessed as well as lemmatized. The resultant text doesn't simply encompass lemmas, on the other hand as well word forms, for instance amounts or typing errors, that couldn't be lemmatized. These items are known as terms. Provided the frequency of incidence, that is to say that the term frequency (TF), is computed for each distinctive term from these credentials, eliminating words in the stop list, the preprocessed set of input documents. Next, the occurrence of every term is weighted by its inverse article frequency (IDF).These IDF values are computed in this manner:

$$IDF(1) = log \frac{|D_b|}{|\{d_b \in D_b : l \in d_b\}|} \tag{4}$$

Here$|D_b|$ is known as the total amount of training in credentials in the surroundings corpus$|\{d_b \in D_b : l \in d_b\}|$ is known as the amount of background documents encompassing the term l. The term document matrix A is built for the specified weighted term frequency values; here weighted term frequency of each vector of the document is denoted by each column. So, here, the amount of each and every distinctive terms in all input documents is denoted by t,the size of A is t d and the total amount of input documents is denoted as d. Next, the LSA is carried out. This technique applies the Singular Value Decomposition (SVD) to the matrix A like this:

$$A = U \Sigma V^T \tag{5}$$

Here t * m column-orthonormal matrix of left singular vectors is depicted by U, an m * m diagonal matrix is denoted by $\in= diag(\alpha_1, \alpha_2, \alpha_3, ..., \alpha_m)$ whose diagonal elements signify non-negative singular values kept in downward order, and an m*d orthonormality matrix of right singular vectors is denoted by V. It is exposed that the matrices U,Σ and V^T signify a concept (semantic) space of the input documents: the matrix U defines a mapping of concepts to the space of terms, V^T seizes how notions are plotted to documents, and the singular values of Σ signify the importance of individual notions.

3.8 Concept space dimension reduction

Actually, since individual input documents encompass a) synonyms or b) partly or entirely dissimilar words as well as word forms, the term-document matrix A is sparse. The later issue happens mainly for inflective languages for instance Czech. The subsequent problem is that A comprises noise, which is denoted by general and/or meaningless terms. These problems could partly be eradicated by means of:
i) stop list of meaningless terms
ii) processing module for text normalization and lemmatization
iii) minimum occurrence threshold for terms
In this research, we make use of the primary two choices for condensing the sparsity of A. From the arithmetical standpoint, this issue is stated by low-rank estimate, which decreases the amount of dimensions of the concept space from m to k. regrettably, the issue of identifying the appropriate value of k is not insignificant and its solution is dependent upon a heuristic knowledge of the provided task.

3.9 Hierarchical clustering

It is dependent upon the supposition that the documents be in the identical cluster must contain related notions (topics). So, subsequent to dimensionality reduction, clustering is carried out for vectors of A_K or for vectors of

reduced matrix V^T that defines mapping of ideas to documents. We carry out clustering in an agglomerative manner, in which every document signifies one cluster at the start. At that time, duos of the most alike clusters are combined in successive steps till the required amount of groups is attained. The result of clustering is a group of clusters in which each cluster encompasses a document list, which must contain alike concept (topic).

3.10 Retrieval of top most matching forum post

The retrieval of top most matching forum post is adapted and refined as defined in [19]. The process is explained below. With the aim of carrying out the document matching, that is to say, to recognize the documents in a set that are associated with a reference document d_q, one method is to consider the document d_q as a request and after that processing of the likeness of every other document d_0 to that request in a means identical to how IR methods perform. Like previously stated, our position is that those tasks must not take every document as a complete on the other hand must be particular on every goal separately, and after that merge the outcomes.

3.11 Single Intention Matching

Every cluster is the group of each document on the particular goal that the cluster signifies. Therefore, it is sufficient to gauge the similarity of the corresponding segment s_0 of d_0 in the cluster I, to the corresponding segment sq of d_q in that similar cluster, to gauge the similarity of a document d_0 to the reference document d_q regarding a particular intention I. In order to calculate this likeness any text comparison, for instance, language models, paraphrasing, or IR methods might be applied. TF/IDF is a well-known IR method. The basic of the real TF/IDF technique and its probabilistic variance encompasses a term weighting technique that calculates the weight of a term in a document regarding the amount of its presences in association to the number of its presences in all the other documents.

3.12 Complete Intentions Matching

The top-n lists produced crosswise the diverse goals, that is to say, the set M stated previously, are utilized for producing the k most associated documents to the corresponding document referenced (d_q). A novel list R is produced that encompasses each document, which seems as a minimum in one among the lists in M. A score is related to every such document that is the summation of the scores and this document seems in the numerous lists in M. The k elements in R with the maximum score are considered as response to the query of the corresponding documents to the reference document d_q

4 Experimental Results

For the provided data set encompassing diverse forum posts from numerous areas such as social, politics, games and so on, the experimentation of the presented system is implemented in the java simulation environment with numerous topics. This assessment is performed amid presented technique called Concept aware Related Forum Post Retrieval Framework (CR-FPRF) and the previous techniques that is to say Content Similarity over Intention based Segmentation (CS-IBS) algorithm. Performance assessment of the presented technique is carried out dependent upon Precision, Recall, and Accuracy and F-measure parameters. Precision is calculated by this formula

$$Precision = \frac{f_p}{t_p + f_n} \tag{6}$$

Recall is calculated by this formula

$$Recall = \frac{t_p}{t_p + f_n} \tag{7}$$

Accuracy is computed by this formula

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{8}$$

F-Measure is calculated by this formula

$$F = 2.\frac{Precision.Recall}{Precision + Recall} \tag{9}$$

Here t_p – True Positive (Correct result), f_p - False Negative, t_n – True Negative (Correct absence of result), f_n – False Negative (Missing result). The simulation outcomes for the assessment of the presented technique in contradiction of numerous performance measures for instance Precision, Recall and Accuracy. This technique yields superior accuracy than the previous method. Accuracy is the rightness of the presented technique to repossess the exact documents that fine fulfill the user requisite professionally. The accuracy is computed for the assessment of presented technique CR-FPRF and previous technique CS-IBS is depicted in Figure 2.

The accuracy evaluation outcomes are depicted in Figure 2. In which CR-FPRF gives 89.28% accuracy when compared to the previous technique CS-IBS with the increased percentage of 2-6% for accuracy parameter. The precision evaluation outcome was presented in Figure 3. In which, presented CR-FPRF gives 89.28% precision that carries out superior with improved percentage of 2-4% compared to the previous research approaches. The recall evaluation outcome was depicted in Figure 4. Here, the presented CR-FPRF gives 89.36% recall value that carries out superior with improved percentage of 2-4% for recall parameter. The F-Measure evaluation outcome was presented in Figure 5. Here the presented CR-FPRF

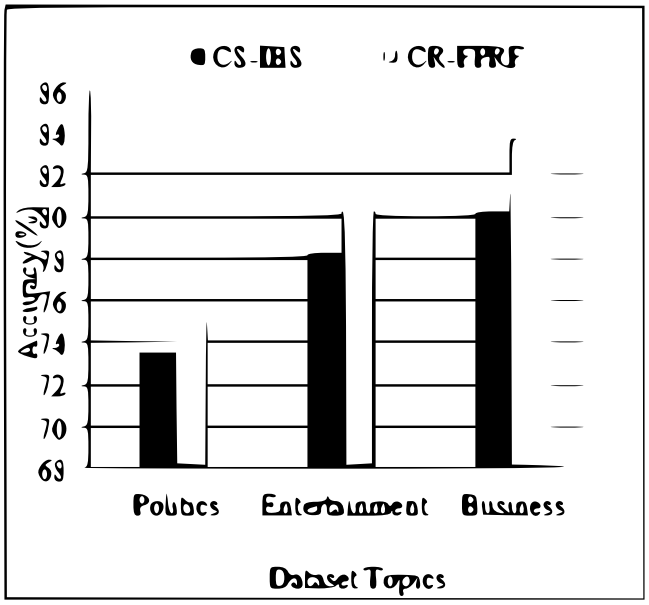

Fig. 3: Accuracy comparison

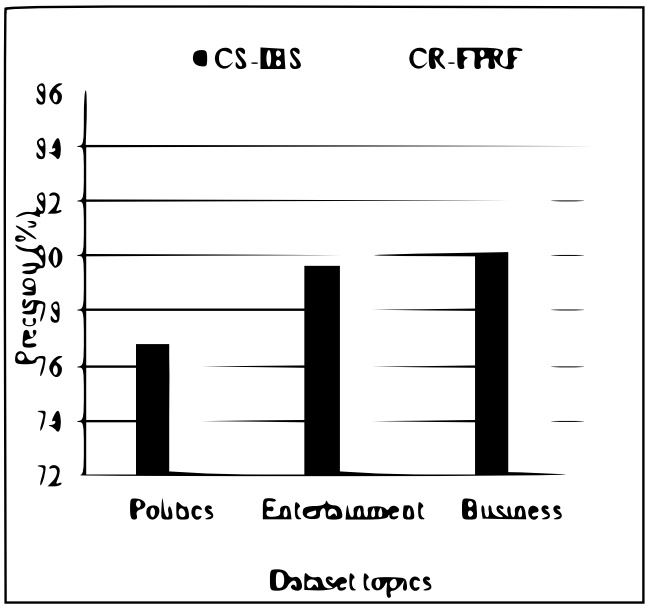

Fig. 4: Precision comparison

depicts 84% performance enhancement when matched up with the previous research techniques in regard to improved F-Measure value

5 Conclusion

In this proposed work, Forum post similar content retrieval is concentrated for enhancing the user fulfillment

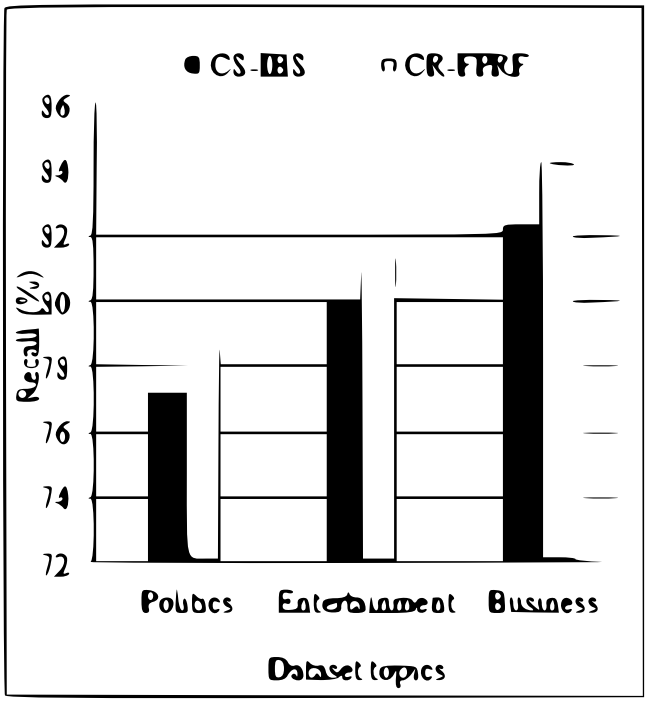

Fig. 5: Recall comparison

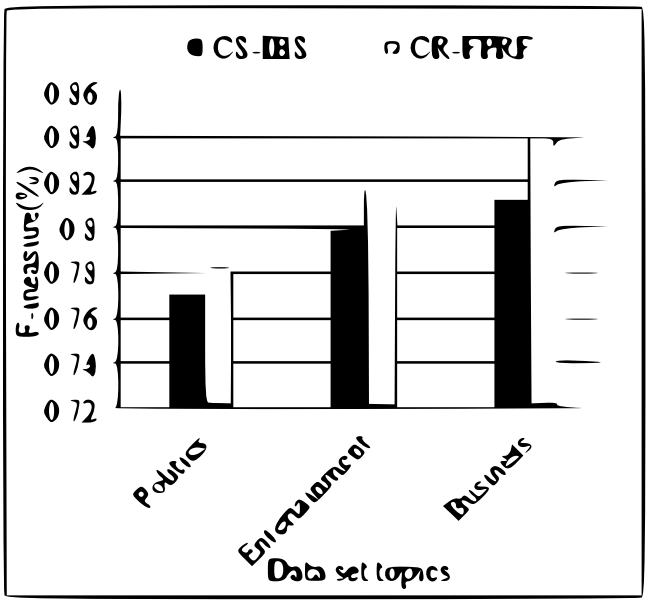

Fig. 6: F-measure comparison of similarity measures

level. This is attained in the presented technique by presenting, Concept aware Related Forum Post Retrieval Framework (CR-FPRF). According to this research, primarily forum post segmentation is carried out. Subsequently, in order to eliminate the recurrent words, start and stop words from the form associated posts,

preprocessing is performed. Later preprocessing, clustering of similar forum posts-based concept similarity is performed by means of latent semantic analysis technique is presented for meaning extraction of the terms existing in the forum posts. Similar content retrieval is assured, depending upon meaning taken out from the forum posts. The complete framework is implemented in the java simulation environment and it is confirmed that the presented system does superior compared to the previous methods.

References

[1] Manning, C. D., Raghavan, P., Schutze, H. (2008). Introduction to information retrieval (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press.

[2] Korfhage, R. R. (2008). Information storage and retrieval.

[3] Mikkola, M., Lempiö, J., Santti, V. (2003). U.S. Patent No. 6,529,143. Washington, DC: U.S. Patent and Trademark Office.

[4] Levy, M. (2003). U.S. Patent No. 6,556,997. Washington, DC: U.S. Patent and Trademark Office.

[5] Bueno, D., David, A. A. (2001, July). METIORE: A personalized information retrieval system. In International Conference on User Modeling (pp. 168-177). Springer Berlin Heidelberg.

[6] Sugiyama, K., Hatano, K., Yoshikawa, M. (2004, May). Adaptive web search based on user profile constructed without any effort from users. In Proceedings of the 13th international conference on World Wide Web (pp. 675-684). ACM.

[7] S.Julier, M.Lanzagorta, Y.Baillot, L.Rosenblum, S.Feiner, T.Höllerer, and S.Sestito, "Information filtering for mobile augmented reality". In Proc. ISAR '00 (Int.Symposium on Aug-mented Reality), pages 3-11, Munich,Germany, October 5-6 2000.

[8] Seth Insley, "Obstacle to General Purpose Augmented Reality," http://islab.oregonstate.edu/koc/ece399/f03/final/insley2.pdf.

[9] jones,Brown, "Context-aware retrieval for ubiquitous computing environments," Invited paper in Mobile and ubiquitous information access, Springer Lecture Notes in Computer Science, Vol. 2954, pp. 227-243, 2004.

[10] Patrick A.S.Sinclair, Kirk Martinez, David E. Millard, and Mark J. Weal, "Augmented Reality as an Interface to Adaptive Hypermedia Systems", In New Review of Hypermedia and Multimedia, Special Issue on Hypermedia beyond the Desktop. Vol. 9, pp.117-136, 2003.

[11] SimonJulier, Yohan Baillot, Marco Lanzagorta, Dennis Brown, Lawrence Rosenblum, "BARS: Battlefield Augmented Reality System", NATO Symposium on Information Processing Techniques for Military Systems, 9-11 October 2000, Istanbul, Turkey.

[12] Bell, S. Feiner, and T. Hollerer, "Information at a Glance," IEEE Computer Graphics and Applications, Vol. 22, No. 4, pp. 6-9, July/August 2002.

[13] Croft, Cronen-Townsend, S.,and Lavrenko, V. (2001, June). Relevance Feedback and Personalization: A Language Modeling Perspective. In DELOS Workshop: Personalization and Recommender Systems in Digital Libraries (Vol. 3, p. 13).

[14] Ruthven, I., Lalmas, M., and Van Rijsbergen, K. (2002). Combining and selecting characteristics of information use. Journal of the American Society for Information Science and Technology, 53(5), 378-396

[15] Cho, J. H., Sondhi, P., Zhai, C., and Schatz, B. R. (2014, September). Resolving healthcare forum posts via similar thread retrieval. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics(pp. 33-42). ACM.

[16] Duan, H., and Zhai, C. (2011). Exploiting thread structures to improve smoothing of language models for forum post retrieval. Advances in Information Retrieval, 350-361.

[17] Weerkamp and de Rijke, M. (2012). Credibility-inspired ranking for blog post retrieval. Information retrieval, 15(3-4), 243-277.

[18] Weerkamp, W., Balog, K., and de Rijke, M. (2009, August). A generative blog post retrieval model that uses query expansion based on external collections. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 1057-1065). Association for Computational Linguistics.

[19] Papadimitriou, D., Koutrika, G., Velegrakis, Y., and Mylopoulos, J. (2017). Finding Related Forum Posts through Content Similarity over Intention-based Segmentation. IEEE Transactions on Knowledge and Data Engineering.

[20] Ranasinghe, W., Bandaragoda, T., De Silva, D., and Alahakoon, D. (2017). A novel framework for automated, intelligent extraction and analysis of online support group discussions for cancer related outcomes. Bju International, 120(S3), 59-61.

[21] Gao, J., Liu, N., Lawley, M., and Hu, X. (2017). An interpretable classification framework for information extraction from online healthcare forums. Journal of healthcare engineering, 2017.

[22] A. I. Kadhim, Y. N. Cheah and N. H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering," 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, Kota Kinabalu, 2014, pp. 69-73

[23] M. Rott and P. Cerva, "Investigation of Latent Semantic Analysis for Clustering of Czech News Articles," 2014 25th International Workshop on Database and Expert Systems Applications, Munich, 2014, pp. 223-227

L. Rasikannan L.Rasikannan, received B.E degree from Thiagarajar College of Engineering, Madurai, Tamilnadu, India and M.E. degree from Anna University, Coimbatore Tamilnadu, India. He is pursuing PhD degree in ICE at Anna University, Chennai. His current research interests include handling Search Engine Issues, information retrieval, machine learning and Sentiment Analysis.

P. Alli received the B.E degree in Electronics and Communication and Engineering from Madras University and M.S. degree in Birla Institute of Technology and Science and Ph.D. degree in Image Processing from Madurai Kamaraj University. She has authored more than 22 publications in journals and 14 publications in conferences. Her research interests include Image Processing, Networking, and Information Security.