International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Architecture for Presence Factor-Oriented Blog Summarization

Rinki[1], Aakash Gupta[2]
[1] M.Tech, Scholar, [2]Assistant Professor
[1,2]Department of Computer Science & Engineering,
[1,2]Gateway Institute of Engineering & Technology (GIET), Sonepat

Abstract:-Numerous approaches for identifying important content for automatic text summarization have been developed to date. Topic representation approaches first derive an intermediate representation of the text that captures the topics discussed in the input. In the summarization through online summarizer tools, we find that only a set of representative sentences could be drawn out. Title and comments are important part of a blog post. It has been found out that the blog post contains the content related to the title of the blog post and the comments which helps the user to know the quality of that blog post. In this paper, we compare "Presence Factor-Oriented Blog Summarization" and "Novel Technique for Relevant Content Extraction from Blog Pages" and find out the best.

Index Terms:-Blog, Indexing, Search Engines, WWW, Crawler, Summarizer Tools.

I. INTRODUCTION

WWW is a hub of information from which we can find necessary information about a topic. Search Engines have made the job easy to find the information required. To search data on WWW search engines uses a special technique called crawler. Crawlers crawls the whole web by matching the keywords and download all the required information. After that, indexing comes into play and performs its function with the help of indexer. Indexer is a tool which is used to map the keywords and provides all the relevant links on the basis of their preferences.

The richest source of information i.e. Blogosphere is a part of WWW which consists of weblogs & blogs. Blogs are the user's personal journals managed either by a single person or a group of persons on a variety of topics. In this way, it allows the users to share their opinion and feeling. On a particular blog there could be several links which may be interrelated to other web pages. The blog follows the same format throughout. These web pages can be updated regularly and stored accordingly. Some of the blogs have their dates of modification which could be accessed by anyone. But, for some other blogs the user has to log in to access the relevant details.

Blogs are of different types. It could be a personal blog, current affairs and comments. Blog is a medium by which we can share our ideas with distinct users and helps us in getting the views of the other users. A Blog can be written on any topic i.e. Entertainment, Games, Music, Movie, Education, Health, Agriculture etc. As we all know that in today's world most of the things are dependent on internet by which we can search anything without leaving the comfort of the chair. Due to the huge data available on the WWW, the search engines are needed to summarize the blog pages so that only relevant information can be stored before indexing. Therefore, in this paper we will know how to extract the relevant data from the blog pages by using the title and valid comments.

It has been organized in five parts as given below:

Introduction, Description of the latest research, Explaination of the proposed blog, summarization system using both of these techniques, Description of the experimental work and the comparison of performance of the proposed system with other researches, Conclusion.

II. LITERATURE REVIEW

As per the growing demand and usage of internet; it has become the important area of research. Some of the researches in this field are as follows:-

Xiaodan Song et. al. in [7] proposed a system that helps us in finding the most prominent blogs. In this, Blog networks are found where nodes represent the blogs and edges represents the links among different blogs. Then, it uses an algorithm named eRank algorithm which is used to rank the blogs according to their priorities.

Another research was done by ShamimaMithunet. Al in [8], have targeted to resolve specifically Question Irrelevancy and Discourse Incoherency problems which have been found to be the most frequently occurring problem in opinion summarization. To overcome from these problems a hybrid approach has been used by combining text schema and rhetorical relations to exploit intra-sentential rhetorical relations.

In [9], Beaux Sharifi, done a research which gives the summary by taking a phrase either trending phrase or any other phrase specified by a user which collects all the posts containing these phrases and created summary of the posts related to that phrase.

Shuang Sun et. Al. [10], worked on it and give it in the form of sentence extraction and sentence ranking problem. In this method, he included three things i.e. important sentences, blog tags and blog comments. Furthermore, calculated the salience scores for these words. After assigning scores; ASS method is used to select sentences based on these salience scores.

In [11], Aixin Sun et. Al., in this, he extracted those sentences which are best discussed among its comments. In this proposed system, firstly representative words are derived from blog posts and then finds sentences containing these words.

The following gaps in the area of blog summarization are found in available literature:

In [7], it uses blog networks which are very complex in nature because it is very difficult to find blog networks which is time consuming.

The study of Aixin Sun et. Al. [11], it only includes the comments for blog summarization and left another important features of blog pages, while all the comments in the blog may or may not be valid. Hence it will effect the performance of Summarization system.

Therefore, in this paper we compare “Presence Factor-Oriented Blog Summarization” and “Novel Technique for Relevant Content Extraction from Blog Pages” techniques and find out which one is better.

Comparison between “Novel Technique for Relevant Content Extraction from Blog Pages” and “Presence Factor-Oriented Blog Summarization”

As we have discussed earlier, blogs are the user’s personal journals written either by a single person or a group of persons on a variety of topics through which they can share their opinions and feelings. It has following main parts:-

Blog Title

Blog Post

Visitor’s valid comments

According to “Novel Technique for Relevant Content Extraction from Blog Pages” :

Each blog contains a title related to that blog content which contains the opinions or emotions related to that post and visitor’s comments. But the current research in the area of blog summarization strongly ignored the role of title and visitor’s comments. Normally, a visitor starts reading any blog after reading its title & comments and after reading it he/she leaves his/her opinion about the post in the comment section. As, both the blog title and visitor’s comments are the inbuilt part of any blog page and play an important role in finding out the summary of a blog page, therefore, in this study, the problem of blog summarization has been tackled by using both of these i.e. the title and the valid comments. It may also used in many fields such as:-

Blog search, blog presentation, reader feedback, marketing research and others.

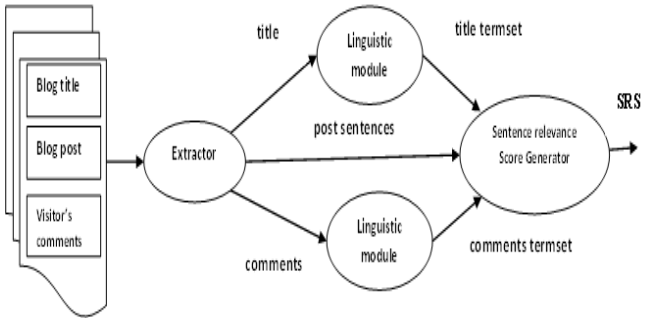

Fig.1 Proposed system

Extractor: Extractor module extracts the title, post sentences and the visitor's valid comments from the blog page.
It is also responsible for separating the sentences and comments that exists in the blog post and creates a set of sentences (S) & Comments (C).

S= {S1, S2, S3,…..,Sn}

C= {C1, C2, C3, …..Cn}

Where, S is the set of all sentences and C is the set of all comments in a blog post.

A. *Linguistic Module:* Linguistic Module performs the following functions-
B. *Normalization:* This process is used to normalize the sentences.
C. *Stopword Removal:* It is used to remove the stopwords like- of, for, in, a, the etc. from each sentence.
D. *Lemmatization/Stemming:* Lemmatization is the process which is used to perform Lemmatization. For eg:- Replacing cars to car etc.

E. *Stemming process* uses Porter's Stemmer. For eg:reducing natural to nature.

The linguistic module will take the title, post sentences and comments as inputs and then generates the title termset and comment termset as given:-

F. *Title Termset:* Title termset can be represented as T. It contains all the terms included / present in the title.
T= {t1, t2, t3, …..,tn}

Where, tn is the nth term in the title termset.

G. *Comment Termset:* It is represented as CiT and contains all the terms present in the comments and title.
C= {c1, c2,c3, ……cn}

Comment Termset can be written as:-

C1T= (c1t1, c1t2, c1t3, …..c1tn)

CiT= (cit1, cit2, cit3, …..citn)

H. *Sentence Relevance Score Generator:* SRS generator takes three types of termsets i.e. title termset, post sentences and comments termset as its input and then computes the sentence relevance score for each sentence.

Firstly, SRS generator consider the title termset and sentences and then generates a matrix in which each column consists of a sentence (Si) and each row consists of a term from the title termset. And each entry in the matrix is represented by TSM[i, j] i.e. the

frequency of term Tiq blog title in the sentence Si of the blog post.

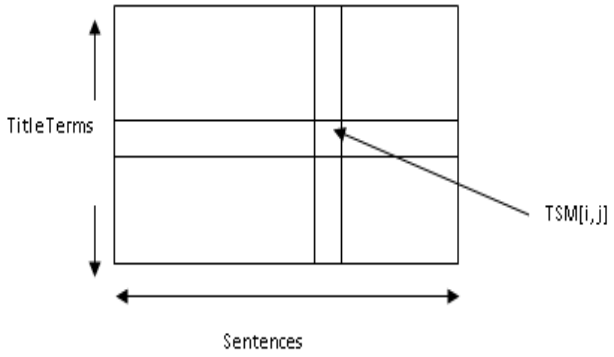

Fig.2 Title- Sentence Matrix

In second part, SRS generator consider all the sentences and the termset of each comment and generates matrices in which a row consist of term (C (k, i)) i.e. the ith term of kth comment and a column consist of the sentences. Hence, the entry in the matrix can be done as:-

CSM[C (k, i), j] …..... (i)

i.e. the frequency of the ith term of kth comment in the jth sentence of the blog post. In this way, the number of matrices created is equal to the number of comments in the comment section i.e. for each comment Ck, a separate matrix is generated which stores the frequency of each term of the comment in each sentence of the blog post. For eg:- if there are three comments in the comment section then three CSM will be generated.

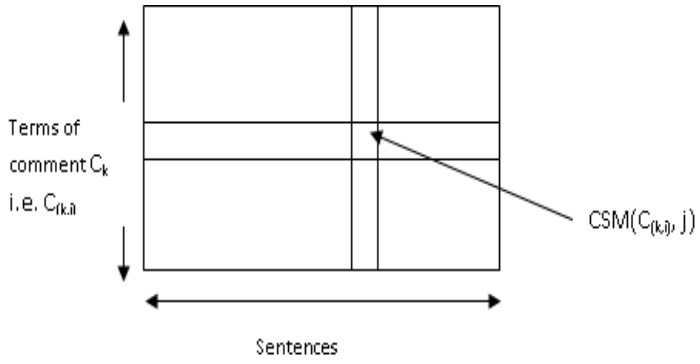

Fig. 3 Comment Sentence Matrix

By this method, a single TSM and multiple CSM's will be generated and after that the SRS generator computes the SRS by the given formula:-

$$RSS(Sj) = \alpha.(\Sigma i=1nTSM(i, j)) + \beta.(\Sigma i=1nCSM(C(k,i), j))$$

Where, RSS (Sj) is the Relevance Sentence Score for jth sentence. TSM (i, j) is the frequency of the ith term of blog title in the jth sentence. CSM (C (k, i), j) is the frequency of ith term of kth comment in the comment section in the jth sentence andα, β arethe weights assigned to title terms and comments terms respectively.

This System works as follows:-

Step 1. The Extractor module extracts the title, post sentences and visitor's comments from the blog page and creates two sets named as:-

Title termset and Comment termset.

Step 2. Then these title and comments are taken as inputs by Linguistic module that performs lemmatization, stemming, stopword removal and normalization and generates title termset and comment termset.

Step 3. In third step, SRS generator assigns the sentence relevance score to each sentences and sorts them according to their relevance score.

Step 4.Then these sentences are arranged according to their ranks and finds the blog summary. The top K sentences are selected as the blog summary. The next section describes the experimental evaluation that justifies the proposed mechanism.

According to "Presence Factor-Oriented Blog Summarization" :

Each blog page contain title related to the blog post, blog post mainly contains the opinion or emotions related to the title and visitor's comments are the comments given by blog visitors. Since, blog title is integral part of a blog page and can play important role in finding out summary of a blog page, therefore, in this research, the problem of blog summarization has been tackled by using the blog title. This work can be used in many areas such as blog search, blog presentation, reader feedback, marketing research and others.

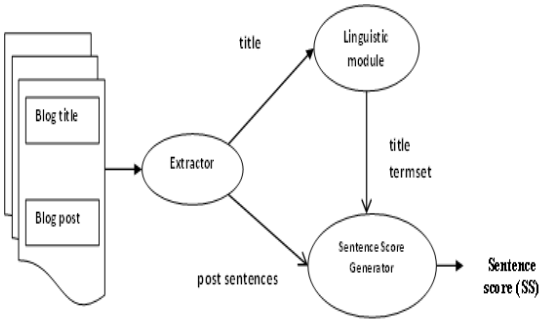

Fig. 4 Proposed System

Extractor:- In order to find out summary of a blog page, Extractor module extracts title from the blog page and post sentences that have been given by the visitors on the blog page.

This module is responsible for separating the sentences that exists in the blog post and generates a set of sentences S.

S = {S1, S2, S3…….Sn}

where, S is the set of all sentences in a blog post and the sentences are referred to as S1, S2, S3…….Sn res. in the order in which they appear.

Linguistic Module:-Linguistic Module performs the following functions:-

" Normalization: This process is used to normalize the sentences.

" Stopword Removal: It is used to remove the stopwords like- of, for, in, a, the etc. from each sentence.

" Lemmatization/ Stemming:- Lemmatization is the process which is used to perform Lemmatization. For eg:- Replacing cars to car

etc.

Stemming process uses Porter's Stemmer. For eg:- reducing natural to nature.

The linguistic module will take the title and post sentences as inputs and then generates the title termset containing all the terms in the title and it is represented by T.

$$T = \{t1, t2, t3,.....tn\}$$

whereti is the ith term in the title termset T.

" Sentence Score Generator: This module takes title termset and post sentences as its input and computes the sentence score (SS) for each sentence. At first, SS generator considers the title termset and sentences and generates a matrix in which each row consists of a term from the title termset and each column consist of a sentence (Si). Each entry in the matrix is represented by TSM[i, j] i.e. the frequency of term Ti of blog title in the sentence Si of the blog post.

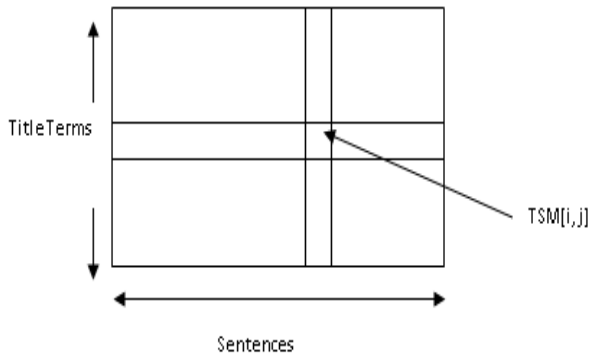

Fig.5 TSM[i,j] i.e. Title-Sentence Matrix

Along with TSM [i, j], a matrix called Presence Factor Matrix PFM [i, j] is also maintained that contains title terms as rows and sentences as the column. Each entry in the matrix represented by PFM [i, j] contains presence of term (in the title) in a sentence (Si). The presence of each term in the sentence Si is represented by '1' and the absence of each term in the sentence Si is represented by '0'.

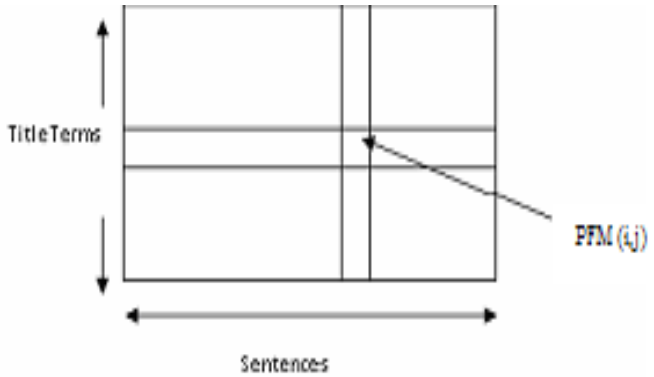

Fig.6 PFM[i,j] i.e. Presence Factor Matrix

After generation of TSM and PFM, the SS Generator computes the Sentence Score by using the following formula:-

$SS(Sj) = \Sigma i=1\ nTSM(i, j) * PFM(i, j)$ ..... (ii)

Where, *SS (Sj)* is Sentence Score for jth sentence, *TSM(i, j)* is the frequency of ith term of blog title in jth sentence, *PFM(i, j)* is the presence of ith term of blog title in jth sentence. The presence is represented by '1' and absence is represented by '0'.

Maintenance of the Presence factor matrix is a key feature used in the approach. This feature ensures that the proposed technique assigns greater score to those sentences, for summary that consists of each term present in the title termset. Also, a higher score is assigned to those sentences in which more number of terms existing in the title termset is present. Keeping this in mind, SS is computed for all the sentences in set S and thereafter, the sentences are ranked according to their Sentence Score. The top k sentences are selected as the Blog Summary. The next section discusses the experimental evaluation that justifies the proposed mechanism.

This System works as follows:-

Step 1.The Extractor module extracts the title, and post sentences from the blog page and creates a set named as sentence set.

Step 2.Then the title is given as input to Linguistic module to performs various functions like:- lemmatization, stemming, stopword removal and normalization and generates title termset.

Step 3.In the third step, Sentence Score generator assigns the sentence score to each sentences and sorts them according to their scores.

Step 4.Then these sentences are arranged according to their ranks and used to find the blog summary. The top K sentences are selected as the blog summary.

Similarity:-Both of these techniques are used to find the summary of the blog post.

Differences:-The Presence Factor-Oriented Blog Summarization technique gives the summary of the blog post by using the title of that blog post. It uses a Presence Factor to match the keywords of the title with post sentences which indicates the presence of each term of the title in each sentence of the blog post. It creates a title termset and generates a matrix using the title termset and post sentences named as Title-Sentence matrix. Along with TSM[i, j], a matrix called Presence Factor matrix PFM [i, j] is also maitained. The presence of each term is represented by '1' and absence of each term is represented by '0'. This feature ensures that the proposed technique assigns greater score to those sentences, for summary that consists of each term present in the title termset. After generating TSM and PSM; SS Generator calculates the Sentence Score for each sentence and then arranges according to their ranks and top k sentences are selected as the Blog summary.
The Novel Technique for Relevant Content Extraction from Blog Pages gives the summary by using the title and valid comments of the blog post. It creates two termsets i.e. Title termset and Comment termset. In this, Sentence Relevance Score generator takes three types of termsets as its input and calculates SRS for each sentence. It also generates two matrices i.e. Title-Sentence matrix and Comment- Sentence matrix. In Title-Sentence matrix, it takes title termset and post sentences as its input and creates the matrix i.e. TSM [i, j]. In Comment- Sentence matrix, it takes comment termset and post sentences as inputs and generates a matrix i.e. CSM [i, j]. It generates a single TSM and multiple CSMs after that SRS is computed for each sentence and then these sentences are arranged according to their ranks and top k sentences are selected as the Blog summary.

III. CONCLUSION

Hence, through this study we found out that Presence Factor-Oriented Blog Summarization approach is better because of its less complexity and consumes less memory. It uses only title termset and compares the post sentences with the title termset which consumes less time as compared with the Novel technique for content extraction from Blog pages.

REFERENCES

[1] T.A. Brooks, "Web Search: How the Web has changed infor-mation retrieval", Information Research, Vol. 8 No. 3, April 2003.
[2] S. Brin, and L. Page, "The anatomy of a large-scale hypertextu-al Web search engine", Journal: Computer Networks and ISDN Systems Archive, Vol. 30 Issue 1-7, April 1, 1998, Pp 107-117 (Proceedings of the Seventh International Conference on World Wide Web 7 (WWW7)).
[3] D. Sharma, A.K. Sharma, and K,K. Bhatia, "Search engines: a comparative review", Proc. of NGCIS, 2007.
[4] M. Burner, "Crawling towards Eternity: Building an archive of the World Wide Web", Web Techniques Magazine, May 1997.
[5] F. M. Facca, and P. L. Lanzi, "Mining interesting knowledge from weblogs: a survey", Journal: Data & Knowledge Engineering archive, Vol. 53 Issue 3, June 2005 Pp 225 – 241.
[6] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace", Proceedings of the 12th International Conference on World Wide Web, New York, NY, USA, 2003. ACM Press.
[7] X. Song, Y. Chi, K. Hino, and B. L. Tseng," Summarization System by Identifying Influential Blogs", ICWSM, 2007(Conference Paper: ICWSM 2007), March 26-28, 2007, Boulder, Colorado, U.S.A.
[8] S. Mithun, and L. Kosseim, "Discourse Structures to Reduce Discourse Incoherence in Blog Summarization", Conference Paper: In proceeding of: Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria.
[9] B. Sharifi, M.A. Hutton, and J. Kalita," Summarizing microblogs automatically", 2010, Proceeding HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Pp 685-688.
[10] S. Sun, "A New Approach to Blog Post Summarization Using Fast Features", Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008.
[11] C.M. Hu, A. Sun, and E. Lim, C.M. Hu, A. Sun, and E. Lim, "Comments-Oriented Blog Summarization by Sentence Extraction", CIKM, 2007, Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Pp 901-904.
[12] R. Madaan, A.K. Sharma & A. Dixit "Presence Factor-Oriented Blog Summarization" , International Journal of Advances in Computing & Information Technology (IJACIT), ISSN: 2277-9140 ,Vol. 2 Issue-2, May 2013. (The paper is online on DBLP).
[13] "A Novel Technique for Relevant Content Extraction from Blog Pages" by Rosy Madaan, A.K.Sharma&Ashutosh Dixit, 2013, International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013; ISSN 2229-5518.