Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Sparse Matrix Representation for Web Opinions

Dileep kumar B R[1] and Dr.L.Venkateswara Reddy[2]

[1] Sree Vidyanikethan Engg College.

Received: 16 December 2011 Accepted: 1 January 2012 Published: 15 January 2012

Abstract

Due to the advancement of Web 2.0 technologies, a large volume of Web opinions is available on social media sites such as Web forums and Weblogs. These technologies provide a platform for Internet users around the world to communicate with each other and express their opinions. Web opinions are short and sparse text messages with noisy content. In this paper, we are using a sparse matrix representation for web opinions and defining a preprocess way for it. Here, we are proposing an algorithm for matrix generation from vector of thread?s. Due to this representation, we use opinions in efficient way.

Index terms— Web opinion, sparse matrix representation.

1 Introduction

eb opinions are usually less organized and sparse messages. Web users who want to express their opinions on political and social issues, religion, consumer products, traveling experiences, movies, music, sports, health, technology or any topics of interests, they will submit a message to a Web forum platform, a Weblog platform or an individual Weblog site to share their opinions with others. A Weblog or Web forum is a channel for Web users to share their personal details to a circle of friends, amplify their voices and sentiment, establish online communication in a topic of interest, and promote an ideology.

The frame work for web opinion project are proposed by C. C. Yang and Tobun D. Ng [1]. The framework of the web opinion project is depicted in Figure **??**.The framework has five major components: web opinions discovery and collection, web opinions analysis, web opinions evolution and understanding, and interactive information visualization.

Web opinions having some properties ,They are(1) the messages are less focused, (2) the messages are usually short with the length ranged from a few sentences to a couple paragraphs, Fig. **??**: The framework of the web opinions analysis and understanding project

(3) different users may use different terms to discuss the same topic, therefore, the terms used in the messages are sparse, (4) the messages contain many unknown terms that do not exist in typical dictionary or ontology, e.g. iPhone, Xbox, (5) there are many noises, many Web opinions do not fall into any categories, (6) the volume of Web opinion messages is huge and it is expanding in an enormous rate, and (7) the topics in these messages are evolving.

2 II.

3 Related Work

In our preliminary studies [2], [3], it is found that over 50% of Web opinions are noise. Due to the sparseness of terms being used in Web opinions, the distance measured by document vectors are usually large although the corresponding documents are related. These reasons cause the poor performance of Web opinion. The representation of Web opinions is not satisfied or applicable because of the Web opinion properties.

Sparse matrix: A sparse matrix is a matrix populated primarily with zeros. The sparsity corresponds to systems which are loosely coupled. The concept of sparsity is useful for which we have a low density of significant data. When storing and manipulating sparse matrices on a computer, it is beneficial and often necessary to use

specialized algorithms and data structures that take advantage of the sparse structure of the matrix [5] [6].(D D D D)

The Object of sparse matrix (Coordinate list) is, a set of triples, <row, column, value>, where row and column are integers and form a unique combination, and value comes from the set item.

For example, consider a matrix A as given below A.

The sparse matrix representations as shown below R C V R=row C=Column V=Value III.

4 Preprocessing

In the preprocessing, we have collected some web opinions with respect to threads. In this paper, we are using three steps [4] to make data ready to represent as sparse matrix:

Step 1: In this step, we exclude some words that are commonly used in conversation or casual online discussions and at the same time to use the most important set of terms to represent each thread for similarity comparison in the clustering process or some other process. After tokenizing a document, commonly used terms or stop words are first removed from the term set of each document.

Step 2: In this step, we are finding the statistics of term frequency tf for all terms.

Step 3: In this step, we create the document vector for each thread after applying above two steps. After this step we get vector of all threads and it use bigrams or two-word terms as part of the document vectors. Natural language processing is an ideal tool to identify noun and verb phrases, which carry higher specificity than single words or monograms and employ a method to form bigrams by joining two adjacent words without any punctuation or stop word between them.

5 IV.

6 Representation in Sparse Matrix

First of all, we have to create a matrix for all opinions. The creation of matrix involved the following steps: 1. Vector Gathering. 2. Matrix Generation.

7 Vector Gathering :

In this step, we are collecting the threads which are pre-processed. After the preprocessing, we have the necessary terms with their Term Frequency and they are defined in a vector from. This step repeats until all the threads are completed.

8 Matrix Generation:

In the Matrix generation, we have already collected the vectors of all threads. Now we store them in the matrix form. a) Column defines the terms occurring in threads. b) Row defines the thread's TF (Term Frequency) according to the term.

Algorithm for GenMatrix( ): 1. Initializing the row_Size=0, col_Size=0, th_Term={" "}. 2. for all threads 3. for each and every term in thread 4. if (term not in th_Term) { 5. add (term) to th_Term; 6. add its TF 7. } 8. for all threads 9. for all term in th_term 10. if (term not thread) { 11. add its TF=0 12. } 13. for all threads 14. { 15. col_size=0; 16. for all trem in th_trem 17. { 18. mat [row_Size][col_Size]=trem's TF; 19. col_Size++; 20. } 21. row_Size++; 22. } In above algorithm, th_Term is a String array which store all the terms from all threads.

From line 02-07: This performs a Searching operation to find new term in thread and store it.

From line 08-12: In this module, adding TF as ZERO. For not having term in their thread.

From line 13-22: Finally, we are storing the value of TF of each thread into a matrix form. Now, we have created a matrix and it having sparse data. The sparse matrix representation is same as we defined in the related work. Consider a matrix of 5 columns and 6 rows as shown in the below. The above define is sparse matrix for matrix A. The matrix is having row, column and it's value. For example, take a value a 62 is value at 6 th -row and 2 ndcolumn. Normal matrix representation take a 6X5=30 unit of memory and sparse matrix takes 39 units. Disadvantage of proposed system is time taking to create and its take more space. In advantages side, it gives good results in different functionality and in this modern days space is not a problem.

V.

9 CONCLUSION

In this paper, we have proposed an algorithm for generating a matrix from vectors. From matrix, we represent it into sparse matrix. This is the best way to represent the opinions. It has different applications in data mining and gives the basic idea of functionality like clustering and etc. From the way of representation is easy to find the term frequency of terms and we can efficiently find the trending topics in discussion. [1] [2]

[Yang and Ng (2011)] 'Analyzing and Visualizing Web Opinion Development and Social Interactions with Density-Based Clustering'. C C Yang , T D Ng . *IEEE Transc. On Sys. Man and Cyb* Nov.,2011. 41 (6) p. .

[Yang and Ng ()] 'Analyzing Content Development and Visualizing Social Interactions in Web Forum'. C C Yang , T D Ng . *IEEE International Conference on Intelligence and Security Informatics*, 2008. p. .

[Golub and Van Loan ()] Gene H Golub , Charles F Van Loan . *Matrix Computations*, (Baltimore: Johns Hopkins) 1996. (3rd ed.)

[Yang and Ng ()] 'Terrorism and Crime Related Weblog Social Networks: Link, Content Analysis and Information Visualization'. C C Yang , T D Ng . *IEEE International Conference on Intelligence and Security Informatics*, 2007. p. .

[Tewarson (1973)] Reginald P Tewarson . *Sparse Matrices (Part of the Mathematics in Science & Engineering series*, May 1973. Academic Press Inc.

[Yang and Ng ()] 'Web Opinions Analysis with Scalable Distance-Based Clustering'. C C Yang , T D Ng . *IEEE International Conference on Intelligence and Security Informatics*, 2009. p. .