International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 4, July - August 2016 **ISSN 2278-6856**

Semantic Analysis of Tweets using LSA and SVD

S. G. Chodhary[1], Rajkumar S. Jagdale[2], Sachin N. Deshmukh [3]
[1]R.D.I.K. & K.D. College, Badnera-Amravati,
Maharashtra, India

[2, 3] Department of CS & IT,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad, Maharashtra, India

Abstract: ***This paper presents experimental result on tweeter data using information retrieval technique called as LSA used to improve searching in text. LSA consist of matrix operation known as SVD, main component of LSA which is used to reduce noise in text. The proposed methodology has given accuracy of 84 % on twitter data.***
Keywords: Semantic Search, Latent Semantic Analysis, Singular Value Decomposition.

1. INTRODUCTION

The aim of semantic search is to enhance search accuracy by realizing the intent of search and related meaning of the word in the searching document or in a large text file. A simple example can explain it in a better way, if someone ask to computer operator "do you have net?" its natural meaning is to enquiring about internet. But if the same question was asked to sports dealer its meaning is changed drastically i.e. net is equipment used in games such as badminton, tennis etc. The aim of semantic search is to provide that intent to the search engine.

In Natural Language Processing (NLP), LSA is used as technique to bring words/terms which are used and occur in same environment or circumstances have tendency to intent similar meaning, known as distributed semantics. LSA is applied to show that the words/terms that have related meaning found in some part or portion of the text document. From a large piece of text, a matrix containing terms frequency per document is obtained, here terms and document are represented by row and columns respectively. Now using Singular Value Decomposition (SVD) the dimension of the above matrix is reduced. The terms are compared by dot product of row and column, if this value is 1 then it represents very similar terms and if near to 0 it represent distinct terms. This helps to remove noise from the data; the noise is a data which is described as rare and less important usages of certain terms [1].

2. LITERATURE SURVEY

2.1 Semantic Search

Semantic search are not limited only to contextual search or search based on the intent of the question. But it also includes several other factors such current trend, location of search, concept matching, variation of words, synonyms, general and special queries and natural language queries to provide accurate search results [2]. Most of the advanced semantic search engine integrates some of the above factors to provide most relevant result.

There are two types searching, navigational and information search [3]. Navigation search intended to find particular document. While in information search, the search engine is provided with broad topic for which there may be large number of relevant results. This approach of semantic search is closely related with exploratory search.

Also there is a list of semantic search systems which analyze other uses of semantics in the searching action to achieve result. These approaches are related to searches / quires, reference results, semantic annotated result, full text similarity search, search on semantic / syntactic annotation, concept search, semantic web search, ontology search, faceted search, clustered search and natural language search [4].

Semantic search may include other technique to retrieve the precise information from structured data sources like ontology and XML found in the semantic web [5]. Other approaches present domain specific ontology's and retrieval framework using vector space model to specify the user intent in detail at query time [6].

2.2 Latent Semantic Analysis (LSA)

LSA is a statistical method used to derive meaning from a text. It was developed in late 1980s at Bell Core / Bell Laboratory by Launder and his team. They defined "LSA is theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text" [7]. They claim that LSA can play important role to gain extra knowledge for human being. The LSA derive more precise method to extract and represent meaning of text than the previous information retrieval applications.

LSA is used to extract and derive relations of expected contextual use of terms in a portion of text. It uses human constructed dictionaries, knowledge base, semantic network, grammar, parser as input. Then texts are passed into words consisting of unique character string and divide them into sentence or paragraph [8]. LSA uses

following steps:

- **Table Creation:**

The first step of LSA is to create frequency table for a given text document as term document matrix, here rows are represented by terms and columns are represented by documents/ tweet [9]. Term is part of tweet may have single word and tweets can be sentence or paragraph. These terms are assign weight to number of times each term appeared in each document/tweet.

- **Singular Value Decomposition:**

Next step of LSA is to apply SVD to the term document matrix, which breaks down term document matrix into three matrices say, an orthogonal matrix U, a diagonal matrix S and the transpose of the orthogonal matrix V. The SVD of term document matrix X can be defined as:

$X_{mn} = \{U\}^{mm} \{S\}^{mn} \{V^T\}^{nn}$ .....(1)

Where,

U is an orthogonal matrix of size m*m,

S is a diagonal matrix called a singular matrix of size m*n,

V is an orthogonal matrix of size n*n.

- **Reduced Singular Value Decomposition:**

We can reduce the dimension of matrix U, S and V obtained in equation 1 to lower rank, by ignoring first some K columns of U and V and first K singular values of S. We get the following equation:

$$X = X^{mn} \cong \{X_K\}^{mn} = \{U\}^{mk} \times \{S\}^{kk} \times \{V^T\}^{kn} \quad \text{.....(2)}$$

This reduction of K-dimension provided by SVD reduces the noise in text and used to capture latent structure in text. The value of K should be selected carefully as it represents latent structure of the data.

Dimension reduction provides benefits that the terms that share substructure become more similar to each other and the terms that are dissimilar begin to become more dissimilar. This means that the tweets/documents about a specific topic become more similar even if exact term/word doesn't appear in all of them.

3. METHODOLOGY

1.1 Process to Retrieve Twitter Data

Prerequisites to retrieve twitter data is installed R tool and Rstudio on windows system, need a twitter account, Use twitter login ID and password to sign in at Twitter Developers. After successful creation of twitter application four secret keys are generated i.e. consumer key, consumer secret key, access token and access token secret keys which are important to put in R script. R comes with a standard set of packages. For twitter data extractions specially twitteR, ROauth, Rcurl etc these packages are needed. We used direct authentication so after using above four keys and loading specific packages direct function are used to extract tweets by hashtag or username. Then we set two variables, one for the search string, which could be a hashtag or username, and the second variable is the number of tweets we want to extract for analysis.

Use searchTwitter () function to search Twitter based on the supplied search string and returns a list. The "lang" parameter is used below to restrict tweets to the "English" language. We extracted 500 tweets of each by hashtag i.e. #cricket and #education but only 10 tweets of each has been used for experimental analysis.

1.2 Cleaning of Corpus

Twitter data first extracted to give meaningful information by converting it to lowercase, removing punctuation, web address, numbers, whitespace etc. then remove stop words from the data. Stop words are those words which do not have distinctive meaning and context, user may add his own list of stop word if required which is followed by indexing. Sometimes instead of making indexing of words/terms as they appear in tweets, the procedure of stemming should be followed. Stemming is the process to reduce the words/terms to their base or root for example term/word education, educating, educated can be stemmed to educate. For better document retrieval first eliminates stop words then stemming of words followed by creation of frequency table [10].

1.3 Frequency Table Creation

The frequency table or term document matrix so obtained consist of 147 rows and 20 columns means that 147 unique terms are found from 20 tweets. The row represent unique terms in the tweets and column represent tweet titles as c1,c2,…,c10 & e1,e2,...,e10 representing tweets of cricket and education respectively. The corresponding term document matrix X is shown below:

Table 1: Snap shot of frequency matrix size of 147 terms From 20 tweets (documents)

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10
action	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
airplan	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
also	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
amp	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
appl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
appli	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
atleast	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
back	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
bat	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bed	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
came	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
center	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
certain	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chanc	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

1.4 SVD Creation

The linear decomposition of term document matrix X, as given in (1) produces three matrices U, S and V. The dimension of U is an orthogonal matrix of size 147 x 147, S is a diagonal matrix of size 147 x 20 and V is an orthogonal matrix of size 20 x 20.

3.5 Term Reduction Matrix

The dimension reduction of three matrices U, S and V obtained above can be produced by considering K=2 i.e. considering only first two columns of U and V and first two singular values from S, stored them, U' as matrix of size 147 x 2, V' as matrix of size 20 x 2 and S' as matrix of size 2 x 2 only. Now we obtained X'= U' x S' x (V^t)' as given in (2).

This dimension reduction produces a matrix in which terms in same context accumulate higher or lower frequency and other terms that are not appearing previously in original matrix appeared with some fraction. The SVD of frequency matrix obtained from Table 1 is shown below:

Table 2. SVD of frequency table as given in Table 1.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10
action	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
airplan	0.01	0	0.03	0.02	0.01	-0.02	0	0	0.01	0	0.09	0.06	0.06	0.14	0.17	0	0	0.02	0	0.01
also	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
amp	0.03	0	0.07	0.07	0.04	0.05	0	0.02	0.03	0.02	0.2	0.13	0.13	0.31	0.37	0	0.01	0.05	0.01	0.02
appl	0	0	0.01	0.01	0.01	0.02	0	0	0.01	0	0.02	0.01	0.01	0.03	0.03	0	0	0	0	0
appli	0.02	0	0.04	0.04	0.02	0.03	0	0.01	0.02	0.01	0.12	0.08	0.08	0.19	0.22	0	0	0.03	0	0.01
atleast	0	0	0.01	0.02	0.01	0.06	0	0	0.01	0.01	0.01	0	0	0.01	0	0	0	0	0	0
back	0.04	0.01	0.11	0.15	0.08	0.3	0.01	0.03	0.07	0.04	0.23	0.14	0.14	0.37	0.38	0	0.01	0.05	0.01	0.02
bat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bed	0	0.01	0.1	0.24	0.13	0.83	0.02	0.05	0.12	0.07	0.01	-0.01	-0.02	0.05	-0.11	-0.01	0	-0.01	0	0
came	0	0	0.04	0.08	0.04	0.25	0.01	0.02	0.04	0.02	0.03	0.01	0.01	0.06	0.02	0	0	0	0	0
center	0	0	0.01	0	0	-0.01	0	0	0	0	0.02	0.02	0.02	0.04	0.05	0	0	0.01	0	0
certain	0	0	0.01	0	0	-0.01	0	0	0	0	0.02	0.02	0.02	0.04	0.05	0	0	0.01	0	0
chanc	0	0	0.02	0.04	0.02	0.13	0	0.01	0.02	0.01	0.02	0.01	0.01	0.03	0.01	0	0	0	0	0

3.6 Correlation

Correlation matrix uses tweets title name down the first column and across the first row. Correlation matrix is always symmetric matrix, we consider only lower triangular matrix for calculation. Correlation is the single number that describes the degree of relationship between two titles of tweets.

3.7 Correlation of Frequency Matrix

We obtained correlation of frequency matrix of size 147 terms for 20 titles (document) as given in Table 1 and result is stored as Table 3 as shown below:

Table 3. Correlation of frequency matrix given in Table 1.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	e1	e2	e3	e4	e5	e6	e7	e8	e9
c2	-0.04																		
c3	-0.04	-0.07																	
c4	-0.05	0.00	0.10																
c5	-0.04	-0.07	0.16	**0.21**															
c6	-0.06	-0.10	-0.03	0.00	-0.02														
c7	-0.04	-0.06	-0.06	-0.08	-0.06	-0.09													
c8	0.06	-0.08	-0.08	-0.02	0.02	-0.11	**0.22**												
c9	-0.03	-0.06	0.08	0.05	0.09	0.11	0.10	0.05											
c10	-0.04	-0.07	-0.07	0.00	-0.07	-0.03	-0.06	-0.08	-0.06										
e1	-0.05	-0.08	-0.08	-0.10	-0.08	-0.06	-0.07	-0.09	-0.06	-0.08									
e2	-0.04	-0.06	0.07	-0.07	-0.06	-0.09	-0.05	-0.06	-0.05	0.07	0.16								
e3	-0.04	-0.06	-0.06	-0.08	-0.06	-0.09	-0.06	-0.07	-0.05	-0.06	0.14	**0.23**							
e4	0.15	-0.08	0.00	0.05	0.01	-0.06	-0.07	-0.09	0.05	-0.08	0.13	**0.24**	**0.21**						
e5	-0.03	-0.06	0.02	-0.07	-0.05	-0.08	-0.05	-0.06	-0.04	-0.06	0.14	0.13	0.12	0.18					
e6	-0.03	-0.06	-0.06	-0.07	-0.06	-0.08	-0.05	-0.06	-0.04	-0.06	-0.07	-0.05	-0.05	-0.07	-0.05				
e7	-0.04	-0.07	-0.07	-0.08	-0.07	-0.10	-0.06	-0.07	-0.05	-0.07	-0.08	-0.06	-0.06	-0.08	-0.05	0.09			
e8	-0.03	-0.06	0.03	-0.07	-0.05	-0.08	-0.05	-0.06	-0.04	0.03	-0.07	-0.05	-0.05	-0.06	0.02	-0.05	0.03		
e9	-0.04	-0.07	-0.07	-0.08	-0.07	-0.10	-0.06	-0.07	-0.05	-0.07	-0.08	-0.06	-0.06	-0.08	-0.05	-0.06	0.17	0.03	
e10	-0.04	-0.06	-0.06	-0.07	-0.06	-0.09	-0.06	0.02	-0.05	-0.06	0.02	-0.05	-0.06	-0.07	-0.05	-0.05	0.05	0.03	**0.37**

3.8 Correlation of SVD

We obtained correlation of term reduction matrix in Table 2 and result is stored in Table 4 as shown below:

Table 4: Correlation of SVD matrix as given in Table 2

Variables	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	e1	e2	e3	e4	e5	e6	e7	e8	e9
c2	0.12																		
c3	**0.65**	**0.77**																	
c4	**0.21**	**0.93**	**0.87**																
c5	**0.24**	**0.93**	**0.89**	**1.00**															
c6	-0.17	**0.90**	**0.63**	**0.93**	**0.91**														
c7	**-0.25**	**0.86**	**0.55**	**0.87**	**0.85**	**0.97**													
c8	**0.22**	**0.92**	**0.87**	**0.99**	**0.99**	**0.91**	**0.86**												
c9	0.10	**0.94**	**0.82**	**0.99**	**0.99**	**0.96**	**0.91**	**0.99**											
c10	0.12	**0.94**	**0.82**	**0.99**	**0.99**	**0.95**	**0.90**	**0.98**	**1.00**										
e1	**0.99**	0.13	**0.67**	**0.23**	**0.27**	-0.15	**-0.23**	**0.24**	0.13	0.14									
e2	**0.99**	0.05	**0.61**	0.15	0.18	**-0.24**	**-0.31**	0.16	0.04	0.05	**1.00**								
e3	**0.99**	0.01	**0.58**	0.11	0.15	**-0.27**	**-0.34**	0.12	0.01	0.01	**0.99**	**1.00**							
e4	**0.99**	0.19	**0.72**	**0.29**	**0.33**	-0.09	-0.17	**0.30**	**0.19**	**0.20**	**1.00**	**0.99**	**0.98**						
e5	**0.98**	-0.06	**0.51**	0.03	0.07	**-0.35**	**-0.41**	0.04	-0.07	-0.06	**0.98**	**0.99**	**1.00**	**0.96**					
e6	**0.55**	**-0.63**	**-0.22**	**-0.64**	**-0.61**	**-0.86**	**-0.87**	**-0.62**	**-0.71**	**-0.70**	**0.54**	**0.59**	**0.63**	**0.48**	**0.68**				
e7	**0.94**	0.03	**0.57**	0.14	0.18	**-0.22**	**-0.29**	0.16	0.04	0.06	**0.94**	**0.94**	**0.94**	**0.93**	**0.93**	**0.56**			
e8	**0.99**	0.02	**0.58**	0.11	0.15	**-0.27**	**-0.34**	0.12	0.01	0.02	**0.99**	**1.00**	**1.00**	**0.98**	**0.99**	**0.62**	**0.93**		
e9	**0.94**	0.03	**0.57**	0.14	0.18	**-0.22**	**-0.29**	0.16	0.04	0.06	**0.94**	**0.94**	**0.94**	**0.93**	**0.93**	**0.56**	**1.00**	**0.93**	
e10	**0.99**	0.10	**0.64**	**0.20**	**0.24**	-0.18	**-0.26**	**0.21**	0.10	0.10	**0.99**	**0.99**	**0.99**	**0.99**	**0.98**	**0.55**	**0.93**	**0.98**	**0.93**

3.9 Accuracy Measures

Now to evaluate the accuracy of the result calculated above we can use most popular measure called as precision and recall. In information retrieval, the positive predictive value is called as precision and sensitivity is called as recall [11]. Precision is the percentage of retrieved tweets which are related and recall is the percentage of relevant tweets that are retrieved. Both together produce measure of relevance in tweets [12].
The prediction model for correlation of frequency table given as Table 3 is:

Table 5: Confusion Matrix obtained from Table 3

	Prediction Model	
	Positive	Negative
Truth: Positive	34	56
Truth: Negative	13	87

From the above confusion matrix, we got Sensitivity 38 %, Specificity 87 %, Accuracy 64 %, Precision (p) 72 %; Recall (r) 38 % and F measure 50 %.

The prediction model for correlation of term reduction matrix as given in Table 4 is as under (considering correlation of .80 and above)

Table 6: Confusion Matrix obtained from Table 4

	Prediction Model	
	Positive	Negative
Truth: Positive	69	12
Truth: Negative	18	91

From the above confusion matrix, we got Sensitivity 85 %, Specificity 83 %, Accuracy 84 %, Precision (p) 79 %; Recall (r) 85 % and F measure 82 %.

4. CONCLUSION

From experimentation, it is found that when LSA with SVD is used for semantic matching of the twitter data. It works well, in spite of the short length of the tweets. The accuracy is 84 % with precision of 79 % and recall of 85 %. Hence we proposed this method for semantic analysis.

REFERENCES

[1.] Dumais, Susan T. "Latent semantic analysis." Annual review of information science and technology 38.1 (2004): 188-230.

[2.] John, Tony (March 15, 2012). “What is Semantic Search?". Techulator. Retrieved June 9, 2016.

[3.] Guha, R.; McCool, Rob; Miller, Eric (May 24, 2003). “Semantic Search”. WWW2003. Retrieved June 9, 2016.

[4.] Grimes, Seth (January 21, 2010). “Breakthrough Analysis: Two + Nine Types of Semantic Search”. InformationWeek. Retrieved June 13, 2016

[5.] Dong, Hai, Farookh Khadeer Hussain, and Elizabeth Chang. "A survey in semantic search technologies." 2nd IEEE International Conference on Digital Ecosystems and Technologies. 2008.

[6.] Ruotsalo, Tuukka. "Domain specific data retrieval on the semantic web."Extended Semantic Web Conference. Springer Berlin Heidelberg, 2012.

[7.] Landauer, Thomas K., and Susan T. Dumais. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." Psychological review 104.2 (1997): 211.

[8.] Hull, David. "Improving text retrieval for the routing problem using latent semantic indexing." SIGIR’94. Springer London, 1994.

[9.] Deerwester, Scott, et al. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391.

[10.] Zaman, A. N. K. "Stop Word Lists in Document Retrieval Using Latent Semantic Indexing: an Evaluation." Journal of E-Technology Volume 3.1 (2012): 17.

[11.] Tóth, Erzsébet, and Béla Lóránt Kovács. "Technical relevance of keyword searches in full text databases." Qualitative and Quantitative Methods in Libraries (QQML) 2 (2014): 477-484.

[12.] Walters, William H. "Comparative recall and precision of simple and expert searches in Google Scholar and eight other databases." portal: Libraries and the Academy 11.4 (2011): 971-1006.

AUTHORS

S G. Choudhary working as Associate Professor in Department of Computer Science at R.D.I.K. & K.D. College, Badnera, Amravati which is affiliated to Sant Gadge Baba Amravati University, Amravati. His area of research is Semantic search and Text mining.

Rajkumar S. Jagdale has received M.Sc. (Computer Science) From Department of Computer Science and IT, Dr Babasaheb Ambedkar Marathwada University, Aurangabad and pursuing Ph.D. in same department and He got DST Inspire fellowship for his Research Work. His research area is Sentiment Analysis Opinion Mining.

Dr Sachin N. Deshmukh is currently working as Professor in Department of Computer Science and IT, Dr Babasaheb Ambedkar Marathwada University, Aurangabad and having experience of around twenty years in teaching for post graduate (M. Tech, M.Sc. and MCA) and graduate B.E., B. Tech courses. University Authorities also have given the responsibility as Director (University Network Information Center), Director (Center for Vocational Education and Training), Chief Coordinator of Spoken Tutorial Project of IIT, Mumbai. He also worked on research projects of UGC and AICTE. Apart from University, worked in AICTE New Delhi on deputation as Deputy Director (e-Governance) and as Associate Professor at COEP Pune on Lien. Working as Member, Standing Complaint Scrutiny Committee at AICTE, and Member of Peer Committee of NAAC for accreditation. His area of research is Text mining, Social Web mining and Intension Mining.