IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Social Media Assist Bot

Devata Anekar[1], Aniruddha Madurwar[2], Payal Hemane[3], Chandrakant More[4], Rahul Porwal[4]

Department of Information Technology, Sinhgad Academy of Engineering, Pune, India[1,2,3,4]

Abstract: A social media consist of a wide range and variety of information or data in the form of Pictures, Videos, Audios or Text, which are encapsulated in a said structure called posts. A social media platform would always have a number of posts in its incoming and outgoing data traffic so managing and maintaining the efficiency of being relevant specifically to a certain user is very difficult. So we propose on developing a bot to handle the traffic. The bot will assist user in filtering the new post update and letting user know which post are of interest to the user by classifying the feed updates using data mining and Machine Learning, thus helping user to navigate through refined posts according to user's interest. It will be plug in to our social media platform, which will store all the posts that are updated on news feed, and then we will use machine learning algorithms to filter the post data to obtain the post that are relevant to the interest stated.

Keywords: Machine Learning; Text Analysis; Semantic Analysis.

I. INTRODUCTION

Social media is a web-based technology for human communication, which occurs through an electronic device. It is the relationship that exists between networks of people. People exchange ideas, feelings and personal information with each other. Social media allows users to create and share textual posts with comments, photos and videos or messages. Social media is designed to allow people to share content quickly, efficiently and in real-time. The ability to share photos, opinions, events, etc. in real-time is transforming the way we do business. There are number of social media websites like Facebook, Twitter, Linked-In, Google+, etc. These social media websites have more than 100,000,000 registered users. Social media has had a profound effect on recruitment and hiring. Many employers use social networking sites to research job candidates. People searching for new job can come to know about new job opportunities using social media. Professional social networks allow people to create and market a personal brand. Social media is also used in learning. Social media has contributed to the increase in long-distance online learning.

Users are connected to many other users, group of users and organizations. User receives many official as well as unofficial posts from different users. Some times user does not have enough time to go through all the posts so it may happen that user skips some important posts. To solve this problem we are creating a social media bot. Social media bots are automated software's having the capability to execute commands on receiving instructions. This bot will save users interest and filter the incoming post. It will also send notification to user when it receives post related to user's interest. The rest of paper is organized as follows: Section 2 explains latent semantic analysis and section 3 mentions some previous work. Section 4 introduces proposed system. Section 5 explains methodology used for post classification and in section 6, we conclude with the expected results and future scope.

II. BACKGROUND

A. Latent Semantic Analysis Introduction:

Latent Semantic Analysis (LSA) was first introduced more than a decade ago as a technique to improve information retrieval. The main idea was to reduce the dimensionality of the information retrieval problem as a means of overcoming the synonym and polysemy problems observed in standard vector space and probabilistic models. A technique from linear algebra, singular value decomposition (SVD), was used to accomplish the dimension reduction. One of the major advantages of LSA in information retrieval and filtering applications is that documents can be retrieved even when they do not match any query words[1].

Latent semantic analysis is a fully automatic statistical approach to extract relations among words. This process analyzes relationships between a set of document and terms they contain. Latent semantic analysis reduces dimensionality of information retrieval problem. LSA is unsupervised learning technique. LSA uses term-document matrix, which describes occurrences of terms in document[7].

LSA analysis consist of four main steps:

a) Term-Document Matrix:

DOI 10.17148/IJARCCE.2018.7356

A large collection of text is represented as matrix. Rows are words and columns are documents, passages, or statements. Individual cell represents frequency of word in document.

b) Transformed Term-Document Matrix:
Entries in term document matrix are transformed. Based on frequencies cells are transformed.

c) Dimension Reduction:
Singular value decomposition method is used to reduce number of rows of matrix by preserving similarity structure among column. A singular value decomposition (SVD) retainsk largest singular values, and sets the remainder to 0. The resulting reduced-dimension SVD representation is the best k-dimensional approximation to the original matrix, in the least-squares sense. Each document and term is now represented as a k-dimensional vector in the space derived by the SVD.

d) Retrieval in Reduced Space:
Similarities are computed between entities in the reduced-dimensional space than in the original term-document matrix.

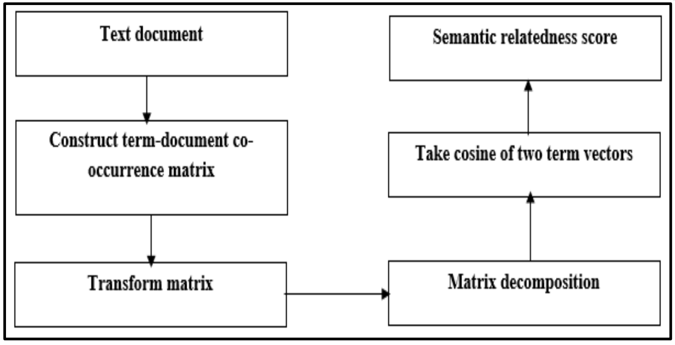

Text document
Construct term-document co-occurrence matrix
Transform matrix
Matrix decomposition
Take cosine of two term vectors
Semantic relatedness score

Figure 1. LSA steps

In order to make difficult problem solvable, LSA introduces some dramatic simplifications:

- Documents are represented as a “bags of words”, where the order of the words in a document is not important, only how many times each word appears in a document.
- Concepts are represented as patterns of words that usually appear together in documents
- Words are assumed to have only one meaning. This is clearly not the case but it makes the problem tractable.

III. LITERATURE REVIEW

ELIZA is an informatics program design in 1966 by Joseph Weizenbaum who was trying to keep a coherent conversation with the user [6].ELIZA searches key words within the text written by the user and it replies with a phrase from its database.

ALICE (Artificial Linguistic Internet Computer Entity) is an Internet project, part of the Pandora Project. This project involves the development of many types of bots especially chat ones [6]. In ALICE’s webpage, the user can chat with an intelligent conversation program, which simulates a real talk. This way, the user may have problems to realize that they are talking with a robot. Dr. Richard S. Wallace develops this technology in Java.

Salto Martinez Rodrigo, Jacques Garcia Fausto Abraham gave implementation of a chat bot on Twitter. This chat bot was developed for entertainment and advertising purpose [6]. They have created chat bot using a database and simple algorithm. This algorithm is divided into three parts: 1) Firstly, bot receives tweets and removes all the punctuation marks and special characters. 2) After formatting, bot starts looking for suitable reply in database. Database contains information stored into a table. This table contains phrases and their suitable replies. 3) If bot finds suitable answer then it retrieves that row. Similarly, in this project punctuation marks and special characters are removed and after that this formatted post is stored in database.

Peter W. Foltz’s theory states how LSA may be used in text-based research [5]. He proposed three applications of LSA for it. He did one experiment. Same text was given to some people and they were asked to write its summary. Then he used LSA: 1) To match each summary with given text. 2) To characterize quality of summary. Two measures of quality of an essay were computed. The first examined amount of semantic overlap of summary with original text. The second measure determined the semantic similarity between summary and 10 sentences chosen by the expert grader. 3) Measurements of text coherence to predict comprehensibility of text. Accordingly amount of semantic overlap of given post text is examined with keywords. Also, comprehensibility of post is predicted by measuring text coherence.

Susan T. Dumais is the first to introduce latent semantic analysis as a technique for improving information retrieval[7]. Mathematical details of LSA approach to information retrieval is presented in Berry, Dumais and O'Brien in 1995. He has mentioned various applications of LSA, which are Information Retrieval, Information Filtering, Cross-Language Retrieval, etc. LSA approach stated in the paper is used for determining comprehensibility of given post. In future applications of LSA stated by author can be used for performing more complex semantic analysis tasks accurately. S. L. Bangare et al have done similar research in terms of software metrics [9] [10].

IV. PROPOSED SYSTEM

The Entire System can be divided into two main partitions, which are Social Media Platform and its Interface and the Bot.

a) Social Media Platform:
The Social Media Platform is the Frontend for the user to interact with the system and share various posts, messages and information through the platform. The platform has a database at the backend, which stores sorts and manages the user's credentials and the posts he shares as well as his connections with other users.

b) Database Management:
The platform generates many unstructured data and to deal with such type of data we use a NoSQL database MongoDB which stores the data in the form of documents. Each user has his details stored in a collection of users and the posts that he shares, his connections and the requests the user gets to establish connections with other users is managed in a different collection.

c) Bot:
The basic work of the bot is to accept inputs from the users to gain insights into the interests the user has and would like to be notified about or would like to take action on the posts shared by his connections. The bot would take note of user's interests stated by the user on the platform when asked for and process them to find associated keywords that match with the buzzwords provided by the user and store them into its database for a specific user. When a post is shared by one of his connections, the post is collected by the bot and it is broken down to find similar keywords and its relevancy with the interests of the user specified by him. If the post is found to be above a certain threshold of relevancy which can be calculated in points or percentage then the bot will notify the user about the said post or will take action on it by liking the post or commenting on it on users behalf, whatever the bot is customized to do by the user.

Thus, the bot will do the same thing with all the posts shared by the connections of a certain user and process them to check whether some of them are of interest to the user or not. The bot will also monitor user's clickstream activities to look for what kind of posts the user goes through and process them to add keywords in its database for a specific user.
The bot will be continuously active in the background to perform these tasks. The bot will be customized as per the user's interest. To process the posts users share we plan to use an unsupervised machine-learning algorithm like Latent Semantic Analysis (LSA) as stated, alternatively Semantic Hashing and various algorithms can also be used for the same.
The Proposed System has the following process:

1. Input from User
The Media platform asks the user to enlist his the kind of posts he would like to go through and the user is asked to provide keywords or buzzwords as input, which in turn are processed by the bot to obtain keywords to be stored into the dictionary for the user.

2. Data Extraction
This process deals with the acquiring of the data, which consists of the posts shared by the user's connection from the database of the platform. This can be done simply by firing a couple of queries on the database server.

3. Data Stemming
The data is obtained by the bot and processed to be broken down and parsed to remove stop words from the posts acquired to create a vector space that can be processed further by using LSA.

4. Acquiring Keywords
After the data is processed with the help of LSA the keywords which exhibit the highest relatedness are acquired and stored into the dictionary of the bot for a specific user which consists of the words from previous iterations of the same process.

DOI 10.17148/IJARCCE.2018.7356

5. Computing Relevancy

After obtaining the keywords, a post in the form of a document is processed to find the accuracy or the relevancy of the said process to the list of interests provided by the user and then the post is picked and stored into the user's database.

6. Notifying User

The user is notified about the obtained set of posts through mail or the bot will take an action on such posts like commenting on the post with a predefined text or like the post if the user has customized the bot to do so.

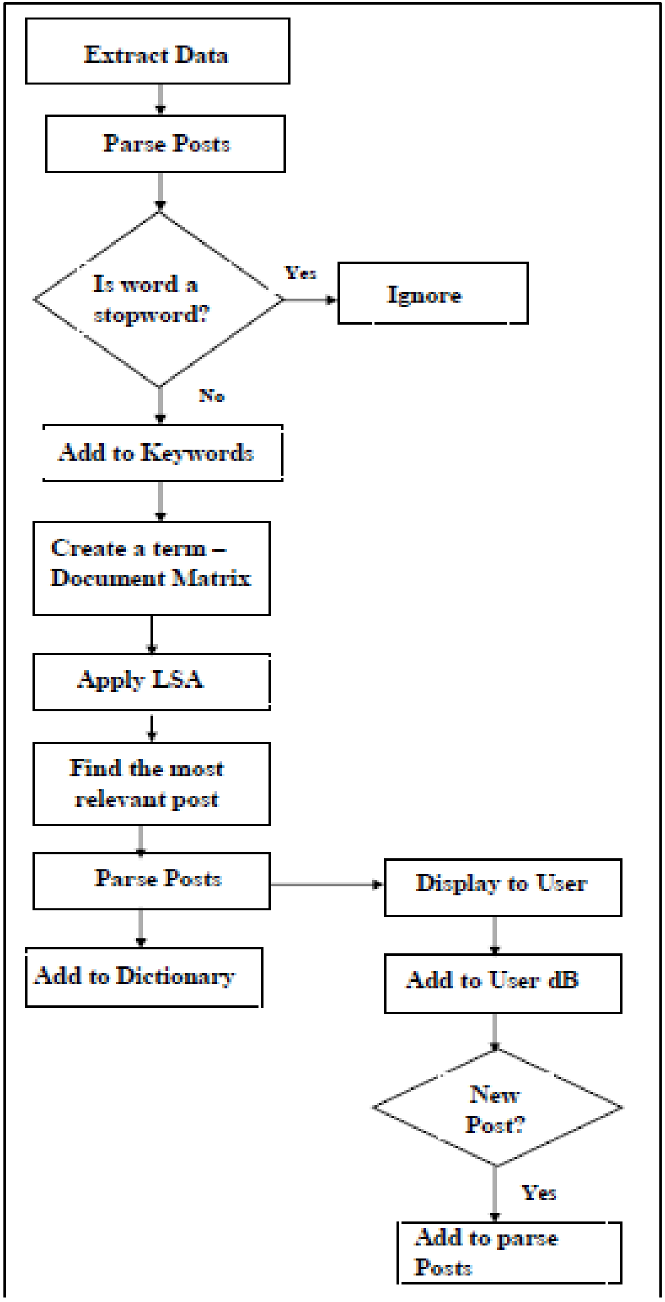


Figure 2. System flow chart

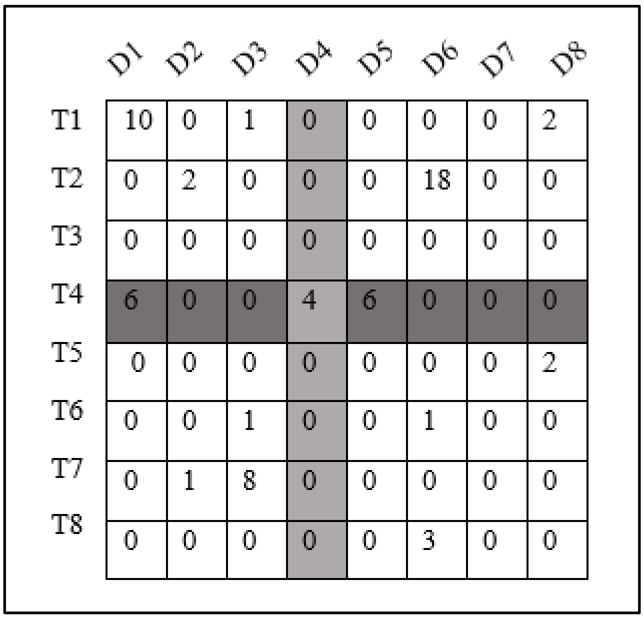
V. DESIGN METHODOLOGY

LSA being a complex algorithm can be divided into three main parts of working which are as follows:

Part 1 – Creating the Count Matrix

The first step in Latent Semantic Analysis is to create the term document matrix (word-title), where each index word is a row and each title is a column. Each cell contains the frequency that word occurs in that title [3]. In the following matrix, 0's have been removed to reduce clutter.

DOI 10.17148/IJARCCE.2018.7356



	D1	D2	D3	D4	D5	D6	D7	D8
T1	10	0	1	0	0	0	0	2
T2	0	2	0	0	0	18	0	0
T3	0	0	0	0	0	0	0	0
T4	6	0	0	4	6	0	0	0
T5	0	0	0	0	0	0	0	2
T6	0	0	1	0	0	1	0	0
T7	0	1	8	0	0	0	0	0
T8	0	0	0	0	0	3	0	0

Figure 3. Term-document matrix

a) Using Python

You need to have installed the Python NumPy and SciPy libraries for the process.

b) Importing various Functions

We need to import a few functions from Python libraries to handle some of the math needed. NumPy is a numerical library, and zeros, is a function that creates a matrix of zeros that we use when building our words by titles matrix. We import the svd function from the package (scipy.linalg) that actually does the singular value decomposition (SVD), which is an essential part of LSA.

c) Defining Data

We define the data that we are using. Documents hold the paragraphs or sentences, “stop words” holds the common words that are to be ignored when we count the words in each title, and “ignore chars” has all the punctuation characters that are to be removed from words.

d) Define LSA Class

The LSA class consists of methods for initialization, parsing documents, building the matrix of word counts, and calculating.

e) Parse Documents

The parse method takes a text, divides it into words, removes the stop word and turns everything into lowercase so the words can easily be compared to the stop words. If a word is a stop word, it is ignored and we move to the next word. If it is not a stop word, we put the word in the dictionary, and we need to append the current document number to keep track of which documents the words appears in.

f) Build the Count Matrix

When all the documents are parsed, all the words (dictionary keys) that occur in more than one document are extracted and sorted, and a matrix is built with the number of rows equal to the number of words (keys), and the number of columns equal to the number of documents. Finally, for each word (key) and document pair the corresponding matrix cell is incremented.

Part 2 – Modify the Counts with TFIDF

In some of the sophisticated LSA systems, the raw matrix counts are generally modified so that rare words are weighted more than common words. The most popular weighting technique is TF-IDF (Term Frequency – Inverse Document Frequency). Under this method, the count in each cell is replaced by the following formula [3].

$TFIDF_{i,j} = (T_{i,j} / T_{*,j}) * \log(D / D_i)$ where

- $T_{i,j}$ = the number of times word i appears in document j (the original cell count).
- $T_{*,j}$ = the number of total words in document j (add the counts in column j).

- D = the no. of columns i.e. the number of documents.
- D_i = the number of documents in which word i appears (the no. of non-zero columns in row i).

In this formula, words that concentrate in certain documents are emphasized (by the $T_{i,j}$ / $T_{*,j}$ ratio)and words that only appear in a few documents are also emphasized (by the log(D / D_i) term).

Part 3 – Using the Singular Value Decomposition

Once we have built our matrix, we call upon a technique called Singular Value Decomposition or SVD to analyze the matrix. The reason SVD is useful, is that it finds a reduced dimensional representation of our matrix that shows the strongest relationships and disregards the noise. In other words, it makes the best possible remake of the matrix with the least possible information [3].

The SVD algorithm is a little involved, but Python has a library function that makes it simple to use. The U matrix provides us the coordinates of each word on our concept space, the Vt matrix gives the coordinates of each document , and the S matrix of singular values gives a clue as to how many dimensions or “concepts” are to be included.

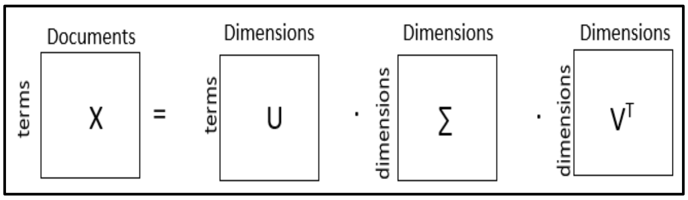
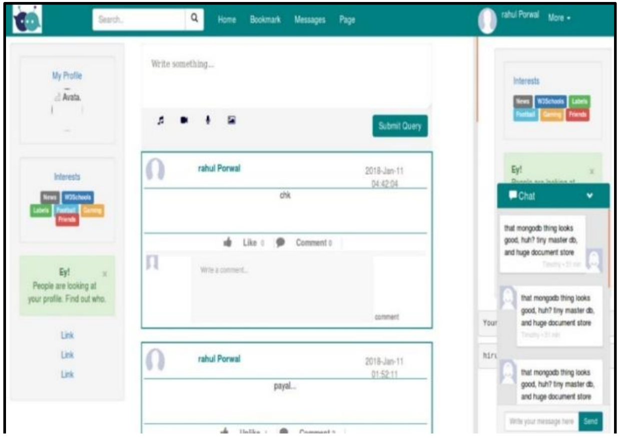


Figure 4. SVD

The platform enables the user in sharing various articles over the internet, which are to be noticed by user’s connections. The data collected from the posts generated by user’s connection is fed to the bot at the backend which is customized according to the user’s preferences and posts which are the most relevant to user’s interest are to be displayed onto his home page which has his news feeds.

The user interface for the social media platform looks as shown in the following figures:



Write something...

rahul Porwal 2018-Jan-11 04:42:04

chk

Write a comment...

rahul Porwal 2018-Jan-11 01:52:11

payal...

Ey!
People are looking at your profile. Find out who.

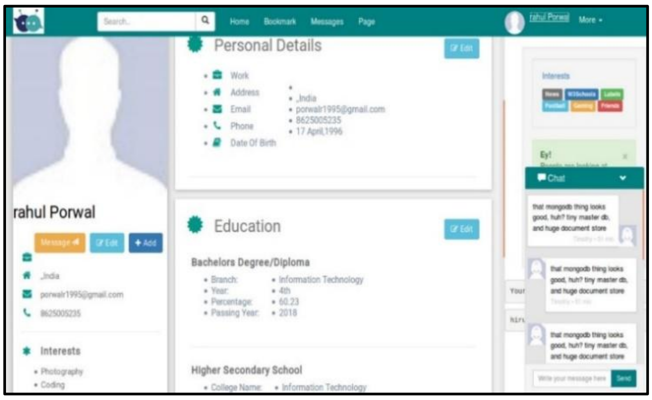
that mongodb thing looks good, huh? tiny master db, and huge document store

that mongodb thing looks good, huh? tiny master db, and huge document store

that mongodb thing looks good, huh? tiny master db, and huge document store

Write your message here

Figure 5. Home page



Personal Details

- Work
- Address: ,India
- Email: porwalr1995@gmail.com
- Phone: 8625005235
- Date Of Birth: 17 April,1996

rahul Porwal

,India
porwalr1995@gmail.com
8625005235

Interests
- Photography
- Coding

Education

Bachelors Degree/Diploma
- Branch: Information Technology
- Year: 4th
- Percentage: 60.23
- Passing Year: 2018

Higher Secondary School
- College Name: Information Technology

Ey!

that mongodb thing looks good, huh? tiny master db, and huge document store

that mongodb thing looks good, huh? tiny master db, and huge document store

that mongodb thing looks good, huh? tiny master db, and huge document store

Write your message here

Figure 6. Profile

DOI 10.17148/IJARCCE.2018.7356

VI. CONCLUSION AND FUTURE WORK

Hence the bot will assist user in filtering the new post update and letting user know which posts are of interest to the user by classifying the feed updates using Machine Learning(Semantic Analysis), thus helping user to navigate through refined posts according to user's interest on the social media platform developed by us. Thus bot will make social media post easily accessible and it will save users time. A continuous process like this lets the bot learn and build its dictionary for a specific user and increase its vocabulary and in turn get more intelligent in determining the relevance of the posts to the user's interest In future, this project can be modified to include talkback feature in bot.To make the social media platform faster and more interactive. To make the bot learn about user interests through user browsing activities.Auto-generated reply on a post on behalf of the user.

REFERENCES

[1] https://en.wikipedia.org/wiki/Latent_semantic_analysis

[2] https://en.wikipedia.org/wiki/Social_media

[3] https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/

[4] Madhuri Dewangan, Rishabh Kaushal "SocialBot: Behavioral Analysis and Detection" Springer Nature Singapore Pte Ltd., P. Mueller et al. (Eds.): SSCC 2016, CCIS 625, pp. 450–460, 2016.

[5] Peter W. Foltz New Mexico State University, Las Cruces, New Mexico "Latent semantic analysis for text-based research"

[6] Salto Martínez Rodrigo, Jacques García Fausto Abraham "Development and Implementation of Chat Bot in Social Network" 2012 Ninth International Conference on Information Technology- New Generations

[7] Susan T. Dumais, Microsoft Research, Redmond, Washington "Chapter 4: Latent Semantic Analysis" Annual Review of Information Science and Technology.

[8] Winterstein, Daniel Ben & Joe Halliwell, "System for organizing social media content to support analysis, workflow and automation".

[9] S. L. Bangare, A. R. Khare, P. S. Bangare, "Code parser for object Oriented software Modularization", International Journal of Engineering Science and Technology, ISSN: 0975-5462, Vol. 2 (12), 2010, 7262-7265.

[10] S. L. Bangare, A. R. Khare, P. S. Bangare, "Quality measurement of modularized object oriented software using metrics", ACM-International Conference ICWET-2011 at Mumbai, ACM 978-1-4503-0449-8/11/02, ISBN: 978-1-4503-0449-8.

DOI 10.17148/IJARCCE.2018.7356