IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 8, August 2020

DOI 10.17148/IJARCCE.2020.9804

Recommendation System based on Soci-Similarity using Social Media Data

Lakshmi Shree K[1], Mythili M[2]

Assistant Professor, Department of Information Science Engineering, Vemana Institute of Technology, Bangalore[1,2]

Abstract: Twitter, Facebook are few prominent Social media platforms to share information. On the other hand, the Internet being an open and free source forum is attracted by news media which have utilized online platforms to publish news articles. An attempt to examine the trending topics in Twitter, Facebook and News Media is the focus of our study. In this paper, we propose a Recommendation System based on Soci- Similarity based on trending topics shared by people in different social media. Compare the effect of trending topics in Media Focus and User Interaction of people. On the predefined days we examine the trending topics of similarity by applying K – Means Clustering to cluster the topics. And we further measure Media Focus and User Interaction using the performance metrics such as Precision, Recall and F1-Score.

Keywords: Recommendation System, Similarity, K-Means Clustering.

I. INTRODUCTION

Media Focus (MF) and User Interaction (UI) are focussed to perform the similarity check on the trending topics which shall be used for building Recommendation System. In the past Media Focus (MF) and User Interaction (UI) is analysed for the Ranking [1]. But in this work Soci-Similarity is performed effectively to identify news topics that are prevalent in both social media and the news media. The degree of similarity between Media Focus (MF) - online news media and User Interaction (UI) - Twitter, Facebook is explored. Soci – Similarity can be achieved by integrating several techniques, such as keyword extraction, cosine similarity and K Means clustering to cluster based on topics. Trending Topics clustering is identified for only a specified period of time to examine the coverage of the topics over this period. Clustering clearly groups distinct topics-based frequency at which the keyword /topic is trending.

II. RELATED WORK

Authors [2] suggest recommendations to a target user consisting in people who publish tweets based on interest level. The same is evaluated and compared using recommendation approaches: the first selects a set of candidate recommendations using only the network topology and the second exploits the user-generated content available in their tweets. Topology of Twitter network. Starting with a target user (the user to whom we wish to recommend new followers) connection a set of candidate recommendations is selected. And those candidates are ranked according to a scoring function involved with three factors of the most influential properties of the Twitter network. The second algorithm creates a vector of terms describing the interests of the target user based on his/her follower's tweets. Also the same vector shall discover new users who may not belong to the target user neighbourhood though they are similar to him/her. The experiment was evaluated by 26 users. These users, 20 males and 6 females. These 26 users were asked to follow at least 20 Twitter users who publish information or news about a set of particular subjects of their interest. These 2 approaches are evaluated and compared for performance of both algorithms. P@5 ("*precision at five*") percentage of relevant recommendations among the first five, averaged over all runs. The results prove the content-based algorithm to always position relevant users earlier in the ranking than the topology-based algorithm.

Author [3] uses a Twitter- LDA model to discover topics and compare the content of Twitter with New York Times news articles. The relation between the proportions of opinionated tweets and retweets for each of the tweets. Some findings identified were though both Medias cover similar topics the distributions of these categories vary. Twitter users tweet less on global events and more of personal life experiences.

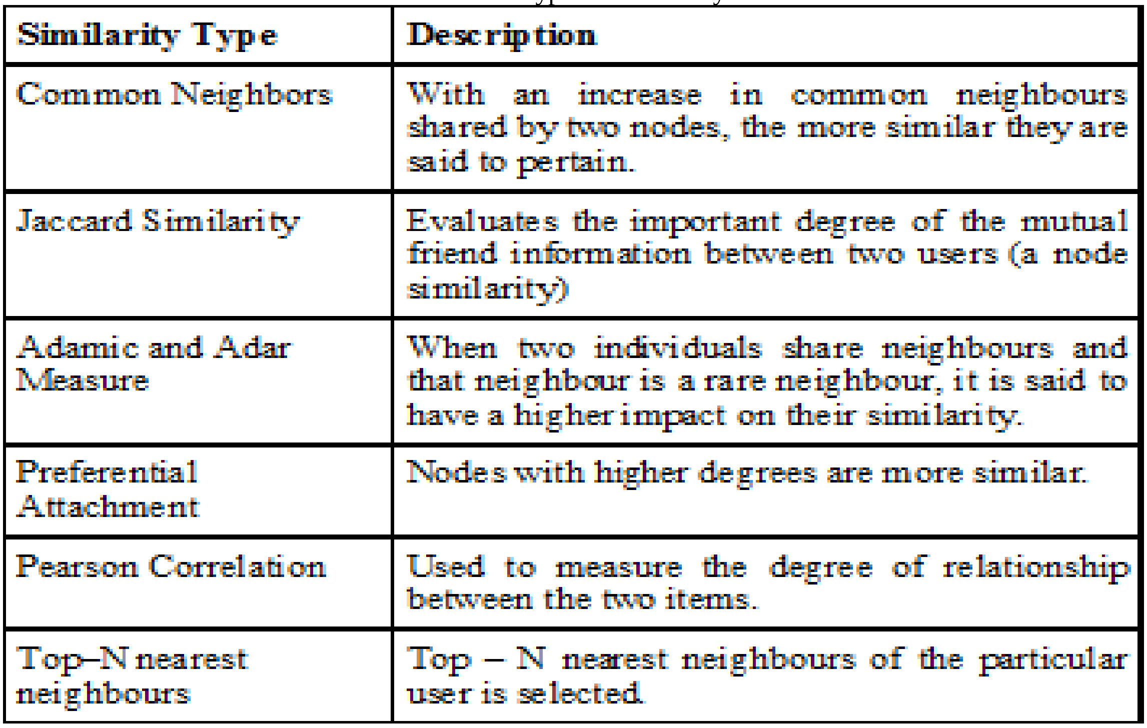
The authors [4] use Automatic Content Linking Device (ACLD) – A real time document and web page retrieval speech based retrieval system. This system analyses spoken input from one or more speakers using automatic speech recognition (ASR). The documents are retrieved using keyword-based search semantic similarity measure. This measures the similarity between documents and the words obtained from automatic speech recognition. Also support access and retrieval of multimedia documents and web pages, using a robust semantic search method in real-time from a variety of repositories.

Authors [5] propose a semi-supervised sentiment analysis method based on topic modelling with Additive Regularization. This is an Unsupervised approach based on different sorts of generative methods such as LDA (Latent Dirichlet Allocation) or dictionaries. LDA is a widely used natural language processing which allows regularizes to adjust distribution. Topics such as Cooking, Films, Sound, Health and Family datasets which also have sentiment labels are available which are used to evaluate the efficiency of this method. The study shows promising results in terms of f1-score with minimal human involvement.

Types of Similarity Measure

In table 1, various types of similarity measures summarized. Few other similarity measures are Similarity–By–Count, Correlation-based similarity, Nearest Neighbourhood, Adjusted cosine similarity, Cosine similarity.

Table 1: Types of Similarity



Similarity Type	Description
Common Neighbors	With an increase in common neighbours shared by two nodes, the more similar they are said to pertain.
Jaccard Similarity	Evaluates the important degree of the mutual friend information between two users (a node similarity)
Adamic and Adar Measure	When two individuals share neighbours and that neighbour is a rare neighbour, it is said to have a higher impact on their similarity.
Preferential Attachment	Nodes with higher degrees are more similar.
Pearson Correlation	Used to measure the degree of relationship between the two items.
Top–N nearest neighbours	Top – N nearest neighbours of the particular user is selected.

III. SOCI- SIMILARITY IMPLEMENTATION

In modern times, social media services such as Twitter provide an enormous amount of user-generated data, which have great potential to contain informative news-related content. This information is similar to the news media which is considered valuable. In this paper we attempt to analyse the role of Media Focus (MF) and User Interaction (UI) on Recommendation System. In this process we consider the occasion of Ratha Yatra 2018 held at Orissa, India. Ratha Yatra has people across the globe exchanging greetings. Media interaction is collected from the internet to create a Media Dataset. User Interaction is created from Twitter and Facebook. From both Media interaction and User Interaction top k news topics are extracted. The entire process of Soci – Similarity implements keyword extraction, similarity measure, K Means clustering. Other attributes present are – Tweet Text, Date, the length of the text, location of tweets.

Dataset Construction: These three datasets are pre-processed to extract the terms. The Twitter Application Programming Interface (API) is used to collect tweets during the period 13/07/2018 to 15/07/2018. The tweets, news data and Facebook posts will be processed between the given time period. Twitter facilitates users to tweet with Hashtags (#) identify messages on a specific topic. Other symbols used are address (@) to mention other users in tweets. The person mentioned shall be notified with that message and Retweet (RT) users can repost to quickly share the tweet with others. The hashtags which were trending were #Puri, #LordJagannath, #RathYatra, #RathaJatra, #JaiJagannath and #MayLordJagannath. The #RathYatra hashtag was one of the top trending hashtags with 42,000 tweets and lakhs of mentions. The posts related to RathYatra from Facebook posts were collected and the count of such posts were minimal in number. Therefore these posts were merged with the Twitter posts.

Pre-Processing of Dataset: After the initial collection of tweets and news articles text the pre-processing takes place. Pre-processing involves conversion of the terms to lowercase, removal of the special symbols such as Retweet (RT), Mention (@), Hashtag (#) and other stop words. The processing of datasets involves extraction of News Term and Tweets Terms. News Term Extraction involves keywords extraction from the news data source from all the queried articles. We have applied the Latent Dirichlet allocation (LDA) model - a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The latent variables aim to capture abstract notions such as topics. [6] . We implement a variant of the popular TextRank algorithm to extract the top k keywords from each news article. The selected keywords are then lemmatized using the WordNet lemmatizer. The lemmatized terms are added to set N. Tweets Term Extraction involves extraction of Tweets from Twitter. The language of each queried tweet is identified in English. From all the tweets words which are less than three characters are eliminated. The POS (Part of Speech) of each term in the tweets is identified using POS. Using Stanford POS tagger, a category to which a word is assigned in accordance with its syntactic functions is carried out. In order to extract the noun called as terms from the sentences POS tagger is applied.

Next N-gram technique methodology is used to find the co-occurrence of the words in the sentences of Tweets, Facebook posts as well as media news. We are implementing two gram and three-gram techniques. Let us consider the tweet "Celebration of Yatra Begins". Here the key words are Celebration, Yatra and Begins. Number of keywords N = 3. We use the parts of speech to extract the keywords in the sentences and using the n-gram technique we can extract the keyword category.

Further the process of term document frequency is calculated. The document frequency of each term 'n' in set N and each term t in set T is calculated. It is represented as df (n) and df (t). Here the document frequency is referred to as occurrence. Thus df (n) is the occurrence of term n and df (t) is the occurrence of term t. For instance "RathYatra for Harmony, Celebration of RathYatra Begins" here the term df (n) for RathYatra is 2.

Clustering and Similarity Measure: In order to perform a Recommendation System based on Soci- Similarity we must identify the similar terms across Media Focus and User Interaction. This methodology we chose K Means Clustering to create the Clusters and Similarity among the terms in Clusters. Clustering is defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we say we are finding homogeneous subgroups within the data such that data points in each cluster are as similar to each other [7] [8].

By clustering we group terms together based on some criterion for similarity. We define similarity using a distance function on terms. Distance here indicated distance between distributions associated with terms by counting (co)occurrences in documents. Using Cosine similarity measures the distance between the data points can be calculated. In the k-means algorithm we specify the number of clusters (K). And initialize centroids by first shuffling the dataset. And further randomly select the K data points for the centroids without replacement. These two steps are iterated until there is no change to the centroids or when the assignment of data points to clusters shall not change.

Relevant Key Term Similarity Estimation is carried out. This process involves topics that are prevalent in both news media and social media extraction. And the Key Term Similarity is estimated using Cosine Similarity. The relationship between the previously selected key terms called co-occurrence is calculated. Cosine similarity explores the relationship between sentences shared by users. It is a measure of how alike two sentences are said to be and generally the score is in the range [0 ~ 1].

Similarity measure with Cosine Similarity is measured as

$$sim(\mathrm{u_u}, \mathrm{u_{v,}}) = \frac{\sum_i \ \mathrm{r_{u.i}} \ \mathrm{r_{v.i}}}{\sqrt{\sum_i \mathrm{r_{u.i}}^2}\sqrt{\sum_i \mathrm{r_{v.i}}^2}} \text{ -------- (1)}$$

Uv and Vv are 2 users whose similarity with the terms are measured.

Cosine Similarity score of two tweets or news article sentence is 1 means, then the two users are 100% similar, if it is 0.98 means 98% similar, this is useful to find where the tweets or news article sentences are related with the same terms.

Each time we count the number of times each of these words appears in each text. Using the cosine similarity, we can get the co-occurrence count and similarity clustering to group them based on topics. We shall also carry Outlier Detection to detect and subsequently extract outliers from the given set of data.

Now we have clustered with respect to the terms from the tweets as well as media news. By this methodology, we will get the count of tweets and media news which are laid in the cluster, by that we can achieve the Media Focus (MF) and User Interaction (UI). We then cluster the topics to depict their co-occurrences in social media. The clustering clearly identifies distinct topics forming topic clusters (TCs). After obtaining well-separated topic clusters (TCs), the factors that signify their importance are calculated.
The Clustering results with Most Similar topics are identified. In this work we have two main topics identified - devotion and celebration of people towards Rath-Yatra and people tweets which highlighted prosperity, unity across the nation. Thus we have K = 2 implies two clusters – Topic 1 and Topic 2. Topic 1 keywords are related to devotion and celebration sentiments of people towards RathYatra which is a famous event. Topic 2 deals with people's emotions in sharing their sentiment as a citizen praising the country. Few identified keywords were prosperity, unity across the nation during RathYatra. For instance, tweet “May our country be happy and prosperous”. In Table 1 we have Topic 1 and Topic 2 instances illustrated.

Table 2: Illustration of Keywords Topics 1 and Topic 2

Topic – Name	Identified Keywords
Topic 1	Blessings Celebration Congratulations Greetings
Topic 2	Country proud Land Peace United Citizens

From the results we have captured observations that extracted topics after Clustering prove that posts, tweets or the news article terms published around the same time are more likely to have the same topic. That is pre -event, post event and during RathYatra has similar topics discussed in User Interaction and Media Interaction.

IV. RESULTS

The news media highlights the temporal prevalence of a particular topic, giving insights of mass media popularity. Whereas the Facebook and twitter media gives us the insights of the user interest. Though the posts related to RathYatra from Facebook posts were collected the count of such posts were minimal in number. Therefore, the analysis focused on tweets of RathYatra. The twitter media and mass media are in line with few trending topics for instance, Natural Disaster such as earthquake or accidents - Flight crash indicated the strength of the community discussing the same topic. The same topics discussion happens to be limited in Facebook media. Facebook Media consisted of user’s inclined towards sharing user interests. We have examined similarity only during the event days when the RathYatra was conducted.

Table 3: Cosine Similarity Measure for both the Topics

Topic – Name	Cosine Similarity- Media Focus	Cosine Similarity- User Interaction
Topic 1	0.94	0.96
Topic 2	0.91	0.921

Based on equation 1 Cosine -Similarity of Media Focus and User Interaction is calculated and compared in Table 3. The study shows promising results in terms of User Interaction compared to Media Focus. Also it is clear in Table 1, Topic 1 has better scores compared to Topic 2.

In Table 4 and 5 Performance Metrics such as Accuracy, F1 – Score and Precision are measured for Media Focus and User Interaction across both the Topics. In this context, precision can be defined as the share of correctly predicted parameter values compared to the total number of predicted parameter values. Furthermore, recall can be defined as the share of correctly predicted parameter values compared to the total number of relevant parameter values [9]. In Table 4 and 5 we observe Topic 1 to perform better compared to Topic 2 in both the Media Focus and User Interaction.

Table 4: Performance Metrics Measured on Topics for Media Focus

Topic – Name	Accuracy	F1-Score	Precision
Topic 1	0.93	0.88	0.91
Topic 2	0.91	0.82	0.85

Table 5: Performance Metrics Measured on Topics for User Interaction

Topic – Name	Accuracy	F1-Score	Precision
Topic 1	0.93	0.90	0.91
Topic 2	0. 89	0.86	0.85

V. CONCLUSION

Recommendation system based on Soci -Similarity analysis the Similarity in social media such as Facebook, Twitter and online News Articles on RathYatra 2018. Topics are extracted from both media focus and social media to analyse the Topic 1 identified contains terms related to RathYatra festival mood and on other hand Topic 2 is related to people sharing terms relating to the nation, peace unity during RathYatra. Based on K-Means Clustering the topics are clustered and the Cosine Similarity approach identifies the similarity of the extracted terms. Topic 1 has better similarity compared to Topic 2 in both Media Focus and User Interaction. Other performance metrics such as Accuracy, F1-Score and Precision also show better results with Topic 1 keywords.

REFERENCES

[1]. Sumit Nihalani, Vaishali Malik, G.K.Sandhia “Ranking News Based On Social Media and News Channel Sources Using Social Media Factors”, International Journal of Pure and Applied Mathematics, Volume 118 No. 22 2018, 663-666 ISSN: 1314-3395

[2]. Marcelo G. Armentano, Daniela Godoy and Analía Amandi, “Towards a Followee Recommender System for Information Seeking Users in Twitter”, PY-2011/01/01, VL-730, IEEE.

[3]. W.X.Zhoa, Wayne, Jiang, “Comparing Twitter and Traditional Media using Topic Models”, PY-2011/04/18, IEEE.

[4]. Andrei Popescu Belis, Majid Yazdani, Alexander Nanchen, Philip N Garner, “A Speech-Based Just-In-Time Retrieval System”, PY-2011/06/06, VL-85, IEEE.

[5]. Timur Sokhina, Nikolay Butakov, “Semi-automatic sentiment analysis based on topic modelling”

[6]. David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03.

[7]. Christian Wartena & Rogier Brussee, “Detection by Clustering Keywords”, October 2008 DOI: 10.1109/DEXA.2008.120 · Source: IEEE Xplore.

[8]. Qiming Diao, Jing Jiang, Feida Zhu, Ee-Peng Lim, “Finding Bursty Topics from Microblogs”, January 2012.

[9]. C.Trattner, A.Said, L. Boratto and A. Felfernig, “Evaluating Group Recommender Systems”, Published in the book Group Recommender Systems.