© 2020 | International Journal of Scientific Research in Science and Technology
IJSRST | Volume 7 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
DOI : https://doi.org/10.32628/IJSRST207226

Article Recommendation and Comics Story Representation for Twitter User Based Preferences

S. Vivekanandan*[1], Swathi. N[2]

*[1]Department of CSE, Velalar College of Engineering and Technology, Erode, Tamil Nadu, India

[2]PG Scholar, Department of CSE, Velalar College of Engineering and Technology, Erode, Tamil Nadu, India

ABSTRACT

Twitter is an interesting platform for the dissemination of news. The real-time nature and brevity of the tweets are conductive to sharing of information related to important events as they unfold. Numerous consumer reviews of topics are now available on the Internet. Automatically identifies the important aspects of topics from online consumer reviews. Our method provides an efficient way to accurately categorize comic topic recommendation without need of external data, enabling news organizations to discover breaking news in real-time, or to quickly identify viral memes that might enrich marketing decisions, among others. We filter the stream of incoming tweets to remove junk tweets using a text classification algorithm. We also compare the performance of different supervised SVM text classification algorithms for this task. This study concentrates on analyzing potential and dynamic user correlations, based on topic-aware similarity and behavioral influence, which may help us to discover communities in social networking sites.

Keywords : Microblogging, Recommendation System, Comics Story Analysis, Comics Genre Classification, Recommendation Algorithms

I. INTRODUCTION

A.DATA MINING

Data mining (sometimes called data or knowledge discovery) is the progression of analyses data from special perspectives and abbreviation into useful data information that can be used enlarge the revenue, reduce cost and both. Data mining software is individual a number of logical tools for analyzing the data's.It allows the users to analyzed data from various dimensions or angles and review the associations recognized. Technically, the data mining is the process of decision correlations or patterns between fields in huge relational databases.

B. TWITTER

Twitter is an online news and social networking service where users post and interact with messages, known as "tweets." These messages were originally restricted to 140 characters, but on November 7, 2017, the limit was doubled to 280 characters for all languages except Japanese, Korean and Chinese. Registered users can post tweets, but those who are unregistered can only read them. Users access Twitter through its website interface, Short Message Service (SMS) or mobile device application software ("app"). Twitter, Inc. is based in San Francisco, California, United States, and has more than 25 offices around the world.

Twitter was created in March 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams and launched in July of that year. The service rapidly gained worldwide popularity. In 2012, more than 100 million users posted 340 million tweets a day, and the service handled an average of 1.6 billion search queries per day. In 2013, it was one of the ten most-visited websites and has been described as "the SMS of the Internet". As of 2016, Twitter had more than 319 million monthly active users. On the day of the 2016 U.S. presidential election, Twitter proved to be the largest source of breaking news, with 40 million election-related tweets sent by 10 p.m. (Eastern Time) that day.

C. NEWS ARTICLE

The Social networks witness a rapid growth development and recommendation on social networks is becoming essential. Recommendation of quality and useful information to the users is an issue here. How to identify user's interest and provide customized recommendation for each user becomes a challenging problem.Where it aims to investigate a framework to combine tag correlation and user social interest for recommendation. News related tweet contents are extracted and tags for the contents are generated using context inference and user preference analysis.A user tag retrieval strategy is developed to assign tags for users and a user-tag matrix is created to provide the initial weights for user's tags. Every user will have different or similar tags based on his/her social interest. News articles relating to the users tag are recommended for high quality and interest. A user-tag matrix if formed for every user. The tags with higher ranking is considered for recommendation. N-gram extraction is used to find the occurrence of similar tags. RankSVM algorithm is used to rank the different tags of the same user and Rank Aggregation algorithm is used for ranking the overall tags hence a customized recommendation of news articles is provided to the users.

D. COMICS STORY

Comics is usually classified into broad categories called "genres" according to its contents such as comedy, horror, science fiction, etc. Because a genre expresses a comics story briefly, people read comics which has contents based on their interest, by relying on comics genres. However, giving only one genre to one comic cannot express the detailed difference of the story. In this paper, we propose a system for generating comics story representation as a sub-sequence of genres. Our comics story representation can be applied to a new search engine based on stories or to recommendation system which analyzes the tastes of the user's favorite comics by finding comics with similar story representation using Latent Drichlet Allocation. We use a LDA to classify each page into the corresponding genre on twitter. Experimental results confirm the advantage of the proposed system.Comics is a popular entertainment culture all over the world. It is a medium that conveys the idea of authors using drawings and text.

Nowadays, the number of produced comics is too huge to read all of them, even in a lifetime .In order to represent the characteristics of the stories, comics is categorized into genres; for example, battle, fantasy, romance, etc.However, using only one genre for one volume cannot express the detailed difference of the story from other comics in the same genre.Therefore, some comics will not be found by some readers who could be interested in them, although recent comics search systems can find well-known comics based on a genre for each volume.Our ultimate goal is to realize a comics search engine that allows readers to find interesting comics for them quickly.

II. RELATED WORKS

A.A Density-Based Spatial Clustering of Application with Noise

Henrik Bäcklund, has proposed today data is received automatically from many different kinds of

equipments. Satellites, x-rays and traffic cameras are just a few of them. To make this information/data understandable for us, it has to be processed. When working with large data sets it is in most scenarios useful to be able to separate information by dividing the data into smaller categories, and eventually, to do class identification. Not least is this important when treating large spatial databases. A satellite, for example, gathers images as it travels around our earth. It is desired to classify what parts of the images are houses, cars, roads, lakes, forests, etc. Since the image database is big, a good classification algorithm is needed.

B. Beyond trending topics: Real-world event identification on twitter

Hila Becker, has proposed User-contributed messages on social media sites such as Twitter have emerged as powerful, real-time means of information sharing on the Web. These short messages tend to reflect a variety of events in real time, making Twitter particularly well suited as a source of real-time event content. In this paper, we explore approaches for analyzing the stream of Twitter messages to distinguish between messages about real-world events and non-event messages. Our approach relies on a rich family of aggregate statistics of topically similar message clusters. Large-scale experiments over millions of Twitter messages show the effectiveness of our approach for surfacing real-world event content on Twitter.

C. Text Categorization with Support Vector Machines: Learning with Many Relevant Features

Thorsten Joachims, has proposed This paper explores the use of Support Vector Machines (SVMs) for learning text classifiers from examples. It analyzes the particular properties of learning with text data and identifies why SVMs are appropriate for this task. Empirical results support the theoretical findings. SVMs achieve substantial improvements over the currently best performing methods and behave robustly over a variety of different learning tasks. Furthermore, they are fully automatic, eliminating the need for manual parameter tuning.

D. A Comparison of Event Models for Naive Bayes Text Classification

Andrew McCallum, has proposed recent approaches to text classification have used two different first-order probabilistic models for classification, both of which make the naive Bayes assumption. Some use a multi-variate Bernoulli model, that is, a Bayesian Network with no dependencies between words and binary word features (e.g. Larkey and Croft 1996; Koller and Sahami 1997). Others use a multinomial model, that is, a uni-gram language model with integer word counts (e.g. Lewis and Gale 1994; Mitchell 1997). This paper aims to clarify the confusion by describing the differences and details of these two models, and by empirically comparing their classification performance on five text corpora. We find that the multi-variate Bernoulli performs well with small vocabulary sizes, but that the multinomial performs usually performs even better at larger vocabulary sizes—providing on average a 27% reduction in error over the multi-variate Bernoulli model at any vocabulary size.

F. A Survey on Transfer Learning

Sinno Jialin Pan, has proposed A major assumption in many machine learning and data mining algorithms is that the training and future data must be in the same feature space and have the same distribution. However, in many real-world applications, this assumption may not hold. For example, we sometimes have a classification task in one Trends of interest, but we only have sufficient training data in another Trends of interest, where the latter data may be in a different feature space or follow a different data distribution. In such cases, knowledge transfer, if done successfully, would greatly improve the

performance of learning by avoiding much expensive data labeling efforts. In recent years, transfer learning has emerged as a new learning framework to address this problem. This survey focuses on categorizing and reviewing the current progress on transfer learning for classification, regression and clustering problems. In this survey, we discuss the relationship between transfer learning and other related machine learning techniques such as Trends adaptation, multitask learning and sample selection bias, as well as co-variate shift. We also explore some potential future issues in transfer learning research.

III. METHODOLOGY

A. EXISTING SYSTEM

In existing work that focus more significant in connection to web and news portals,where the quality of the news portal is commonly measured by amount of news added to the site.Then the most renowned news portals add hundreds of new articles daily.The classical solution usually used to solve the information overloading is a recommendation.In this existing work that present an monopoly approach for fast but misclassified recommendation may appear.Theperformance for a new class of data analysis software called “recommender systems”. Recommender systems apply knowledge discovery techniques to the problem of making personalized product recommendations during a live customer interaction. The tremendous growth of customers and topics in recent years poses some key challenges for recommender systems. These are: producing high quality recommendations and performing many recommendations per second for millions of customers and topics. Singular Value Decomposition (SVD)-based recommendation algorithms can quickly produce high quality recommendations, but has to undergo very expensive matrix factorization steps. In this work, we propose and experimentally validate a technique that has the potential to incrementally build SVD-based models and promises to make the recommender systems highly scalable.

DRAWBACKS

- Limited Content Analysis and Overspecialization.
- Opinions of a user do not match with any group and therefore, is unable to get benefit of recommendations.
- The availability of huge size of data about tweets the catalogue and the disinclination of users to rate tweets make a dispersed profile matrix leading to less accurate recommendations.
- The sparse rating in CF systems makes it difficult to make accurate predictions about tweets.
- Fewer ratings make it computationally hard to calculate neighbors trends

B. PROPOSED SYSTEM

In news article recommendation extracting opinion targets/words as a co-ranking process called Support vector Machine .We assume that all nouns/noun phrases in sentences are opinion target candidates, and all adjectives/verbs are regarded as potential opinion words, which are widely adopted by previous method.The given data is possibly of any modality such as texts or images, while it can be treated as a collection of documents. SUBJECT wise and TOPIC wise Opinion analysis is also possible.we formulate opinion relation identification as a word alignment process. We employ the word-based alignment model to perform monolingual word alignment, which has been widely used in many tasks such as collocation extraction and tag suggestion. SVM method that starts out with a base classifier that is prepared on the training data. A second classifier is then created behind it to focus on the instances in the training data that the first classifier got wrong. The process continues to add classifiers until a limit is reached in the number of models or accuracy.

For comics story representation the Behavioral influence correlation Algorithms recommends tweets by matching users with other users having similar interests. It collects user feedback in the form of ratings provided by user for specific tweets and finds match in rating behaviors among users in order to find group of users having similar preferences. One of the main features on the homepage of Twitter shows a list of top terms so-called comic topic recommendation at all times. These terms reflect the topics that are being discussed most at the very moment on the site's fast-flowing stream of tweets. In order to avoid topics that are popular regularly (e.g., good morning or good night on certain times of the day), Twitter focuses on topics that are being discussed much more than usual, i.e., topics that recently suffered an increase of use, so that it trended for some reason. Here, a user profile represents user preferences that the user has either explicitly or implicitly provided.

An example is Twitter uses GB approach, which suggests tweets based on the purchase patterns of its users as well as user ratings. Respectively each user has a list of tweets that are rated either explicitly or implicitly. This way a user-tweets rating matrix 'R" is generated, where user preferences about tweets are represented. For finding missing ratings, different techniques are used including finding "nearest neighbor" for new users in recommending tweets to them by considering ratings provided by their nearest neighbors.

IV. ARCHITECTURE

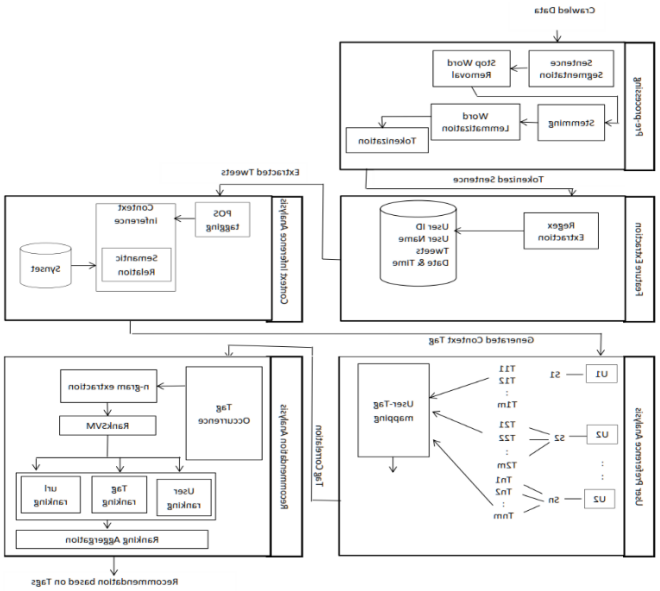


Figure 1: News article recommendation

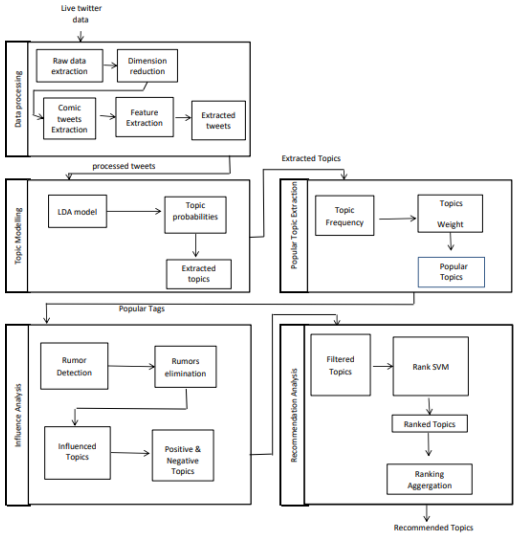


Figure 2 : Comics story representation system based on genre

V. MODULE DESCRIPTION

A. PREPROCESSING

Pre-processing of the acquired data is first carried out. The sentence are segmented and the stop words are removed from the sentence. Now the sentence is

stemmed and lemmatized. Tokens for every sentence of the tweet is achieved at the end of pre-processing.

B.FEATURE EXTRACTION AND TWEETS RATING PREDICTION

The tokens are taken as input for feature extraction. The required features are user names, user id, date&time and the tweet contents. Those features are extracted using regular expression extraction.

C.CONTEXT INFERENCE ANALYSIS

The extracted tweets are POS tagged. For every tweets the tags are generated and using the wordnet or synset dictionary the semantic relation of the tags are found by context inference.

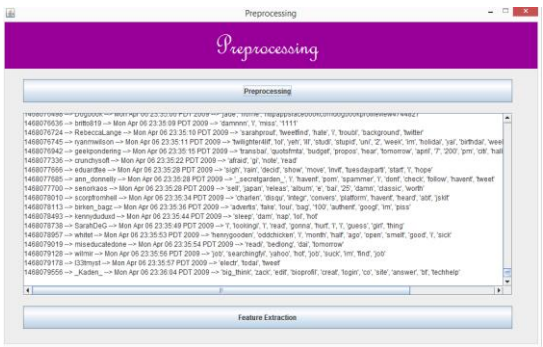
D.USER PREFERENCE ANALYSIS AND COLLABORATIVE FILTERING

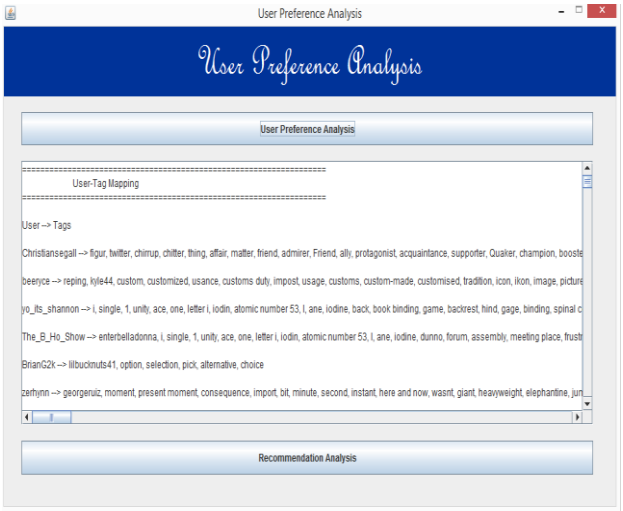
Generated context tags are taken as input. Every user will have multiple tags, for every tag a user-tag matrix is generated to analyze the weights of the tags.Only the tweets with high degree of similarity to user's preferences are would get recommended.

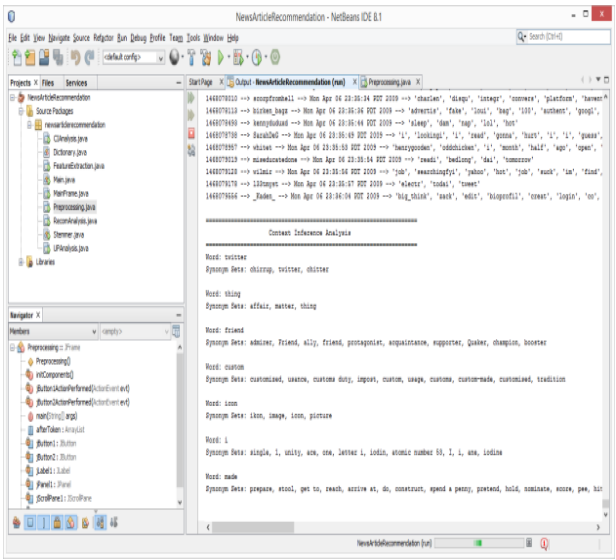
E.RECOMMENDATION ANALYSIS

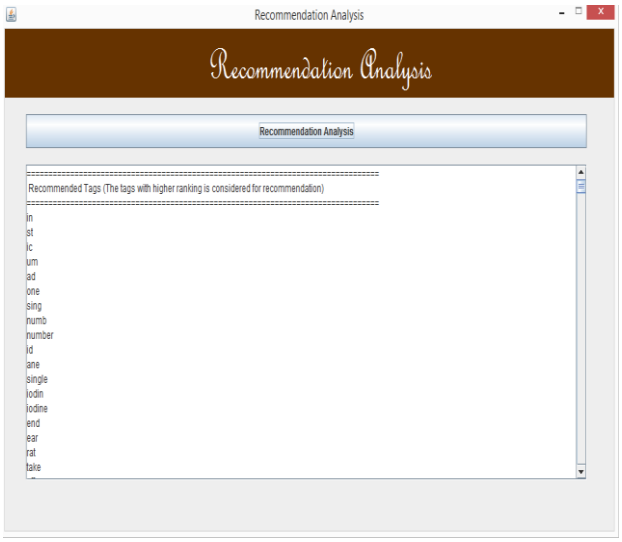
User-tag correlation is taken as input. The tag occurrence is evaluated using n-gram extraction strategy. RankSVM is used for ranking of the tags. The overall ranking is determined using rank aggregation algorithm.

VI. SCREENSHOTS FOR OUTPUT









VII. CONCLUSION

In the last few decades, twitter asynchronous systems have been used, among the many available solutions,

in order to mitigate information and cognitive overload problem by suggesting related and relevant tweets to the users. In this regards, numerous advances have been made to get a high-quality and fine-tuned twitter asynchronous system. Nevertheless, designers face several prominent issues and challenges. In this work, we have touched variety of topics like natural Language Processing, Text Classification, Feature selection, Feature ranking, etc. Each one of these topics was used to leverage the massive information flowing through twitter. Understanding twitter was as important as knowing the topics in question.The results of the previous experiments, led us to the conclusion that feature selection is an absolutely necessity in a text classification system. This was proved when we compared our results with a system that uses the exact same dataset without feature selection. We were able to achieve 33.14% and 28.67% improvement with bag-of-words and TF-IDF scoring techniques correspondingly.

We also mentioned recognition and some opportunities that our work provides in the fields of news media, marketing and businesses in general. We hope that our work can provide a good foundation to the future of text classification in social media and to the opportunities that comes with it.

VIII. REFERENCES

[1]. B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Found. Trends Inf. Retrieval, vol. 2, no. 1/2, pp. 1–135, 2008.

[2]. J. Bollen, H. Mao, and A. Pepe, “Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena,” in Proc. Int. AAAI Conf. Weblogs Social Media, 2011, pp. 17–21.

[3]. B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From tweets to polls: Linking text sentiment to public opinion time series,” in Proc. Int. AAAI Conf. Weblogs Social Media, 2010, pp. 122–129.

[4]. M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2004, pp. 168–177.

[5]. T. Chen, R. Xu, Y. He, Y. Xia, and X. Wang, “Learning user and product distributed representations using a sequence model for sentiment analysis,” IEEE Comput. Intell. Mag., vol. 11, no. 3, pp. 34–44, Aug. 2016.

[6]. Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu, “OpinionFlow: Visual analysis of opinion diffusion on social media,” IEEE Trans. Vis. Comput. Graph., vol. 20, no. 12, pp. 1763–1772, Dec. 2014.

[7]. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: Sentiment classification using machine learning techniques,” in Proc. ACL Conf. Empirical Methods Natural Language Process., 2002, pp. 79–86.

[8]. A.Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” Stanford Univ., Stanford, CA, USA, Project Rep. CS224N, pp. 1–12, 2009.

[9]. F. Wu, Y. Song, and Y. Huang, “Microblog sentiment classification with contextual knowledge regularization,” in Proc. 29th AAAI Conf. Artif. Intell., 2015, pp. 2332–2338.

[10]. J. Blitzer, M. Dredze, and F. Pereira, “Biographies, bollywood, boom-boxes and blenders: Trends adaptation for sentiment classification,” in Proc. 45th Annu. Meeting Assoc. Comput. Linguistics, 2007, vol. 7, pp. 440–447.

Cite this article as :

S. Vivekanandan, Swathi. N, "Article Recommendation and Comics Story Representation for Twitter User Based Preferences", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 7 Issue 2, pp. 105-111, March-April 2020. Available at doi : https://doi.org/10.32628/IJSRST207226

Journal URL : http://ijsrst.com/IJSRST207226