A SURVEY ON: HASH TAG RECOMMENDATION SYSTEM BASED ON SEMANTIC TF-IDF

Priyanka Shrivastava [#1], Shailendra Gupta [#2,] Pankaj Richhariya [#3]
# Department of Computer science & Engineering, BITS, Bhopal, M.P., India
[1] pspriyankacse@gmail.com

ABSTRACT Semantic measure and algorithm among the text generated in various social media platform plays a major role while automated data generation. Twitter is a social media platform where a large number of tweets and data generate in every hour where as the similarity major text concern is important factor. Various example such as #friend, #frd puts a similar message but unable to identify as similar text with similarity major system. In existing research author discussed various related technique to measure it such as Shortest path, Wu & Palmer, Lin, JiangConrath, Resnik, Lesk, LeacockChodorow, and HirstStOnge. In base paper algorithm TF-IDF based approach is given which is described as best among available approach in recommendation system. The algorithm further does not find similarity in between hash tags, while it exhibit similarity measure only in between tweets. Thus enhancement of existing work is going to generate an algorithm extension to exhibit similarity measure in between Hash tags occurrence in tweets, which make more reliable and accessible in micro blogging system.

Keywords– Twitter extraction, web mining, recommendation system, Hash Tag generation.

I. INTRODUCTION

In the recent era, a large amount of raw data is being gathering day by day and storing in databases anywhere across the world, which is mainly collecting from different industry and social media sites**.** There is a requirement to extract and determine useful data and knowledge from such a data that is being collected. Data mining is an interdisciplinary field of computer science. It is referred to as mining knowledgeable data from large databases. It is the process of performing automated extraction and generating the predictive information from a large database. It is the process of searching the hidden information from the repositories .The fields that use Data mining techniques include medical research, marketing, telecommunication, and stock markets, health care and so on. In information retrieval, tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

Data mining consists of the different technological methods including machine learning, statistics, database system etc. The aim of the data mining process is to discover knowledge from large databases and transform into a human understandable format. Data Mining with knowledge discovery are important parts to the organization due to its decision making strategy. Classification, clustering and regression are three methods of data mining. In these methods instances are grouped into identified classes. Classification is a popular task in data mining especially in knowledge discovery. It gives an intelligent decision making. Classification is not only studies and examines the existing sample data but also predicts the future behaviour of that information. It maps the data into the predefined class and groups. It is used to predict group membership for data instances.

A semantic TF-IDF based weighting method is proposed in the current paper. The vector is used for redefining semantic weights and thus the similarity of tweets. For a given tweet, T, the tags of the Top N similar tweets are recommended. The classical metrics of data mining is used for evaluating current approach. Semantic similarity and relatedness algorithms are compared and results showed significant improvement than normal TF-IDF weighting schema.

Usually tags which are semantically related to the terms are used not semantically similar. Consider “plasticsurgery” tag as an example, some of the terms with this tag are: surgery, body, arm, health, beauty. Where they are not semantically similar but are related. Semantic similarity algorithms usually takes a shortest path method on a IS A like graph, in order to calculate semantic similarity while semantic relatedness algorithms uses a graph with Has Part, Kind of, and Opposite edges. This is why HirstStOnge (as semantic relatedness algorithm) has better results than other semantic similarity algorithms.

Twitter as a microblogging system, allows users to share posts each containing maximum of 140 characters, known as tweets. Each tweet is enriched with content-based and context based tags.

II. LITERATURE REVIEW

1. Mir Saman Tajbakhsh, Jamshid Bagherzadeh,2016

In this Paper TF-IDF approach for the recommendation system and similarity measure algorithm in between tweets and twitter dataset is being introduce. This paper also exhibits the difference and effect of approach in Micro-blogging and social media. A relevant approach of finding data similarity is given by the paper. They have performed experiment using Java 8 platform and execute the parameters as precision, recall and F-measure. Further an introduction of pseudo code with efficiency is given by author [1].

2. Otsuka, E., Wallace, S.A., and Chiu, D. 2014

In this paper describe Weblogs one of fundamental components of internet have complexity in searching relevant blogs. Recommender 2.0 and other a lot of unskilled bloggers and visitors who systems are a solution to the information overload problems. In this research a web log recommender system based on link structure of weblog graph is introduced. Here we treat links between weblogs as some kind of rating [2].

3. Givon, S., and Lavrenko, V. 2009

In the paper author review the key decisions in evaluating collaborative filtering recommender systems: the user tasks being evaluated, the types of analysis and datasets being used, the path in which the quality of prediction is measured, the evaluation of prediction attributes other than quality, and the user-based evaluation of the system as a whole [3] .

4. Godoy, D., Rodriguez, G., and Scavuzzo, F., 2014

In this work, the goal is to support system developers in rapid prototyping recommender systems using Case-Based Reasoning (CBR) techniques. In this research article author describes how jcolibri can serve to that goal. jcolibri is an object-oriented framework in Java for building CBR systems that greatly benefits from the reuse of previously developed CBR systems [4].

5. Sigurbjornsson, B., and Zwol, R.v. 2008

In this work author investigates some approaches to exploit context in Recommender Systems. It provides a general architecture of a context aware recommender system and analyzes separated components of the model. The main focus is to investigate new approaches that can bring a real added value to users. In this research article also describe my initial results on item selection and item weighting for context-dependent Collaborative Filtering (CF). Moreover, Author shall present my ongoing research on CF hybridization using context [5].

6. Bart P. Knijnenburg, Martijn C. Willemsen, Alfred Kobsa 2011

In this work, paper concentrates on two classes of phenomenon's, which are decoy effects and serial position effects. Tightly coupled to these phenomenon's is the problem of getting the utility function of a recommender right, as this function serves both as the basis of result set calculation as well as the fundament of exploitation of above mentioned phenomenon's [6].

7. Erich Christian Teppan, 2008

In this work author concentrates on two classes of phenomenon's, which are decoy effects and serial position effects. Tightly coupled to these phenomenon's is the problem of getting the utility function of a recommender right, as this function serves both as the basis of result set calculation as well as the fundament of exploitation of above mentioned phenomenon's. Putting all these aspects together an extended architecture for recommender systems have been performed[7].

8. A. S., and Yang, Y. 2008,

In this work author report on Presented work to date concerning the development of a course recommender system for University College Dublin's on-line enrolment application. Author outlines that influence student choices and propose solutions to address some of the key considerations that are identified. Author empirically evaluates Presented an approach with historical student enrollment detail and show that promising performance is achieved with Presented initial design.

III. CONCLUSION

Twitter and other micro-blogging platform are major resources of communication in social media. They opt out a number of features and thought sharing model in between the available user and their interaction. In this survey paper, different approach for the recommendation system and web mining is described which gives their best over algorithm. An existing base paper approach TF-IDF algorithm is also discussed which exhibit proper result in finding similarity measure in between the input tweets. A further discussion on limitations in them and further approach which can opt out for finding similarity measure in hashtags to make more appropriate and reliable system is given by us. Thus a further work is going to exhibit in finding the accurate approach with more enhance parameter in similarity measure between hashtags available in twitter dataset. Microblogging and their similarity measure approach with high accuracy, precision, F-measure, and MAE is can be performed in further research work.

REFERENCES

[1]. Mir Saman Tajbakhsh, Jamshid Bagherzadeh, "Microblogging Hash Tag Recommendation System Based on Semantic TF-IDF", IEEE 2016 4th International Conference on Future Internet of Things and Cloud Workshops.

[2]. Otsuka, E., Wallace, S.A., and Chiu, D.: 'Design and evaluation of a Twitter hashtag recommendation system'. Proc. Proceedings of the 18th International Database Engineering & Applications Symposium, Porto, Portugal2014 pp. Pages.

[3]. Givon, S., and Lavrenko, V.: 'Predicting social-tags for cold start book recommendations'. Proc. Proceedings of the third ACM conference on Recommender systems, New York, New York, USA2009 pp. Pages.

[4]. Godoy, D., Rodriguez, G., and Scavuzzo, F.: 'Leveraging Semantic Similarity for Folksnomy-Based Recommendation', IEEE Internet Computing, 2014, 18, (1), pp. 48-55.

[5]. Sigurbjornsson, B., and Zwol, R.v.: 'Flickr tag recommendation based on collective knowledge'. Proc. Proceedings of the 17th international conference on World Wide Web, Beijing, China2008 pp. Page.

[6]. Bart P. Knijnenburg, Martijn C. Willemsen, Alfred Kobsa" A Pragmatic Procedure to Support the User-Centric Evaluation of Recommender Systems", RecSys'11, October 23–27, 2011, Chicago, Illinois, USA. ACM 978-1-4503-0683-6/11/10 (pp 321-324).

[7]. Erich Christian Teppan," Implications of Psychological Phenomenons for Recommender Systems", RecSys'08, October 23–25, 2008, Lausanne, Switzerland. ACM 978-1-60558-093-7/08/10 (pp 323-326).

[8]. A. S., and Yang, Y. (2008). Personalized active learning for Collaborative filtering. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on the research and development in the information retrieval, Singapore (pp. 91–98). New York: ACM.

[9]. G. Adomavicius and A. Tuzhilin.Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transaction Knowledge Data Eng., 17(6):734–749, 2005.