International Journal of Scientific Research in Computer Science and Engineering
Survey Paper
Vol.6, Issue.3, pp.86-92 , June (2018)
E-ISSN: 2320-7639

Social Hash Tag Techniques Using Data Mining- A Survey

M. Vidhyalakshmi[1*], P. Radha[2]

[1] Department of Computer Science, Government Arts College, Coimbatore, India
[2] Department of Computer Science, Government Arts College, Coimbatore, India

Available online at: www.isroset.org

Received: 07/Jun/2018, Revised: 13/May/2018, Accepted: 21/Jun/2018, Online: 30/Jun/2018

***Abstract*—** The increase in reputation of microblogging utilities like Twitter has advanced to the enhanced use of content explanation approaches like the hashtag. Hashtags offer users with a tagging methodology to facilitate categorize, cluster, and generate visibility for their posts. This is an easy perception but can be tough for the user in order to perform which directs to rare usage. In this paper, a survey has been taken for various methods of recommending hashtags as latest posts are generated to encourage more extensive recognition and procedure. Hashtag recommendation appears with frequent disputes comprises processing enormous quantity of streaming data and content which is tiny and noisy. In this paper, an effective method of hashtag can be recommended along with the approaches applied to which the recommendation can be suggested.

***Keywords*—**Social Tags, News, Hash Tag Recommendation, Twitter Hash Tags.

I. INTRODUCTION

Twitter is a result of promising technology for social media platform which plays a vital role in the distribution of news. There are nearly 9 in 10 Twitter users utters that they utilize Twitter for news. Twitter has originated from, “daylong brainstorming session” which was the idea of an individual using an SMS service to communicate with a small group. The code service was previously known as, “twttr”. News spreading on Twitter has names that embody in the form of “hashtags” that milieu the stories. Tweets are openly noticeable by avoidance, but senders can restrict message delivery to their followers. Users can tweet via the Twitter website, compatible external applications (such as for Smartphone’s), or by Short Message Service. Twitter makes the users to modernize their profile, which is done either by their smart phones or by the app installed. Users can post by means of hash tags which can be prefixed with a “#” sign. Likewise, the “@” sign preceded by the username can be used to send a reply or to repost a message from another. In order revert back to one’s post, user can access “retweet” button within the Tweet. Trending topic in twitter refers to those words, phrase or topic that is mentioned at a greater rate than others. Popularity of trending topic depends upon the careful attempt by the users. This will make the users to talk about fastidious topics. Trending topic becomes very common in nowadays and these results in such a way to make the users understand what is happening in the world and what does people think over it. Twitter helps to work on trending topics in a better way by revealing the topics in the side bar of the home page. Twitter removes the trending topics with hash tags in a case when a topic is offensive.

Twitter Censors are broadly used to thwart over illegal tweets. Twitter works in a recovered mode for verifying the twitter account. Verification of the twitter account has become mandatory for ensuring the twitter account. A major advantage of this verification is, open verification is possible for the users; meanwhile allows to verify the more facts themselves. A blue tick mark followed by the user name indicates that the twitter account has been verified. Tweet is also possible for the users by forwarding the SMS. Twitter pioneered “Twitter Lite”, a progressive web app which has been developed for the provinces with slow internet connections, with a size of less than one megabyte and also for the devices with limited storage capacity. For enhancing and to provide better security mechanisms, twitter uses “OAuth”, an authentication mechanism which works in such a way that user doesn’t need to enter their password while entering the authenticated application. This has been widening in order to increase the security and to enhance the user experience. Added features of the Twitter will be resided at the permalink of the Twitter page, for the users to understand the working of the twitter. Twitter expands its services by combining partnerships for its streaming video services at the event. Twitter has ranked as a third most used social network which is based on the count of 6 million unique monthly visitors and 55 million monthly visits. Twitter has its impacts on, instantaneous short & frequent communication, emergency use, education, public figures, world leaders, religion. Any information can be easily reached out by means of the Twitter. Hashtag-written with a # symbol can be used to direct keywords or topics in Twitter. Hash tags are mainly used for the people to follow topics they are interested in. By means of this hashtag, people can easily search on specified topics and also just by tapping people can find out the other topics with the hash tag. The keywords are nothing but the words based tag, which directs the matter of a tweet and is the unique way to tweet for new stories. Hashtags materialize breaking news or developing news stories, and is the way to connect to a specified story or to a community. Hashtags mainly spots on the updates in real-time. For approval of any information, organizations use hashtags by which the target is set to support the content to readers. Even though many media has invented new hashtags, Twitter crew is one that regularly creates the hashtags and constrains many competing hashtags. Hashtags has unmitigated to television that has began in augment of importance. Broadcasters has demonstrated a hashtag as an on-screen bug in order to persuade the viewers for such discussions via social media. Hashtag bugs have emerged on the screen or at the end. The correspondents exploit by the hashtags along with their usernames in order to promote obtain reveal the posts, by using the "branded" hashtags along with the Twitter usernames. Hashtags can be used to systematize the real-world events and maintains the ad hoc lists for discussion and endorsement among participants.

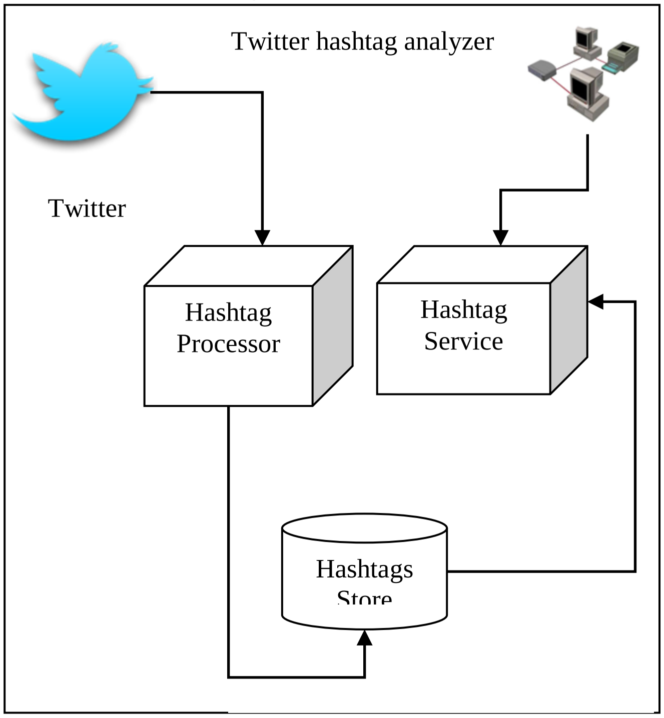

Figure 1 Working Structure of Hash Tag

The figure 1 shows the working steps of hash tag. The Hashtags can also be used as beacons for the sake of event participants which can be used to locate the arrangement either on the Twitter or some other physical events. Vital part of the hashtag is for encouraging a product, services or campaigns. Shadow banning which is also known as stealth banning, ghost banning or comment ghosting works in such a way in order to block a user or their content from an online community by which the user doesn't realize that they have been banned. In this survey, the outline and conclusion based on the exertions of hashtags in Twitter is specified.

II. LITERATURE REVIEW

Bourlai, Elli et al (2017) studied the working mechanism of Tumblr, which is a familiar content sharing websites that exerts in such a way that perceptible of posts is high-flying. Major part of Tumblr is that, there is no split up for comment section to the posts and so, tag section may also be used for comments. A comment mainly includes the way of articulating the reader's perceptions. Tumblr investigate the live out of comments in tags by considering the technological features of tumblr's tagging method and also part of communities within the website. This has been examined for variations subjected to use, structure and sentiment. This examination has resulted by demonstrating the comment tags and traditional keyword tags. Reaction has become one of the most fashionable chatting functions of comment tags; meanwhile this can be specifically eminent in quote, link, video and photo posts, and the musical acts. Tumblr tends to lug on sentiments but fewer to take in the special characters. Conversely, they are petite both the stipulations of characters and words than the tags expressing opinions.

Wei, Jianliang, FeiMeng, and N. Arunkumaret al (2018) proposed a method of recommending the personalized information of users. Social tagging of user's information has been discussed, by which the quality and quantity of user's authority can be calculated from a user co-incidence connection. This calculation can be done by means of "Degree centrality" which is considered as credence for tag voting. Finally, summation of tags from each user and their credence's are considered to categorize the collection. Recommendation of user's is done by the cosine similarity. Tags can be added to the authoritative users by coalesce of traditional ranking algorithms such as HITS and PageRank methodology by offering the higher quality. Contributions are done on three ideas namely, (i) Construction of tag-based and resource-based user networks, (ii) Merging user authority and user vote, along with the weights of tags added by users, which results in a accretion of resource model algorithm, (iii) Diverging the ratios between quality authority and quantity authority.

Andris, Clio, Xi Liu, and Joseph Ferreira Jr et al (2018) intended on the technique to built environment by means of social and interpersonal connections. In such cases, people oblige physical communications in order to meet and telecommunicate in order to colonize the infrastructures with movement and information dynamics. This can be formulated with the GIS analysis by which stroke is represented as a unit of spatial information termed as social flow directs the individuals to connect places through travel. It has been resulted that the flows differ from traditional spatial networks and distinct networks in operation systems. Extension of social flow data includes the mechanism of formal definitions for data type, new-fangled topology crafting, tackling new issues and finally rehearing social distance as the appearance of social flows. Pertaining GIS technology will be healthier to lodge the complex systems.

Zappavigna, Michele, and J. R. Martin et al (2017) incorporated the methodology of social metadata, which plays a vital role in the facet of social media communication while, this has been slam correlated and heighten the values included in the twitter posts concerning about depression. This paper makes use of promulgating the user's feelings. This method doesn't require the users to directly communicate with others. This is possible by subjecting the excursive system, emphasize association, and also construe how specified values are located as mend able in the diffusive environment. This is possible by the direct enact of social relations that employs the hashtags. Meanwhile, the affordances tender the tweet unearthed and therefore, more mend able as a form of "searchable talk".

Impacts of Crowd computing, which is intended by Jabeur, Nafaâ, Ahmed Nait-Sidi-Moh, and SheraliZeadally(2017), elucidates the crowd computing technique by the social media. In order to enhance the revenues social media producers invest their innovative ideas by fascinating new user activities to emblematize contents and services. This can be achieved by the working mechanism which employs the united exertion of human intelligence and computer systems in such a way tackle the problems which will be thorny for the individuals to do by their own. Crowd computing has included the new mapping based methods which focus on seven characteristics namely goals, content, audience, computing platforms, usage, data, and Return of Investment. The ultimate concern was applied on Return of Investment. Pertaining techniques of crowd computing is done on the social media ecosystem.

Xu, Zheng, et al (2017) determined a fascinating concept of mobile phones usage i.e., Mobile crowd sensing is explicated. This can be accomplished by the mobile devices to form participatory sensor networks, which focus on the carving up of local knowledge by their sensor-enhanced devices. Mobile crowd sensing techniques deals with the inclusion of social sensors, social sensor receiver platform, and mobile crowd sensing paradigm, which compiles the operational methods of physical sensors present in mobile devices namely GPS, by which conjuring of social relationships and human activities is possible. Mobile crowd sensing applications are promising and are possible by three sections such as public security, smart city, and location based services. Most applications are fit to one of the factions by which social sensors and social receiver platforms can be functioned.

After analyzing social media data in précised, Nguyen, Hoang Long, and Jai E. Jung (2018) said that it is made by utilizing the performances for social media intelligence. This can be achieved by applying a simple topology of social media analytics for enterprises and also thrashes out various analytics methods for social media data. Behaviors of real-world consumer review data will be considered. Social media analytics has been applied to scrutinize and eavesdrop to the word-of-mouth that extends in social media platforms and analysis is done based on consumer opinions on products and services. Examination is done on real and non-real time consumer analytics along with real and non-real time competitive analytics. Four challenges were recognized in the social media analytics. Social media analytics has become a part of huge business data analytics, which engrosses the non-social Meta data for entire business astuteness.

To understand internet of social knowledge, a concept of SocioScope has been implemented in a research paper of Sreenivasan, Sameet, et al (2017).This works in such a way by gathering information on social data from various resources in a proficient manner. This paper deals with the method of investigating the amalgamation to the system there by, augments the social data. This mechanism includes the risk of discerning the hidden patterns among the social data along with the process of comprehend our society. Authors had employed the SocioScope as a scaffold for shrinking the endeavor time with the chores such as collecting data, pre-processing data and analyzing data.

As Stai, Eleni, et al (2018) proposed a study on the effect of the scope of user consideration on the possibility of cascade has taken onto account. Cascades tend to be viral with the process for different concentration spans and different possibilities. Both the forwarding probabilities and the attention span have an inverse relationship. Branching factors are considered for the assessments of message generation rates, message forwarding rates and attention spans. By the divergent of hashtags in Twitter it has been analyzed and proved that the estimated branching factors co-relates well with the cascade-size distribution.

The paper of Van Deursen, Davy, et al (2010) absorbs the Information diffusion a promising method for the temporal dynamics. This deals out with the topics enveloped in Twitter. Trending topics are communicated by means of hashtags. The ultimate goal of this paper is to concentrate on the methodology of information spread in Twitter. Along with this, authorization of real data with numerous hashtags is considered. This exerts in the method of considering the superset of Twitter abusers who have seen/produced/reproduced tweets with a specific hashtags. Examined results proved persuaded results for all the hashtag categories. This method has its impacts on several factors specifically time-varying infection rates that mainly depends on the hashtag type. The table 1 shows the overall comparison between the techniques associated with the hash tag.

Table 1 Hashtag Generation and Recommendation Techniques Comparison Table

Paper ID	**Technique**	**Advantages**	**Disadvantages**
1	Linear geographic feature	Method of how humans, information and thoughts spread between and within places. GIS analysis with the evidence of individual's decision connects places through travel, telecommunications.	Accuracy and listens to the user's needs instead of providing a fixed path that the user cannot adapt.
3	Tumblr	Enhances search ability and visibility of posts, Analysis differences regarding use, structure and sentiment of comments.	No separate area for comments and so along with tags, comments has to be posted.
5	Fierce competition by combining human intelligence with computer systems applying on social eco-system	Solves problems that individuals can do alone, layer-based model applied specifically on Return of Investment (ROI).	Coordination of human and machine forces using machine learning techniques.
6	Analytics of social media on social meta data	Views on customer and competitive analysis on both real and non-real time methods. Deals with sentiment analysis, social network analysis, statistical methods, and image and video analytics	Limitations on improving supply chain efficiency and effectiveness and for the direct competitors.
7	Working methodology of SocioScope by collecting social data from multiple sources	Automatic structure in reducing collecting data, pre-processing data, and analyzing data	Other analyzing features (e.g., transformation, denoising, and feature extraction) has to be implemented for better processing of social data.
8	Cascading process applied for different spans and forwarding probabilities	Feed-based networks for dealing the finite attention span of users, message generation rates and message forwarding rates, branching factors correlate well with the cascade-size distributions.	Must focus on empirical data obtained from Twitter.
9	Temporal dynamics of information distribution	Considers as informed a superset of Twitter users who have seen/produced/reproduced tweets with a specific hashtags.	Concentrate on time-varying parameters of the proposed realistic epidemic model, exploring their fit for properly addressing context based applications.

11	Information recommendation, Degree centrality	Defines user authority in tagging system, combining user authority and user vote, differing ratios between quality authority and quantity authority.	Improvement of authority calculation method & analyzing the resource quality.
12	Crowd sensing applications on social sensors based on social receiver platform.	Share local knowledge acquired by their sensor-enhanced devices, application adapts to one of these categories namely public security, smart city, and location based services.	Local analytics, data storage in database, standardization of sensing interfaces, data delivery.
13	Convoking communities of feeling around people and employing the discursive system	No need of any interactions directly, Verbal and non-verbal communications are negotiated, paves way for three major functions namely convoking, Extreme performance is notified.	Discrete conversational turns cannot be assumed

III. RELATED WORKS

3.1 Learning-to-Rank [L2R] approach

Learning to rank also termed as machine-learned ranking (MLR) has been applied in machine learning [11] in order to administer or to strengthen learning methods which is the vital part in the edifice of ranking models for the information retrieval systems. Training data comprises of list of items possessing some "partial order" which is given among the items in each list. Order in ranking can be given as a numerical or ordinal score or a binary judgment for each item. This rank can be given by mentioning either relevant or irrelevant values. The ultimate goal of the ranking model is to rank, which is resulted by fabricating the permutation of items which is presented in new, unseen lists which is termed to be analogous to rankings in the training data. An efficient L2R algorithm works in real-time brook environments. The choice of L2R modeling in Hashtagger+, in disparity to multi-class classification (MCC) modeling, which enables us to concentrate on the following disputes:

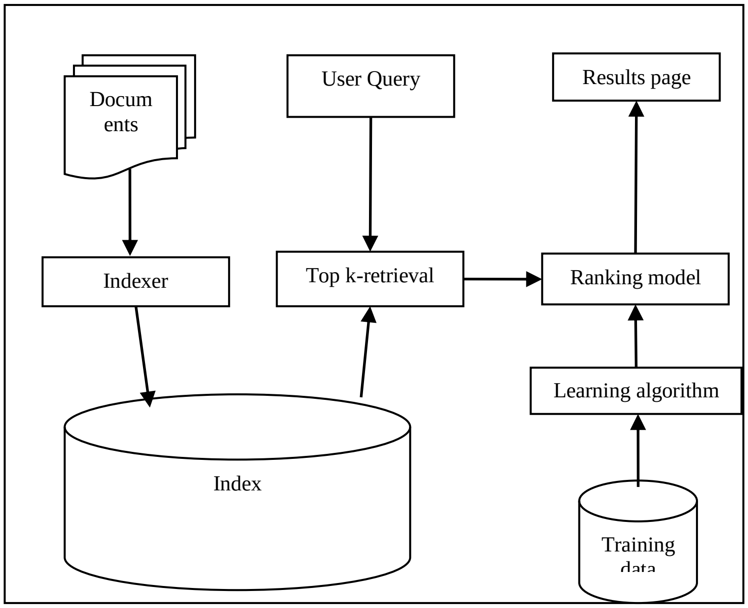

Figure 2.0. Learning-to-Rank [L2R] approach

– Many Classes: In MCC modeling, training data is detached based on hashtags (where a hashtag is interpreted as a class), and a model is trained for each single hashtag. As there are thousands of hashtags rising every few minutes, we need to instruct thousands of models. Furthermore, by dividing the obtainable training data, MCC needs to collect enough labeled data for each class. In our L2R model, we can use all the labeled data to train a single consequence model.

– Energetic Classes: Hashtags are very forceful, thus if modeled as classes, previously qualified models will not be helpful for predicting on new data, since old hashtags are discarded and new hashtags appear quickly. This means an MCC model has to be reinstructed often, while our L2R model does not need reinstructing.

– Perception Drift: The procedure and significance of a hashtag may change over time, thus its content profile will be exaggerated by perception drift. This means pre-instructed MCC models will not work well if the importance of the hashtag amends in new data.

3.2 Multiclass classification

In machine learning, multiclass or multinomial classification defines the problem in pigeonholing occurrences into one of three or more classes, whereas categorizing occurrences into one of the two classes has termed to binary classification [12]. In such cases classification algorithms naturally allow the access of more than two classes. Conversely, some of them can also be revolved into multinomial classifiers by assortment of tactics. Classification can be achieved by trying to categorize tweets in classes which can be defined by the hashtag clusters. Spontaneously, the classification algorithm will imply what hashtags would contain. Because of clustering the hashtags, implications doesn't include what hashtags the tweet may restrain, but will imply on which cluster the hashtag will actually fall. Ultimate goal of classification algorithm is to identify the topic of tweets by means of classification. Allusions to hashtags are broadly classified based on four categories namely, (i)Description of feature vectors and classes, (ii) Classification using Dimensionality Reduction, (iii) Classification using SVM algorithm optimized for sparse vectors, (iv) Majority Vote classification.

-Description of feature vectors and classes: Regards each tweet a document and use standard document classification techniques in this step. This method will generate a set of all unique words (removing common stop words and all hashtags) that occur in all the tweets. While creating the training data, if a tweet has multiple hashtags, we regard as that tweet to be a part of multiple classes and so we add the same training point multiple times with different classes as the intention. Figure 3 shows the multiclass classification technique and its working process.

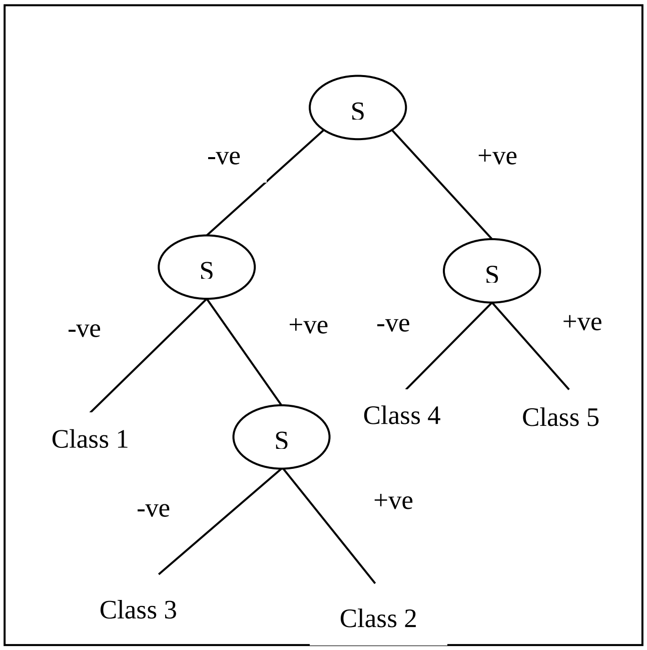

Figure 3 Multiclass classification

- Classification using Dimensionality Reduction: This is achieved by applying the PCA tools from the Python library scikit-learn in order to diminish the dimensionality in the data. This is possible by the concept of scikit-learn's implementation of Bernoulli naive Bayes, which duals the data by replacing all positive values in the feature vector with the value 1 and all negative values with the value 0.

- Classification using SVM algorithm optimized for sparse vectors: An unconventional approach to covenant with the high dimensionality of the feature vectors is to exploit an algorithm that was designed to use the sparse representation of a vector. By representing the feature vectors as a sparse vector, the problem becomes biddable and we don't need to do any dimensionality declination. This SCM classification can be subjected on the PCA-reduction data.

- Majority Vote classification: The majority vote classification accuracy can be utilized as a baseline measure. The majority vote accuracy is defined as the percentage of data points whose class label is the most common class label.

3.3 Fragment Identifier

A fragment identifier can be termed as a short string of characters which means the contributions that is subsidiary to another, chief source [13]. This chief source can be recognized by means of Uniform Resource Identifier (URI), along with the fragment identifier which summits to the subsidiary sources. Fragment identifier has been invented by a hash mark (#) containing a URL at its last part. This can be used to discover the segment of that document. The hash mark separator in URI's doesn't belong to the fragment identifier. The four categories of fragment identifier lies onto, (i) A fragment URL stipulates a location within a page, (ii) Fragment URL's cannot be parsed as a request, (iii) Anything after a first # is the fragment identifier, (iv) Changing the fragment Id doesn't reload the page, but will create a history.

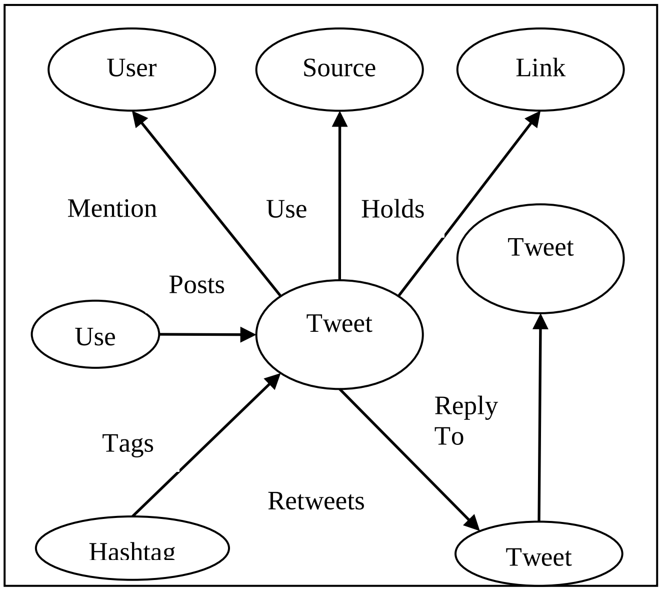

Figure 4 Fragment Identifier

The figure 4 describes the basic functionalities of the fragment identifier in the hash tag identification system. The functionalities of fragment identifier reclines on, * Particular syntaxes for symbolizing fragments in text documents by line and character series, or in graphics by synchronization, or in structured documents using hierarchy, are appropriate for standardization but cannot be defined. * The fragment-id pursues the URL of the entire object from which it is divided by a hash sign (#). If the fragment-id is null and void, the hash sign may be omitted: A void fragment-id with or without the hash sign means that the URL refers to the entire object. * While this hold is permitted for specifying the fragments, the difficulty of tackling the parts of objects, or of clustering the objects and relationship flanked by continued and containing objects, aren't be considered by this document. * Fragment identifiers do NOT deal with the query of objects which are different versions of a "living" object, nor of articulating the affairs between different versions and the living object. * There is no proposition that a fragment identifier pass on to anything which can be haul out as an object in its own right. It may, for example, refer to an inseparable point within an object.

IV. CONCLUSION

Hash tag mechanism is a most popular approach in twitter. With the help of data mining techniques, the hash tag creation and recommendation makes the application better. Hash tag techniques are widely used in the story telling and news related applications. The paper provides a brief introduction about hash tag techniques in twitter and the recent data mining techniques to solve the issues of hash tag recommendation. From the analysis and comparison, many techniques such as L2R and multi class classification approaches having more future directions. The selection of dataset and the application will play a unique role in future research. Comparison of various other technologies has been made. The improvement on such techniques will lead to a successful hash tag recommendation application.

REFERENCES

[1]. Andris, Clio, Xi Liu, and Joseph Ferreira Jr. "Challenges for social flows." Computers, Environment and Urban Systems (2018).
[2]. B. Shi, G. Ifrim, and N. Hurley, "Learning-to-rank for real-time high-precision hashtag recommendation for streaming news," in Proc. 25th Int. Conf. World Wide Web, 2016, pp. 1191–1202.
[3]. Bourlai, Elli E. "'Comments in Tags, Please!': Tagging practices on Tumblr." Discourse, Context & Media (2017).
[4]. Hong, Liangjie, Ovidiu Dan, and Brian D. Davison. "Predicting popular messages in twitter." Proceedings of the 20th international conference companion on World Wide Web. ACM, 2011.
[5]. Jabeur, Nafaâ, Ahmed Nait-Sidi-Moh, and SheraliZeadally. "Crowd social media computing: Applying crowd computing techniques to social media." Applied Soft Computing (2017).
[6]. Lee, In. "Social media analytics for enterprises: Typology, methods, and processes." Business Horizons (2017).
[7]. Nguyen, Hoang Long, and Jai E. Jung. "SocioScope: A framework for understanding Internet of Social Knowledge." Future Generation Computer Systems 83 (2018): 358-365.
[8]. Sreenivasan, Sameet, et al. "Information cascades in feed-based networks of users with limited attention." IEEE Transactions on Network Science and Engineering 4.2 (2017): 120-128.
[9]. Stai, Eleni, et al. "Temporal Dynamics of Information Diffusion in Twitter: Modeling and Experimentation." IEEE Transactions on Computational Social Systems (2018).
[10].Van Deursen, Davy, et al. "Implementing the media fragments URI specification." Proceedings of the 19th international conference on World Wide Web. ACM, 2010.
[11].Wei, Jianliang, FeiMeng, and N. Arunkumar. "A personalized authoritative user-based recommendation for social tagging." Future Generation Computer Systems (2018).
[12].Xu, Zheng, et al. "Mobile crowd sensing of human-like intelligence using social sensors: A survey." Neurocomputing(2017).
[13].Zappavigna, Michele, and J. R. Martin. "# communing affiliation: Social tagging as a resource for aligning around values in social media." Discourse, Context & Media (2017).