A User Query-Centered Recommender System from Public Repository

[1]V. Kakulapati, [1]D. Vasumathi
[1]SNIST, Yamnampet, Ghatkesar, Hyderabad, Telangana-501301
[2]JNTUCE, Kukatpally, Hyderabad, Telangana
[1]vldms@yahoo.com, [2]rochan44@gmail.com,

Abstract: Query-based User Information Categorization and Extraction (QICE) methods allow the classical query extraction with its knowledge obtained from useful resources. Open Data encodes machine-readable user facts from different sources, including third-party, that play a vital role in this QICE. Mining techniques from documents available in free sources and constructing the user text based on the user query with its knowledge and analysis are the core research problems in the public repositories and query understanding tasks such as query pattern analysis and user information requirements. However, the public repositories encode the user information through Wikipedia and web pages which are static, and these do not understand the user requirement perspectives. These static web pages have many quality issues with user query information, such as information extracted, complete data representing, time of query retrieval, and correctness of the information categorized based on the user query. QICE methods are, therefore, facing problem user query variances and type of user query confusability. In this paper, a query recommender system proposes developing a technique for user-query-centered knowledgeable integration and addressing the challenges of knowledge mining of the Twitter social network by extracting the knowledge from query log data. The proposed User-Query Centered Recommender System (UQCRS) is applied to exploit different measures to demonstrate the efficiency of recommendations delivered. The proposed algorithm exhibits an effective result to the search shortcuts issues. And to provide a performance comparison of the proposed system, the comprehensive evaluation is compared with well-known methods and demonstrates the impact of the results.

Keywords:mining, public open data, extraction, user query, knowledge retrieval

1. Introduction:

Social networks provide a means for different types of users over the globe to establish communications with among them for their interest in technical and professional, mostly the user's interest [1] [2] [3] information requirements. There is a huge amount of information in data and messages received from different zones and types of users from diverse parts of this social network. Researchers have major challenges in designing a model to ascertain and provide a regularized traffic for these social network information flow through their capacity and conjunction. In this paper, on the Twitter social network, the analysis is made. Initially, the Twitter messages in the social network have arrived from the traffic considered communications based on the user's population in a certain message category. The process of Twitter tweet's analysis process through a precise pattern, a model-based approach is as follows:

1. This model makes the tweets regulate based on the pattern behavior, and the variations of the tweet's patterns are observing.
2. The tweet's patterns are estimating for the parameters based on the observed user tweet outcomes.
3. The observed outcomes of the tweets are collectively complete into possible events.
4. From these possible events, the precise pattern model from the observed outcomes is esteemed and generated as the model outcomes.

To develop such a precise pattern-based model method on the Twitter social network has to get characterized by its available data. For that, either a small or large big sample should be detected early on during the interactive period between the sending and receiving message. This is made possible through the relation between Twitter users. The messages are made public among the user or friends or shared among the Twitter users. This enables a large amount of data collection, i.e., tweets in less time from the Twitter users, making the researchers develop a precise pattern model-based approach to characterize the Twitter tweets.

Twitter tweets are about sixty million a day, of about 1,500 % since 2009. In the literature, the model-based approaches have mail or message-based [4] and content-based [5] methods, which could not recommend tweets arrival in social networks. Because of using open data repositories' unavailability and their design architectures, may not be modelled on future tweet statistics.

In this paper, tweets recommender system architecture demonstrates user tweet data analysis through the collection, detection, and extraction to develop such precise pattern models of user Twitter information. This

paper focuses on the knowledge-based approach, where the tweets and their linked tweets are extracting from the log data. Through Twitter prevision through knowledge keyword, the tweets are collected and correlated because the tweets from different web pages are differing. So, the data collection of tweets in the proposed system is based on the Twitter APIs through Python [6] with MySQL [7].
The proposed method contains:

1. Mining the Twitter background knowledge from the public repositories by the tweet's quality issues and adopting the QICE technique to extract this tweet's knowledge by maximizing the User-Query Centered Recommender System.
2. Perceive and execute Twitter Information Depository strategy for the built framework on tweets through the QICE portion.
3. The User-Query Centered Recommender System evaluation and illustration through the discussion of the presented system and its importance on the QICE is evaluating.

In this work, the proposed method is organizing as section II describes the user-query analysis of the related works. Section III outlines the proposed work and the collected Twitter content analysis. Section IV is describing our implementation results and discussions.

2. Related Works

Mining methods workflow contains a group of mining data and models, with an utmost data operator work to set the parameters of the mining model used. The data is not articulating an indirect form in mining, but it is unseen in the model connectors. The user provides the indirect form of data and applies a model on the indirect form of data, generating the direct data formation. During this process, mining techniques should distinguish between components: data model, operators, and parameters. In order to enable user design, include web mining, there is a need to develop online data workflow through concepts and categories. Online data refers to a frequent visitor group or several web pages in social networks to cover and gather the complete user-required information by locating the web page and fetching the desired user valid information. Web pages are application-specific to fetch the user-selected target information through the user-defined key-words using a user-constrained specific web application to provide up-to-date information through the Online Social Networks (OSN). OSN is protecting billions of active and passive web user's knowledges. The rapid change in social networking sites has proven exponential growth in user information and knowledge exchange rate. As per (8), the social network or other e-commerce portal is browsed by two-third of online users, with an average of 10% of their Internet time. By covering such a large amount of useful information exchange, OSNs through social media become a great platform for mining techniques and research in the context of data analysis.

Twitter is a huge amount of information social network, to perform an analysis on Twitter, a keyword-based search for possible and relevant posts [9], where such search keywords cover all the likely tweets of the user [10], which is a lengthy and time-consuming process. Typically, to reduce the complexity of searching the posts from Twitter data sources, a user search keyword identification is made [11] to reduce the manual effort. User search keywords extraction is made based on the target keywords instead of the general word phrase of the keyword selected. This keyword extraction process is iterative. The regular user interaction in the social network through Web search and advertising requires a query recommendation and query expansion system to improve the keyword extraction process from Twitter tweets and provide recommendations to Twitter users. The keyword's frequency statistics and machine learning models are recommending [11] [12] [13] to classify the Twitter tweets keywords and to extract them from the Twitter tweets. Search based dataset is given in [14] to find the keyword topics and search the keywords in the dataset, but for the huge tweets, this method results from relevant tweets or empty tweets because of the huge Twitter dataset. Therefore, keyword recommendation in search based through query suggestion is recommending [15] [16]. In keyword recommendation, the query system is designed based on the relevant keywords in the Twitter tweets through the query log mining and search query suggestions [17]. The query expansion [18] from the original query expands and improves query ranking for the searched tweets through query suggestions.

The most common types of recommender systems are Collaborative Filtering and Content-based recommender systems. Based on the properties and descriptions of items and user interest information, the content-based recommender system is our subject of interest. In the Twitter recommender system, the basic nature of Twitter tweets is noisy and with less content for understanding because of the exact use of posting users [19]. Construction of actual, the relevant content of the recommendation system with the account authority is considered in [20], with the learning to the rank algorithm is considered. The relevant content is a similar type of information retrieval [21] using the tweet contents posted by the user and user friends, which provides

recommended set of tweets to the user. Tweets to the recommendation identification scheme [22-23] are selected to form the question in the query base system for the substance of the Twitter application.
The main challenges in the Twitter content query-based recommendation are:

1. Retrieving the collection of Twitter tweets that match with one or more content keywords of user-query.
2. Ranking the query within the text.
3. Develop a method for user-query centered knowledgebase integration.
4. Predict the outcome of the automatic query-analyzer of Twitter tweets concerning the recommendations.

The above challenges can be solved using query expansion and semantic models. In query expansion, the reformulation of the query is made based on the vocabulary mismatch among the query and content retrieved. Through semantic models, the similar words of the user query are retrieving.
In this paper, in comparative results, semantic models are compared with the proposed model through query expansion analysis. In this paper, a framework of the User-Query Centered Recommender System is proposing. The results show potential requirements to apply to the semantic models for the tweet's analysis tasks: query expansion and tweet's verification.

3. Proposed Recommender System

In this paper, UQCRS proposes a concept of the recommendation system at the user-query level that mainly aims to find the right tweet's information extraction through content to user requirements. Unlike the prior works of the recommendation system, the proposed UQCRS is a system that can perform Twitter mining from a huge Twitter database through query logs and user tweets to understand user interaction in Twitter tweets. UQCRS provides the search-based tweets content recommendation through the found user-query-centered content in the short tweets and long tweets to depict user tweets. In the proposed system, the workflow is: firstly, the Twitter background knowledge retrieves for user-query-centered knowledgebase integration. Secondly, implementing the strategy of UQCRS on the Twitter knowledge repositories. And finally, evaluation and illustration through the discussion of the proposed URCRS system are made.

3. A. Content-based Twitter Tweets Detection:

Twitter users communicate with the recommender systems through a user interface, a web portal, or a mobile app. Based on user availability, the user interacts with the social network to extract the information in the tweets on user interest, which predicts by the tweets ranking method to provide the list of proposed tweets based on user content query. The UQCRS data system is creating on the database. It stores and updates the tweets based on the content and ratings of tweets through the query search. UQC recommendation system on tweet content architecture is formed on the profile and database of user twitter profile that store query information and updates the entities continuously through Twitter user customer recommendation is as represented in fig 1.

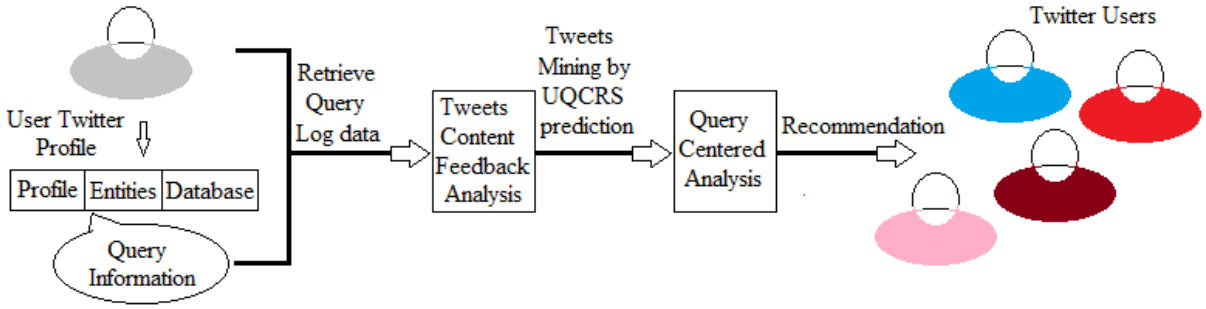

Fig 1: Architecture of the UQCRS implementation

With the content feedback and query-centered analysis, recommender systems are implemented in Figure 1, used within the e-commerce websites, to guide the Twitter customers by retrieving log data. Twitter mining originates from the users themselves.

The tweets clustering process and the filtering of tweets are performing, shown in fig 2, of the tweets given by the user recommendation, which analyses at every instant of user interaction.

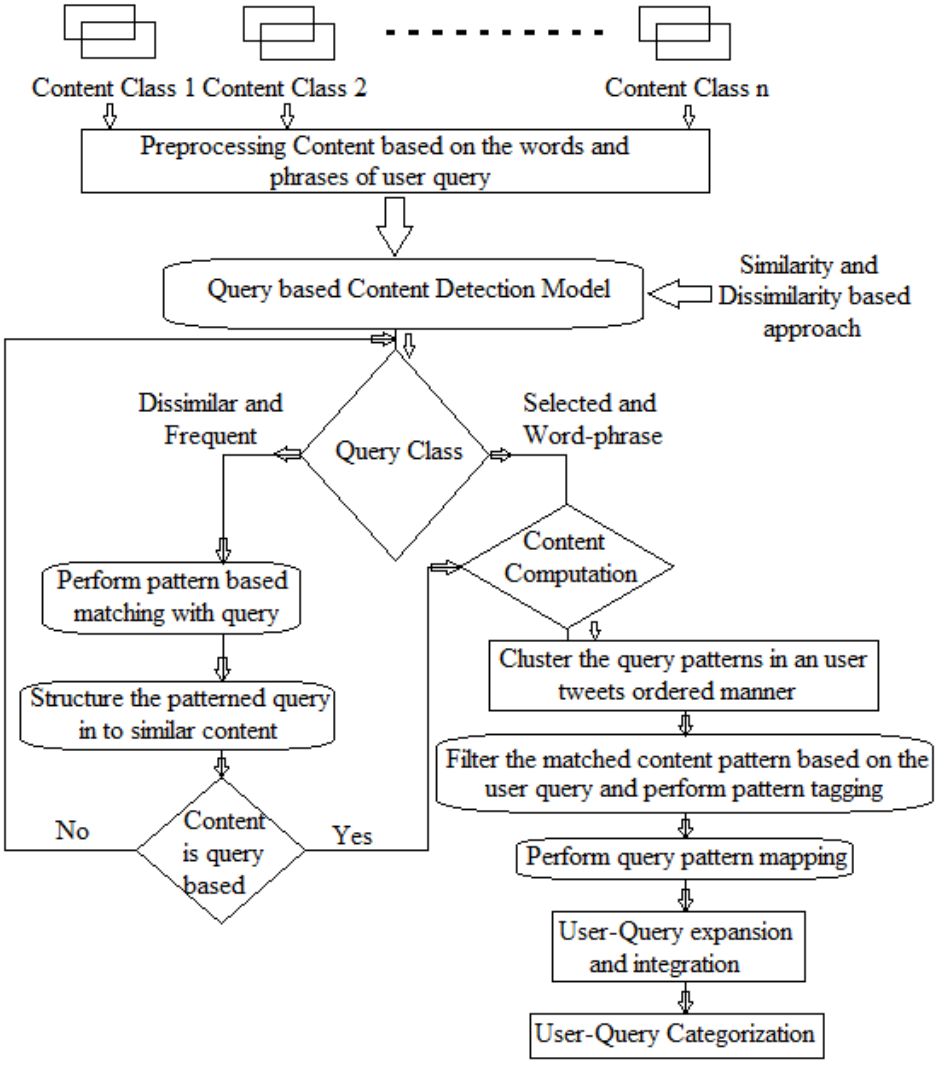

Fig 2: User-query pattern categorization

In the proposed system, each analysis is a service, and the operation of the Content Detection Model algorithm is explaining in fig 3, which is defining in two parts. The first content phrase built as a final set of query phrases per each query log brings together many ordered patterns that make each filtered pattern generate enough to the matched tweets. In the second, query mapping is intending, which shares similar tweets during consecutive query logs, exceeding the maximum ordered pattern.

Algorithm 1: Content Detection Model
input : *set of content class* $C[t_i]$ *for query log* q_i
output : *set of similar tweets* T_s
T_s *= Null*
index.pattern($C[t_i]$*, p), p is the content phrase*
for *all non-empty class* $C_{x,y} \epsilon$ *index do*
$S_r = C_{x,y}$
$C_s=[C_{x-1, y-1}, ..., C_{x+1, y+1}]$
if $\lvert C_s \rvert \geq \tau$, τ *is the ordered pattern*
for *each* $s_r \epsilon S_r$ *do*
f*=Filter (*s_r*, p),* $\lvert f \rvert \geq \tau$*, map (*$s_r$*,* c_s*)* $\leq p$*,*
$c_s \epsilon C_s$
for *each* $c_i \epsilon f$ *do*

```
if {s_r, c_j} is matched then
        compute query mapping
        {q_1, q_2,..q_n}
        given by q_i& number of p
for each query mapping
        q_k ϵ {q_1,q_2} do
            q = q_k Ω f
if |q| ≥ τ
            then T_s.Add(q)
end
end
end
end
end
end
```

Figure 3: User Twitter Tweets Detection.

3.B. User-query centered knowledgebase integration:

User query analysis is complete by extracting knowledge from query log data, shown in figure 4. Here, the scenario is that the successive user tweets are retrieved based on the query log data, and through subsequent extraction, the matched query gets into correspondingly. The accessed query content related to a Twitter account focuses on extracting the respective accounts of the tweets account. The latter tweet's accounts focus on the generic tweets with similar tweet content, having a similar phrase of tweets query. But during relevance query knowledge, similar user query search with less effective, and provide user tweet with high similarity with different text types.

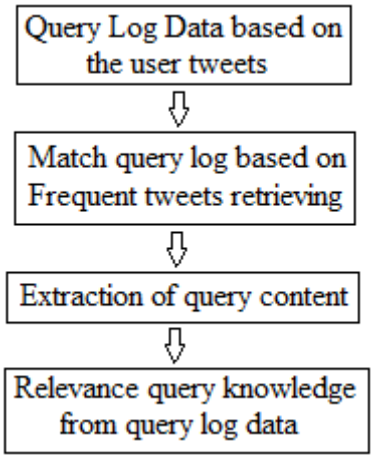

Figure 4: Extracting the knowledge from query log data.

The scenario of user-query knowledgebase construction is showing in figure 5. Here, from the previous systems, with the selected user-query method used as a category of tweet query, user integration history is proposed, with knowledge extraction at every query feedback history loop-back. For query integration, the projected tweets from one day to many days are structured for continuous knowledge construction through the constant tweet's attribution and trace the future knowledge analysis, which evolves in the future based on the tweeted user query.

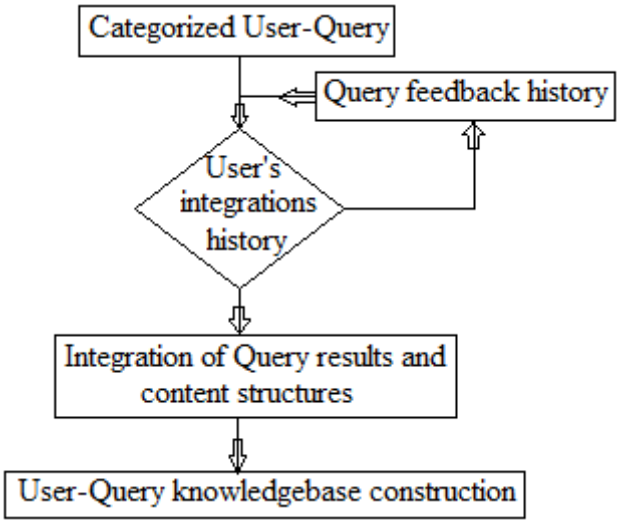

Figure 5: User-query centered knowledgebase integration

The model of the proposed recommendation system is shown in figure 6, consists of:

a. a set T_U of N users, $T_U = \{ T_U1, T_U2, T_U3,...., T_UN\}$
b. a set C of M items, $C=\{c_1, c_2, c_3,.....,c_M\}$
c. a query cluster matrix Q_C, $Q_C=[qc_{mn}]$ where $m \epsilon T_U$ and $n \epsilon C$
d. a set f of N feature query sets, $f=[f_{mn}]$
e. a tweet knowledge weights $K_\omega=\{\omega_1, \omega_2,... \omega_N\}$

The user-item set is associated with the number of feature vectors representing the tweet customers with different tweet phrases assigned to the user-query content model. The decision ranking prediction gives the order and decision of the similarity between the users, and Twitter queries in the categorized user-query item set and tweet's weights in the recommender content model.

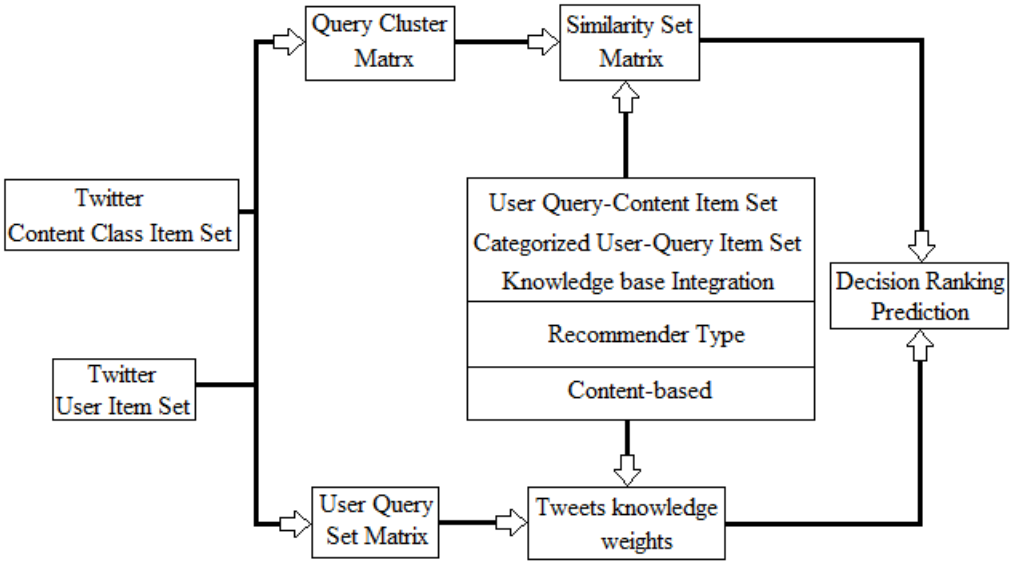

Figure 6: The proposed User-Query Centered Recommender System

4. Results and Discussions:

For experiments, a random public user Twitter dataset and real-time data using the API of Twitter are complete. The Twitter tweets containing the keywords "basket," "pencil," "work," "enter," and "formal"

from the general domain are taking as the standard bag-of-words approach. Used this dataset for classification and collected 300 documents each of the general environment.
The content detection and query analysis using Java SDK, WordNet Version 3.0, and Intel (R) Core i5-3230M CPU @2.60 GHz with 8 GB RAM the 64-bit Operating System. For content categorization, the Weka tool of version 3.6 is using.

For the classification of tweets, the true +ves, true -ves, false +ves, and false -ves constraints are utilized to equate the consequences of the classifier under the test with investigation techniques, which is illustrating in figure 7.

	Actual Class (Expectation)	
Predicted Class (Observations)	True +ve (TP) Correct Result	False +ve (FP) Unexpected Result
	False -ve (FN) Missing Result	True -ve (TN) Correct absence of result

Figure 7: Classification Matrix Model for metric analysis.

The relations between TP, FP, FN, and TN are:

a. The relation $TP/[TP + FP]$ is describing as precision, which is the correctly classified metric.
b. The relation $TP/[TP + FN]$ is describing as recall, which is the actual classified metric.
c. Relation $2 * [{}^{precision * recall}/_{precision + recall}]$ is describing as the F-measure, which is a measure of precision and recall.

Fig 8 shows the accuracy of the classification of a user query tweet in the users defined the recommended system. The highest accuracy is achieved through the proposed work, with a good number of word phrases, depending on the content of the user query and compared with the Naive Bayes classifier (NB) [24].

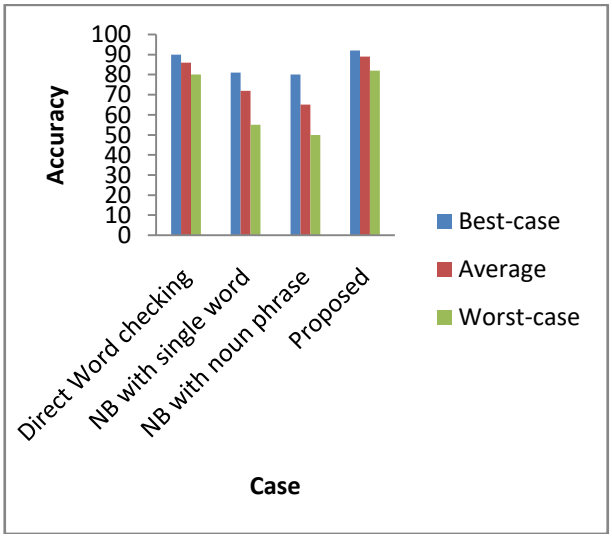

Fig 8: Accuracy Comparison of a classified tweet

Fig 9 compares the proposed system with different dataset approaches in terms of the F1-score and the exact match. Because of the phrase-based content mining is made and tweets analysis is made accurately on two other datasets of different methods [25-27] and proposed.

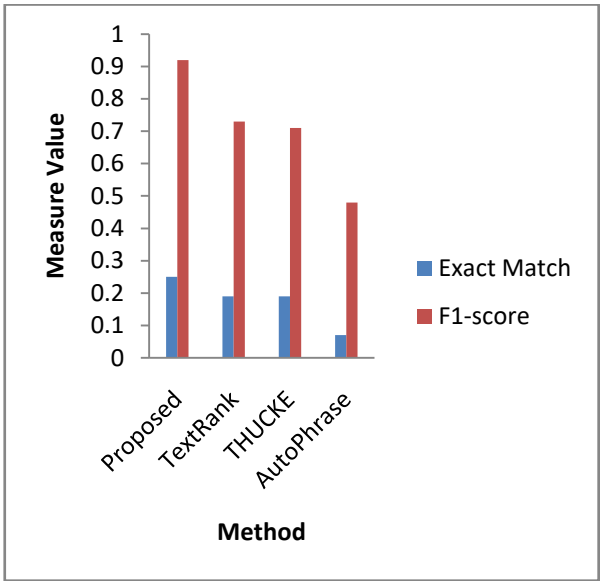

Fig 9: Comparison of different algorithms for measured values.

5. Conclusions:

In this paper, for Twitter tweets, a user query-centered recommendation system is designed to improve user query and tweets analysis. The proposed and implemented query categorization gives a satisfactory exact match performance in tweets categorizing. The content model is the most important model for the majority of the tweets. Apart from finding the tweets, the phrases are identified and located in an accuracy performance metric. As discussed, content integration is useful for tweet match retrieval. For the given user tweet query, if the aim is to retrieve similar tweets from the public repository, the accuracy of retrieving the keywords is improved with the use of words and phrases. Also, a novel algorithm based on content detection is using to extract the tweets using the bag-of-word's method. The proposed recommendation system avoids the dissimilar tweet pattern identification problem using the tweet's knowledge weights. The experiments on the above three parameters are complete, indicating that the proposed approach produces better accuracy results than the other methods.

6. Future enhancement

As future work, the recommendation system on user behavior and history-based tweet analysis can be complete. Users and their retweets are occupying as the source of behavior analysis of the Twitter tweets.

References:

[1]. [Online]. Available: http://www.linkedIn.com
[2]. [Online]. Available: http://www.facebook.com
[3]. [Online]. Available: http://www.twitter.com
[4]. R. D. Malmgren, J. M. Hofman, L. A. N. Amaral, and W. D, J, "Characterizing individual communication patterns," Proc., Intl. Conf. on Knowledge, Discovery and Data Mining (SIGKDD'09), Oct. 2009.
[5]. P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: A view from the edge," ACM SIGCOMM Conf. on Internet Measurement, Oct. 2007.
[6]. Python home page. [Online]. Available: http://python.org
[7]. Mysql home page. [Online]. Available: http://www.mysql.com/
[8]. Global Faces and Networked Places, A Neilsen report on Social Networking's New Global Footprint, March 2009. Neilsen company.
[9]. Turney, P. D. 2000. Learning algorithms for keyphrase extraction. Information Retrieval 2(4):303–336.
[10]. Zhao, W. X.; Jiang, J.; He, J.; Song, Y.; Achananuparp, P.; Lim, E.-P.; and Li, X. 2011. Topical keyphrase extraction from Twitter. In ACL, 379–388.

[11]. El-Kishky, A.; Song, Y.; Wang, C.; Voss, C. R.; and Han, J. 2014. Scalable topical phrase mining from text corpora. VLDB 8(3):305–316.
[12]. Danilevsky, M.; Wang, C.; Desai, N.; Ren, X.; Guo, J.; and Han, J. 2014. Automatic construction and ranking of topical keyphrases on collections of short documents. In Proceedings of MINING.
[13]. Zhao, W. X.; Jiang, J.; He, J.; Song, Y.; Achananuparp, P.; Lim, E.-P.; and Li, X. 2011. Topical keyphrase extraction from Twitter. In ACL, 379–388.
[14]. King, G.; Lam, P.; and Roberts, M. 2014. Computer-assisted keyword and document set discovery from the unstructured text—copy at http://j.mp/1qdVqhx 456.
[15]. L¨uke, T.; Schaer, P.; and Mayr, P. 2013. A framework for specific term recommendation systems. In SIGIR, 1093– 1094.
[16]. Bhatia, S.; Majumdar, D.; and Mitra, P. 2011. Query suggestions in the absence of query logs. In SIGIR, 795–804.
[17]. Zhang, Y.; Zhang, W.; Gao, B.; Yuan, X.; and Liu, T.-Y. 2014. Bid keyword suggestion in sponsored search based on competitiveness and relevance. Information Processing & Management 50(4):508–523.
[18]. Hahm, G. J.; Yi, M. Y.; Lee, J. H.; and Suh, H. W. 2014. A personalized query expansion approach for engineering document retrieval. Advanced Engineering Informatics 28(4):344–359.
[19]. K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical clustering of tweets," Proceedings of the ACM SIGIR: SWSM, 2011.
[20]. Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum, "An empirical study on learning to the rank of tweets," in Proceedings of the 23rd COLING, 2010, pp. 295–303.
[21]. M. Pennacchiotti, F. Silvestri, H. Vahabi, and R. Venturini, "Making your interests follow you on Twitter," in Proceedings of the 21st CIKM, 2012, pp. 165–174.
[22]. A. Pal and S. Counts, "Identifying topical authorities in microblogs," in Proceedings of the 4th ACM WSMINING. ACM, 2011, pp. 45–54.
[23]. J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic sensitive influential twitterers," in Proceedings of the 3rd ACM WSMINING, 2010, pp. 261–270.
[24]. Daniel Lowd, Pedro Domingos: Naïve Bayes Models for Probability Estimation. (2005). Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany
[25]. Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In EMNLP,
[26]. Zhiyuan Liu, Xin Xiong Chen, Yabin Zheng, and Maosong Sun. 2011. Automatic keyphrase extraction by bridging the vocabulary gap. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning. ACL, 135–144.
[27]. Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering 30, 10 (2018), 1825–1837.