We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,300

Open access books available

172,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Knowledge Extraction from Open Data Repository

Vijayalakshmi Kakulapati

Abstract

The explosion of affluent social networks, online communities, and jointly generated information resources has accelerated the convergence of technological and social networks producing environments that reveal both the framework of the underlying information arrangements and the collective formation of their members. In studying the consequences of these developments, we face the opportunity to analyze the POD repository at unprecedented scale levels and extract useful information from query log data. This chapter aim is to improve the performance of a POD repository from a different point of view. Firstly, we propose a novel query recommender system to help users shorten their query sessions. The idea is to find shortcuts to speed up the user interaction with the open data repository and decrease the number of queries submitted. The proposed model, based on pseudo-relevance feedback, formalizes exploiting the knowledge mined from query logs to help users rapidly satisfy their information need.

Keywords: Data Mining, Query, Public Open Data, Social Network, Knowledge Extraction

1. Introduction

SNS (Social networking services) is online services, platforms, or sites designed to support the development of Internet-based communities or community links between, for instance, individuals who often regularly interact with hobbies, experiences, and emotional interactions. SNS includes an account of every member and its community ties and a range of additional capabilities (typically a biography). A significant number of SNS are social media and allow consumers to communicate Online like e-mail and automatic messages. While SNS is often an individual-centered service in a broad context, social media facilities are a team. Social media platforms may be regarding constitute SNS. Online Communities enable individuals inside individual unique systems to exchange opinions, tasks, experiences, and goals.

Social networks communicate to interact in many innovative approaches, such as shows, hashtags, perform and engages electronically, revealing further cooperation and projected benefit that could scarcely imagine only a short time before. Online communities can play a significant role in the organizational processes as well as helping to develop company concepts and emotional responses and give up different prospects for the examination of social interactions and social behavior.

Presently, people rely on social media and its vast and diverse wealth and have progressively penetrated each human living area. Increasingly individuals prefer to engage valuable time on social media to develop a significant social entertaining community and again try to communicate with each other so often that the interaction around them is robust. POD repository analytics is perhaps a commonly used scientific and commercial approach for investigating the social media of interpersonal, organizational, and corporate links. The necessity for solid knowledge in DPO analysis has lately increased with ready availability to computational power and the rise of social popular social networking platforms such as LinkedIn, Twitter, Netlog, and more.

Twitter social network by study the contents of the tweets and the links between the tweets to extract knowledge from log data. By selecting buzzwords, began the 'Twitter review and then collecting all Twitter posts (Tweets) correlated to the keywords. It is a social-economical problem in India. Mining the query log based on social networks like Facebook, Twitter, etc. Study and address the discovery, access, and citation of POD repositories like Twitter data sets; and strengthening educational programs of academics of current and future generations specializing in such areas. This is an auspicious time for extracting useful information from social media query log data. Substantial efforts to decipher large amounts of data are steps towards complete search log records integrating POD repository analysis. From these data sets, we extract valuable knowledge.

The search log obtained by user actions with the Public Open Data (POD) database is an excellent data collection for improving its efficiency and the effectiveness of the online community. The data in the user input logs are gathering from individuals who communicate on online platforms. The search log assessment is complicating due to the variety of customers and diverse resources. As a result, numerous scientific articles written about query log analysis.

The word "data set" can also describe the data in a set of specifically relevant tables that correlate to a specific investigation or occurrence. Records generated by satellites testing hypotheses using devices aboard communications satellites are one instance of such a category. A data source is the standard measurement for data provided in a POD repository in the open data domain. Over a quarter a million datasets are gathering on the European Open Data platform. Alternative interpretations were presented in this area, although there is presently no accurate statement. Various difficulties (relevant data resources, non-relational datasets, etc.) make reaching a compromise more challenging. The utilization of query logs for knowledge discovery improves the speed of the POD repositories and improves the use of open data source capabilities.

POD repository analysis and mining for valuable extract knowledge from query log data. We perform on knowledge discovery, ML or similar approaches, challenges connected to pre-processing and model assessment, for data sets (web usage log files, query logs, collection of documents), and collaborative data (images, videos, and their explanations multi-channel handling data). We summarize the fundamental results concerning query logs: analyses, procedures used to retrieve knowledge, the outstanding results, most practical applications, and open issues and possibilities that remain to be studied. We discuss how the retrieved knowledge can be utilized to progress different social media class features, mainly its effectiveness and efficiency.

In addition, several concurrent inquiries of multiple distinct users are addressing by business social networks. The query stream has simultaneously been defining by a stop-time rate, making it impossible for the POD repository to generate massive query load times without over-sizing [1]. Web Search engines Query Logs Social

indirect form of data, generating the direct form of data. During this process, mining techniques should distinguish between components, are data model, operators, and parameters. Enable the user in designing such mining for the web. There is a need for the development of online data workflow through concepts and categories. Online data refers to a frequent visitor of a group or several web pages in social networks to cover and gather the complete user required information through locating the web page and fetching the desired user valid information.

Web pages are complete application-specific to fetch the user's desired target information through the user-defined keywords using a 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 an exponential growth in user information and knowledge exchange rate. According to [18], two-thirds of the online users browse a social network or an eCommerce website, with an average of 10% of all internet utilization time. By covering such a large amount of helpful information exchange, OSNs through social media become an excellent platform for mining techniques and research in data analysis.

The method allows data on social media user tweets for goods and commodities. In [19], a RES approach is recommending to give a level of precision compared to the previous approaches used in the tweet assessment of the consumer. Fasahte et al. [20] has presented the method to anticipate tweets by utilizing the Online Reviews dataset sorting procedures. It investigates the search engine extraction and training algorithm to collect data from the unstructured text in the available online content. In addition to the keyword-based evaluation, the data model on the Internet is connecting with complex searches. They utilized to locate tweets on various tweets while maintaining the surfing data operational inside the account location. The data collection, processing of data, and data sampling are all three aspects of tweets availability. Verma et al. [21] developed a dynamic analysis classification technique by implementing ML and evaluated the different variables in these learning approaches. A public repository response assessment technique [22] discussed huge data volumes on Twitter to create the emotional state of every message. Rosenthal et al. [23] describes the user opinion mining system used to extract similar users' views from the person's view using a moderate data analysis method. Ibrahim et al. [24] established an online emotional assessment that supplied many functional tweets out of interest to identify comparable personal data. The decision relates to extracting features, extensive conversion, and different recognition using machine learning techniques in many tweet solutions for the clustering techniques, correlating the query response pattern, relationship regulations for Twitter tweet extraction, and visualization in the Tweet API application.

The phrase retrieval from several texts is provided by [25–27] since the words should be user-specific, and the searching procedure should be preserving. Due to the powerful conventional method to all these, it is possible to investigate a method that relies on the recommendation, utilizes iteratively in the search engine and advertising searching.

Optimization techniques provide AI (Artificial Intelligence) and NLP (Natural Learning Processing) capabilities in order to deliver necessary assessed user suggestions interpretations in different networks/services of social network applications [28, 29]. Interface design such as mobile web apps permits various movies, cuisine, literature, YouTube, healthcare and more information related material. Films, culture, and entertainment are communal societies. Depending on the user's awareness of the material, the recommender system [30] has problems with confidentiality and protection. Thus, classic recommenders are becoming inevitable for current user ratings and Twitter posts to evaluate user-generated content [31, 32].

4. Results and discussions

For experiments, a random public user Twitter dataset and real-time data using the API of Twitter is complete. The Twitter tweets containing the keywords “basket,” “pencil,” “work,” “enter,” and “formal” from the public domain are taking as the standard bag-of-words approach. Used this dataset for classification and collected 300 documents in each of the public domains.

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**.

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.

Predicted Class (Observations)	Actual (Expectation) Class	
	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.

Figure 8.
Accuracy comparison of a classified tweet.

Figure 9.
Comparison of different algorithms for measured values.

c. Relation $2 * \left[\frac{precision * recall}{precision + recall} \right]$ is describing as the F-measure, which is a measure of precision and recall.

Figure 8 shows the accuracy of classifying a user query tweet in the users defined the recommended system. The highest accuracy is achieved through the proposed work, with an incredible number of word phrases, depending on the content of the user query and compared with Naive Bayes classifier (NB) [41].

Figure 9 compares the proposed system with different dataset approaches in terms of 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 [42, 43] and proposed.

5. Conclusions

Social media connects connected individuals in numerous ways. For instance, online people can interact, communicate, collaborate, and socialize, showing new types of integration and interaction that were difficult to imagine only a short period before. Online communities also play a significant role in entrepreneurship, influence company feelings, and concepts and create many opportunities for unique investigation of person and team performance. In this chapter, a user query-centered recommendation system is deliberate to improve user-query analysis and tweets analysis for Twitter tweets. 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 tweets. Apart from finding the tweets, the phrases are identified and located in accuracy performance metric. As discussed, content integration is helpful for tweets match retrieval. For the given user tweet query, if the aim is to retrieve similar tweets from the public repository, 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 bags-of-word method. Using the tweet's knowledge weights the proposed recommendation system avoids the dissimilar tweet's pattern identification problem. The above said three parameters are complete, which indicates that the proposed approach produces better accuracy results than the other methods.

6. Future scope

Future work focuses on The DTSR System may be extended along with additional user profiles such as film playlists, community groups, social media tweets, user emotion, user posts, and feature tweets to better the method recommended.

Author details

Vijayalakshmi Kakulapati
Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India

*Address all correspondence to: vldms@yahoo.com

IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

- [1] References: Wayne Xin Zhao, et al. "Incorporating Social Role Theory into Topic Models for Social Media Content Analysis," *IEEE Transactions On Knowledge And Data Engineering*, Vol. 27, No. 4, April 2015.
- [2] Metin Turan, et al. "Automatize Document Topic and Subtopic Detection with Support of a Corpus," *Social and Behavioral Sciences*, Published by Elsevier, DOI: 10.1016/j.sbspro.2015.02.373, 2015.
- [3] K. D. Rosa, et al. "Topical clustering of tweets," *Proceedings of the ACM SIGIR: SWSM*, 2011.
- [4] Y. Duan, L. et al. "An empirical study on learning to the rank of tweets," in *Proceedings of the 23rd COLING*, 2010, pp. 295-303.
- [5] M. Pennacchiotti, et al. "Making your interests follow you on Twitter," in *Proceedings of the 21st CIKM*, 2012, pp. 165-174.
- [6] A. Pal et al., "Identifying topical authorities in microblogs," in *Proceedings of the 4th ACM WSMINING*. ACM, 2011, pp. 45-54.
- [7] J. Weng, et al. "Twitter rank: finding topic sensitive influential Twitterers," in *Proceedings of the 3rd ACM WSMINING*, 2010, pp. 261-270.
- [8] Turney, P. D. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval* 2(4):303-336.
- [9] Zhao, W. X et al. 2011. Topical keyphrase extraction from Twitter. In *ACL*, 379-388.
- [10] El-Kishky, A.; et al. 2014. Scalable topical phrase mining from text corpora. *VLDB* 8(3):305-316.
- [11] Danilevsky, M.; et al. 2014. Automatic construction and ranking of topical keyphrases on collections of short documents. In *Proceedings of MINING*.
- [12] Zhao, W. X.; et al. 2011. Topical keyphrase extraction from Twitter. In *ACL*, 379-388.
- [13] King, G.; et al. 2014. Computer-assisted keyword and document set discovery from the unstructured text—copy at <http://j.mp/1qdVqhx> 456.
- [14] Luke, T.; et al. 2013. A framework for specific term recommendation systems. In *SIGIR*, 1093– 1094.
- [15] Bhatia, S; et al., 2011. Query suggestions in the absence of query logs. In *SIGIR*, 795-804.
- [16] Zhang, Y.; et al. 2014. Bid keyword suggestion in sponsored search based on competitiveness and relevance. *Information Processing & Management* 50(4):508-523.
- [17] Hahm, G. J.; et al. 2014. A personalized query expansion approach for engineering document retrieval. *Advanced Engineering Informatics* 28(4):344-359.
- [18] Global Faces and Networked Places, A Nielsen report on Social Networking's New Global Footprint, March 2009. Nielsen company.
- [19] Z. Tan et al. "An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle," *IEEE Access*, vol. 5, pp. 27211-27228, 2017.
- [20] U. Fasahte, et al. "Hotel recommendation system," *Imperial Journal of Interdisciplinary Research*, vol. 3, no. 11, 2017.
- [21] A. Verma et al. "A hybrid genre-based recommender system for movies

using genetic algorithm and kNN approach," *International Journal of Innovations in Engineering and Technology*, vol. 5, no. 4, pp. 48-55, 2015.

[22] H. Jazayeriy et al. "A fast recommender system for the cold user using categorized items," *Mathematical and Computational Applications*, vol. 23, no. 1, p. 1, 2018.

[23] Rosenthal, S., et al. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 502-518).

[24] M. Ibrahim et al. "Design and application of a multivariant expert system using Apache Hadoop framework," *Sustainability*, vol. 10, no. 11, p. 4280, 2018

[25] Turney, P. D. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval* 2(4):303-336.

[26] Zhao, W. X.; et al. 2011. Topical keyphrase extraction from Twitter. In *ACL*, 379-388.

[27] 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.

[28] H. Li, D. et al. "Disaster response aided by tweet classification with a domain adaptation approach," *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 16-27, 2018.

[29] H.-T. Zheng, et al. "A deep temporal neural music recommendation model is utilizing music and user metadata," *Applied Science*, vol. 9, no. 4, p. 703, 2019.

[30] Java, A., et al.: Why we twitter: understanding microblogging usage and

communities. In: *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. pp. 56-65 2007.

[31] Joachims, T., et al.: Accurately interpreting clickthrough data as implicit feedback. In: *Proc. of the 28th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'05)*. pp. 154-161 2005.

[32] Kwak, H., et al.: What is Twitter, a social network or a news media? In: *Proc. of the 19th Int. Conf. on World Wide Web (WWW'10)*. pp. 591-600 2010.

[33] Liben-Nowell, D., et al.: The link prediction problem for social networks. In: *Proc. of the 12th Int. Conf. on information and knowledge management*. pp. 556-559. CIKM '03, ACM, New York, NY, USA 2003.

[34] M. Gao et al. A movie recommender system from tweets data, January 2019, http://cs229.stanford.edu/proj2015/299_report.pdf.

[35] M. Ibrahim et al. "Design and application of a multivariant expert system using Apache Hadoop framework," *Sustainability*, vol. 10, no. 11, p. 4280, 2018.

[36] M.-Y. Hsieh, et al. "Building a mobile movie recommendation service by user rating and APP usage with linked data on Hadoop," *Multimedia Tools and Applications*, vol. 76, no. 3, pp

[37] Muhammad Ibrahim, et al. " A Neural Network-Inspired Approach for Improved and True Movie Recommendations," *Hindawi, Computational Intelligence and Neuroscience*, Volume 2019, pp-1-19.

[38] O. Loyola-Gonzalez et al. "Fusing pattern discovery and visual analytics approach in tweet propagation,"

Information Fusion, vol. 46, pp.
91-101, 2018.

[39] Phelan, O. et al.: Using Twitter to recommend real-time topical news. In: Proc. of the 3rd ACM Conf. on Recommender Systems (RecSys'09). pp. 385-388 2009.

[40] S. Kumar, et al. "Movie recommendation system using sentiment analysis from microblogging data," 2018, <https://arxiv.org/abs/1811.10804>.

[41] Rada Mihalcea et al. 2004. Textrank: Bringing order into text. In EMNLP,

[42] Zhiyuan Liu, et al. 2011. Automatic keyphrase extraction by bridging the vocabulary gap. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning. ACL, 135-144.

[43] Jingbo Shang, et al. 2018. Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering 30, 10 (2018), 1825-1837.