International Journal of Engineering & Technology, 7 (4.36) (2018) 471-474

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Efficiently Identification of Misrepresentation in Social Media Based on Rake Algorithm

Prof. Devendra P. Gadekar[1]*, Dr. Y. P. Singh[2]

[1]Research Scholar, Department of CSE Kalinga University Raipur, India
[2]Professor, Department of CSE Kalinga University Raipur, India
**Corresponding author E-mail: devendraagadekar84@gmail.com*

Abstract

The social organizations offer an extensive variety of extra data to advance standard learning calculations; the most difficult part is separating the applicable data from arranged information. Fake conduct is indistinctly disguised both in nearby and social information, making it considerably harder to define valuable contribution for expectation models. Beginning from master learning, this paper prevails to efficiently join interpersonal organization impacts to identify misrepresentation for the Belgian legislative standardized savings foundation, and to enhance the execution of conventional non-social extortion expectation undertakings. Finding the semantic reasonable subjects from the colossal measure of rational points from the substantial measure of User Generated Content (UGC) in online networking would encourage numerous downstream uses of shrewd processing. Subject models, as a standout amongst the most effective calculations, have been broadly used to find the inactive semantic examples in content accumulations. In any case, one key shortcoming of point models is that they require archives with certain length to give dependable measurements adversary producing intelligent themes. In Twitter, the clients' tweets are for the most part short and loud. Perceptions of word events are immeasurable for theme models. The RAKE algorithm shows better performance than TextRank, Supervised Learning.

Keywords: *UGC, RAKE algorithm, TextRank, Supervised Learning*

1. Introduction

The important techniques of the data mining have under development for pervious ten year; in innovation regions are AI (artificial intelligence), ML (machine learning) and statistics. Nowadays, the maturity of the model, integrates with huge data incorporate efforts and high-performance RDBMS engine, make these technique experimentally for represent data warehouse environments. According to Berry and Line off [1] privacy is a composite difficulty, since of technology, is rising becoming a social difficulty. In the year of 2004, the Cambridge Advance Learner's Dictionary, explain the concepts of security and how to manage the personal or private data and related secret data. Presently, every formation of trading business and electronic devices, and begin that examines the private and stored for further references. These are the important problems to analysis the innovation work and privacy. We introduce that many numbers of difficulties on the security.

- Restricted are pervious on security by the social influence, and the problem is really how much data should be combined and who is manage these data.
- Each and Every customer has various objectives on privacy.
- Various levels of responsibility with view to data regarding them being possible to others. These techniques work an import works in defined the security; protecting these security.

The data mining is a capability that represents the crucial need of business to handle their consumer relationship and calculates more easily. Data mining are mostly used in the marketing domain. Following represent the two important analysis of data mining in marketing:

- Privacy contravention may sustain legal responsibility that could output in exorbitant law suits.
- Privacy violations may output in bad press that can do analysis destruction to corporate or brand image.

1.1. Machine Learning Technique

The ability to link multiple data sources, analyze large volumes of data, and apply newer algorithms on the transactions, provide organizations an opportunity to capture, and sometimes predict, fraud in a more efficient manner. More recent analytics based approaches include the use of descriptive and predictive analytics, machine learning, and social network analysis methods for fraud detection. The classical approach to fraud identification relies on creation of explicit rules (IF-THEN-ELSEIF) based on the recommendation of experts. These rules are developed and modified through their collective field experiences. Nevertheless, over time, due to the dynamic and sophisticated nature of the frauds, the rules become complex and difficult to maintain and implement (unless they are very regularly updated). This is also a very labour intensive approach requiring human intervention at every stage of evaluation, identification, and monitoring [30]. The present study investigates various issues involved in designing a new model of web mining tools and other mining services for social networking websites, with an aim to suggest some hypothetical solutions in the form of measures and management standards from User Generated Content (UGC) in online social networking. These will help to overcome the some major threats in social network and websites technology

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Motivation

Traditional data mining techniques rely on the statistical patterns used for identifying fraud. Yet, given the uncommon, time-evolving, and carefully concealed nature of fraud, these methods are often unable to detect various types of frauds. Application of a number of graph algorithms can help in identifying such patterns by utilizing relationship information in addition to the user level attribute information.

3. Literature Survey

Business database are expands at unmatched rate. The current META Group study of data warehouse project investigate that 19% of increments after the fifty GB level, while the fifty nine percent anticipate that second quarter of 1996.1 in several industry, thus are retail. Attend necessary for increase execution engine can be reach in a cost-effective direction with analogous computer technology. The data mining methods shown that pervious last 10 years are, implements application is mature, dependent, understandable tools that regularly working existing statistical techniques. The developments form business data to the business information, present phase has implement on technique.

The important techniques of the data mining have under development for pervious ten year; in innovation regions are AI (artificial intelligence), ML (machine learning) and statistics. Nowadays, the maturity of the model, integrates with huge data incorporate efforts and high-performance RDBMS engine, make these technique experimentally for represent data warehouse environments. In [1], authors attempt to examine a Web page as data with social aspects. Each web page is the outcomes of hiding social communication. This communication among various groups of people converts into a confirmed unification of Web page production. The external sign of this unification are the attribute of the web page that achieves the user's anticipations. Through analysis of the attributes, the authors can acquire data that can normally explain the web page. This simple explanation consist powerful data regarding the social group the page is intentional for. If the user utilized this data to clarify the search, then he detection himself as a part of a social group. For the simplification of the social feature of web page we utilized the concepts MicroGenre. In this research paper author innovated the basic theory of MicroGenre and also demonstrates practical for the identity and usage of MicroGenres. In [2], Author addresses the problem of privacy conservation data mining. Particularly, we examines a structure in which two parties possess secure databases want to developments a data mining techniques on the incorporate of their database, without reveal any redundant information. Our research is prompt by the necessary to both preserve privileged data and allowing it's utilized for innovation or other objectives. The above issue is a particular instance of secure multi-party execution and as thus, can be solved using called generic protocols. Data mining technique are typically ambiguous and ahead the input normally inside of huge datasets. The generic protocols in thus a case are of no experimental use and therefore more efficient protocol is necessary. The author concentrates on the problem of decision tree learning with the useful ID3 algorithm. Ours protocol is greatly efficient than generic solutions and requirement both some round of interaction and reasonable bandwidth.

Aiello et al (2002), in [5] multiple numbers of huge graph (for example, WWW graph and Call graph) share definite universal attribute which can be report is known as "power law". In this innovation paper, we will firstly deeply study the existing research paper on power law graph. Then we will provide 4 evolution system for produce power law graphs by contributed one node/edge at a time. Author also represents the any provided edge density and wanted for in-degrees and out-degrees the output graph will are content the power law and the in/out-degree situations. In contributed, author analysis another key aspects of huge graph is known as "scale-free" in the sense that the frequency of sampling is freely of the arguments of the output of the power law graphs.

Yuan He et al(2017) [4], in this innovation work investigate that, finding semantic comprehensive concepts from the large amount of UGC (-Generated Content) in social media would help multiple downstream application of intelligent computing. The concepts approach, as one of the large powerful method (algorithms), to calculate the dormant semantic scheme in text collections has been mostly utilized. However, one key blames of ideas approach is that they require archives with certain length to provisions reliable measurements for creating sound ideas. In Twitter, the client's tweets are predominantly short and loud. Perceptions of world advancement are confused for ideas approach. His calculation pre-learns two sorts of premium data from the dataset: the premium word-sets and a tweet-premium inclination network. In addition, a devoted foundation demonstrates is acquainted with judge whether a word is drawn from the background noise. Practical on two real life twitter datasets represent that his model got significant improvements above state-of-the-art baselines.

4. Methodology

As the marketing manger you have explosion to a huge data regarding all of your clients: their age, credit card history, usage and sex etc. The well news is that you also have a huge data regarding your probable client: their credit history, sex and age. Our issue is that we don't know the long distance knowing usage of this expectation (since they are most likely now customers of your competition). Client likes to concentrate on these expectations that have large amounts of large distance usage. Mining the result of a check market representing a large not associated low sample of anticipates can supplies a foundation for recognize better anticipate in the entire market. The table I represent another general structure for creating system; anticipates what will go to happen in the future.

Table i. Data mining for predictions

	Yesterday	Today	Tomorrow
Static Information and current plans(e.g demographics data, marketing plans)	Known	Known	Known
Proprietary information (e,g, Customer transaction)	Known	Known	Target

4.1. System Architecture

The architecture is help to the OSN services is a three-tier strategy in figure. 1. The first layer of architecture is SNM (Social Network Manager), generally main goal to supply the fundamental functions are profile and relationships, since the second level aided the SNAs. The helps SNAs may in revolve needs a contributed layer for their necessary GUI. According to this liking scheme, the investigation model discus in the previous two layers.

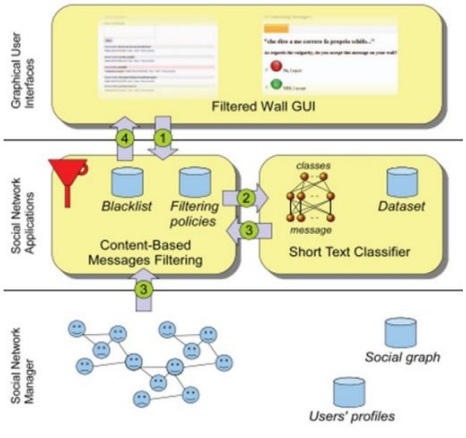


Fig. 1. Filtered wall theoretical scheme and the direction messages follow, from writing to communication.

4.2. Data Cleaning

Data cleaning is nothing but the data scrubbing or cleansing, deal with deleting and finding errors and inconsistency from information in order to increase the data quality.

Algorithm 1: Noise Removal Algorithm
Input: P here P is Document file
d1 = EscapingHtmlCharacters(P)
d2 = DecodingData(d1)
d3 = ApostropheLookup(d2)
d4 = RemovalOfStopW ords(d3)
d5 = ApostropheLookup(d2)
d6 = RemovalP unctuations(d5)
d7 = RemovalExpressions(d6)
d8 = SplitAttachedW ords(d7)
d9 = Slangslookup(d8)
d10 = StandardizingW ords(d9)
d11 = RemovalOfUrl(d10) Stop

4.3. Bag of Words (BoW)

Bag of Words (BoW) is an algorithm that computes how multiple times a word applying in a document. . In the table III where the documents and words adequately becomes vectors are saved, each and every row is a word, and column is a document and each cell is a word is computed. The documents in the collection are stored by column of identical length. The word computed vectors, outcomes deprived of context. Each word TF-IDF applicability is a normalized data scheme these are combined to another one.

$$W_{i,j} = tf_{i,j} X \log \frac{N}{dfi} \quad (1)$$

Where, tfi,j = numberofoccurrencesof iinj, df i = numberof documentcontainingi and N = totalnumberof document
Algorithm:
Step 1: Collect Data Below is a piece of the first some lines of text from the book ”A Tale of Two Cities” by Charles Dickens, taken from Project Gutenberg. It was the best of times, It was the worst of times, It was the age of wisdom, It was the age of foolishness, For this littlie example, let’s consider each line as a different ”document” and the 4 lines as our entire corpus of documents.

Step 2: Design the Vocabulary Now we can make a list of all of the words in our system vocabulary. The unique words here (ignoring case and punctuation) are:
• “it” “was” “the” “best” “of” “times” “worst” “age” “wisdom” “foolishness” That is a vocabulary of 10 words from a corpus containing 24 words.

Step 3: Create Document Vectors
In this step we calculate score the words in each document. The aim is to turn each and every document of free text into a vector that we can use as input or output for a machine learning model. Because we know the vocabulary has 10 words, we can utilize a constant-length document representation of 10, with one position in the vector to score each word. To mark the presence of words as a Boolean value the simplest scoring method, 0 for absent, 1 for present Using the arbitrary ordering of words listed above in our vocabulary, we can go through the first document (”It was the best of times”) and change it into a binary vector. The scoring of the document would look as follows:
• “it” = 1 “was” = 1 “the” = 1 “best” = 1 “of” = 1 “times” = 1 “worst” = 0 “age” = 0 “wisdom” = 0 “foolishness” = 0

As a binary vector, this would look as follows: [1, 1, 1, 1, 1, 1, 0, 0, 0, 0] The other three documents would look as follows:
“Itwastheworstof times” = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
“Itwastheageofwisdom” = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
“Itwastheageoffoolishness” = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]
All series of the words is regularly not helpful and we have a fixed directions of discover features from any types of files in our entity are disposed for utilize in modeling. The recently produce files that overlap with the language of known words, but it not assured about words from of the language, can be quiet be encoded, where only the existence of known word are save and the unknown word are eliminating.

4.3. Bag of Words (BoW)

With regards to grouping, there are three noteworthy ways to deal with make a classifier for blackmail position: To manage the data, to analysis the distinctive error value and to detecting the Error. In this task represent what fraud is and what is the difficulty see by characterization determination when trying to distinct and prediction blackmail. Along with this explanation, we represent the group of mostly used techniques to solve this difficulty and comment the uniqueness of few of the business regions that are impacted by deception.

5. Results and Discussion

The performance of proposed method (RAKE + TF) is better than existing methods for text as well as Features Extraction. Refer diagram 2.

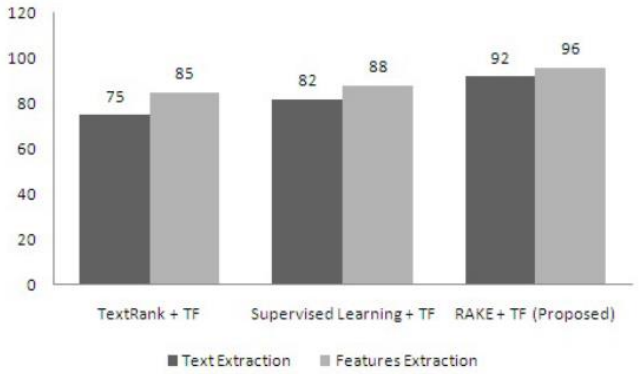


Fig. 2. Text and Features Extraction

Accuracy: It is the larger reflex output determine and it is generally a ratio of properly anticipated outcomes to the whole analysis.

$$Accuracy = \frac{tp + tn}{tn + tp + fp + fn} \quad (2)$$

Recall: In data recovery, recollect is the ratio of the associated files that are favourably recovery.

$$Recall = \frac{Relevantdocuments \Pi Retriveddocument}{Relevantdocuments} \quad (3)$$

$$Recall = \frac{tn}{tn + fn} \quad (4)$$

Precision: It is large associated if false positive rate is small. We have achieved 97.45 recognize are too better. The field of data collecting recognizes is the fraction of collections files that are associated to the query:

$$Precision = \frac{Relevantdocuments \Pi Retriveddocument}{Retriveddocuments} \quad (5)$$

$$Precision = \frac{tp}{tp + fp} \quad (6)$$

F-Measure: It is 0.701.A measure that merges recall and precision is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

6. Conclusion and Future Work

This system is developed to filter redundant message from OSN walls. The wall that controlled the redundant message (can be considered as fraud) called as the FW (Filtered Wall). This system extends to denote the thought concerning the filtering system. In contributed, this system extend to survey and analysis the approach and techniques restricting the interferences that user will do on the required to filtering protocols with the objectives of going the filtering model and therefore the variable rules of existing based model. This system extends to allow the text filtering algorithm at pre-processing phase thus on reason the message and believe computation for message. In this innovation,proposed system builds in a manner in which each and every amp and message supplies the trust goodness of user. The project has implemented almost entire requirements. Moreover, requirements and enhancements that requirement can simply be complete since the coding is mainly scheme or modular in form. It has represented a system to filter out redundant content such as text messages from OSN walls.

TABLE III
ACCURACY, RECALL, PRECISION AND F-MEASURE COMPARISION

# Text Sent	TextRank + TF	Supervised Learning + TF	Proposed Method (RAKE + TF)
		Accuracy	
100	82.23	85.49	88.45
200	79.56	81.34	84.78
500	76.51	79.89	81.67
1000	75.64	77.09	80.16
		Recall	
100	78.11	81.22	85.33
200	76.56	80.34	83.78
500	75.21	79.89	82.02
1000	74.05	77.09	81.16
		Precision	
100	91.73	94.21	97.45
200	88.56	89.89	94.82
500	86.21	88.89	88.67
1000	82.64	85.09	86.16
		F-Measure	
100	65.35	68.52	70.35
200	64.01	66.21	68.32
500	62.13	65.89	66.59
1000	61.3	64.25	65.87

TABLE IV
TEXT CLASSIFICATION COMPARISON

#Message	TF	random walk
100	90.1	83.3
200	88	81.3
500	88.23	85.5
1000	90.6	89.3

To compulsory customizable content-based filtering rules the system utilizes a soft classifier. Furthermore, the scalability of the model in theory of filtering selection is good through the industry of BLs. It is needed to provide security to blacklist management system and Filter rules. This research innovation system have overcome the drawback of existing system to instead of blocking user to notify message to that user by using mail. For future, image can be filter in online social network by using OCR

TABLE V
EXPECTED AND CORRECTED KEYWORD EXTRACTION

Method Name	Expected Keywords	Correct keywords
RAKE	7815	2116
TextRank	5284	2037
Supervised Learning	4788	1973

TABLE VI
PERFORMANCE MEASURE PARAMETERS

Extraction Method	Precision	Recall	F-Measure
RAKE	33.7	51.7	37.2
TextRank	31.2	43.1	36.2
Supervised Learning	29.7	42.2	33.9

References

[1] Berry MJA and Linoff GS, "Mastering Data mining: The art and science of customer relationship management" Canada Wiley, 2000.

[2] New W. , "Pentagon failed to study privacy issues in data mining effort", IG says 2004.

[3] Verykios, VS; Bertino, E; Fovino, IN; Provenza, LP; Saygin, and Theodoridis, Y. "State-of -the-art in Privacy Preserving Data Mining", SIGMOD Record. Volume 33, Issue 1:50-57 2004.

[4] Yuan He, Cheng Wang and Changjun Jiang, "Mining Coherent Topics with Pre-learned Interest Knowledge in Twitter", DOI 10.1109/ACCESS.2017.2696558, IEEE Access, 2017.

[5] Aiello, W., Chung, F., Lu, L., "Random evolution of massive graphs", M.G.C. (eds.) Handbook of Massive DataSets, pp. 97-122. Kluwer, Dordrecht (2002).

[6] Viaene, R. Derrig, G. Dedene, "A Case Study of Applying Boosting Naive Bayes to Claim FraudDiagnosis", IEEE Transactions on Knowledge and Data Engineering, vol. 16, pp. 612- 620, 2004.

[7] Hongyu Gao, Jun Hu, Tuo Huang, Jingnan Wang, Yan Chen "Internet Computing", IEEE 15(4), July-Aug. 201,1 56 - 63, (journal article).

[8] W. Lee, D. Xiang, "Information-theoretic Measures for Anomaly Detection", In proceedings of 2001 IEEE Symposium on Security and Privacy, pp. 130-143, 2001.

[9] Wen-jun, S., and Hang-ming, Q, "A Social Network Analysis on Blogspheres", In Proceedings of the 15th IEEE International Conference on Management Science and Engineering, 2008, pp.1769 - 1773, Long Beach, CA, USA.

[10] Bandar Alghamdi, Jason Watson, Yue Xu. "Toward Detecting Malicious Links in Online Social Networks through User Behavior", IEEE/WIC/ACM International Conference on Web Intelligence Workshops, 2016.

[11] Anna Leontjeva, Konstantin Tretyakov, Jaak Vilo, Taavi Tamkivi,"Fraud Detection: Methods of Analysis for Hypergraph Data",Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, 2012.

[12] Michalis Faloutsos, Thomas Karagiannis, and Seung-Hyun Moon,"Online social networks", Network, IEEE, 24(5):4-5, 2010.

[13] Shariati, B., Soraya, S., Seddigh, R., Keshavarz-Akhlaghi, A.-A., Azarnik, S. "Comparison between the personality traits of the parents with children addicted to methamphetamines and that of the parents of the healthy children", (2018) International Journal of Pharmaceutical Research, 10 (3), pp. 122-125.

[14] Hongyu Gao, Jun Hu, Tuo Huang, Jingnan Wang, and Yan Chen, "Security issues in online social networks", Internet Computing, IEEE, volume, 15(4):56-63, 2011.

[15] Prateek Joshi and CC Jay Kuo, "Security and privacy in online social networks: A survey", In Multimedia and Expo (ICME), 2011 IEEE International Conference on, pages 1-6. IEEE, 2011.