Keyword Extraction Using Machine Learning Approaches

BhavneetKaur
Department of Computer Science and Engineering
Thapar University
Patiala, India
bhvsidhu@gmail.com

Dr.Sushma Jain
Department of Computer Science and Engineering
Thapar University
Patiala, India
sjain@thapar.edu

Abstract— **Data Mining is the procedure to separate concealed prescient data from database and change it into a reasonable structure for sometime later. The varying areas in information mining are Web Mining, Text Mining, Sequence Mining, graph Mining, temporal Data Mining, Spatial Data Mining, Distributed Data Mining and Multimedia Mining. A portion of the utilizations of information mining, it is utilized for the budgetary information examination, retail and media transmission ventures, science and designing and interruption location and counteractive action. In this exploration paper a calculation is proposed for ensembling and catchphrase extraction in content mining. In the proposition the above methods are upgraded which in turns, enhance the exactness, review, f-measure and accuracy.**

Keywords*— *NLP, Keyword Extraction, Ensembling Models

I. INTRODUCTION

The reality of today could beillustratedas the computerized world and we all are beingsubject to the advanced/automatic type of information. It is the condition neighborly in light of the fact that we are utilizing less measure of paper. In any case, again this reliance brings about a substantial measure of information. Indeed, even any little movement of humanoidcreates digital information. Like, while any individual purchases thecoupon on the web, then his subtle elements are put away into the storage. Today, almost 80% of the electronic information is as content [1]. This tremendous information is unclassified and unstructured (or semi-organized) as well as containing helpful information, pointless information, logical information and business particular information, and so forth. As per a review, 33% organizations exist working with the enormouscapacity of information i.e. approximately at least 500terabyte. In this situation, to separate intriguing and already concealed information design procedure of content mining is utilized. Normally, information is put away as content [3]. And after that, taking after three stages happen: a) Data is pre-processed. b) The system of content mining is connected. c) The consequences of uncovered broke down. Content mining and information mining are comparable, with the exception of information mining takes a shot at organized information while content mining deals with semi-organized and unstructured information. Information digging is in charge of extraction of certain, obscure and potential information and content digging is in charge of unequivocally expressed information in the given content. Then again potential data withdrawal is straightforwardtogether. Fewdissimilar terms, for example, Content Study, Content,Documents Mining orLearning Revelation in Content, is likewise utilized as a part of place of content mining infrequently [4].

II. RELATED STUDY

The article Self–Organizing Map Representation for Clustering Wikipedia Search Results introduces a way to deal with computerized association with content information. The assignments have been performed on Wikipedia, which is chosen subset. The portrayal of vector space demonstrates in view of terms has been utilized to fabricate gatherings of comparable articles extricated from Kohonen Self-Organizing Maps with Database examine (DBSCAN) bunching. We performed direct measurement diminishing of crude information utilizing Principal Component Analysis to legitimize effectiveness of information preparing. The grouping technique has been utilized as a part of the execution of the framework that bunches consequences of watchword based inquiry in Polish Wikipedia. The measure of data given as records written in a characteristic dialect requires exploring techniques for viable substance recovery. One method for enhancing recovery proficiency is performing archives classification which sorts out records and permits find significant substance less demanding. The technique distinguishes huge relations in the information and consolidates associated highlights into one simulated trademark which permits diminishing component space essentially. In the decreased element space they develop Kohonen SOM (Self-Organizing Map) which permits 2D introduction of topological comparability relations between articles (here archives). Utilizing DBSCAN bunching we remove from the SOM gatherings of the most related papers. changing the Self-Organizing Map granularity we develop various leveled classifications that compose records set. Content mining or data disclosure from content (KDT) — shockingly determined in Feldman et al. deals with the machine maintained examination of substance [4]. It uses methodologies from information extraction, information recuperation also NLP (typical lingo taking care of) and partners them with the counts and technique of data disclosure

978-15090-6403-8/17/$31.00 © 2017 IEEE

database (KDD), machine learning, estimations and data mining. a neat system is picked as with the data divulgence database get ready, whereby not data when all is said done, yet rather the examination of substance reports are in purpose of the meeting. Thusly, imaginative request for the new data mining procedures happen. One issue is that we now need to oversee issues of from the data exhibiting perspective unstructured educational files. We can insinuate interrelated research domains if we describe content mining. For each of them, we can give a substitute significance of substance mining [5], which is induced by the specific perspective of the zone. The central approach acknowledges that substance mining fundamentally thinks about to information extraction - the extraction of realities from compositions. Content mining can beadditionally characterized like information mining as the use of calculations and techniques from the field's machine learning and measurements to writings with the objective of discovering helpful examples [5]. For this reason it is important to pre-handle the writings therefore. Data extraction strategies are utilizations many creators, NLP (common dialect preparing) likewise some straightforward preprocessing ventures to concentrate information from the writings. The information mining calculations can be apply for removing information. The finding of records which contain answers for inquiries is a data recovery and not itself finding of answers. With a specific end goal to accomplish this objective geometric measure and strategies are utilized for the general handling of content information and judgment to the given question. Data recovery in the more extensive judgment skills manages the whole scope of data preparing, for information recovery to learning recovery [5]. The general objective of NLP is to accomplish a superior comprehension of regular dialect by utilization of PCs. Others incorporate likewise the work of basic and strong methods for the quick preparing of content [6]. The scope of the doled out procedures comes to from the basic control of the strings to the programmed preparing of regular dialect request. The objective of data extraction strategies is the extraction of particular data from content archives. These are put away in information base-like examples and are then accessible for further utilize [7]. Keeping in mind the end goal to get all words that are utilized as a part of a given content, a tokenization procedure is required, i.e. a content record is part into a flood of words by evacuating all accentuation marks and by supplanting tabs and other non-content characters by single void areas [7]. Bunching is a division of information into gatherings of related articles. Each gathering is called bunch, which having objects that are comparable amongst themselves and distinctive two objects of further gatherings. Speak for information by few bunches basically loses certain amazing points of interest, however gives disentanglement. It indicates numerous information protests by few bunches, and thus, information is shown by its groups.

III. TECHNIQUES OF CONTENT MINING

For mining the content variety of techniques are available discussed as follows:

a. Summarisation
b.Info Retrieval
c.Classification
d.Picturing
e.Grouping
f.Theme Discovery
g.Queryresponding
h.Sentimentality examination

a. Summarisation: It's the most superior technique among all the basic techniques. In basic terms rundowns are the procedures for making synopsis of the records enclosing extensive measure of data, however topic and fundamental thoughts of archive is kept up. It helps the client to comprehend whether a specific record is valuable for him or not. Pressure is additionally identified with outline, however, is not in intelligible form. For a case of synopsis, "dynamic" some portion of an exploratory paper depicts the fundamental thought of researches directed by researchers.

b. Information Extraction: It uses relationships between the documents thatutilizes design coordinating on behalf of it. For instance, in rdbms put away information is accessible as tables. At the point when the information is in unstructured shape, data can't be separated effortlessly (no settled reference). In IE regular dialect record is changed over into organized one and afterward the information is removed. IE prepare extricates little lumps/pieces (individuals, put, time, date, address) of content by coordinating examples. This method gives noteworthy outcomes when it is connected to the exceptionally immense volume of content. By utilizing machine taking in an IE framework can be developed, yet it can't give completely exact outcomes. That is the reason mistakes can be identified in the yield of consequently separated database.

c. Classification: It is a regulated learning procedure that assigns records according to the classes defined. Archive order is to a great extent utilized as a part of libraries. Report order or content classification has a few applications, for example, call focus steering, extraction of the data about data, disambiguation, emails,spam identification, sorting out and keeping up vast inventories of Web assets, news articles arrangement and so forth. For content arrangement many machine learning methods have been utilized to advance standards (which appoints specific record to specific class) consequently. There aretwo ways for record characterization which are Content Taxonomy andScript Miner. Preceding couldevaluate, break down as well as remove content though last one relies on upon recurrence of events.

d. Clustering: This is a record's literary closeness based on an unsupervised system (that need not prepare information) that is utilized by the information experts for the isolation of the content in fundamentally unrelated gatherings. Content mining has less utilization of grouping. Grouping can be separated into 2 sections various leveled bunching and partitions bunching. Various leveled bunching yields a solitary grouped level and again sectioned in the base progressive bunching and topmostdowncast progressive bunching. In partitioned bunching report sets are separated into k number of disjoint points [16].

e. Theme Discovery: As per the name mentions theme following stands a procedure for taking after a particular subject in view of necessity or intrigue. It predicts conceivable point of enthusiasm for the use of the premise of theme intrigue past. Theweb crawler of is using this approach successfully. For instance, when a client pursuit a thing and afterwards it would go for another thing, then the person gets advertisements related leading thing on various sites opened therefore of second inquiry. On the off chance that an understudy scans for a few colleges and after some time he open long range informal communication site like Facebook. For this situation person would be shown the supporting pages and promotional pages in relation with other colleges. Then again, it likewise has confinements. On the off chance that a client subscribes for "counterfeit consciousness" at that point person wouldacquirefresh data on itsflapnot as manyof them will have a genuine utilization of the client. Point following has extensive useful in the field of restorative, study and training.

f. Query Responding: question taking note of is careful to choose the mode for dealing for searching anadditional proper reaction intended for thecertain inquiry. Such as, while customer requestsGoogle forany inquiry andinterestedto specific thing, at that point Google provides a person withfinest organized replies or associations for the requested inquiry via looking for watchwords for questions into the database. It could be used for more than one digging strategies immediately such as it could use info retrieval and questions, requests to think components then to consign the inquiry independently. Questions and answers can be utilized as a part of a few web applications, therapeutic field and in addition to training.

g. SentimentalExamination: - Sentiment Analysis which is otherwise called assessment, mining is arranged of client's feeling, for the most part into a few modules that are certain, negatives, impartial as well, as blended. The situation is essentially useful for getting individuals' views or state of mind in the direction of anything which incorporates administrations as well as items. Such as illustration, Amazon's site gives storage to client's remark on their everything the offering items. By these people groups can express their demeanor towards a particular thing which is again useful for different purchasers since they can read audits from past purchasers. From the organization's (Amazon) point of view, this onecomforting them to enhance their item's superiority by constructingessential alterations into it. Presently through the developing prominence and range of informal communication destinations any association could receive a substantial amount of information (audits) identified with their items.

IV. METHODOLOGY

In the proposed work the following steps are included, in which pre-processes are the basic noise removal technique in which the tokenization of a sentence is done. After that the process of keyword extraction is started, in which the proposed algorithm is applied, which use an ensemble technique extract keyword using NLP and score the documents on the basis of cosine similarity.

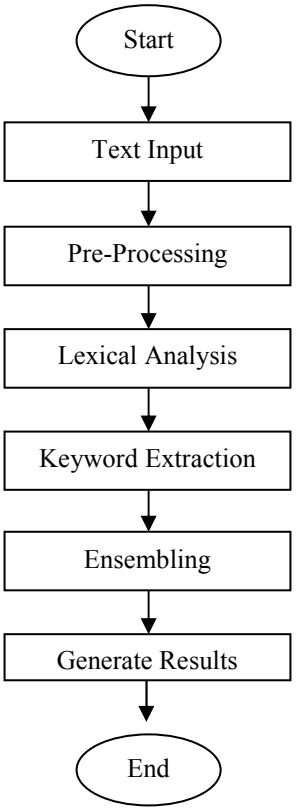


Figure 1: FlowChart

A. Text Input

In the very first step a Natural Language String is feed into the algorithm. This string is fed will be pre-processed, i.e. lexical and semantic analysis phase is applied to that string. In this step various data sets and ACM text is considered.
For e.g.: Keyword Extraction in NLP.

B. Preprocessing

- In this step is to clean data for example removing outliers,missing values.
- Then documents are converted into token using tokenization.
- Stop words are removed, Stop words are the words which occur many of the times in the document and have very less importance such as a, an, the, of, into which are just used as helping verbs in the documents.
- Stemming is done, which means to chop off the words which have same meanings but written in various forms. Such as in singular plural or in different tense forms. For eg. system and systems.

C. Lexical Analysis

- In this step the string is processed using lexical phase. Lexical phase is used to convert the string into various tokens. From these tokens the keywords may be extracted.
- In this phase the tokens are generated form NLP string. Hence there will be four tokens as KEYWORD, EXTRACTION, IN, NLP

D. Keyword Extraction

In this step using unsupervised classification algorithms keywords are extracted based upon cosine similarity between documents .The documents are ranked according to the similarity.

E. Ensembling

In the last steps the ensembling of algorithms is applied and results will be generated and evaluated.

V. Results and Discussion

Performance Metrics considered:
True positives (TP) = Total no. of outputs that are correctly identified are true
False positives (FP) = Total no of outputs that are incorrectly identified are true
True negatives (TN) = Total no. of outputs that are correctly identified are false
False negatives (FN) = Total no of outputs that are correctly identified are false

A. Accuracy

It is defined as the method to discover the measure of correctly scored documents from the given input. For calculating the accuracy, we need to specify the proportion of true negatives and true positives.

$$Acc = \frac{True\ P+True\ N}{TP+FN+FP+TN} (1)$$

B. Precision

This measure is used calculate the quality, it is the most commonly used performance metric, this measure is used to measure the standard deviation which means that how much proportion of defined measures are different from each other.

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} (2)$$

C. Recall

This measure is used to calculate the quantity, it is defined as the proportion of useful instances that are retrieved over the total number of present useful instances.

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negative} \quad (3)$$

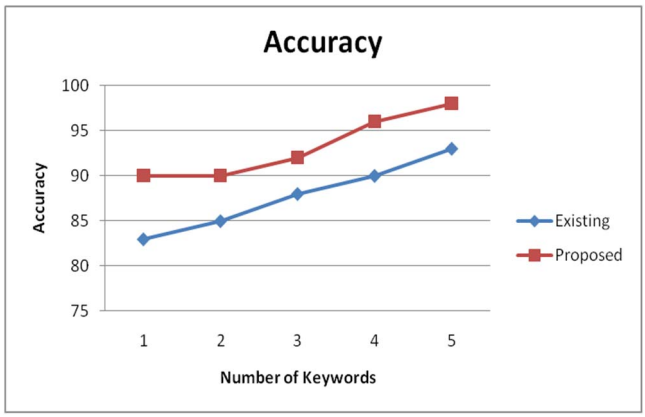
D. F-Measure

This measure is used to calculate the combined value of precision and recall known as the accuracy of both precision and recall. This can also be defined as an evaluation of missing or added value. It calculates the performance of the models and its values are 1 or 0.

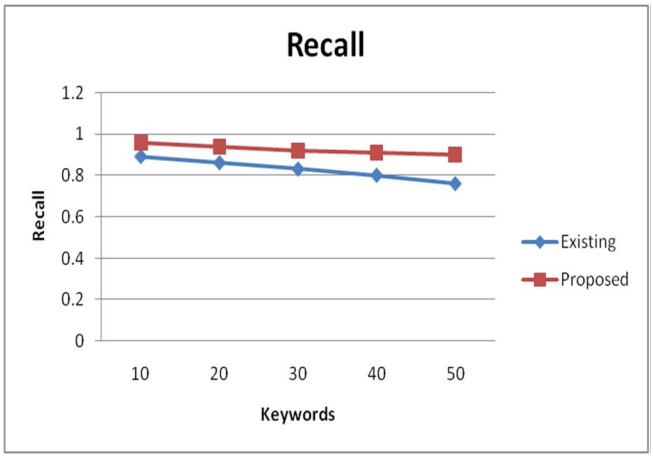
$$F\ measure = 2.\frac{Precision*Recall}{Precision+Recall} (4)$$

Table1: Confusion Matrix

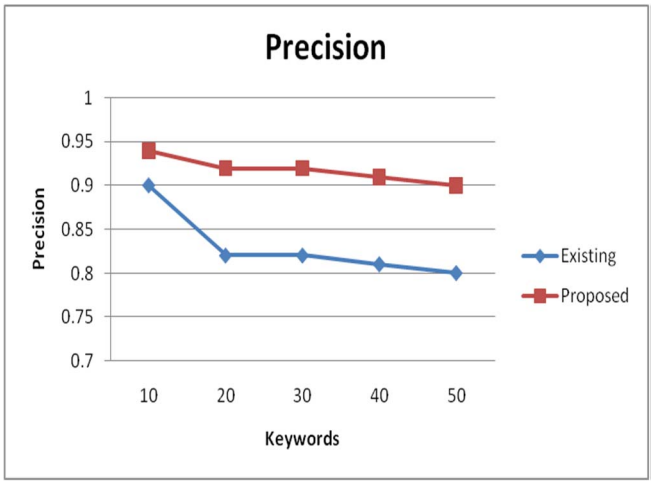
	P' (Predicted)	N' (Predicted)
P(Actual)	True Positive	False Negative
N (Actual)	False Positive	True Negative



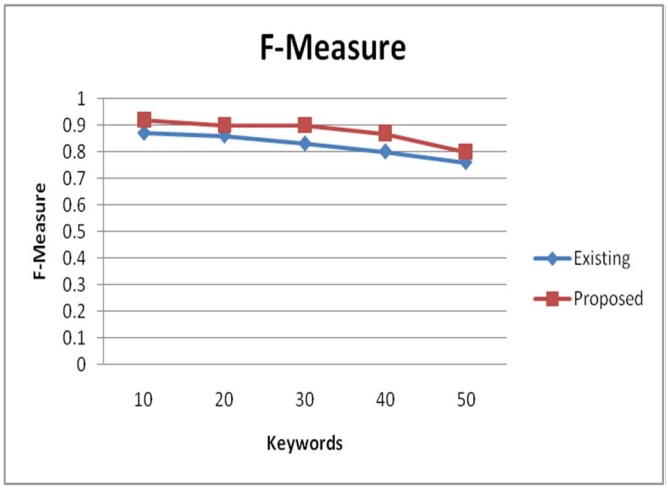
(a)



(b)



(c)



(d)

Figure (a) Demonstrates thecomparative study of accuracy between existing that is statistical and proposed that is using machine learning approach for keyword extraction (b) comparative study of recall (c) comparative study of precision (d) comparative study of f-measure.

From the above figures it is clear that the machine learning approaches give better results than statistical approaches for keyword extraction.

VI. CONCLUSION

Nowadays fastest developing field is text mining, With the Content mining is one of the quickest developing field from recent years. By the evolution of period its significance will increase day by day because the online data is present in enormous amount due the digitalization. Extracting useful information from text has a far way to go. In recent couple of years content mining (feeling investigation) is to a great extent being utilized to foresee the aftereffects of races at all levels that is most important improvement in this field as of late. By developing the benefit of connecting the content mining to other fields such as machine learning, perception, normal dialect preparing, it could be conceivable to sketch more effective and helpful content mining frameworks. Content mining is very useful for industry to utilize and developing the way of learning that can't be devour by the people. In this paper we attempted to introduced the application of content mining i.e the extraction of keywords from text with the use of ensembles approach, NLP ,instruments and applications.

VII. REFERENCES

[1]. VairaprakashGurusamy, SubbuKannan: "Preprocessing Techniques for Text Mining," October 2014

[2]. Dr. S. Vijayarani, Ms. J. Ilamathi, Ms.Nithya: "Preprocessing Techniques for Text Mining – An Overview," International Journal of Computer Science & Communication Networks, Vol.5(1), pp.7-16, 2014

[3]. C. Ramasubramanian, R.Ramya: "Effective Pre Processing Activities in Text Mining using Improved Porter's Stemming Algorithm," International Journal of Advanced Research in Computer and Communication EngineeringVol.2(12), 2013.

[4]. ParthSuthar, Prof. BhaveshOza: "A Survey of Web Usage Mining Techniques," International Journal of Computer Science and Information Technologies, Vol.6(6), pp.5073-5076, 2015.

[5]. "Data Preprocessing Techniques for Data Mining", Winter School On "Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets, Vol 5(7), 2010

[6]. Vikram Singh and BalwinderSaini "An Effective Pre Processing Algorithm For Information Retrieval Systems," International Journal of Database Management Systems (IJDMS) Vol.6(6), 2014.

[7]. Moral, C., de Antonio, A., Imbert, R. &Ramírez, J. (2014). A survey of stemming algorithms in information retrieval/ Information Research, 19(1) paper 605. [Available at http://InformationR.net/ir/191/paper605.html]

[8]. Shilpa Dang, Peerzada Hamid Ahmad, "A Review of Text Mining Techniques Associated with Various

Application Areas,"International Journal of Science and Research, Vol.4(2), pp. 2461-1466,2015.
[9]. Danqi Chen, Richard Socher, Christopher D. Manning, Andrew Y. Ng, "Learning New Facts From Knowledge BasesWith Neural Tensor Networks and SemanticWord Vectors," Proceedings of the International Conference on Learning Representations (ICLR, Workshop Track), 2013.
[10]. Will Y. Zou, Richard Socher, Daniel Cer, Christopher D. Manning, "Bilingual Word Embeddings for Phrase-Based Machine Translation," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013.
[11]. Karl Pichotta, Raymond J. Mooney, "Statistical Script Learning with Multi-Argument Events," Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014.
[12]. Stephen Roller, Sabine Schulte imWalde, "A Multimodal LDA Model Integrating Textual, Cognitive and Visual Modalities," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), pp.1146--1157, 2013.
[13]. Sergio Guadarrama, NivedaKrishnamoorthy, GirishMalkarnenkar, "YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition," Proceedings of the 14th International Conference on Computer Vision (ICCV-2013), pp.2712--2719, 2013.
[14]. Karl Pichotta, John DeNero, "Identifying Phrasal Verbs Using Many Bilingual Corpora," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), pp.636--646, 2013.
[15]. ShrutiBhosale, Heath Vinicombe, Raymond Mooney, "Detecting Promotional Content in Wikipedia," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), pp.1851--1857, 2013.
[16]. Dan Garrette, Jason Mielens, Jason Baldridge, "Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages," Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013), pp.583--592, 2013.
[17]. NivedaKrishnamoorthy, GirishMalkarnenkar, Raymond Mooney, Kate Saenko, Sergio Guadarrama, "Generating Natural-Language Video Descriptions Using Text-Mined Knowledge," Proceedings of the NAACL HLT Workshop on Vision and Language (WVL '13), pp.10--19, 2013
[18]. SindhuRaghavan, Raymond J. Mooney, "Online Inference-Rule Learning from Natural-Language Extractions," Proceedings of the 3rd Statistical Relational AI (StaRAI-13) workshop at AAAI '13, 2013.
[19]. VarunChandola, Eric Eilertson, LeventErtÄoz, GyÄorgy Simon and Vipin Kumar, "Data Mining for Cyber Security," Data Warehousing and Data Mining Techniques for Computer Security, Springer, 2006.
[20]. Mahesh Kumar Kond Reddy, Sujeeth .T, "Data Mining Tool using Clustering Technique on Exploration Engine Dataset", Int. Journal of Engineering Research and Application, Vol.3(5), pp.2032-2036, 2013.
[21]. HamzehAgahi, A. Mohammadpour, S. Mansour Vaezpour, "Predictive tools in data mining and k-means clustering: Universal Inequalities," Vol. 63(3), pp.779-803, 2013.
[22]. Ahmed Elgohary, Ahmed K. Farahat, Mohamed S. Kamel, and FakhriKarray, "Embed and Conquer: Scalable Embeddings for Kernel k-Means on MapReduce", arXiv:1311.2334v2 [cs.LG], 2013.
[23]. Breiman, L.: Random forests.2001. Mach. Learn. 45(1), 5–32, 2001.