www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 11 November, 2014 Page No. 9116-9118

Analysis of Supervised algorithm for Voice Query Classification

Amol Kamble[1]

[1]PICT Engineering College, Pune University,
Amolkamble32@gmail.com

Abstract: ***Categorization of voice queries is useful for the finding the intent of the user. By finding intent it is easy to monitor that user. By performing classification search engine can find the class of the query. By using that class it becomes easy for the search engines to retrieve the results which are already classified in that class. This classification is useful for the targeted advertisement depending on search queries. For that classification two algorithms are used and analyzed. This analysis is based on the different parameters.***

Keywords: Classification, Supervised Techniques, Naïve Bayes, Rough Set Theory**.**

1. Introduction

Voice search queries are growing due to use of voice interface. These queries are short and that's why it is challenging to classifying these queries. [1] Intent query classification shows that use of multistage algorithm and POS, and domain keywords for the accurate classification of the quires. Two Supervised algorithms are used for the classification in two different stages. In first stage Naïve bayes is used and possible classes of given search query are retrieved. [2] Mentions classification of the text documents by using Rough Set Theory and analyze the results by considering accuracy, precision. Rough Set theory finds the upper bound and lower bound for given text document. [4] Compares four different algorithms for the text classification. [5] Classifies words by using KNN, Naive Bayes. This finds the root of the given words. These queries are classified using two different classification techniques. Naïve bayes and Rough Set theory are these two techniques used for the classification of the voice queries. Both algorithms are supervised techniques for the classification. Naïve Bayes is the techniques which finds the possible probability for inputted query and decides the probable class for the query. Rough Set theory is used technique which finds the upper bound and lower bound for the given query and then finds the respective class of the query. These queries are classified in to five different classes Map, Music, Sports, News, and Travel.

Classification is the technique which is used for the categorization of the uncategorized data. Classification has three types supervised classification, unsupervised classification and semi supervised classification.

Supervised Classification: - Supervised classification is the technique used for the classifying the uncategorized data. For that it uses supervised dataset for the training. Supervised categorization has two main steps training and testing. For training purpose already classified data is used. This data may be manually classified.

Unsupervised Technique: This technique classifies the testing data without help of training phase. It does not use training dataset for the classification.

Semi supervised Classification:-It is technique which used for the classification of information. This technique used for the classification of information by using very less classified dataset. It is in between supervised techniques and unsupervised techniques.

In this paper supervised technique is followed. Naïve Bayes and Rough Set Theory these are two different algorithms used for the classification of voice query. Naïve bayes is simple classification algorithm. It is made up from Bayes rule. It can be used on the given test data. It finds probability of classification given test data in to particular class. It finds probability of test data with each class and at last class of which probability is greater it chosen as final probability.

Rough set Theory is supervised technique of classification. It relates with fuzzy system and genetic algorithm, Artificial intelligence. Rough Set Theory has Information Set for representing the dataset.

S = {U, A} [1]

Where U= Nonempty finite set of objects

A= Nonempty finite set of attributes

In this method upper bound and lower bound for the given query is calculated with respect to required classes. Let X is a subset of the U for which lower bound and upper bound is to be calculated. One can found upper bound and lower bound as follows,

Lower Bound = {e belongs to U | [e] is a subset of X} [1]

Upper Bound = {e belongs to U | [e] intersection with X is not null} [1]

From lower bound and Upper bound accuracy is calculated as follows.

Accuracy= |Lower Bound RX| / |Upper Bound RX|. [1]

From Accuracy class of X can be decided.

$Accuracy_{class}$ = (lower bound)/(upper bound).

This gives accuracy for different classes for a given test data. Then class for which given value of accuracy is maximum that class is chosen as class of given test data.

These two different classification algorithms are further analyzed by using same training and testing dataset.

2. System Architecture

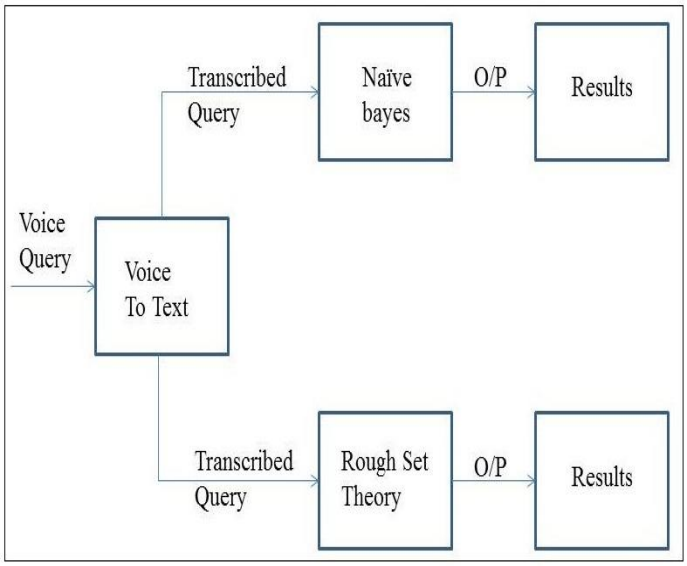

Figure 1: System Architecture

In this experiment voice query is inputted via microphone and that is first transliterated into the text. This transliterated query s inputted to the two different algorithms separately. Both algorithms are applied separately on the Transcribed Queries. Before that both algorithms are trained on the same dataset. This dataset is already labeled. Dataset is made up of five classes Tourism, Music, Sport, News, Map. All the related queries related to each class are kept together. Every entered query is classified in above given classes. Each query is classified using both classification techniques and results are compared.

Algorithms use fully labeled dataset for the classification of given data in a partic ular class. For the analysis of both algorithms three parameters are considered Time required for the classification, Accuracy, Precision.

Time required for the classification is the total time required for the classification of the transliterated voice query in one of considered class.

Accuracy= (no of true positive + no of true negative)/(no of true positive+ no of true negative+ no of false positive + no of false negative) (1)

Precision= no of true positive/ (no of true positive + no of falls positive) (2)

3. Experimental Setup

For the voice based search query classification voice is converted into text. Both classifiers are trained on the same dataset. For the comparison time required for execution, accuracy, precision these parameters are considered. 2800 supervised queries are used as dataset. These queries are manually classified into five different classes that are Tourism, Map, Music, Sport, and News. Four different datasets are used for the analysis of both algorithms.

4. Result Analysis

Four different datasets are used for the analysis of both algorithms. For analysis execution Time, Accuracy and Precision these three parameters are considered.

4.2 Execution Time Analysis

We have generated four different datasets with different test queries. These queries are varies in number. These datasets are feed to both algorithms separately. For each dataset execution time is calculated for each algorithm. Following graph shows Execution Time analysis for both Naïve bayes and Rough Set Theory algorithms. Above graph gives execution time required for the classification of transcribed query. In the analysis Naïve Bayes requires less time than the Rough Set theory algorithm for the classification. That is Naïve Bayes is faster than Rough Set theory.

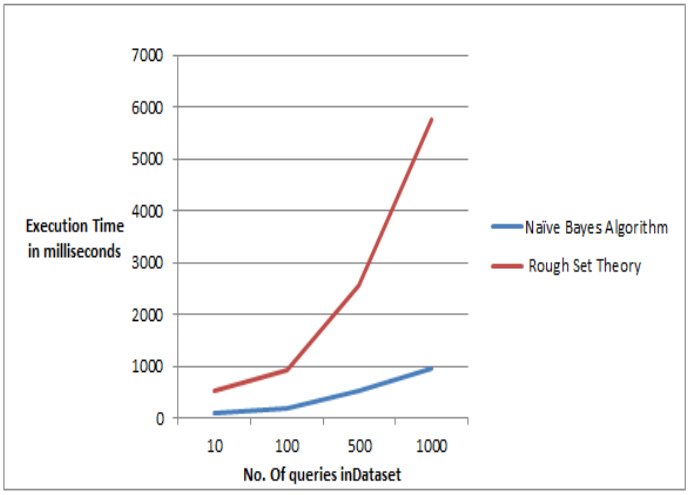

Figure 2: Execution Time.

4.2 Accuracy Analysis

For this analysis four different datasets are generated and feed to both the algorithms and results are analyzed. All the datasets are having mixed unlabeled queries. To check accuracy of algorithms above mentioned formula is used.

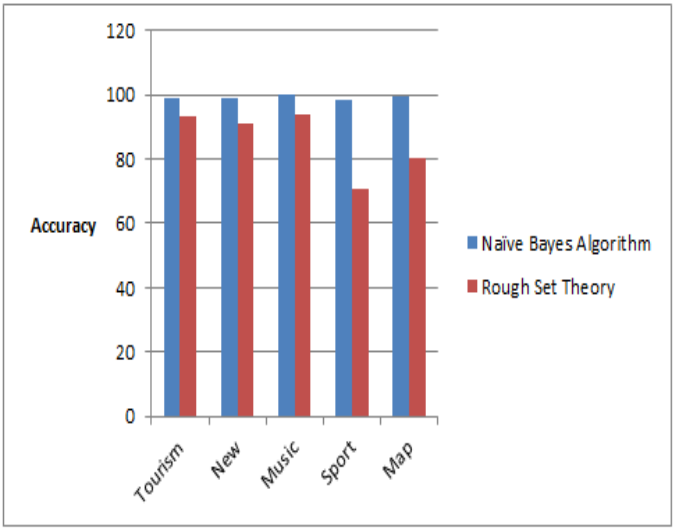

Figure 3: Accuracy Analysis

4.3 Precision Analysis

Following graph shows Precision parameter analysis for both the algorithms. Precision is calculated from above mentioned formula.

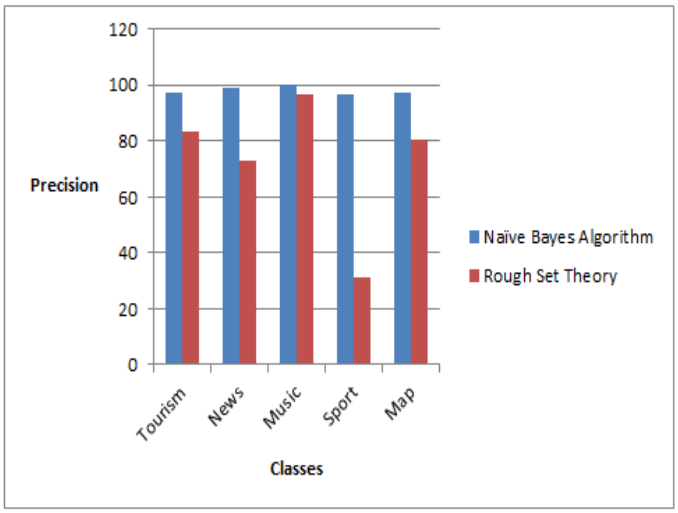

Figure 4: Accuracy Analysis

5. Conclusion

Both algorithms are analyzed by using same parameters. Training dataset is similar for the training of both the algorithms. While analyzing performance of these algorithms it can be concluded that Execution time for the naïve bayes is less than the rough set theory. Naïve bayes is faster than the Rough Set theory. The Naïve Bayes algorithm is more accurate than the Rough Set theory, because the accuracy of the Naïve Bayes algorithm is between 98% to 99% and accuracy of Rough Set theory is in between 80% to 93%.

References

[1] Subhabrata Mukherjee, Ashish Verma, Kenneth W. Church "Intent Classification of Voice Queries on Mobile Devices". In Proceedings of the 22nd international conference on World Wide Web companion, pp. 149-150., May 13–17, 2013.

[2] Dr. Ahmed T. Sadiq, Sura Mahmood Abdullah, "Hybrid Intelligent Techniques for Text Categorization." International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 2, Page: 23-40, April 2013.

[3] Munezero, Myriam, Maxim Mozgovoy, Tuomo Kakkonen, Vitaly Klyuev, and Erkki Sutinen. "Antisocial behavior corpusfor harmful language detection." In Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on, pp. 261-265. IEEE, 2013.

[4] Lee, Sangno, Jeff Baker, Jaeki Song, and James C. Wetherbe. "An empirical comparison of four text mining methods." In System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pp. 1-10. IEEE, 2010.

[5] Warintarawej, Pattaraporn, Anne Laurent, Pierre Pompidor, Armelle Cassanas, and Bénédicte Laurent. "Classifying Words: A Syllables-based Model." In Database and Expert Systems Applications (DEXA), 22nd International Workshop on, pp. 208-212. IEEE, 2011.

[6] D. Y. Choi, I. K. Ra ―Toward a Voice Interface and Personalized Local Web Search in Smart Phones‖, In Research Challenges in Information Science (RCIS), 2010 Fourth International Conference (IEEE), 2010, pp. 641-646.

[7] J. Yi, F. Maghoul, J. Pendersen, ―Deciphering Mobile Search Patterns: A Study of Yahoo Mobile Search Patterns‖, In Proceedings of the 17th international conference on World Wide Web (ACM), 2008, pp. 257-266.

[8] Munezero, Myriam, Maxim Mozgovoy, Tuomo Kakkonen Vitaly Klyuev, and Erkki Sutinen. "Antisocial behavior corpus for harmful language detection." *In Computer Science and Information Systems (FedCSIS)*, 2013 Federated Conference on, pp. 261-265. IEEE, 2013.

Author Profile

Amol Kamble received the B.E. in Computer Science and Engineering from Shivaji University in 2008. He is currently pursuing Masters Degree in 'Computer Engineering' from 'Pune Institute of Computer Technology, Pune'.