ISSN 2395-1621

An Efficient Analysis of SMS String Classification Process

[#1]Prof. A.L.Salunke, [#2]Spiti Sawant, [#3]Rashmi Sharma, [#4]Janani Iyer

[1]spiti.myworld@gmail.com,
[2]rashmi9943@gmail.com,
[3]iyerjanani140@gmail.com

[#1234]Department of Information Technology, JSPM's Bhivarabai Institute of Technology & Research, Wagholi

ABSTRACT

With SMS classification system, it has become effortless and feasible to provide required answers to the student's queries related their department. Without the proposed system, students usually need to contact their teachers or fellow mates in order to get the required information. The soft wares available for these concepts are acronym handling: replacing acronym with full words, preprocessing: removal of redundant words and fussy classification: classifying SMS according to the department. Different methodologies used in above mentioned systems are apriori algorithm, naïve Bayes and Ensemble learning. One of the biggest flaw in the above mentioned system is that although spam filtering techniques are now available in the market today, it is inevitable to deny that these solutions cannot guarantee 100% effectiveness at eliminating the problems of spam because a variety of these filters have weaknesses and strengths. For proper traffic handling, the proposed system puts forward the retrieval of information from SMS string using preprocessing technique and SMS classification. This paper presents an idea of SMS classification using the concepts of proper acronym identification, preprocessing and then applying fussy logic for answer extraction.

Keywords-- Short message service (SMS), fuzzy logic, SMS classification, information retrieve, mobile gateway

ARTICLE INFO

Article History

Received :8th March 2016

Received in revised form :

10th March 2016

Accepted : 13th March 2016

Published online :

18th March 2016

I. INTRODUCTION

Short Message Service (SMS) has been widely exploited in day-to-day communication. Short Message Service (SMS) is a communication service standardized in GSM mobile communication systems; it can be sent and received simultaneously with GSM voice, text and image. Using communications protocols such as Short Message Peer-to-Peer (SMPP). It allows the interchange of short text messages between mobile telephone devices. A general concept of Campus Short Message Service (CSMS) is to receive the query of any user and send the appropriate reply to the same user related to that particular query about any department of the organization. Here the main concept in this project is about SMS Classification for an organization, especially educational institutes. After receiving SMS of user the system categorize the SMS according to the SMS acronym and forward it to the respective department, then the respective information will retrieve from department database and forward it to the main server and from main server to the respective user through mobile gateway.The main purpose of this proposed system is to extract the answer for the SMS Query sent by the student on the college server to fulfill the time quotient and provide best information in lesser time.This project provides a model way to get the answer for the student queries, Proper Acronyms identification, Proper Preprocessing, Proper Applying of Fuzzy Logic, and answer Extraction.The system performs following functions like, It Identifies the Acronyms properly, it performs Preprocessing properly, It Classifies the SMS based on departments, and it sends a reply to the Student via SMS.

ACRONYM HANDLING

It is the extraction of abbreviated words from the text message and converting them into original text. Acronyms are the special form of abbreviations. The first step includes Initialization:

1. In initialization, removal of stop words is performed. Words which are often insignificant parts of the acronym (E.g.the, is, for, a) are removed from the document in order to identify the particular department with accuracy. Distinguishing stop words from the main text is important for the algorithm to make good matches with the department.

2. Removal of rejected words- This algorithm makes a list of rejected words such as (E.g., Table, Figure) are removed because they are irrelevant to the document and do not make any contribution in the required contents of the document. This list is optional but distinguishing them can make the document more efficient and in turn fewer coincidental matches.

3. Creation of a database with their definition- this is done to override program's searching routine especially when the search goes fruitless. This database is also optional yet efficient, if used.

PREPROCESSING

The input is pre-processed to disregard lines of text that are all uppercase (e.g., titles and headings). Upon identifying an acronym candidate, the reject word list is consulted before successive processing. If the candidate does not appear in the reject list, then an appropriate text window surrounds the acronym is searched for its definition.

Word parsing: In order for this algorithm to find a reasonable number of acronym definitions, a precedence has to be assigned to different types of words. Currently, these types are limited to (1) stop words, (2) hyphenated words, (3) acronyms themselves, and (4) ordinary words that do not fall into any of the above categories. These abstractions simplify the main mechanism since it becomes unnecessary to scan the text strings. We can systematically search through the text windows for matches of the letters of the word.

LOGIC FUSSY

Fuzzy logic is an approach to computing based on "degrees of truth" rather than the usual "true or false" (1 or 0) Boolean logic on which the modern computer is based.

Fuzzy logic includes 0 and 1 as extreme cases of truth (or "the state of matters" or "fact") but also includes the various states of truth in between so that, for example, the result of a comparison between two things could be not "tall" or "short" but ".38 of tallness." Fuzzy logic seems closer to the way our brains work. We aggregate data and form a number of partial truths which we aggregate further into higher truths which in turn, when certain thresholds are exceeded, cause certain further results such as motor reaction.In this research paper, section 2 is dedicated for background work and section 3 for conclusion.

II. LITERATURE SURVEY

To present a novel idea about SMS classification, this paper analyses various aspects in depth as described below:

Paper [1] narrates a baysian approach to filter junk email. We constructan automatic filter to eliminate the problem of junk email from the user stream. We use a probabilistic learning method for filtering the messages. Using baysian classifier it is possible to learn effective filters which eliminates a large scale of junk and supper vector machine is used for text categorization which helps to classify whether it is junk email or not.

Paper [2] represents two algorithms, Apriori and AprioriTid, which discover association rules between items present in large database of transactions. These algorithms arecompared to previous AIS andSETM algorithms. We come to aconclusion that thesealgorithms are far better than AIS and SETM.One of the drawbacks of this technique is, as the number of transactions increases, the execution time also increases linearly.

Paper [3] proposes the idea of labeling unlabeled data using a small set of labeled data and constructing a classifier for unlabeled data based on set of labeled data. Here naïve Bayes classifier is used for classification of unlabeled data.
Its major drawback was that Robotics, vision and information extraction are the three domains which generate large amount of unlabeled data and the labeled data is small and expensive thus degrading the classification results.

Paper[4] narrates the idea of using graph cuts for classification of unlabeled data .The similarities between the data are measured to construct a graph and then partitioning the graph for further classification . We come to a conclusion that Graph Mincut algorithm is better than previous learning algorithms which use small amount of unlabeled data. It is also robust to noise.

Paper [5] expresses the idea of using two redundant views of data for classification of unlabeled data from previously known labeled data. It makes use of co-training strategy for classification. The experimental results show that this technique of classification is more efficient than the previous learning algorithms.

Paper [6]In this paper, we propose a new co-training strategy for classification of unlabeled data from labeled data. This strategy does not make use of two redundant views of classification of data as proposed by Tom Mitchell in his findings. Drawback of this system was, Estimation of the confidence interval must be improved. Significant increase in the no. of iteration of co-training rounds, results in high error rates.

Paper [7] states that Tri-training: exploiting unlabeled data using three classifiersproposes tri-training algorithm which is a semi supervised learning algorithm. Unlabeled data are mainly available in data mining applications in order to get labeled data it is bit costly to get. From the original labeled set of data the algorithm produces three types of classifiers and these classifiers can be refined in the tri-training process using unlabeled set of data. It has a broader application from the previous co- training algorithm. To enhance the learning performance tri-training can be used to exploit the unlabeled data by using the UCI data sets and web page classification task.

Paper [8] "PEBL: Web page classification without negative examples" narrates an overview called positive example based learning (PEBL) classifying web page which helps to eliminates the need to collect negative training data sets. For a pre-processing technique such as to collect positive and negative training data sets a classifier is been constructed. In order to avoid bias negative training data sets are collected. To achieve high accuracy PEBL provides a mapping convergence algorithm which supports theoretical and experimental features. Mapping - Convergence algorithm has two phase to run the code Mapping stage and Convergence stage. In mapping stage weak classifier is used whereas Convergence uses internal classifier ie SVM.

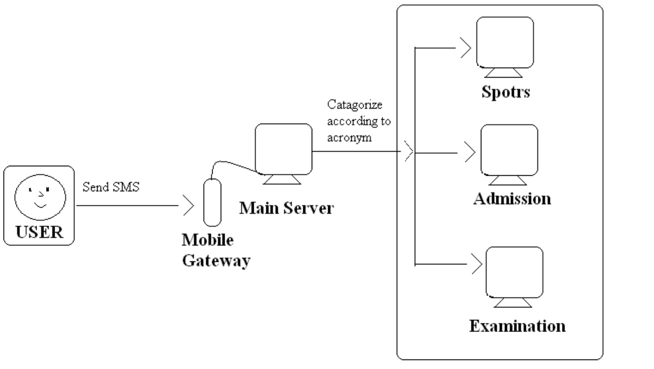
Paper [9] describes that PAC Learning from Positive Statistical Queries describes PAC learning model which learning model for various features set of data. From the given positive and the data set which are unlabeled PAC learning model can be described.in order to make learning model from positive data we have to collect extra information and given to the learner. Various relationships among PAC model, classification noise model and statistical query model are learned. In the above model k-DNF and k-decision list are studied with less information.

Paper [10] narrates a theory of the learnable expresses the methodology from a computational viewpoint which consists of proper set of information which proceeds with that type of mechanism, its nothing but a learning protocol. From the information provided we can derive the algorithm complexity in order to set the limitations to the set of range of learning concepts. For designing the leaning system we use various results and methodology concepts which seems to be realistic.

Paper [11] explains learning from positive and unlabeled examples" narrates the design of algorithms related to learning methodology from various unlabeled data set. To evaluate statistical queries the various algorithms are used like decision tree, data mining and naïve Bayes algorithm. It is difficult to collect labeled data whereasunlabeled data are available in huge quantity. We design a decision tree which is a induction algorithm which uses positive and unlabeled data set on the basics of experimental results for the following algorithm. The drawback of this paper is that in case of imbalanced class the learning class remains open.

III.PROPOSED SYSTEM

In recent years, Short Message Service (SMS) has been widely exploited in day-to-day communication. A general concept of Campus Short Message Service (CSMS) is to receive the query of any user and send the appropriate reply to the same user related to that particular query about any department of the organization. Here the main concept in this paper is about SMS Classification for an organization, especially educational institutes. After receiving SMS of user the system categorize the SMS according to the SMS acronym and forward it to the respective department, then the respective information will retrieve from department database and forward it to the main server and from main server to the respective user through mobile gateway.



Different departments in organization

In this above system a user will send the SMS to the public number which is provided by the college, where the SMS is received by the mobile gateway of the college, and then it sends the SMS to the web server. Here in the web server this SMS is separated according to the acronym in the SMS and forward it to different departments.

Then the information get retrieved from the department database and forward again to the web server or main server and forward the data to the respective user through the mobile gateway.

Now to develop software as specified above we will require classifying the receiving SMS at the server side. We need to define some acronyms which will define the respective department such as admission, sports, examination, etc. Here, SMS classification remains the important task which leads to send the SMS to respective department and will generate the correct reply.

IV.CONCLUSION

This paper puts light on many different methodologies and aspects of SMS classification. So as a generalized view of this cumulative study shows no any methodologies are perfect in providing a solution for SMS classification. Therefore as an initiative process, this paper puts forward an idea of SMS classification using the concepts of properacronym identification, preprocessing and then applying fussy logic for answer extraction.

REFERENCES

[1] "Time to confirm some mobile user numbers: SMS, MMS, Mobile internet, M-News", Toni T. Ahonen, Blog retrieved, September 16, 2013.

[2] "Fast algorithms for mining association rules in large databases", proceedings of the 20th international conference on very large databases, VLDB, Santiago Chile, September 1994.

[3]"Learning to classify text from labeled and unlabeled documents", K Nigam, A McCallum, S Thrun, T Michell, AAAI/IAAI, 1998.

[4] "Learning from labeled data using graph mincuts", A Blum, S Chawla, 2001.

[5] "Combining labeled and unlabeled data with co-training", A Blum, Tom Mitchell, 1998.

[6] "Enhancing supervised learning with unlabeled data", S Goldman, Y Zhou, In the preceding of ICML 2000.

[7] "Tri-training: exploiting unlabeled data using three classifiers", Zhou, Z-H and Li, IEE Trans knowledge and data engineering 17(11):1529-1541., 2005.

[8] "Pebl: Web page classification without negative examples", H. Yu, J. Han, C.C. Chung, IEEE transactions on knowledge and data engineering, 2004.

[9]"PAC learning from positive statistical queries", Denis, F, in the preceedings of 9th international conference on algorithmic learning theory, 1998.

[10] "A theory of the learnable communications of the ACM", Valiant, L.G, 1984.

[11] "Learning from positive and unlabeled examples", Letouzey, F., Denis and Gilleron R., 2000.