An Emphirical Study On Ticket Classification Using Machine Learning In OSTicket System

Indrakumar S S
Research Scholar
VTU
Belgaum

Prof. M S Shashidhara
Dept. of Computer Applications,
The Oxford College of Engineering,
Bangalore-560068,India

Venu C P
Associate Architect
Aptean India Pvt Ltd
Bangalore-560010

ABSTRACT

In today's world, everyone is trying to get support online. An issue raised by the end user need to be addressed by the Support team. Support team has to play a key role in verifying issue and classifying the issue to get support from other teams. Classification needs to be done properly to provide improved end user satisfaction. In many organizations, ticket classification still done manually, which is time consuming and requires human effort. There may be human errors, which leads to wrong classification. Manual classification also increases the response time, which results in decrease end user satisfaction.

There are automatic multiple choice phone support systems which provide the user to choose the related categories, but these systems are not much useful because users have never used the system before, usually have no idea about the which option to choose from the menu. In web-based support system, End users do not want to fill long forms, which are needed to identify the issue. In this study, Support team needs classification of the ticket in ticketing tool automatically is proposed. In this system, we have used bag of word approach and machine learning techniques. This method helps the support person to classify the ticket and transfer to the relevant team. It reduces manual efforts and human errors. It helps to improve the end-user satisfaction, also helps in improving the SLA.

Index Terms - Tracking System, Automatic Assignment, Ticket Classification, Machine Learning

1. INTRODUCTION

Now days every one trying online shopping instead of shopping from physical stores which saves their travelling time, searching for shops. Online shopping users needs help or assistance in case of a problem. Customers interact with support staff virtually and they have to use call centers or online web based system. Most of the companies provide these options to their customers. Companies try to improve customer satisfaction by means of better support service quality, quick response and resolution of issues with minimal procedural steps. All of these reveal the fact that virtual customer support systems play a critical role in organization's support operations

End user issues can be collected in many ways, such as e-mail, web system, mobile application, call center, monitoring systems, social networks like Twitter and Facebook also used as issue capturing sources.

A typical Support desk can effectively perform several functions. It provides a single or multiple point of contact for users to gain assistance in troubleshooting, get answers to questions, and solve problems. A support desk generally manages its requests with software such as issue tracking systems. In standard, practice all support systems support process, all requests are analyzed and assessed by the organizations support team. In large organizations, better allocation and effective usage of the valuable support resources is directly results in substantial cost cuts.

In this study, automatically assigning the tickets to the relevant department is proposed, using machine-learning techniques, the recommended extension, which is capable of responding to the needs of the large organizations, reduces manual efforts and human errors while ensuring high quality service levels and improved end-user satisfaction.

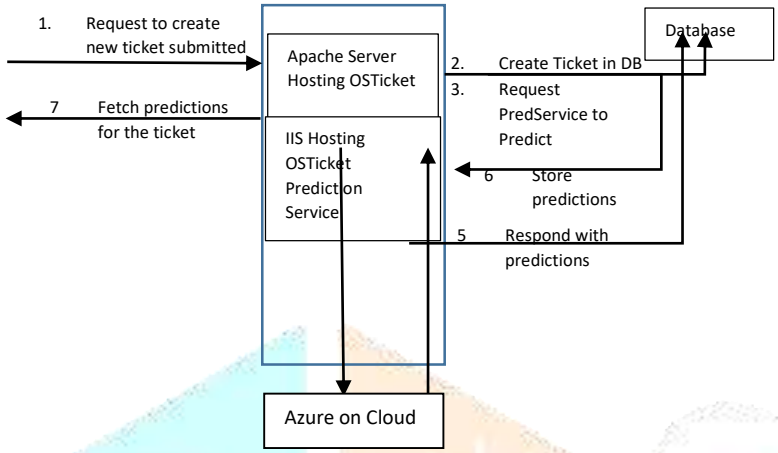
2. RELATED WORKS

Wei et al, 2007, proposed CRF method, which demonstrated 82% accuracy. Diao et al, 2009, suggested rule based crowd-sourcing approach. Sakolnakorn et al proposed a framework of automatic resolver group assignments of IT service desk in banking business based on the text mining. Weiss et al, 2002, suggested method to choose only m best keywords of each is used. Bruno et al, 2005, used Web Service Description Language documents to classify with accuracy of 83% using support vector machine and term frequency–inverse document frequency (tf-idf) weighting factor. Sebastian, 2002 (Yang & Liu, 1999), proposed the document classification work in agglutinative languages using modified statistical methods (Tantug, 2010), the study of filtering undesired mails for agglutinative languages (Ozgur et al, 2004), the work which aims to estimate sex and style writer from unstructured natural text (Amasyali & Yildirim, 2006), text document classification study using shorter root of word (Cataltepe et al, 2007).

3. PROPOSED SYSTEM

Ticket classification is text classification problem. This problem is a widely studied problem in which various algorithms and feature extraction techniques can be used. However, the proposed system is language independent, the implementation of the systemmay require additional language preprocessing steps.

The proposed system allows Support technician to choose appropriate department by giving list of department based on the text in the OSticket system. For example, an issue ticket describing network connection problem must be directed to network department since the proposed system is semi-automatic, if the prediction confidence of each classification is greater than the predetermined threshold value. An operator to assign to related category performs manual classification of ticket. According the classifications results, the issue tickets are assigned to the support staff whom has the right expertise with the issue described in ticket in order to return a response to end user. Figure 2 shows the proposed system architecture.



4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

The recommended system is implemented using previously labeled data of ostciket System. In this section, implementation stages are discussed respectively.

4.1 Dataset

To conduct our experience a dataset consisting of approximately ten thousand issue tickets in collected from OSTicket System, which is a web application that users can request on various issues to different departments within the organization. Each issue tickets contain user name, Email Address, Telephone, user, Help topic subject, Details of the problem. A typical example of issue ticket is given in Figure 3. The problem definition of each ticket is defined in unstructured natural language text. In this study to categorize help tickets, category, free form ticket content and ticket details are used. The rest of attributes such as First Name, Email and Telephone number of tickets are ignored.

4.2 Pre-Preparation

In the preparation step, purifying of tickets from html and numerical expression tags was carried out.

4.3 Feature Extraction

A bag-of-words model is a way of extracting features from text for use in modeling, such as with machine learning algorithms. The approach is very simple and flexible, and can be used in a myriad of ways for extracting features from documents.
A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

1. A vocabulary of known words.
2. A measure of the presence of known words.

It is called a "bag" of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

Bag of word technique needs following steps to be performed

Step 1: Collect Data
Step 2: Design the Vocabulary
Step 3: Create Document Vectors

4.4 Managing Vocabulary

There are simple text cleaning techniques that can be used as a first step, such as: Ignoring case, Ignoring punctuation, Ignoring frequent words that don't contain much information, called stop words, like "a," "of," etc., Fixing misspelled words and Reducing words to their stem using stemming algorithms.

4.5 Scoring Words

Once a vocabulary has been chosen, the occurrence of words in example documents needs to be scored. Some additional simple scoring methods include Counts. Count the number of times each word appears in a document. Frequencies. Calculate the frequency that each word appears in a document out of all the words in the document.

4.6 Word Hashing

We can use a hash representation of known words in our vocabulary. This addresses the problem of having a very large vocabulary for a large text corpus because we can choose the size of the hash space, which is in turn the size of the vector representation of the document.

Words are hashed deterministically to the same integer index in the target hash space. A binary score or count can then be used to score the word. This is called the "hash trick" or "feature hashing". The challenge is to choose a hash space to accommodate the chosen vocabulary size to minimize the probability of collisions and trade-off sparsity.

4.7 TF-IDF

A problem with scoring word frequency is that highly frequent words start to dominate in the document, but may not contain as much "informational content" to the model as rarer but perhaps domain specific words.

One approach is to rescale the frequency of words by how often they appear in all documents, so that the scores for frequent words like "the" that are also frequent across all documents are penalized.

This approach to scoring is called Term Frequency – Inverse Document Frequency, or TF-IDF for short, where: Term Frequency: is a scoring of the frequency of the word in the current document. Inverse Document Frequency: is a scoring of how rare the word is across documents. The scores are a weighting where not all words are equally as important or interesting. The scores have the effect of highlighting words that are distinct in a given document.

4.7 Limitations of Bag-of-Words

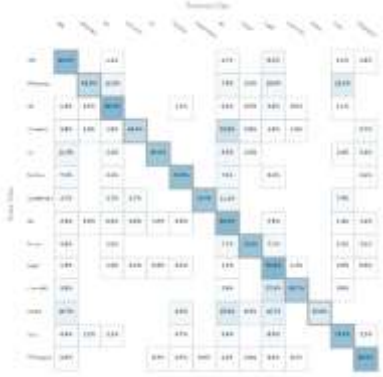
The bag-of-words model is very simple. It has been used with great success on prediction problems like language modeling and documentation classification also, it suffers from some shortcomings, such as: Vocabulary, Sparsity and Meaning.

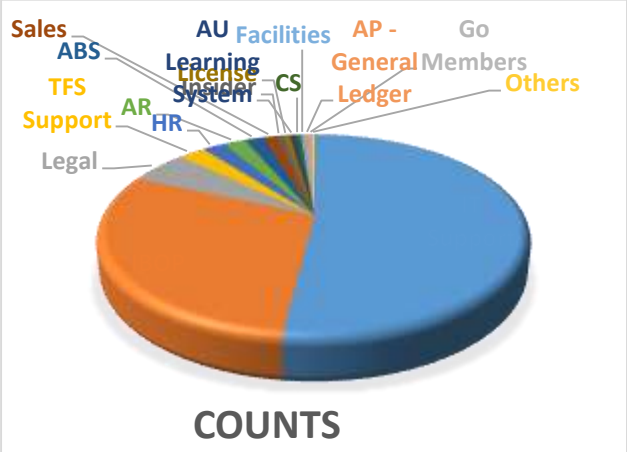
5. CLASSIFICATION

The number of classes in a dataset affects negatively the classification accuracy. Therefore, class number must be small as possible. If the data set contains a large number of classes, classification process should be divided into appropriate sub-classifications and be carried out sequentially.

Metrics

Overall accuracy	0.828029
Average accuracy	0.975433
Micro-averaged precision	0.828029
Macro-averaged precision	0.824719
Micro-averaged recall	0.828029
Macro-averaged recall	0.70344





6. DATABASE CHANGES

Tables added

Table Name: ost_prediction_category_map: This table is a master table, which holds category name and the default department for which the ticket will be assigned.

Column Name	Description
ost_Prediction_Category_Map_Id	Primark Key
Category_Name	Name of the category. This name gets displayed in the "Transfer Ticket" section of the view ticket page for staff
Default_Department_Id	Default department, which the ticket will be transferred to.
Included_Departments	All related departments that gets grouped under one category.

Table Name: ost_ticket_dept_prediction: This table stores the predictions done for every ticket.

Column Name	Description
ost_TicketDeptPrediction_id	Primary Key
Scored Probability	Probability percentage of ticket belonging to the category
Ticket_Id	Ticket Id for which predictions are made.
Prediction_Category_Map_Id	Foreign key for table ost_Prediction_Category_Map_Id
department_id	Default department ID

Rank	Rank of the scored probabilities per ticket

7. CONCLUSION

The manual assignment of issue tickets to appropriate unit or person in support team is not feasible sufficiently for large organizations. It is time consuming and there may be mistakes due to human errors. In this study, to assign tickets automatically, a model based on supervised machine learning algorithms is proposed. Dataset consisting of previously categorized tickets are used to train classification algorithms. Bag of words approach is utilized to extract features vectors. Morphological analysis of terms is performed to avoid data sparseness problem and decrease the vector size. Proposed approach reduces manual efforts and human errors while ensuring high service levels and improved end-user satisfaction. In addition, the proposed system provides to a large organization better allocation and effective usage of the valuable support resources.

8. ACKNOWLEDGMENT

I would like to thank Aptean IT and Research Teams for their support in implementing.

9. REFERENCES

[1] Alpaydın, E. (2010). Introduction to Machine Learning, Second Edition, The MIT Press, ISBN-10: 0-262-01243-X, ISBN-13: 978-0-262-01243-0

[2] Amasyalı, M. F., & Diri, B. (2006). Automatic Turkish text categorization in terms of author, genre and gender. In Natural Language Processing and Information Systems (pp. 221-226). Springer Berlin Heidelberg. ISO 690

[3] Bruno, M., Canfora, G., Di Penta, M., &Scognamiglio, R. (2005). An approach to support web service classification and annotation. In e-Technology, e-Commerce and e-Service, EEE'05. Proceedings. The 2005 IEEE International Conference on (pp. 138-143). IEEE.

[4] Diao, Y., Jamjoom, H., & Loewenstern, D. (2009). Rule-based problem classification in it service management. In Cloud Computing, 2009.CLOUD'09.IEEE International Conference on (pp. 221-228). IEEE.

[5] Eryigit, G. (2014). ITU Turkish NLP Web Service, 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden

[6] Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques, Informatica 31, 249–268

[7] Ozgur, L., Gungor, T., & Gurgen, F. (2004). Adaptive anti-spam filtering for agglutinative languages:a special case for Turkish. Pattern Recognition Letters, 25(16), 1819-1831.

[8] Sebastian, F. (2002). Machine Learning in Automated Text Categorization, ACM Computing Surveys (CSUR) Surveys Homepage archive Volume 34 Issue 1, 1-47 ACM New York, NY, USA, doi:10.1145/505282.505283

[9] Sakolnakorn, P. P. N., Meesad, P., & Clayton, G. Automatic Resolver Group Assignment of IT Service Desk Outsourcing in Banking Business

[10] Tantug, A. C. (2010). Document Categorization with Modified Statistical Language Models for Agglutinative Languages. International Journal of Computational Intelligence Systems, 3(5), 632-645.

[11] Wei, X., Sailer, A., Mahindru, R., & Kar, G. (2007). Automatic structuring of it problem ticket data for enhanced problem resolution. In Integrated Network Management, 2007.IM'07.10th IFIP/IEEE International Symposium on (pp. 852-855). IEEE.

[12] Weiss, S. M., &Apte, C. V. (2002). Automated generation of model cases for help-desk applications. IBM systems journal, 41(3), 421-427.

[13] Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods, SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval 42-49 ACM New York, NY, USA, ISBN:1-58113-096-1 doi:10.1145/312624.312647

[14] Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization.In ICML (Vol. 97, pp. 412-420).

[15] Youn, S., McLeod, D. (2007). A Comparative Study for Email Classification, Advances and Innovations in Systems, Computing Sciences and Software Engineering, pp 387-391