Future Generation Computer Systems 78 (2018) 41–58

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Signature based trouble ticket classification

Jian Xu [a], Hang Zhang [a], Wubai Zhou [b], Rouying He [a], Tao Li [b,c,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[b] School of Computer Science, Florida International University, Miami, FL, USA
[c] School of Computer Science, Nanjing University of Posts and Telecommunications, China

HIGHLIGHTS

- A trouble ticket classification framework using ticket partition and signature construction is presented.
- The ticket partition and signature construction is regarded as an optimization problem and is solved by a local search strategy.
- A signature based ticket classification algorithm is proposed to identify the problem type of an incoming ticket.
- An empirical study on real world ticket data from a large enterprise IT infrastructure is conducted.

ARTICLE INFO

Article history:
Received 18 December 2016
Received in revised form 22 May 2017
Accepted 21 July 2017
Available online 12 August 2017

Keywords:
Document clustering
Ticket classification
Domain knowledge
Local search
Signature
Semantic similarity

ABSTRACT

When a critical system exhibits an incident during its operation, a ticket is usually generated by the monitoring systems or users to describe its issue and should be fixed by system maintenance teams in an acceptable short period of time to avoid serious economic or reputation losses. Although there are a few works about ticket classification, they suffer from poor performance because of the obvious characteristics of unstructured, short free-text with large vocabulary size, large volume, and so on.

To address this performance issue, this paper proposes a trouble ticket classification framework that automatically and accurately identifies the problem type of an incoming ticket. First, a ticket partition and signature construction algorithm is developed, which integrates domain knowledge to improve the quality of data preparation and applies a local search strategy to simultaneously construct ticket groups and their signatures. And then, a signature based ticket classification algorithm is proposed to identify the problem type of an incoming ticket by finding a group signature with the most similarity satisfying the similarity threshold. To demonstrate the effectiveness of the proposed solution, we empirically validate it on real world ticket data from a large enterprise IT infrastructure. Experiments show that our solution outperforms other alternatives in terms of the overall performance.

1. Introduction

1.1. Motivation

Operations support and customer support are business critical functions that have a direct impact on the quality of service provided to customers. When any event which is not part of the standard operation of a service and which may cause an interruption or a reduction happens, the quality of that service is monitored, and a system-generated ticket (abbreviated to SGT for simplicity) is automatically created by the monitoring system. When using an IT resource, the users sometimes face errors, faults, difficulties or special situations that need attention from system management experts, a user-generated ticket (abbreviated to UGT for simplicity) is also created by help desks or administrators. Both a SGT and a UGT are called trouble ticket, or incident ticket, or problem ticket.

In general, troubles described in tickets in the future are unpredictable and often demand rapid allocation of skilled resources to bring an abnormal service back to normal. For example, management of power failures or dangerous situations is an absolutely critical trouble in the electric distribution sector. Hence, a ticket management system should be a time-bound system, which has strict controls on how fast problem should be resolved. In real world applications, quick problem solving mainly relies on the accurate ticket classification. As an initial step of rapid management, the ticket classification is used to identify problem types of tickets based on the problem descriptions and then route them to suitable maintenance teams for problem solving. In a typical ticket system, the ticket classification is done manually by system

* Corresponding author.
E-mail addresses: dolphin.xu@njust.edu.cn (J. Xu), 1033831484@qq.com (H. Zhang), wzhou005@cs.fiu.edu (W. Zhou), 626937431@qq.com (R. He), taoli@cs.fiu.edu (T. Li).

http://dx.doi.org/10.1016/j.future.2017.07.054
0167-739X/

administrators to assign a problem type such as "capacity", "hardware", "file System/dataset", and so on. This manual process is time-consuming and error-prone, especially when there is a large number of tickets in a short period. Hence, we need an automated approach to classify trouble tickets with a high accuracy in order to reduce resource wastage due to delay in wrong identification and routing of tickets.

1.2. Ticket characteristics

Since ticket classification works on ticket problem descriptions, the ticket characteristics will have a great impact on the classification accuracy. Therefore we have to face with several challenges resulted from obvious ticket characteristics.

Trouble tickets comprise two types of fields: structured fields and free-text fields, as shown in Fig. 1. Note that the listed fields are just a small part of all fields, tightly related to the problem solved in this paper. For a SGT, both a structured field and a free-text field are generated automatically by the monitoring system with the predefined alert templates. Thus, a free-text field is relatively regular. For a UGT, both a structured field and a free-text field are generated from administrators or users by the user interface of a ticket report system, and a free-text field is less regular. Actually, the unstructured and free-text field, problem description, provides a detail view of a trouble.

Based on the above ticket structure and historical ticket samples, we can find four obvious characteristics and their corresponding challenges.

(1) There are many non-English words in ticket problem descriptions, and they need be differentiated. For SGTs, the number of non-English dictionary words in SGTs is very large, and even far more than the number of English dictionary words. The main reason lies in that there are many identifiers such as server names, router names, and alert key names in SGTs. For UGTs, although the number of non-English words is far fewer than the number of English dictionary words, the non-English words mainly source from spelling errors and abbreviation. Besides the large size, non-English words also play different roles in problem descriptions. They may not only be entity identifiers that have no contribution to ticket classification and need be removed, but also be domain-specific words, such as "freeSwapSpace" and "usedMemory", that are related to some specific issues and need be preserved.
(2) For tickets with the same problem type, their problem descriptions may vary greatly, particularly in UGTs. For instance, considering two problem descriptions with the same problem type: "b03cxnd1050—High space used (88%) for /fmc/fmcuser" and "b28aedbp1040: The percentage of available space in the filesystem /var is low", both of them are related to the same issue of low available space, and should be semantically equivalent.
(3) The characteristic of relatively short text with a large vocabulary size in ticket problem descriptions often leads to a poor performance when using the most common used vector space model based text mining algorithms on ticket data. Although a single ticket has only a few terms, the vocabulary size of all tickets is very large. Hence, the vector space would be high dimensionality and very sparse, which has a great influence on the performance of ticket classification algorithms. An effective representation model is needed.
(4) There is an inherent difference in the structure and heterogeneity in problem descriptions for two types of tickets. Generally, SGTs have a well-defined structure that defends on the ticket generation templates of the monitoring system. But there are different in structure among the monitoring systems. On the other hand, UGTs are unstructured and contain free-form text written by end users, or support agents. As a result, typos, spontaneous abbreviation, grammatical errors, templates attached with an agent's conversation, addresses, or over length text cutoff are very common. The different types of information comprised in two types of tickets have significant impacts on classification accuracy.

Currently, there is still lack of some systematic automated ticket classification approaches. Most existing solutions [1–4] are to develop supervised or unsupervised ticket classification algorithms on ticket problem descriptions. As for the first challenge, some filter mechanisms such as an input word list [2], a black list and a white list [3], and an Exclude list and an Include List [1], etc., are proposed to improve the quality of data cleaning. As for the second challenge, some solutions derived from the semantic similarity metric between words are proposed [4]. As for the third challenge, the string kernel [5] can be used to extract deep semantic information (e.g., the order of terms) to improve the performance of algorithms. It maps a string to a high dimensional vector to represent all possible term orders. However, because of the large vocabulary size, the dimensionality of the transformed space would be very high. Although the kernel trick does not have to explicitly create those high dimensional vectors, algorithms would still be influenced by the high dimensionality. As for the last challenge, the existing solution [1] is to develop different techniques for two types of tickets respectively, where a graph clustering algorithm is applied to SGTs based on the word matching similarity, while a keyword based approach is applied to cluster similar descriptions for UGTs. Moreover, some solutions utilize both the format and the structure information as well as words of problem descriptions [6,7]. However, these methods only work well for strictly formatted/structured descriptions, such as the descriptions in SGTs, and their performances heavily rely on the format/structure features of descriptions. Hence, the traditional data clustering methods based on the bag-of-word model cannot perform well when applied to UGTs.

1.3. Contributions

To deal with these challenges, this paper proposes a trouble ticket classification framework for two types of tickets, SGTs and UGTs. In our framework, two critical algorithms, the ticket partition and signature construction algorithm and the signature based ticket classification algorithm, are proposed to identify the problem type of a new incoming ticket. Moreover, ticket representation models, domain-specific data preparation, signature representation approaches, and similarity measures between a ticket and signature are carefully designed according to the characteristics of SGTs and UGTs in order to improve the accuracy of ticket classification.

Our main contributions made in this paper include:

(1) We propose an automated trouble ticket classification framework, where the ticket partition and signature construction algorithm and the signature based ticket classification algorithm are proposed for ticket classification with a high accuracy. The former is regarded as an optimization problem instead of a traditional classification or clustering problem to partition historical tickets into several groups and build group signatures, while the latter can be used not only to assign a problem type for a ticket, but also to discover new problem types.
(2) The domain knowledge extracted from historical tickets, such as the to-keep words, the to-discard words and the synonym library, is incorporated into the framework to further improve the accuracy of ticket classification.