Expert Systems With Applications 60 (2016) 130–140

Contents lists available at ScienceDirect

Expert Systems With Applications

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

System situation ticket identification using SVMs ensemble

Jian Xu[a], Liang Tang[b], 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, Nanjing, China

ARTICLE INFO

Article history:
Received 25 July 2015
Revised 1 February 2016
Accepted 15 April 2016
Available online 3 May 2016

Keywords:
Ticket
Classification
Support vector machine
Ensemble

ABSTRACT

System maintenance for large and complex IT infrastructures highly depends on automatic system monitoring, and the performance of system monitoring depends on their configurations specified by system administrators. Misconfigurations and frequent configuration changes are two main causes responsible for false positives (false alarms) that can consume limited maintenance resources and false negatives (missing alerts) that can cause serious system faults. Thus, identifying situation tickets that are created by humans is a critical task to help system administrators correct and improve the configurations of existing monitoring systems to minimize the false negatives.

To address this issue, this paper proposes a situation ticket identification approach based on an ensemble of Support Vector Machines (SVMs), named STI-E, to discover situation tickets from the manual tickets that are created by humans. A primary advantage of this solution is that it can label the most representative tickets from the imbalanced manual tickets by administrators with minimal labeling effort using the discovered domain words from historical monitoring tickets. The proposed SVM ensemble classification model is also able to identify situation tickets with a higher accuracy than the classical SVM classification model. To demonstrate the effectiveness of the proposed approach, we empirically validate it on real system monitoring and manual tickets from a large enterprise IT infrastructure.

1. Introduction

Modern IT infrastructures are maintained by automating routine maintenance procedures, including problem detection, determination and resolution (Tang, Li, Pinel, Shwartz, & Grabarnik, 2012). Automatic system monitoring provides an effective and reliable means for problem detection in enterprise IT infrastructures. Coupled with automated ticket creation, system monitoring ensures that a degradation of the vital signs, defined by acceptable thresholds or monitoring conditions, is flagged as a problem candidate and sent to supporting personnel as an incident ticket (Tang, Li, Shwartz, & Grabarnik, 2013). Defining monitoring conditions (situations) requires the knowledge of a particular system as well as the relationships among different hardware and software components. Continuous updating of IT infrastructures also leads to a number of system alerts that are not captured by system monitoring (i.e., false negatives). The false negatives eventually cause system faults (such as system crashes and data loss). These faults are extremely harmful to enterprise users. When missed or misconfigured monitoring alerts are discovered, the system administrators would identify and correct monitoring settings.

However, discovering situation tickets from manual tickets is a quite challenging task. First, the manual ticket data is highly imbalanced. Most system alerts are captured by automatic monitoring system. Very few of them are missed but recorded by the system administrators. At the same time, manual tickets are responsible for tracking customer requests and other issues. Hence, most of manual tickets are customer requests. Only very few of them are about system situation alerts. Second, the number of manual tickets is very large and labeled tickets are very limited. The large amount of manual tickets makes manual labeling practically impossible. In addition, most system administrators are only working on certain types of tickets. Only a few senior experts can label all tickets. Thus, it is difficult to obtain a large number of labeled tickets for the situation ticket identification. Third, the manual tickets are often short textual messages with a large vocabulary size. The texts of the monitoring tickets are generated from the monitoring system with a problem-specific vocabulary. Compared with the monitoring ticket classification, situation ticket identification from the manual tickets is more difficult.

The goal of our proposed approach is to improve the configurations of existing monitoring systems to minimize the false

[*] Corresponding author. Fax: +13053483549.
E-mail addresses: dolphin.xu@njust.edu.cn (J. Xu), ltang002@cs.fiu.edu (L. Tang), taoli@cs.fiu.edu (T. Li).

http://dx.doi.org/10.1016/j.eswa.2016.04.017
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

negatives. The proposed approach consists of three steps. In the first step, a domain word discovery algorithm is applied to historical monitoring tickets in order to discover the domain words and enhance the domain knowledge. Afterwards, a selective labeling policy is used to appropriately choose sample tickets by incorporating the discovered domain knowledge. This step is critical to reduce the manual ticket labeling effort while maintaining a good classification quality. In the final step, an ensemble of SVM classification algorithm is utilized to automatically discover the false negatives in the manual tickets, where the combination of under-sampling and over-sampling techniques is used to cope with the imbalance of the manual tickets and the outputs of individual SVMs are combined to make the classification model more robust. We conducted experiments to understand the coverage and precision of the proposed method, as well as the training efficiency in comparison with known methods.

Three main contributions made in this paper include:

(1) We propose a domain words discovery algorithm to obtain domain knowledge from historical monitoring tickets rather than from the traditional sources, such as system administrators, related documents and the taxonomy of system management, which is contributed to accurate situation ticket identification.
(2) We design a selective labeling policy by taking the discovered domain words into consideration, which provides a better chance of selecting the situation tickets in the training tickets, while keeping the manual labeling efforts minimal.
(3) We propose an ensemble of SVM classification model, which integrates the classification results of individual base classifiers learned from the different training ticket sets using the resampling technique.

The rest of the paper is organized as follows: Section 2 provides a description of the problem settings. Section 3 presents our situation ticket identification approach. In Section 4, we present our empirical studies with real IT monitoring tickets and manual tickets. Section 5 summarizes the related work for ticket classification and ticket mining for system management. Section 6 concludes the paper.

2. Problem description

In IT Service management, tickets are often short textual messages which describe system incidents. Those system incidents can be any type of system issues, such as the system alerts and customer requests. The system alerts are generally about high utilization of the disk space, crashes of a database, and so on. The customer requests are about resetting database passwords, installing a new web server, and so on. We call the system alerts related ticket as the situation ticket. Most of the situation tickets are generated from the monitoring system, and they are also called monitoring tickets. But a few situation tickets generated from system administrators, helpdesk or end users are hidden in the manual tickets. The ticket categories and examples are shown in Fig. 1.

The missed monitoring alerts or false negative alerts recorded by the system administrators in the manual tickets are not captured by system monitoring due to misconfigurations and configuration changes. When missed monitoring alerts are discovered by identifying the situation tickets from the manual tickets, the system administrators would improve monitoring settings. Thus, we study the situation ticket identification problem in this paper, i.e., the problem of finding all situation tickets from the manual tickets. Note that discovering the situation tickets from a collection of textual tickets is a binary text classification problem in nature. Given a manual ticket, our approach classifies it into “1” or “0”, where “1” indicates this ticket is a system alert situation and “0” otherwise. Because real-world IT infrastructures are often over-monitored, the customer request is the majority of the manual tickets. As a result, the dataset of manual tickets is highly imbalanced.

3. Semi-supervised situation ticket identification

It is time-consuming for human experts to scan all manual tickets and label their classes. In our proposed approach, we aim to tackle the problem of situation ticket identification while keeping the manual labeling effort minimal. Note that achieving good predictions is as important as minimizing the manual labeling effort. A general overview of our ticket identification approach is as shown in Fig. 2.

In our approach, we only select a small proportion of tickets for labeling. But the selection is crucial for the highly imbalanced data. The situation tickets from the manual tickets are very rare. If the selected ticket set contains none of them, the classification model cannot be trained well. Before we obtain the class labels, however, we do not know in advance if a ticket is related to the monitoring situations or not. We took the classic approach of utilizing the domain words in system management. The domain words can be obtained from existing historical monitoring tickets, which is the first step of our approach. Second, we design a ticket selection policy to select proper tickets for labeling and utilize the domain words to increase the size of the training tickets. We treat each domain word as a pseudo-ticket and put all pseudo-tickets into the training ticket set. All pseudo tickets are considered as situation tickets, so the label of each pseudo ticket is “1”. Although the selected training tickets have a better chance of containing the situation tickets, the set of the situation tickets is still the minority of all selected training tickets. The highly imbalanced training data affects the performance of the SVM classification model (Chawla, Bowyer, Hall, & Kegelmeyer, 2011). To deal with the imbalanced data, the minority class tickets are over-sampled until the number of positive tickets is equal to the number of the negative tickets. Finally, heterogeneous SVMs are trained and then integrated for ticket identification. We choose SVM as the classification model because it is known to perform the best in text classification (Joachims, 1998; Pang-Ning, Steinbach, & Kumar, 2005). We provide the details of each step in the following sections.

3.1. Domain words discovery

Traditionally, the domain words can be obtained from the system administrators and the related documents, and can also be obtained from the catalog taxonomy of the system management. Generally, domain words obtained from these sources are some proper nouns or verbs. Our goal is to identify the situation tickets described by texts from the manual tickets. The intuition is that those words frequently appearing in the problem description, the alert summary, and the corresponding resolution of the historical monitoring tickets may have a better chance to be used in the description text of a manual ticket. This motivates us to discover domain words from historical monitoring tickets. Note that discovering domain words from historical monitoring tickets may be more accurate than those humans’ domain knowledge based methods. Table 1 lists some examples of the domain words with their corresponding causes from different sources.

In this paper, we propose a domain words discovery algorithm from the historical monitoring tickets, based on the idea of frequent pattern mining commonly used in the association rule mining. The algorithm is described in Algorithm 1.

Lines 1 to 5 obtain refined bag-of-words. In the extraction step, we extract domain words only from the “summary” and “resolution” text of a historical ticket rather than all ticket texts. In the stemming step, we reduce domain words to their stem forms. This