Journal of Computer and System Sciences 80 (2014) 27–38

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

WCOID-DG: An approach for case base maintenance based on *W*eighting, *C*lustering, *O*utliers, *I*nternal *D*etection and *D*bsan-*G*means

Abir Smiti *, Zied Elouedi

LARODEC, Université de Tunis, Institut Supérieur de Gestion de Tunis, 41 Avenue de la liberté, cité Bouchoucha, 2000 Le Bardo, Tunisia

ARTICLE INFO

Article history:
Received 19 July 2012
Received in revised form 13 November 2012
Accepted 14 March 2013
Available online 21 March 2013

Keywords:
Case based reasoning
Case base maintenance
Gaussian-Means clustering
Density based clustering
Outliers detection

ABSTRACT

The success of the Case Based Reasoning system depends on the quality of the case data and the speed of the retrieval process that can be costly in time, especially when the number of cases gets bulky. To guarantee the system's quality, maintaining the contents of a case base (CB) becomes unavoidably. In this paper, we propose a novel case base maintenance policy named WCOID-DG: *W*eighting, *C*lustering, *O*utliers and *I*nternal cases *D*etection based on *D*bscan and *G*aussian means. Our WCOID-DG policy uses in addition to feature weights and outliers detection methods, a new efficient clustering technique, named DBSCAN-GM (DG) which is a combination of DBSCAN and Gaussian-Means algorithms. The purpose of our WCOID-GM is to reduce both the storage requirements and search time and to focus on balancing case retrieval efficiency and competence for a CB. WCOID-GM is mainly based on the idea that a large CB with weighted features is transformed to a small CB with improving its quality. We support our approach with empirical evaluation using different benchmark data sets to show its competence in terms of shrinking the size of the CB and the research time, as well as, getting satisfying classification accuracy.

1. Introduction

One of the great aspirations of Artificial Intelligence (AI) is to create smart methods and systems able to understand and emulate human reasoning. Case Based Reasoning (CBR) [1,2] is a diversity of reasoning by analogy. It is a technique to model the human way in reasoning and thinking. It is able to find a solution to a problem by employing its luggage of knowledge or experiences which are presented in form of cases. To solve the problems, CBR system calls the past cases, it reminds to the similar situations already met. Then, it compares them with the current situation to build a new solution which, in turn, will be incorporated into the existing case base (CB).

CBR has been used to create several applications in different domains, such as manufacturing, medicine, law, technical maintenance, quality control, etc. For instance, in the diagnosis [3,4], CBR is used to build intelligent system which increasingly finds entry into the industrial practice. Recently, CBR approaches have been applied in the medical domain, the diagnosis of CBR is performed to solve new problems by remembering solutions to problems that are similar to the current problem. Moreover, for the Data Mining domain, the natural ability of the CBR is used to access corporate data structure (Corporate Memories), and to search for important relationships within medical, financial, marketing, and commercial data [5]. In addition, in the e-commerce fields, CBR has been considered as an assistant in e-commerce stores and

* Corresponding author.
E-mail addresses: smiti.abir@gmail.com (A. Smiti), zied.elouedi@gmx.fr (Z. Elouedi).

0022-0000/$ – see front matter
http://dx.doi.org/10.1016/j.jcss.2013.03.006

CASE: 1
Title: Envelopes jam laser printer due to glue.

Description: The copies are in bad quality.
We have black traces.
The cleaning of the printer does not regulate the problem...

Action: Check the toner and replace it.

Fig. 1. Example of a case in the CBR system.

as a reasoning agent for online technical support, as well as an intelligent assistant for sale support or for e-commerce travel agents. It uses cases to describe commodities on sale and identifies the case configuration that meets the customers' requirements [6].

Case Based Reasoning system is built to work for long periods of time, it adds cases to the case base through the retain process. As a result, the case base can grow very fast in the sense that it can affect negatively the CBR's quality results and can slow the speed of the query execution time concerning case-research phase. To ensure the system's betterment, maintaining CBR system becomes required. As a result, there has been a significant increase in the research area of Case Base Maintenance (CBM). Its objective is to guarantee a good operating in time of an information processing system and to facilitate future reasoning for a particular set of performance objectives [7].

Various case base maintenance policies have been proposed to maintain the Case Base (CB). One branch of research has focused on the partitioning of case base which builds an elaborate CB structure and maintains it continuously [8,9]. Another branch of research has focused on CBM optimization which uses an algorithm to delete or update the whole of CB [10–12]. Generally, there are some problems when using these CBM methods. In fact, many of them are expensive to run for large CB, and suffer from the decrease of competence especially when there exist some noisy cases, since the system's competence depends on the type of the cases stored. Furthermore, they require many input parameters which are hard to determine but have a significant influence on the results quality.

In a preliminary paper [13], we have presented a WCOID deletion policy for CBM, which suffers from some shortcomings as it is sensitive to the parameters requirement which are difficult to know and can have a negative influence on the results when an ordinary user introduces such parameters.

This paper extends this initial approach to handle this drawback and propose a novel approach for automatically maintaining case base, named WCOID-GM: *W*eighting, *C*lustering, *O*utliers and *I*nternal cases *D*etection based on *D*bscan and *G*aussian means. WCOID-GM decides which cases to delete and adds feature weights to achieve a higher reduction and better competence case bases. It focuses on balancing case retrieval efficiency and competence for a case base. It employs a new efficient clustering technique, named DBSCAN-GM (DG) which is a combination of DBSCAN and Gaussian-Means algorithms. The method could be able to maintain the case bases by giving satisfying accuracy, reducing its size, and consequently reducing the case retrieval time.

This paper is organized as follows: In Section 2, CBR system and its cycle will be presented. In Section 3, some of strategies for maintenance the case base will be approached. Section 4 describes in detail our new approach WCOID-GM for maintaining case base. Finally, Section 5 presents and analyzes experimental results carried out on data sets from the U.C.I. repository [14]. Finally, Section 6 ends this work and presents future works.

2. Case Based Reasoning

Case Based Reasoning (CBR) is defined as follows: "*CBR is a problem-solving paradigm that solves a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation* [1]".

More specifically, CBR uses a database of problems to resolve new problems, called Case Base (CB). The CB can be built through the knowledge engineering process or it can be collected from previous cases.

In a problem-solving system, each case would describe a problem and a solution to that problem. The CBR solves new problems by adapting relevant cases from the library. Moreover, CBR can learn from previous experiences. When a problem is solved, the case based reasoning can add the problem description and the solution to the case library. The new case in general represented as a pair ≪problem, solution≫ is immediately available and can be considered as a new piece of knowledge (see Fig. 1).

The CBR process can be represented by a schematic cycle, as shown in Fig. 2. It can be described typically as cyclical process comprising the four REs:

- RETRIEVE the most similar cases: During this process, the case base reasoning searches the database to find the most approximate case to the current situation.
- REUSE the cases to attempt to solve the problem: This process uses the retrieved case and adapts it to the new situation. At the end of this process, the reasoner might propose a solution.
- REVISE the proposed solution if necessary: Since the proposed solution could be inadequate, this process can correct the first proposed solution.