Information Sciences 336 (2016) 129–144

Contents lists available at ScienceDirect

Information Sciences

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

Using minimal generators for composite isolated point extraction and conceptual binary relation coverage: Application for extracting relevant textual features

S. Elloumi[*], F. Ferjani, A. Jaoua

Qatar University, Qatar

ARTICLE INFO

Article history:
Received 10 October 2014
Revised 8 October 2015
Accepted 8 December 2015
Available online 21 December 2015

Keywords:
Fringe relation
Isolated labels
Minimal generator
Conceptual coverage
Textual feature extraction

ABSTRACT

In recent years, several mathematical concepts have been successfully explored in the computer science domain as a basis for finding original solutions for complex problems related to knowledge engineering, data mining, and information retrieval. Hence, relational algebra (RA) and formal concept analysis (FCA) may be considered as useful mathematical foundations that unify data and knowledge into information retrieval systems. For example, some elements in a fringe relation (related to the (RA) domain) called isolated points have been successfully used in FCA as formal concept labels or composite labels. Once associated with words in a textual document, these labels constitute relevant features of a text. This paper proposes the MinGenCoverage algorithm for covering a Formal Context (as a formal representation of a text) based on isolated labels and using these labels (or text features) for categorization, corpus structuring, and micro–macro browsing as an advanced information retrieval functionality. The main thrust of the approach introduced here relies heavily on the close connection between isolated points and minimal generators (MGs). MGs stand at the antipodes of the closures within their respective equivalence classes. By using the fact that the minimal generators are the smallest elements within an equivalence class, their detection and traversal is greatly eased and the coverage can be swiftly built. Extensive experiments provide empirical evidence for the performance of the proposed approach.

1. Introduction

Formal concept analysis (FCA) [40] and relational algebra (RA) [36] have been successfully used to formalize the data space and discover regularities embedded within it. It is assumed here that data can be mapped into a binary context defined by a binary relation (BR) between a set of objects and a set of properties, attributes, or items. For example, web pages can be related to index terms, a customer with products purchased, diseases with medicines, etc. Moreover, a textual document can be transformed into a binary relation where objects are sentences and properties are words. In all these cases, it is always necessary to extract regular associations between objects, or attributes, to make right decisions in similar cases. Information regularities can in addition be exploited to keep data size to a minimum and to extract frequent and reliable associations considered as pertinent knowledge.

[*] Corresponding author. Tel.: +216 20317870.
E-mail addresses: elloumi.samir@gmail.com, samir.elloumi@fst.rnu.tn (S. Elloumi), fethif@qu.edu.qa (F. Ferjani), jaoua@qu.edu.qa (A. Jaoua).

http://dx.doi.org/10.1016/j.ins.2015.12.013
0020-0255/© 2015 Elsevier Inc. All rights reserved.

As a mathematical background, FCA and RA have been already combined and used to discover regularities in data [22]. In fact, a formal concept represents the atomic regular structure for decomposing a BR. Riguet's difunctional relation (fringe relation) [34], whose elements are defined as isolated points, describes invariant regular structures that may be used for database decomposition. Formal concepts and isolated points are also connected to each other because an isolated point belongs only to a unique formal concept [25]. The connections between formal concepts and isolated points have already been explored by the authors, who have proposed the algorithm "GenCoverage" [13] to build the conceptual BR coverage based on isolated points. GenCoverage extracts the formal concepts associated with isolated points in the fringe relation and performs property compositions (PC) and fringe recalculations using PC until the whole BR is covered. However, the major difficulty with GenCoverage was the combinatorial nature of the PC possibilities. Although the number of PCs was restricted to two, serious performance deficiencies were encountered when handling large dense data sets [13].

Naturally, data mining techniques are the most suitable and efficient tools for processing large datasets [21,37]. A tremendous effort towards building optimized and compact structures such as closed itemsets [30], minimal generators [31], and generic bases [8,16,20] has been undertaken in that field.

This paper relies on the close connection between isolated points and minimal generators (MGs). MGs stand at the antipodes of the closures within their respective equivalence classes. By using the fact that MGs are the smallest elements within an equivalence class [19], their detection and traversal is greatly eased, and the BR coverage can be swiftly built. Hence, a new approach called MinGenCoverage is presented here for isolated point extraction and BR covering. It ensures significant improvements in terms of complexity reduction and output quality compared with other approaches.

This paper is organized as follows. Section 2 presents existing work related to conceptual approaches and their usefulness in knowledge engineering applications. In Section 3, basic notions of FCA and a useful background in data mining that will be useful in the remainder of the paper are recalled. Section 4 presents necessary definitions related to isolated points and proves some useful properties related to the relations between isolated points and minimal generators. Section 5 presents the proposed new algorithm devoted to the coverage computation of a BR based on isolated properties. Section 6 reports the results of the experiments carried out and shows the utility of the proposed approach in terms of coupling–cohesion and relevance of the extracted textual features. The experiments in this study have involved two types of datasets (*i*): one related to the financial sector, including event management changes, transactions, and company performance (with data prepared as part of the "National Research Priority Project" (NPRP) [13]); and (*ii*) a second set of benchmark datasets. Section 8 draws conclusions and points out avenues for future work.

2. Related work

The first attempt to define lattice theory as a mathematic model was made by Birkhoff [17] in the 1940s. An underlying deep concept is the notion of the Galois connection, which emerged in the early 1940s after a long gestation period that started at the beginning of the century. Over the last two decades, research has demonstrated how concept lattices formalize conceptual structures by coding any kind of duality, such as the duality between the intent and the extent of a concept. An application to data analysis using this duality to analyze questionnaire data has been carried out by Barbut and Monjardet in the social science domain [3]. The concept lattice, also called the "Galois lattice", was promoted by Wille [40] and then extended as a discipline called "*formal concept analysis*" (FCA) [14]. FCA is a mathematical tool for analyzing data and formally representing conceptual knowledge. FCA helps to form conceptual structures from data, enabling meaningful and comprehensible interpretations [27,33]. Hence, FCA mathematical settings have recently been shown to provide a theoretical framework for the efficient resolution of many practical problems from data mining, software engineering, and information retrieval [6,23,24,39].

An interesting problem is to find coverage of a formal context by a minimal number of concepts. By associating a label with each formal concept of a text, various possible features of a document can be obtained, depending on the minimal conceptual coverage. In this respect, finding the optimal cover of a BR is known to be an NP-hard problem [15]. Nevertheless, a large number of researchers have shown interest in tackling this problem. Belkhiter et al. [5] introduced an optimal rectangular decomposition of a BR as well as an application to documentary databases. The decomposition introduced is based on the selection of optimal maximal rectangles (or equivalently formal concepts) that achieve a maximal gain in terms of storage space. More recently, Khcherif et al. [25] introduced a rectangular decomposition approach based on the Riguet's difunctional relation [34]. The computation of this difunctional is reduced to the determination of a set of isolated points, enabling the minimal set of rectangles covering a given BR to be determined. Belohlavek and Vychodil [7] tackled the same issue by attempting to solve the *Boolean factor analysis* problem by proposing a new method for decomposing a binary matrix into a Boolean product of factors. Their proposal contains the notion of mandatory concepts in a coverage [7], akin to the isolated concepts used in the present study. Fairly recently, Mouakher and Ben Yahia [28] introduced a new approach, based on a greedy algorithm to extract the optimal cover of a BR. This last approach relies on the formal concept lattice representation. The guiding concept of Mouakher and Ben Yahia's approach is that the cover should not be extracted regardless of the quality of knowledge that may be drawn from it. For this reason, the authors introduced a gain function based on their assessment of the correlation of the intent parts of pertinent formal concepts.

Generally, the number of formal concepts grows exponentially with the size of the matrix (i.e., BR) [18]. Fortunately, in the method proposed here, only an information lossless set of a few relevant concepts in polynomial time is selected. Extracting the coverage of a BR $\mathcal{R}$ relies on the use of a regular embedded relation $\mathcal{R}_d$, also called a *fringe relation*. This