International Journal of Cognitive Informatics and Natural Intelligence
Volume 12 • Issue 2 • April-June 2018

A Semantic Information Content Based Method for Evaluating FCA Concept Similarity

Hongtao Huang, Engineering Research Center of Henan Provincial Universities for Educational Information, Henan Normal University, Xinxiang, China

Cunliang Liang, Engineering Research Center of Henan Provincial Universities for Educational Information, Henan Normal University, Xinxiang, China

Haizhi Ye, Engineering Research Center of Henan Provincial Universities for Educational Information, Henan Normal University, Xinxiang, China

ABSTRACT

Probability information content-based FCA concepts similarity computation method relies on the frequency of concepts in corpus, it takes only the occurrence probability as information content metric to compute FCA concept similarity, which leads to lower accuracy. This article introduces a semantic information content-based method for FCA concept similarity evaluation, in addition to the occurrence probability, it takes the superordinate and subordinate semantic relationship of concepts to measure information content, which makes the generic and specific degree of concepts more accurate. Then the semantic information content similarity can be calculated with the help of an ISA hierarchy which is derived from the domain ontology. The difference between this method and probability information content is that the evaluation of semantic information content is independent of corpus. Furthermore, semantic information content can be used for FCA concept similarity evaluation, and the weighted bipartite graph is also utilized to help improve the efficiency of the similarity evaluation. The experimental results show that this semantic information content based FCA concept similarity computation method improves the accuracy of probabilistic information content based method effectively without loss of time performance.

KEYWORDS

Concept Similarity, Hierarchy, Information Content, Probabilistic, Semantic

1. INTRODUCTION

Formal Concept Analysis (FCA) is becoming critical for data analysis. As a result, FCA is widely required by information retrieval, data mining, software engineering, and so on. Assessing concept similarity is a key issue which is growing in importance within the application of FCA. Similarity graph is able to accurately distinguish the similarity between FCA concepts (Formica, 2006; Formica & Missikoff, 2002; Kang & Miao, 2016). However, this method requires human interaction to build similarity graph, therefore, it is time-consuming and error-prone.

(Aouicha & Taieb, 2016; A. Formica, 2008) introduces an information content based method in order to avoid excessive reliance on domain experts. This method is a refinement of a previous proposal of (Formica, 2006). The improvement is information content is measured by concept occurrence

DOI: 10.4018/IJCINI.2018040106

probability in corpus, which makes it be capable of being independent of similarity graph, because the measurement of information content in this method does not depend on similarity knowledge owned by domain expert, which makes FCA concepts similarity evaluation more automatically and efficiently. However, it takes only the frequency of concepts in corpus as criteria to compute information content between concepts, there is still room for further improvement on the accuracy of FCA concept similarity evaluation.

(Mezghanni & Gargouri, 2017; Otero-Cerdeira, Rodríguez-Martínez & Gómez-Rodríguez, 2015; Pirró, 2009; Pirró & Seco, 2008) proposed an improved measure model for information content, this model relies on the hierarchical ontology organized by cognitive significance principle (Blank, 2003). A concept appears more special when it needs to be distinguished from other existing concepts. Therefore, a concept that has more subordinate words has less information capacity than leaf nodes in the hierarchy, as a leaf node do not need to be distinguished from others. (Sánchez, Batet, & Isern, 2011) pointed out that superordinate and subordinate semantic relationship can be used to improve the accuracy of information content. Inspired by (Sánchez, Batet, & Isern, 2011), we propose a semantic information content based FCA similarity evaluation method, which uses superordinate and subordinate relationship between concepts of ontology to evaluate the similarity between FCA concepts. This approach can get better access to general and specific features of FCA concepts, which makes the accuracy of this method higher than probabilistic information content based approach.

2. RELATED CONCEPTS

Information content describes the basic dimension, size and volume of a concept, it represents the information capacity of a concept in a specific environment. A specific or specialized entity contains more information than a general or an abstract entity, this is the basic principle of information content, and which has been widely used in evaluating semantic similarity between concepts (A. Formica, 2008; Jiang, Bai, Zhang, & Hu, 2017; Resnik, 1997). In this work, information content is associated with the scale of the subordinate tree of concepts. No matter how many internal concepts are introduced into the hierarchy, these concepts can be described by leaf nodes on subordinate tree. Furthermore, they can also be distinguished from other concepts with different leaf node sets. The following is the definition of leaf node.

Definition 1: Let C be the concept set of domain ontology O, then for any concept $c \in C$, a leaf node set $L(c)$ is defined by:

$$L(c) = \{c_l \mid c_l \in C \wedge c_l \in hyp(c)\} \tag{1}$$

where c_l is a leaf node on the hierarchy of *c*, $hyp(c_1)$ is the subordinate word set of c_l, c_l is a leaf node if and only if $hyp(c_1) = \varnothing$. Multiple inheritance of internal nodes on subordinate word tree will lead to multiple path among leaf nodes. In order to avoid this redundancy, each leaf node is counted only once when *L(c)* is created. The more leaf nodes there are on the subordinate tree of a concept, the stronger the generality of this concept. Leaf nodes have the same maximum information content value, because they are specific enough to be distinguished from other nodes.

Definition 2: Let C be the concept set of domain ontology O, for any $c \in C$, the information content of c can be defined as:

Adaptive Multiobjective Memetic Optimization

Hieu V. Dang and Witold Kinsner (2016). *International Journal of Cognitive Informatics and Natural Intelligence (pp. 21-58).*
www.igi-global.com/article/adaptive-multiobjective-memetic-optimization/172532?camid=4v1a