Knowledge-Based Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Automatic selection of a single solution from the Pareto front to identify key players in social networks

Dimas de la Fuente*,[a], Miguel A. Vega-Rodríguez[b], Carlos J. Pérez[c]

[a] *Cátedra ASPgems, Edificio de Institutos Universitarios, Avda. de la Universidad s/n, Cáceres 10003, Spain*
[b] *Instituto de Investigación en Tecnologías Informáticas Aplicadas de Extremadura (INTIA), Universidad de Extremadura, Avda. de la Universidad s/n, Cáceres 10003, Spain*
[c] *Facultad de Veterinaria, Universidad de Extremadura, Avda. de la Universidad s/n, Cáceres 10003, Spain*

ARTICLE INFO

Keywords:
Centrality measures
Key players
Multi-objective optimization
Post-Pareto analysis
Social networks

ABSTRACT

Social networks are increasingly growing and identifying the relevant and influential users within a network is becoming a problem of interest in many contexts. Although multi-objective optimization approaches have been proposed to address this problem, they identify a large number of valid and non-dominated solutions, so selecting a relevant single solution over the others is a difficult task. Several methods for post-Pareto optimality analysis have been considered to reduce the Pareto fronts identified by the multi-objective optimization approach to a single solution. Eleven methods have been implemented, tested, and compared. Most of them have never been used before for a reduction task and/or for a key player context. The highest hypervolume method and the method based on the Euclidean distance to the ideal point combine the best averages and the lowest dispersions, reporting statistically significant differences with the rest of the methods. Improvements up to 60.79% have been obtained. This methodology will be implemented in an e-learning platform in order to identify the most relevant and influential students in the social network of a course.

1. Introduction

In recent years, due to the emergence of technologies and social media, there has been a growing interest in social networks as they are present in many real-world disciplines such as politics, marketing, communication... As a consequence, the aim of identifying a set of individuals from a social network that have a relevant influence (key players) is a matter of interest [1–3].

According to the latest research results in [4], the key players in social networks can be divided into superblockers and superspreaders. In our work, a linear threshold model is supposed for the information spreading and not an independent cascade model (which is the one used in [4]). As a consequence of supposing a linear threshold model, the mapping between influence maximization (superspreaders) and optimal percolation (superblockers) could be considered exact or nearly exact [4,5].

Several approaches have been proposed to identify key players focusing on a single objective of interest, usually based on node centrality measures. Nevertheless, these approaches present deficiencies as they perform well only when their objective is considered as the only characteristic the set of key players should have. For example, regarding the node degree centrality measure [6], its prevailing deficiency when identifying key players can be explained by the fact that the key players identified present a substantial amount of direct relationships, but they are located very close to each other, not being able to reach different areas of the social network. Therefore, it seems natural to consider more than one objective with the aim of covering different areas of the social network.

Addressing properly this problem can be achieved by using a multi-objective optimization approach as it takes into account two or more objective functions to be optimized. Regarding the identification of key players sets and to the best of the authors' knowledge, there is only one multi-objective optimization approach published in the scientific literature [7,8]. However, with multiple objectives, there is generally not a single solution optimizing all the objective functions at the same time, but instead there is a set of different solutions that are good and represent different trade-offs of the objectives. These solutions are known as non-dominated solutions or Pareto-optimal solutions [9]. The Pareto front identified contains a large number of non-dominated solutions, so choosing one solution over the others can be a challenging problem for the decision-maker, specially if no preference criteria are considered. In this paper, an approach from a multi-objective optimization point of

* Corresponding author.
E-mail addresses: ddelafuente@unex.es (D. de la Fuente), mavega@unex.es (M.A. Vega-Rodríguez), carper@unex.es (C.J. Pérez).

https://doi.org/10.1016/j.knosys.2018.07.018
Received 18 December 2017; Received in revised form 19 June 2018; Accepted 6 July 2018

view is also addressed.

The choice of one single solution in an optimization problem is usually required in many contexts. A motivating problem is finding key players in an e-learning platform called NeuroK [10] (https://neurok.es/). NeuroK is a new e-learning platform leveraging the latest technologies and implementing learning analytic tools that support pedagogical principles from neuroscience. The application of this approach could identify influential students within the social network, which are relevant in the teaching-learning process.

In this paper, several methods for post-Pareto optimality analysis are considered to automatically select one relevant solution among the several alternatives found by a multi-objective optimization approach. Specifically, a multi-objective artificial bee colony algorithm has been used to derive the Pareto fronts. Most of the implemented methods have never been used before for a reduction task and/or for a key player context. The approach has been tested with six social networks addressing very different topics.

The outline of this paper is as follows. Section 2 presents a review of approaches related to our study. Section 3 formulates the problem of identifying key players in social networks as a multi-objective optimization problem. Then, Section 4 describes the proposed techniques for automatically reducing the Pareto front to a single solution. In Section 5, the obtained results are presented along with the datasets and the quality metrics used to assess the method performance. Finally, Section 6 focuses on conclusions and future research.

2. Related work

Different approaches have been proposed to reduce the set of non-dominated solutions. These methods have been classified into three main categories: reduction methods based on users' preferences, clustering procedures, and distance-based methods.

Firstly, methods based on users' preferences allows to identify optimal solutions that are acceptable and preferred based on the desired users' criteria. One technique used in several studies is the non-numerical ranking preferences (NNRP) method [11], that is based on iteratively generated weight values for the objective functions. Other techniques have also based its performance on the generation of weights as the Greedy Reduction method [12], the Weighted Stress Function Method (WSFM) [13] or the TOPSIS method [14]. There are other approaches considering a tradeoff-analysis technique capable of identifying Pareto-compromise solutions [15] or following a 2-step Pareto filtering procedure that removes low quality solutions [16]. Also, users' preferences have been considered in the selection of a threshold angle in order to identify areas with desirable solutions within the Angle based with Specific bias parameter pruning Algorithm (ASA) [17]. Sorting procedures, such as the UTADIS method, have been also used to categorize the solutions into preference ordered classes [18]. Specific algorithms based on an arbitrary finite collection of users' information have been also proposed for the Pareto set reduction [19]. In spite of the fact that some of the previous approaches are very interesting, they need the interaction of users, which is not possible in contexts where an automatic selection of one relevant solution is sought.

Regarding the clustering procedures, they base their performance on grouping non-dominated solutions into different clusters such that elements within the same cluster have a high degree of similarity. The most popular clustering technique is, probably, the k-means clustering algorithm [20], a partitioning procedure that calculates the centroid for each group and assigns each observation to the group with the closest centroid. Another similar approach for finding a reduced Pareto subset uses fuzzy clustering techniques (FCM) [21]. Alternative partitioning techniques have been proposed such as the Partitioning Around Medoid (PAM) [22]. Four clustering methods: the aforementioned k-means partitioning, Maximum Split Partitioning (MSP), Minimum Diameter Partitioning (MDP), and p-Median Partitioning (PMP) are compared in [23]. Subtractive clustering has been proposed over the k-means and fuzzy c-means methods in other studies where no initial number of clusters is provided. For example, Zio and Bazzo [24] proposes a subtractive clustering based technique. This technique was assessed and the results were compared with the ones obtained from other two clustering techniques: the Self-Organizing Maps (SOM) and Data Envelopment Analysis (DEA) [25]. Another technique is the Dynamically Growing Self-Organizing Trees (DGSOT) [26,27], where the algorithm optimizes the number of clusters and can rearrange misclustered data. All these clustering approaches prune the Pareto-optimal solutions to a reduced set with more than one solution instead of a single one.

Finally, distance-based methods seem to be suitable approaches to automatically select one relevant solution among the several alternatives within the Pareto-optimal set. These methods calculate the existing distance within the objective space between non-dominated solutions. One common technique identifies the relevant non-dominated solution by using the shortest distance to a given ideal point, which best optimizes all the objectives taken into account. For example, in a manufacturing context, a problem was solved by using a multi-objective optimization approach, and the non-dominated solutions were given to this method with the Euclidean distance (L_2) in order to select a single solution [28]. A comparison with another method based on a user's predecided reference point is presented in [28]. On the other hand, Siwale [29] advocates for the use of Tchebycheff distance from each point to the ideal point to identify a compromise solution. As both distance-based approaches are able to optimally identify a single solution from the Pareto-optimal set, both are included within the automatic methods considered in this paper.

3. Problem definition

A general formulation of a multi-objective optimization problem consists of a set $F(x) = \{f_1(x), f_2(x), \ldots, f_L(x)\}$ with L objective functions, which have to be simultaneously optimized as a function of the vector x subject to some possible constraints.

Let x and x' be two solutions from the decision space, x is called a non-dominated solution if it is not possible to find another solution that improves an objective function in value without worsening some of the other objectives. These non-dominated solutions are considered optimal solutions and they are called the Pareto-optimal set [9].

The aim is to identify a Pareto set that optimizes the objective functions taken into account. This paper considers two objective functions to identify key players sets: eigenvector centrality [30] and the distance between key players within the set based on Dijkstra's algorithm [31], both of them to be maximized. These two objective functions have been previously used in this context [7,8].

Regarding the eigenvector centrality, it measures the relative influence of a node in a network based on the relevance of the nodes that are directly connected to it, i.e., the sum of the centralities of its neighbors [30]. It uses the eigenvector and the largest eigenvalue of the respective adjacency matrix of a graph. Hence, for a given graph $G = \{\mathcal{V}, \mathcal{E}\}$, being $\mathcal{V}$ the set of N vertices (nodes in the network) and $\mathcal{E}$ the edges linking vertices, let A be the resultant adjacency matrix, where $a_{ij} = 1$ if vertex v_i is linked to vertex v_j and 0, otherwise. Thus, the eigenvector centrality of a vertex $e(v_i) \in [0, 1]$ is computed as follows:

$$e(v_i) = \frac{1}{\lambda} \sum_{j=1}^{N} a_{ij} \cdot e(v_j) \qquad i = 1, 2, \ldots, N, \tag{1}$$

where λ is the largest eigenvalue of A and $e = (e(v_1), e(v_2), \ldots, e(v_N))$ is its corresponding eigenvector. Eq. (1) can be expressed as the matrix equation $Ae = \lambda e$. Note that, as the value of a node depends on the value of its direct neighbors, computing the eigenvector centrality must be performed in a recursive way. Power Iteration Method [32] is an alternative approach to iteratively calculate the eigenvector centrality of each node.