EFIICIENT METAHEURISTIC METHODS FOR BIG DATA ANALYTIC IN SOCIAL NETWORKS

1. Motivation

One of the most important challenges related to the social network analysis is dealing with big data daily created within numerous social media sites. For example, both Facebook and Myspace produce more than a petabyte of social data per day. Only Facebook's logging data excess 25 terabytes per-day. For such a large volume of data, it is almost impossible for humans to find useful information from the social networks data in limited amount of time. The processing of social network data in a timely manner remains a big challenge.

Another issue of social network analysis is dealing with the unstructured, and in many cases, inconstant data. Non-text contents in the form of sound, music, pictures and video material, leave the basic data mining techniques ineffective. Another approach used is social network analytics are clustering techniques, which are widely used for big data analysis. Clustering is a process of organizing data into groups according to certain property or similarity. It is used for discovering natural groups or underlying structure of a given data set in many fields. Another issue is the interpretation of the context, which is essentially semantic in nature. Given the level of slang and abbreviations on social networking sites, understanding recovered conversations, and captured content may be quite difficult. An intelligent interpretation and analysis of users' profiles contents and users' communication may contribute to crime prevention and detection, as well as greater public safety.

In this study, we use a novel approach to tackle big data from social media sites, in order to discover different interest groups or underlying structures of the considered social network. We use optimization methods to analyze online social networks by examining linking behaviors and information flow in social media sites. We develop three mathematical models using similar assumptions on the considered social network, but different objective functions that reflect different goals of the search. Since social networks usually involve large number of nodes (users), the proposed models cannot be solved to optimality, due to memory or time limits. Therefore, it was necessary to develop optimization techniques tailored to the problems under consideration, which are capable to solve the cases of large-scale dimensions in acceptable amounts of running time

2. Mathematical models designed for efficient search strategy

Our idea is to use concepts and models from network-based studies in optimization theory and applications to adapt them for research into social networks, such as facility location models, hub networks, network flow models, and covering models.

Since social networks involve large number of user nodes exchanging large amounts of data of different types, it is impossible to implement a searching technique that will provide up to date information on the considered social network. Simple exhaustive search by using given keywords through all nodes and all network flow may take days or months, and when (and if) finally obtained, the information has already lost its timeliness, and becomes worthless. Therefore, it is important to provide

simple, but efficient strategy on exploring the data flow in a social network, in order to provide the valuable information in short amount of time. We want to identify nodes that exchange information containing certain keywords, which further may indicate that identified nodes belong to the same interest group.

Suppose that we dispose the large amount of data that has been collected from an online social network within certain (usually small) amount of time Δt. As it was discussed above, in each second, the users of social networks exchange enormous quantity of data, containing files of various size and type (text, video and music clips, images, etc), which may be classified as big data. Therefore, we are dealing with the problem of searching and analysis of big data collected from an online social network within time period Δt, in order to provide new information on linking behaviors and information flow between user nodes in a timely manner.

Following the concept of the vertex p-center problem, our idea is to choose exactly p nodes in a social network that will serve as control devices and to allocate each node to its nearest control device (in the sense of searching cost or searching time). These devices are searching through the data originating from user nodes looking for certain keywords. For example, a control device may be a computer station or server that is able to collect and process large amount of data in relatively short time. Each of the located control devices is servicing certain number of users with different amount of data flow consisting of various data types. The goal of the considered problem is to minimize the maximal load of an established control device, ensuring that the data search is provided within minimal amount of time. Another goal, which is considered in this study, is to ensure the load balance between the established p control devices, i.e. to minimize the difference between the maximal and minimal load of a control device. In this way, located control devices will be as equally loaded as possible, which will lead to minimizing the searching time. The exploration of carefully chosen sub-network is used in cases when we are dealing with very large graphs and when the exploration of the whole graph would be extremely slow and impractical.

We have developed three mathematical models, which start from the same input data derived from particular online social network. These models represent variants of the well-known vertex p-center problem and, to our knowledge, they have not been considered in the literature up to now. The proposed models have certain common assumptions with vertex p-center problem, which are reflected on several common constraints.

3. Metaheuristic methods as the engine of big data analytic software

Since online social networks usually involve large number of users and large amount of data flow, exact methods fail to provide solution of the proposed models, due to memory or time limits. Three metaheuristic methods are designed for solving large-scale dimensions of the proposed models: a robust Evolutionary Algorithm EA, which was further hybridized with Local Search and Tabu Search methods, resulting in two efficient hybrid metaheuristic approaches EA-LS and EA-TS.

The EA, EA-LS and EA-TS implementations were coded in C# programming language. Binary solution's encoding is used in all three methods and efficient greedy strategies for calculating objective function are implemented. The constructive elements of the proposed metaheuristic approaches are adapted to problems under consideration. All three metaheuristics were subject to broad computational experiments on generated large-scale data set. The obtained results clearly indicate that the proposed hybrids EA-LS and EA-TS represent promising approaches for the three considered problems related to efficient analysis of the big data flow in social networks. The computational results show the stability and efficiency of the proposed methods, which indicate that this approach may be used in for real-time or near

real-time information delivery to allow analysts to quickly spot trends and patterns in users' behavior within a social network.

The proposed models and metaheuristics may be used as a base of data analysis tools. They may further be hybridized with existing software for collecting data from the web or professional databases that already contain necessary data. They can be easily adapted applied for solving similar problems related to large-scale networks with big data flow. The proposed hybrid metaheuristic methods may be used in combination with existing clustering and data mining techniques to achieve even better results in this field. The proposed methods may be also parallelized, which will lead to further improvement of the efficiency.

The proposed mathematical models and metaheuristic methods may be applied for exploration and analysis of big data originating from various social networks. The presented results may be used for the purposes of social behavior studies, market research, political marketing, but also in security purposes, such as discovering sexual harassment, child pornography, mobbing, bullying in the cyberspace etc. Further applications may include efficient exploration of other large-scale networks, especially in telecommunication and transportation systems, computer and satellite networks, etc.

Beyond the commercial usage, social network analysis offers a number of other opportunities, such as improving threat detection capabilities of government agencies. The expanding use of the Internet and social networks caused a recent explosion of data, which can be mined to help in defending from growing threats coming from terrorists, hackers, and criminals in cyberspace. For example, web-organized revolutions and uprisings may be predicted by monitoring what people are searching for and how they are communicating online. By analyzing big data, governments will better understand various threats that they are facing, potential attacks and the actors who might perpetrate them.

References:

[1] **Z. Stanimirović,** S. Mišković „A hybrid evolutionary algorithm for efficient exploration of online social networks*", Computing and Informatics*, accepted for publication in December 2012, in press.

[2] **Z. Stanimirović,** S. Mišković „ Efficient Metaheuristic Approaches for Exploration of Online Social Networks*" (chapter),* In: Wen-Chen Hu, Naima Kaabouch (Eds.); *Data Management, Technologies, and Applications*, Publisher: IGI Global, PA, USA, accepted for publication in May 2013, in press.

[3] **Z. Stanimirović,** S. Mišković, D. Trifunović „Metod optimizacije za efikasno otkrivanje i prevenciju vršnjačkog nasilja na društvenim mrežama", In: „*Reagovanje na bezbednosne rizikeu obrazovno-vaspitnim ustanovama*", Univerzitet u Beogradu, Fakultet bezbednosti, Beograd, ISBN: 978-86-84069-69-8, 2012, pp. 243-260. (in Serbian).