Available online at www.sciencedirect.com

SciVerse ScienceDirect

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

Survey

Data mining of social networks represented as graphs

*David F. Nettleton**

Universitat Pompeu Fabra, Barcelona, Spain
IIIA-CSIC, Bellaterra, Spain

ARTICLE INFO

Article history:
Received 18 October 2012
Received in revised form
20 December 2012
Accepted 20 December 2012

Keywords:
Graphs
Online social networks
Graph mining
Data mining
Statistical analysis
Data modelling

ABSTRACT

In this survey we review the literature and concepts of the data mining of social networks, with special emphasis on their representation as a graph structure. The survey is divided into two principal parts: first we conduct a survey of the literature which forms the 'basis' and background for the field; second we define a set of 'hot topics' which are currently in vogue in congresses and the literature. The 'basis' or background part is divided into four major themes: graph theory, social networks, online social networks and graph mining. The graph mining theme is organized into ten subthemes. The second, 'hot topic' part, is divided into five major themes: communities, influence and recommendation, models metrics and dynamics, behaviour and relationships, and information diffusion.

Contents

* *Correspondence to:* Universitat Pompeu Fabra, Barcelona, Spain. Tel.: +34 93 542 25 00.
E-mail address: david.nettleton@upf.edu.

1574-0137/$ - see front matter
doi:10.1016/j.cosrev.2012.12.001

1. Introduction

The analysis of social networks has recently experienced a surge of interest by researchers, due to different factors, such as the popularity of online social networks (OSNs), their representation and analysis as graphs, the availability of large volumes of OSN log data, and commercial/marketing interests. OSNs also present interesting research challenges such as the search/matching of similar sub-graphs, community analysis/modelling, user classification and information propagation. Hence, OSN data analysis has a great potential for researchers in a diversity of disciplines. However, we propose that OSN analysis should be placed in the context of its sociological origins and its basis in graph theory. Thus, we have devised a survey which firstly presents the key historical and base research ideas and the different associated themes, and secondly presents a selection of the latest research and tendencies taken from international conferences.

Graph Mining of on-line social networks is a relatively new area of research which however has a solid base in classic graph theory, computational cost considerations, and sociological concepts such how individuals interrelate, group together and follow one another.

For the purposes of the survey, we will divide the base themes as follows: graph theory, social networks, OSNs and SN dataset analysis, and graph mining. The ‘graph mining’ theme is divided into sub-themes as is shown in Fig. 1. Then the ‘hot topics’ are divided into five sub-themes, as is illustrated in Fig. 1. The ‘hot topic’ themes were selected by classifying the papers found in recent editions of four major conferences: WWW 2012, ICSWM 2012, WOSN 2010 and WCCI 2012.

The structure of the paper is as follows: Section 2 consists of a survey of the four major ‘base’ themes and ten sub-themes, highlighting the key concepts and authors. Then in Section 3 we present five ‘hot topics’ in which we summarize a selection of the latest research. Section 4 concludes with a summary of the survey and of the identified key tendencies. In Fig. 1 we see a schematic representation of the structure of the complete survey.

2. Base survey

In this section we consider the base themes related to graph mining of OSNs: graph theory, social networks, online social networks and graph mining.

2.1. Graphs

In this section we will summarize some of the key abstract concepts of graphs. We will see that graph mining has a solid basis in classical graph theory.

In general, a graph G is represented as $G(V, E)$ where V is a set of vertices (or nodes) and E is a set of edges (or links) connecting some vertex pairs in V. Statistically, a graph can be characterized by derived values such as the average degree of the nodes and the average path length between nodes. Additional characteristics are the graphs diameter, the number of triangles, the number of isomorphisms and the clustering coefficient, among others.

In Fig. 2 we see an elementary graph with five vertices and five edges. As there are no arrows, we assume it is undirected, and as the edges have no additional information attached we assume it is un-weighted. We see that nodes A, B and D have degree 2, node C has degree 3 and node E has degree 1, hence the degree sequence is $\{1, 2, 2, 2, 3\}$.

In this survey we are more interested in a graph as an abstract data type rather than a mathematical entity, the former being used to represent the latter. Different algorithms exist which perform higher level operations on graphs, such as finding its degree, finding the connectivity between its neighbours (clustering coefficient), finding a path between two nodes, (using depth-first search or breadth-first search), or finding the shortest path from one node to another. Refer to [1] for a general introduction to different types of graph algorithms which are relevant to OSNs.

A list of typical lower level graph processing operations could be the following: ‘adjacent’, tests if the exists an edge between two nodes; ‘neighbours’, finds all the nodes which have an edge with a given node; ‘add’, adds an edge between two nodes; ‘delete’, deletes an edge between two nodes; ‘get’ and ‘set’ values associated with nodes: ‘get’ and ‘set’ values associated with edges.

How to represent a graph in computer memory is a key issue, due to the potentially high computational cost of many of the higher level operators we wish to perform. Two of the most popular data structures are ‘adjacency lists’ and ‘adjacency matrices’. Refer to [2] for more details about these structures.

Two other data structures, ‘incidence lists’ and ‘incidence matrices’ are similar to the former, with the distinction that the information stored indicates if edges and vertices are incident.