A Survey of Applications that use Graph-based Educational Data Mining

João Oliveira¹, Rodrigo Alexandrino¹, Ana Paula Ambrosio ¹

¹Instituto de Informática – Universidade Federal de Goias (UFG)
Caixa Postal 131 – 74001-970 – Goiania– GO – Brazil

{joaosantos, rodrigoalexandrino, apaula}@inf.ufg.br

Abstract. Graph-based Educational Data Mining can be used to analyze educational scenarios that are best represented by complex networks. To manipulate these networks, several algorithms have been proposed and are implemented in tools that facilitate the development of practical applications. This article presents a survey of applications in the educational domain implemented using graph-based data mining, and published from 2000 to 2019. 30 papers were selected. Information extracted from these papers include the research questions they proposed to answer, the adopted graph representation, and the data mining algorithms and tools used. Non-statistical methods were used to evaluate and interpret the findings. The results highlight the domains and types of problems where graph-based education data mining is being used and the graphs/networks used to represent these contexts, as well as the dominant data mining approaches being adopted to extract information from these structures.

1. Introduction

The application of data mining techniques on educational data has developed into a new knowledge domain known as educational data mining [Baker and Yacef 2009]. Traditional techniques work mainly with data structured in record-based formats [Fayyad et al. 1996b]. However, some applications are better represented using complex networks, for example to represent social networks and the relations between
users [Figueiredo 2011]. In order to obtain information from complex networks, graph mining techniques have been used as efficient approaches to knowledge extraction [Chakrabarti and Faloutsos 2006].

With time, the application of complex networks and graph mining techniques has become more popular [London et al. 2015]. Often, these complex systems can be represented by graphs or networks, where the vertices or nodes represent individuals or entities, while the edges or links represent the interaction between pairs of these individuals. The complex network approach is not only useful to simplify and visualize this huge amount of interconnected data, but is also effective in identifying key elements of the system and finding its most important interactions, as well as discovering groupings of network characteristics [Fayyad et al. 1996a]. Several papers describe techniques and theories of network generation and information extraction [Kavitha et al. 2011] [London et al. 2015] [Rehman et al. 2012] [Lima Filho et al. ] [London and Németh 2014].

Graph-based Educational Data Mining (G-EDM) has been the new term used by researchers who extract information from compelling networks of data derived from educational bases [Lynch et al. 2015]. This graphical data includes complex logs of user-system interaction or between users, graphs generated by the student, and conceptual hierarchies that have large amounts of graphical data [F Lynch 2017]. There is a wealth of pedagogical information under these graphical data sets. As a result, graphical data mining techniques, such as grammatical induction of graphs, path analysis, and the identification of necessary prerequisites, have become increasingly important.

To date, three international workshops on Graphics-based Educational Data Mining have been successfully proposed within the more general conference on Educational Data Mining (EDM). The first was held in London, co-located with EDM 2014 [Loureiro and Goussevskaia 2015]. It had 12 publications, of which 6 were full-papers, the remaining were short papers. The second was co-located with EDM 2015 in Spain [Lynch et al. 2015]. 10 authors submitted their work for publication, including 4 full papers and 6 short papers. The third was co-located with EDM 2017 in China [F Lynch 2017].

Although there are more papers related to the subject, many do not use the expression "Graph-based Educational Data Mining (G-EDM)" or similar, making them hard to find. This survey aims to identify these papers and group them under one reference that can be used by other researchers in this domain.

In this study, we survey several papers that address practical applications of information extraction in complex networks in very different contexts. It is intended to highlight which are the most appropriate graph mining tools to analyze them and what kind of additional information can be extracted by their use. Also, what are the analysis trends and what researchers are searching, and what kinds of approaches are more frequent. It aims to show how the Graph Mining domain has expanded and what challenges it still has to overcome.

2. Research methodology

Following the original guidelines proposed by [Kitchenham et al. 2009] the main objective of this study is to elaborate a survey of articles that contain practical research in-
volving G-EDM with a special focus on those articles that do not mention the term "G-EDM" but are linked to the domain. A secondary objective is to elucidate expanding concepts and important techniques that benefit the field of education in order to further explore the field of study.

2.1. Research questions
The following research questions are addressed, and are distinguished into primary (generalized: set to fulfill the goals of the review) and secondary (sub-objectives/specific: refine the primary - explanatory).
- Q1: What is the main focus of studies in Graph-based Educational Data Mining?
  - Q1.1 What are the results and contributions of these studies?
- Q2: Which are the most frequent techniques used to extract information in graphs?

To carry out the literature review, we defined a review protocol, composed of four distinct stages: I) the literature - collect data, II) review and evaluate research results - selection of primary studies, III) analyze, code and synthesize the results, and IV) report the review.

2.2. Search string
For the search process, both automatic and manual search were used on national and international databases. Automatic search was done using string search and keywords. The sites consulted in the research were: ACM Digital Library, IEEE Xplore, Google Scholar, Elsevier(Scopus and ScienceDirect) and SciELO. The search terms included: Graph Mining, Educational Graph Mining, Graph-based Educational Data Mining, Social Network Analysis, Network Analysis, Complex networks analysis, Educational Social network analysis.

Each database has its own search mechanism, so different strings were used to account for these specificities. In the ACM Digital Library database the following string was used:

"graph mining") AND ("educational") OR ("scholar") OR ("academic") OR ("student") OR ("education")

It returned 24 articles of which only 2, published from 2000 to 2019, were selected by the inclusion and exclusion criteria (see Section 2.3).

The IEEE Xplore database was queried using the string:

(((graph mining) AND educat*) OR scholar) OR student*) OR academic)

It returned 12,971 papers. To refine the search, only journal were included, and the total was reduced to 32 results from which 5 papers, published between 2000 - 2019, were selected following the inclusion and exclusion criteria.

Thus the automatic search yielded 7 articles. To complement the research, a manual search for articles was undertaken using related terms (Networks, Educational Graph, and SNA educational), on Google Scholar and SciELO, as well as in the Proceedings of G-EDM. This manual search resulted in 23 selected articles. Figure 1 shows the distribution by year of publishing.
2.3. Inclusion and Exclusion Criteria

Based on the aim of this survey, to identify applications that use G-EDM, inclusion/exclusion criteria were defined.

**Include:**

- Articles published in Journals or Proceedings.
- Articles that include information extraction from complex networks.
- Full papers and short papers will be accepted.

**Exclude:**

- Articles that present some analysis with graphs but do not use an educational database.
- Theoretical articles without practical implementation will be excluded.
- Book chapters.

The G-EDM is an emerging domain and educational studies using graph representation are new and many are published as short papers. A total of 30 articles were selected after systematic analysis and application of the inclusion and exclusion criteria.

3. Results and discussion

Table 1 contains the selected papers. They are classified according to the type of analysis used in the research as defined in the Proceedings of G-EDM [F Lynch 2017]. Articles included in *Graphical behavior models* contain studies that observe the different behaviors of the networks in several temporal stages. *Graph-based log analysis* studies data from interaction in learning environments and virtual interactions. Five papers in this category analyzed LMS logs describing interactions in the environment. *Graphical solution representations* are used to solve problems through network representation. *Novel graph analysis* present new solutions (algorithms or methods). *Tools and technologies for automatic concept hierarchy extraction* focus on graphical solutions of hierarchy (priority...
dependence). *Tools and Technologies for Graph grammar (pattern) recognition* are studies that present grammar solutions or novel representation models. *Relevant analytical tools and standard problems* discuss new analytic tools. *Social network analysis* present solutions based on the concept of Social Networks Analysis (SNA).

<table>
<thead>
<tr>
<th>Point of View</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphical behavior models</td>
<td>[Okamoto et al. 2014], [Brown et al. 2015], [Kumar 2014], [Jiang et al. 2014]</td>
</tr>
<tr>
<td>Graphical solution representations</td>
<td>[Yeung et al. 2005], [Li et al. 2015], [Tang et al. 2008], [Gašević et al. 2013], [Eckles and Stradley 2012], [Fire et al. 2012], [Dekel and Gal 2014]</td>
</tr>
<tr>
<td>Tools and technologies for automatic concept hierarchy extraction</td>
<td>[Chen et al. 2018], [Cateté et al. 2014], [Yang et al. 2010]</td>
</tr>
<tr>
<td>Tools and technologies for graph grammar (pattern) recognition</td>
<td>[Lynch 2014]</td>
</tr>
<tr>
<td>Relevant analytical tools and standard problems</td>
<td>[Sheshadri et al. 2014]</td>
</tr>
<tr>
<td>Social network analysis</td>
<td>[Wejnert 2010], [Kovanovic et al. 2014], [Garcia-Saiz et al. 2014]</td>
</tr>
</tbody>
</table>

The *Directed Acyclic Graph* is the most widely used representation (34%), followed by *Directed Acyclic Multigraph* with 23% (see Table 2). In general Directed Acyclic Graph are graphs are often used in analysis of user interactions in virtual environments.

Aggarwal and Wang [Aggarwal and Wang 2010] define three main approaches in graph mining: classification, clustering, and discovery of frequent patterns (or substructures). Table 3 organizes the selected papers according to the adopted approach. Some studies use more than one approach. Classification is the more frequent approach. After analysis, it was identified that most of the studies are related to the behavior and identification of student profiles.

### 3.1. Discussion of Research Questions

**Q1:** What is the main focus of research in Graph-based Educational Data Mining? In 21 papers, approximately 70% of the selected papers, research focuses on student analysis, their behavior in virtual environments and student interactions. [Okamoto et al. 2014] [Gitinabard et al. 2017] [Dawson 2008] [Li et al. 2015] [Gašević et al. 2013] [Gardner and Brooks 2018] [Brown et al. 2015] [Eckles and Stradley 2012]
### Tabela 2. Seleção dos artigos subdividido pelo tipo de grafo.

<table>
<thead>
<tr>
<th>Types of Graph</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directed Cyclic Graph</td>
<td>[Yang et al. 2010]</td>
</tr>
<tr>
<td>Directed Acyclic Multigraph</td>
<td>[Dawson 2008], [Yeung et al. 2005], [Li et al. 2015], [Brown et al. 2015], [Patel et al. 2017], [Kovanovic et al. 2014], [Ramos et al. 2016]</td>
</tr>
<tr>
<td>Tree</td>
<td>[Gardner and Brooks 2018], [Dekel and Gal 2014], [Mostafavi and Barnes 2014], [Sheshadri et al. 2014]</td>
</tr>
<tr>
<td>Undirected Acyclic Multigraph</td>
<td>[Eckles and Stradley 2012], [Fire et al. 2012]</td>
</tr>
<tr>
<td>Undirected Hipergraph</td>
<td>[Okamoto et al. 2014]</td>
</tr>
</tbody>
</table>

Q1.1 What are the results and contributions of these studies? Several articles concentrate their analysis on the user and aim to promote improvements in the learning process, understand student, teacher and researcher behaviors, build patterns of student profiles and behaviors in social networks, and understand the students’ learning process in virtual environments. Some of the selected studies deserve special attention. For example, the KnowEdu system [Chen et al. 2018], which automatically constructs a knowledge graph for education to facilitate understanding of how students build knowledge in virtual environments. The same approach is adopted by [Ramos et al. 2016] [Dekel and Gal 2014] [Yang et al. 2010]. In studies by [Okamoto et al. 2014] [Dawson 2008] [Gašević et al. 2013] [Gardner and Brooks 2018] [Buluz and Yilma 2017] [Fire et al. 2012] [Kovanovic et al. 2014] [Yang et al. 2013] the contribution involves the techniques of extracting information from graphs. They include i) learning ii) analysis of students’ behavior iii) correlation and association of patterns in students vi) others, through measures of complex networks such as Degree, Betweenness and Centrality degree. In [Zheng et al. 2017] techniques and algorithms of machine learning were applied to identify student behavior in open courses open (MOOCs) as a conclusion it was possible to suggest better techniques of content delivery and distribution of study groups. A similar approach by [Kumar 2014] [Buluz and Yilmaz 2017] [Gardner and Brooks 2018]. The analysis of social networks
Tabela 3. Seleção de artigos pelo tipo de abordagem em mineração de grafos.

<table>
<thead>
<tr>
<th>Graph mining approach</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>[Tang et al. 2008], [Gašević et al. 2013], [Fire et al. 2012],</td>
</tr>
<tr>
<td></td>
<td>[Dekel and Gal 2014], [Kovanovic et al. 2014], [Kumar 2014], [Lynch 2014],</td>
</tr>
<tr>
<td></td>
<td>[Cateté et al. 2014], [Jiang et al. 2014], [Mostafavi and Barnes 2014],</td>
</tr>
<tr>
<td></td>
<td>[Sheshadri et al. 2014], [Costa et al. 2019], [Wejnert 2010], [London and</td>
</tr>
<tr>
<td></td>
<td>Németh 2014], [Zheng et al. 2017], [Yang et al. 2013], [Yang et al. 2010],</td>
</tr>
<tr>
<td></td>
<td>[Ramos et al. 2016], [Gitinabard et al. 2017], [Dawson 2008], [Brown et</td>
</tr>
<tr>
<td></td>
<td>al. 2015], [Grawemeyer et al. 2017], [Chen et al. 2018]</td>
</tr>
<tr>
<td>Clustering</td>
<td>[Yeung et al. 2005], [Yang et al. 2013]</td>
</tr>
<tr>
<td>Discovery of frequent patterns</td>
<td>[Okamoto et al. 2014], [Li et al. 2015], [Brown et al. 2015], [Eckles and</td>
</tr>
<tr>
<td></td>
<td>Stradley 2012], [Buluz and Yilmaz 2017], [Patel et al. 2017],</td>
</tr>
<tr>
<td></td>
<td>[Gardner and Brooks 2018]</td>
</tr>
</tbody>
</table>

is another aspect well studied by the authors [Wejnert 2010] [Yeung et al. 2005] [Tang et al. 2008] where it is possible to extract behaviors and characteristics of distinct or common profiles to improve academic performance.

**Q2: Which are the most frequent tools used to extract information in graphs?**

Most tools focus on analyzing and visualizing the data. NetVis (2D) [Dawson 2008] presents an interface, evaluation metrics and network information extraction algorithms. GEPHI [Gašević et al. 2013] [Costa et al. 2019] is one of the most complete, offering data import, algorithms and data visualization. Neo4J [Grawemeyer et al. 2017] is a graph-based database where it is possible, through its query language Cypher, to construct and visualize graphs from the data stored in the database. LPGraph [Ramos et al. 2016] is a plugin for R Studio. While it is possible to make analysis in graphs, there are few visualization resources and analysis algorithms. Netbeans [Sheshadri et al. 2014] through Invis is a Java tool for viewing and extracting information in graphs. Python NetworkX [Grawemeyer et al. 2017] [Fire et al. 2012] [Lynch et al. 2015] is a Python script for graph analysis and visualization, and offers one of the largest varieties of algorithms for graph analysis. The Rgraph plugin for R [Fire et al. 2012], [Patel et al. 2017] presents some algorithms and metrics for visualization and extraction of information in graphs.

4. Conclusion

This study focused on Graph-based Educational Data Mining applications, but because the term did not exist until 2014, research related to educational data mining using some approach/technique with graphs were included in this survey. Although three workshops have occurred within the more general Educational Data Mining conference, G-EDM is still a little explored domain, that tends to grow with time.

Representation via Graphs has been shown to be an efficient technique of visualization and representation, mainly when it involves databases that clearly present relationships between entities.
The ever growing interest in Education and in understanding the teaching/learning processes are strong promoters of G-EDM. The relevance of social media and online education, where interaction between actors are naturally represented by graph data, lead to research that aim to extract pedagogical information that will contribute to improving academic success. An increase is, therefore, expected for research in this domain.

Referências


Lima Filho, S. P., Cavalcanti, M. C., and Justel, C. M. Modelagem em grafos a partir de bancos de dados relacionais.


