ARPPA: Mining Professional Profiles from LinkedIn Using Association Rules

Paula R. C. Silva, Wladmir C. Brandão
Department of Computer Science
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Brazil
paula.raissa@sga.pucminas.br, wladmir@pucminas.br

***Abstract*—Human resources managers design and develop professional profiles to maximize the organization workforce. These organizations typically maintain extensive static resume databases from where managers extract and analyze professional data, discovering people with appropriate knowledge, skills and experience to fulfill organizational positions. Nowadays, online professional networks, such as LinkedIn, provide a rich, dynamic, and massive scale resume database useful for professional profile analyses. Considering such massive scale databases, while manual analysis is an exhaustive and often prohibitive task, the use of data mining techniques allows managers to effectively the huge volume of data with a lower cost. In particular, for educational institutions focused on the development of persons with knowledge and skills required by organizations, the use of data mining techniques over professional networks is paramount to plan, direct and implement academic activities and curricula. In this article, we introduce ARPPA, a novel approach to discover professional profile patterns from LinkedIn by using association rules mining. Particularly, our approach crawls resumes from LinkedIn uses a multidimensional data model suitable for professional profile analyses to create and load the crawled data to a data warehouse, and extracts relevant patterns from the data warehouse using an Apriori algorithm. Additionally, we evaluate our approach attesting its usefulness to plan, direct and implement academic activities and curricula in educational institutions.**

Keywords-Data-minin; association rules; LinkedIn.

I. INTRODUCTION

There has recently been a proliferation of social networks addressing the information needs of different groups of users with multiple interests, such as relationship, jobs and business [1][2]. In particular, LinkedIn stands out as the largest and most popular professional social network in the world, with more than 250 thousand of users distributed in more than 200 countries [3]. LinkedIn provides a plethora of services, such as the storage of online versions of multi-language users resume, publishing and searching job opportunities, career planning, and partnership recommendation [4].

This rich, dynamic, and massive scale source of professional information has replaced static resume databases being widely exploited by human resources managers to discover people with appropriate knowledge, skills and experience to fulfill organizational positions, and to design and develop professional profiles to maximize the organization workforce. Particularly, for educational organizations, the use of such source of professional information to discover professional profile patterns of their current and former students is paramount to plan, direct and implement academic activities and curricula in order to meet industry requirements. For instance, analyzing professional profile from LinkedIn can help universities to invest in new educational or research lines, such as the creation of new classes or research groups for specific subjects based on trend of the former students' career. However, due to the high cost of manual data crawling, many universities do not maintain a database with updated information on professional profiles, particularly for their former students. Therefore, mechanisms to automatically crawl professional information from LinkedIn is useful for organizations interested in low cost and effective solutions to build and maintain professional profiles databases.

Crawling and mining professional profiles from LinkedIn are challenging problems. Particularly for crawling, there are duplicated data, spam, access restrictions and ambiguity issues that must be overcome. In addition, the massive scale nature of LinkedIn imposes limitations for data extraction, transformation and storage. Moreover, manual analysis is an exhaustive and prohibitive task often demanding the use of data mining techniques for fast, less expensive and more effective analytical processing. In this article, we introduce ARPPA, a novel approach to discover professional profile patterns from LinkedIn by using association rules mining. ARPPA is an acronym for "Association Rules for Professional Profile Analysis". Our proposed approach addresses the professional profile mining problem by providing a multidimensional data model suitable for professional profile analysis and using an *Apriori* algorithm to recognize mutual implications among professional events, lastly retrieving relevant information on professional profiles. Experiments on a professional dataset crawled from LinkedIn attest the simplicity and effectiveness of ARPPA, showing that it can be used to plan, direct and implement academic activities and curricula in educational institutions.

The remainder of this article is organized as follows: In Section II, we review the related literature on data mining. In addition, we review different approaches for professional profile analysis reported in the literature. In Section III, we introduce the ARPPA approach by presenting its architecture and main components. In Section IV, we present the evaluation procedures that serve as the basis for the experiments, and we thoroughly validate our professional

 ISBN: 978-1-61208-386-5

profile analysis approach using a real professional dataset crawled from LinkedIn. Lastly, in Section V, we summarize our main contributions and conclusions, presenting directions for future research.

II. BACKGROUND

In this section, we review data mining from the literature. Additionally, we review the literature on data crawling and mining from social networks for professional profile analysis.

A. Data Mining

Data mining is a set of techniques to discover knowledge in a large amount of data, enabling the extraction of relevant patterns, which cannot be easily detected only by navigation or searching [5]. Typically, data mining algorithms recognize relevant patterns in datasets organized by data models suitable for effective data processing and recovery.

According to Inmon [6], the multidimensional data modeling promotes the organization of datasets using dimensions and facts to describe event occurrences, where facts are numerical measures related to events, and dimensions are properties that describe and classify the events. Multidimensional data models are used to structure and summarize datasets, presenting them in dimensional views to support online analysis. The multidimensional structure, created according to the event of interest, is commonly known as data cube. Data cubes structure data are to be viewed in multiple dimensions, in which each face of the cube represents a dimension, i.e., a perspective of an event of interest [7].

There are different approaches for multidimensional data modeling, each one based on the selection of relevant properties of the event of interest. The most used approach is the *star schema* that organizes event properties in facts and measures, linking a dimensional dataset to each fact [8]. In particular, the *star schema* is composed by one central fact table, which contains the numerical measures related to event occurrences, and a set of dimension tables [7]. A data warehouse (DW) is a n-dimensional data cube often structured using the *star schema* approach and used to support online analysis processing (OLAP) [6], creating perspective by extracting and crossing data [9]. Over DW, we can use scanning techniques to recognize data patterns and show relevant information [10].

A subset of data mining algorithms extracts relevant patterns by recognizing implication rules over data. While clustering and classification algorithms segment data into groups based on features, sequence algorithms identify frequent occurrences over the time, and association rules algorithms use unsupervised learning techniques to identity frequent occurrences in events. The *Apriori* algorithm is an example of an association rule algorithm which extracts relevant patterns by discovering frequent association rules in events [11].

B. Crawling and Mining Social Networks

Crawling data from the Web consists in visit web servers using an automatic mechanism to collect public documents. According to Myllymaki [12], a crawler must request and store documents from web servers, extract links from the collected documents and schedule the next crawling step by using the extracted links.

Despite the inherent difficulty of crawling public data, crawling data from social networks is an even more difficult task, since the social network servers are usually not freely available for crawling [13]. Social network data may be only available for search by using an application programming interface (API), which restricts the crawling scope. For instance, the LinkedIn social network of jobs opportunities and business presents its own “People Search API” to search people, where programmers must use keywords and predefined fields, such as name, company, and school, to submit queries and receive a restricted set of professional profiles in a standardized format [1].

According to Lops [2], the proliferation of social networks generates a massive volume of data useful to learn user interests and tastes. The authors propose an approach to model user interests based on professional profiles extracted from LinkedIn. The proposed approach uses the *LinkedIn API* to get professional profiles freely available by LinkedIn users and has been used to recommend scientific articles to researchers.

In the same line, Hodigere [14] proposes an approach to predict employee careers using professional profiles extracted from a private social network. Data mining algorithms use multiple dimensions, such as positions, courses and schools, as features to rank employees by their potential of professional development.

There are different approaches reported in literature that have been using data mining techniques to discover knowledge from professional social network [15][16][17]. The Pizzato [18] applied data mining, machine learning and social network analysis on raw data extracted from LinkedIn to a people recommender system. The *SimCareers* framework [19] models member similarity over professional networks. The framework use raw data collected from LinkedIn and data mining algorithms to model and compare a sequence of work experiences finding similarities between the professional profiles.

 ISBN: 978-1-61208-386-5

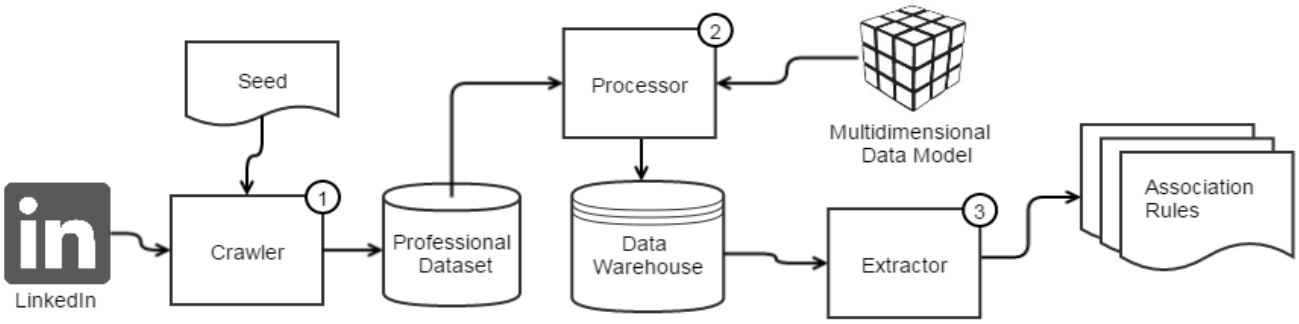

Figure1.The ARPPA architecture and main components.

III. THE ARPPA APPROACH

In this article, we introduce ARPPA, a novel and effective data mining approach to support professional profile analysis. Figure 1 presents the ARPPA architecture, including its main components.

A. *Crawling LinkedIn*

Step 1 in Figure 1 presents the Crawler component responsible to collect LinkedIn data. It is a Hypertext Preprocessor (PHP) component that uses the "People Search API" to retrieve professional profiles in JavaScript Object Notation (JSON) format. The crawling process is divided in two phases.

In the first phase, we use the "People Search API" to retrieve professional profiles, considering the predefined fields name and school, and particular keywords for each field. In preliminary crawling process, we observe that LinkedIn users do not properly use the name field. Sometimes users provide the first and last name but not the middle name, other times user provides the first and middle name but not the last name, increasing ambiguity. Thus, we use different name combinations to improve matching performance.

In the second phase, we collect professional profiles directly from the public LinkedIn profile pages to extract complementary data. Figure 2 presents the professional profile crawling flow used in the second phase.

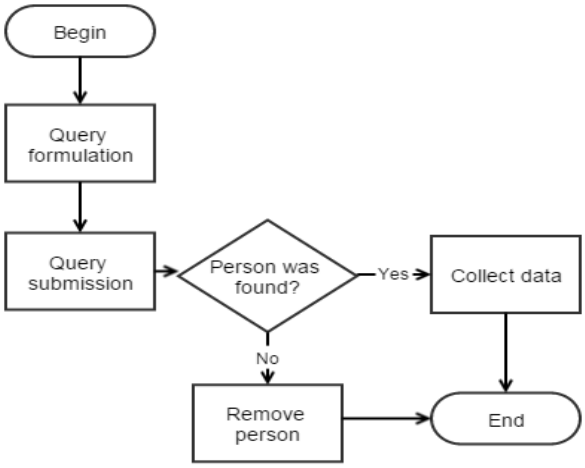

Figure 2. The professional profile crawling flow

This open crawling process is primarily necessary because users frequently disallow the access of their professional profile by the "People Search API". In addition, we need to validate collected data in the prior phase, ensuring that the professional profiles are really related to an educational institution. We particularly process the professional profile collected in the second phase to extract the course name and the school name, comparing them with data collected in the phase one.

B. *Processing Professional Profiles*

Step 2 in Figure 1 presents the Processor component responsible to extract, transform and load (ETL) raw data collected from LinkedIn to a data warehouse, considering our multidimensional data model. The ETL process includes name deduplication and dimensional resolution, i.e., break one transactional entity in multiple dimensions and grouping similar instances of a dimension in one single instance. For example, the transactional entity "location" is broken into "city" and "country" dimensions, the instances of the position dimension "IT Administrator" and "Information Technology Manager" is grouped into "IT Manager", synonyms are reduced to a single form, and typos are removed.

Figure 3 presents an excerpt of our multidimensional data model used by the Processor component to build the data warehouse.

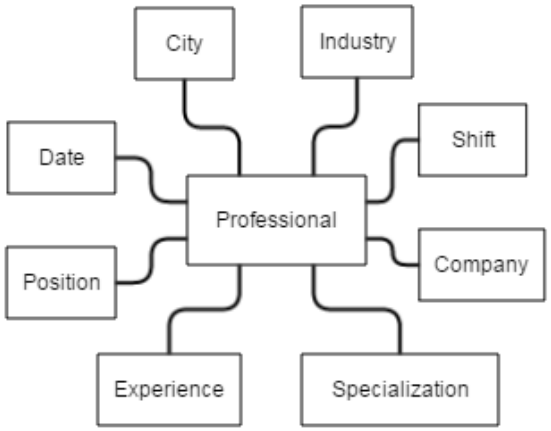

Figure 3. Multidimensional data model for mining professional profiles

 ISBN: 978-1-61208-386-5

From Figure 3, we observe the *star schema* used to organize professional profile data with eight dimension tables and one "Professional" fact table storing professional occurrences, suitable for online analysis processing. These multidimensional data model structures and summarizes the professional profile dataset, creating perspectives by crossing data. Table I describes each dimension from the multidimensional data model.

TABLE I. MULDIMENSIONAL DATA MODEL - DIMENSIONS

Dimension	***Description***
City	The city where people work.
Company	The company where people work.
Date	Semester and year of graduation.
Experience	The level related to the position dimension.
Industry	The activity area business.
Position	Job title, role.
Shift	Shift in which people attended university.
Specialization	Technology in which people are specialized.

C. *Extracting Association Rules*

Step 3 in Figure 1 presents the Extractor component responsible to discover professional patterns from the multidimensional database. In particular, we use the *Apriori* algorithm implemented in Waikato Environment for Knowledge Analysis (WEKA) to extract association rules from the data warehouse.

For each extracted rule, the *Apriori* algorithm computes metrics for analysis: confidence, conviction, leverage, and lift. The confidence measures the accuracy of the rule, i.e., the probability of occurrence of the association pattern [20]. The conviction is an alternative to confidence, also used to measure the accuracy of the rule. The leverage is a measure of divergence from the expected value [21]. Lastly, the lift, also known as interest, is used to access the standard deviation.

IV. RESULTS

To access the usefulness of the ARPPA approach for academic planning, we analyze professional patterns using a professional profile dataset collected from LinkedIn. This professional profile dataset is composed by professional data of 1,847 current and former students from the department of computer science of a major private university in Brazil. For privacy, the dataset were anonymised.

Particularly, we use the dataset as a source of information for the ARPPA approach, which extracts, transforms and loads the raw data to the DW, considering the ARPPA multidimensional data model, and uses the *Apriori* algorithm to discover association rules from DW. The association rules discovered by ARPPA were used to characterize professional behavior. Despite the discovered association rules alone do not characterize professional behavior, they point to nontrivial professional patterns that motivate further investigation.

A. *Professional Characterization*

Figure 4 presents professionals per city, considering only the top 5 cities with more professionals.

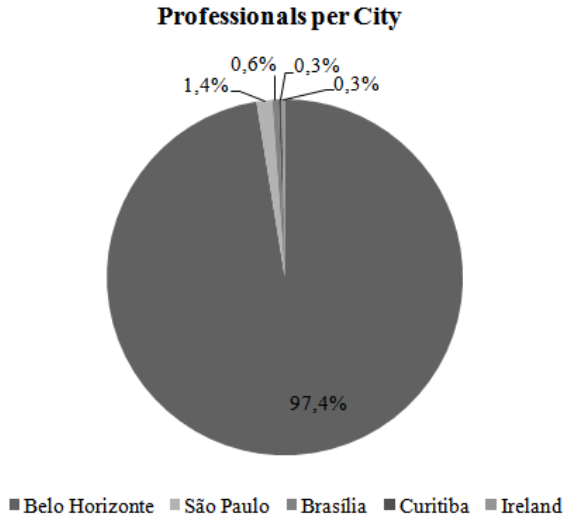

Figure 4. Professionals per city (top 5 cities)

From Figure 4, we observe that the most part of the professionals are working in *Belo Horizonte*. This is an expected behavior since the IT students sample are from the metropolitan region of *Minas Gerais*. They have studied and continue working in this region. We also observe a short percentage, but not negligible, outside Brazil. This behavior can be caused by the Brazilian government's initiatives to do partnership with abroad universities and offer scholarships in these universities. The students have to come back to finish the course in Brazil, but if they found job opportunities abroad, they commonly return to work.

Figure 5 presents the number of professional distributed by industry and shift, considering the top 5 industries.

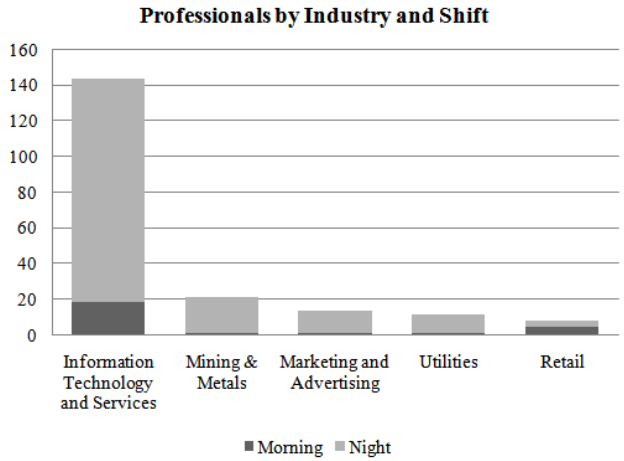

Figure 5. Number of professionals by industry and shift

Starting from the five main industries where the most part of professionals are working, we cross data with the shift that the people studied. We observe that the night shift has more professionals than the morning shift, since most

universities campus offer IT courses at night. Only one campus offers IT courses at morning. In addition, we notice that there is no difference between the industries followed by these students. The main industry where the students have actuated is "Information Technology and Services", as expected. Moreover, there is a greater participation of professionals from the morning shift in the retail industry.

Figure 6 presents the distribution of professional distributed by positions and shift, considering the top 5 positions.

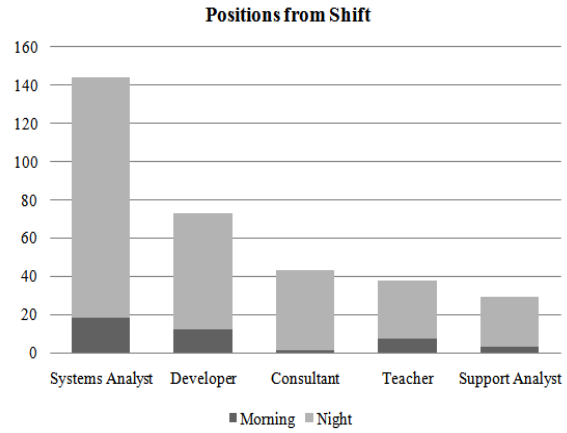

Figure 6. Number of professionals by position and shift

From Figure 6, we can see that the main position that professionals work is "Systems Analyst". But there is a tendency that professionals follow a technical career, except for those following an academic career.

Figure 7 presents the distribution between the top five industries and the top five specializations.

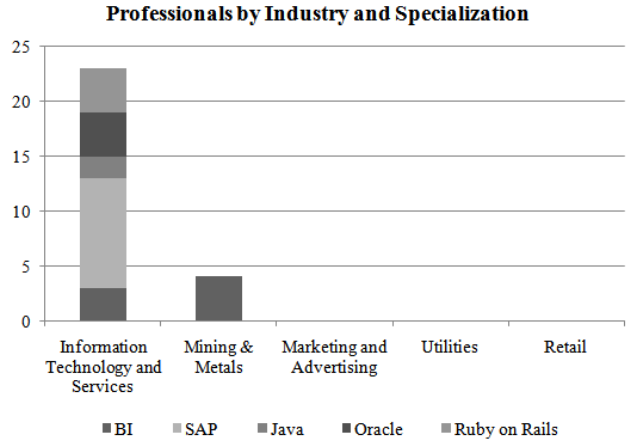

Figure 7. Relation between top 5 industries and top 5 specializations

The specialization dimension represents the technology, which professionals are expert in. From Figure 7, we notice that the main technology is "Java", but there is a little number of professionals that is specialized in this technology. It might be caused by the many kinds of technologies that the companies have adopted. Another explanation for these values is that the computations are based on the data provide by people in their public profile; they might not provide full information about their careers. There is a bigger concentration of technologies in industry "Information Technology and Services". Its main cause may be the large demand for solutions to different problems; each solution is related to certain technology. The fourth and fifth industries are not related with any specialization. These industries are more specific and they demand other technologies.

Figure 8 presents the distribution of professionals by crossing between top five companies and top five positions.

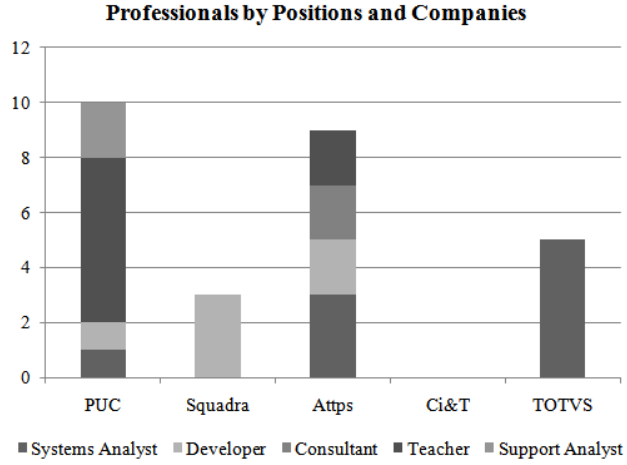

Figure 8. Relation between top 5 companies and top 5 positions

The behavior presents the main industry where the students are working: "PUC Minas" ("Pontifícia Universidade Católica de Minas Gerais"), the university where they studied. The other companies are from the "Information Technology" industry. There is a little number of professionals' concentration in each company and position. This behavior may be caused by the large distribution of the professional in many companies of different sizes. The forth company does not hire professionals from any top five positions. The necessity for positions like "Software Engineer" and "Test Analyst" can be the main cause of this fact.

B. Professional Profile by Rules

A set of association rules show, with a minimum confidence of 0.87, which professionals following the "Systems Analyst" career are working in the "Information Technology and Services". This is an expected behavior because "Systems Analyst" is the main position and "Information Technology and Services" is the main industry. So, we can conclude that former students are following the "Systems Analyst" career in "Information Technology" industry.

A set of association rules show, with a minimum confidence of 0.95, those students, who graduated in 2010, reached the senior level experience in Information Technology and Services industry. With this behavior we can infer that students graduated in 2010 are Senior in "Information Technology and Services" Industry.

 ISBN: 978-1-61208-386-5

With a minimum confidence of 1.0, an association rule shows that professional who are specialist in "SAP" technology are working in "Information Technology and Services" industry. In figure 7, we see the number of professionals who are working in each main specialization by industry, which proves the existence of this pattern of career behavior.

An association rule, with a minimum confidence of 0.97, shows that students who are in "Information Technology and Services" industry are working and living in "Belo Horizonte". This behavior could be caused by the large proliferation of companies from this industry in "Belo Horizonte". There is a region named "San Pedro Valley", which has a large number of "Startups". The development of startups has been caused by the universities and governments projects to encourage the people to undertake in this region.

V. CONCLUSIONS AND FUTURE WORK

In this article, we introduced ARPPA, a novel professional mining approach to discover professional patterns from LinkedIn by using association rules. Our approach provides a multidimensional data model suitable for professional profile analysis and uses an *Apriori* algorithm to recognize mutual implications among professional occurrences, retrieving relevant information on professional behavior. Experiments using a professional dataset extracted from LinkedIn attest the effectiveness of ARPPA showing that it is useful for educational institutions interested in planning academic activities and upgrading curricula for development of people with knowledge, skills and experience required by organizations.

Particularly, we have used ARPPA to analyze professional profiles of current and former students of the major private university in Brazil. The professional profile analyses presented in this article are samples of the possible analyses that can to be performed using ARPPA. Despite the discovered association rules alone do not characterize professional behavior, they point to nontrivial professional patterns that motivate further investigation.

Lastly, considering possible directions for future research directly inspired by or stemming from the results of this article, we plan to investigate and make a comparison with other data mining algorithms to retrieve relevant information on professional behavior, such as supervised learning algorithms and neural networks. Moreover, we plan to enrich the multidimensional data model considering social data crawled from different social networks.

REFERENCES

[1] M. A. Russell, "Mining the Social Web: Data Mining Facebook, Twitter,LinkedIn, Google+,GitHub, and More," O'Reilly Media, Inc, 2003.

[2] P. Lops, M. de Gemmis, G. Semeraro, F. Narducci, and C. Musto, "Leveraging the linkedin social network data for extracting contentbased user profiles," ACM, pp. 293–296, 2011.

[3] LinkedIn, "About us," Mar. 2014. [Online]. Available: http://www.linkedin.com/about-us/

[4] D. Agarwal, "Computational advertising: the linkedin way," ACM, pp.1585–1586, 2013.

[5] R. Elmasri, Fundamentals of database systems. Pearson Education India, 2008.

[6] W. H. Inmon, Building the data warehouse. John wiley & sons, 2005.

[7] J. Han and M. Kamber, Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann, 2006.

[8] P. Vassiliadis and T. Sellis, "A survey of logical models for olap databases," ACM Sigmod Record, vol. 28, no. 4, pp. 64–69, 1999.

[9] S. Chaudhuri and U. Dayal, "An overview of data warehousing and olap technology," ACM Sigmod record, vol. 26, no. 1, pp. 65–74, 1997.

[10] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques, 3rd ed., ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.

[11] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," ACM, pp. 207–216, 1993.

[12] J. Myllymaki, "Effective web data extraction with standard xml technologies," Computer Networks, vol. 39, no. 5, pp. 635–644, 2002.

[13] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, Dbpedia: A nucleus for a web of open data. Springer, 2007.

[14] R. Hodigere and D. Bilimoria, "Constructing professional resource networks from career biographical data," IEEE Computer Society Washington, DC, USA, pp. 1242–1247, 2012.

[15] S. Cetintas, M. Rogati, L. Si, and Y. Fang, "Identifying similar people in professional social networks with discriminative probabilistic models," in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '11. New York, NY, USA: ACM, 2011, pp. 1209–1210.

[16] A. C. d. S. G. d. Santos, T. d. P. Menezes, H. R. M. da Hora et al., "Students' and alumni's profiles analysis through the data mining technique: a case study in the federal institute in rio de janeiro state interior," Research Gate, 2014.

[17] J. Wang, Y. Zhang, C. Posse, and A. Bhasin, "Is it time for a career switch?" in Proceedings of the 22Nd International Conference on World Wide Web, ser. WWW '13. International World Wide Web Conferences Steering Committee, 2013.

[18] L. A. Pizzato and A. Bhasin, "Beyond friendship: the art, science and applications of recommending people to people in social networks," in RecSys'13, pp. 495–496, 2013.

[19] Y. Xu, Z. Li, A. Gupta, A. Bugdayci, and A. Bhasin, "Modeling professional similarity by mining professional career trajectories," in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '14. New York, NY, USA: ACM, pp. 1945–1954, 2014.

[20] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," Information Systems, vol. 29, no. 4, pp. 293–313, 2004.

[21] G. Piateski and W. Frawley, Knowledge discovery in databases. MIT press, 1991. G. Piateski and W. Frawley, Knowledge discovery in databases. MIT press, 1991.

 ISBN: 978-1-61208-386-5