International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 275-280
e-ISSN:2278-621X

A METHODOLOGY FOR MINING LINKEDIN DATA FOR EXTRACTING AND VISUALIZING PROFESSIONAL PROFILES

Mr. Rakesh Kumar B[1], Shyan Aloysius De Abreu[2] & Mohammed Arshad[3]

Abstract- The large volume of data produced by the social networking websites facilitate the need of analyzing it. The wide variety of these kind of data is widely available to the users. This kind of data can be used to extract beneficial information to the corporate world for recruiting people. This paper describes one of the methodology that can be used for inferring desired profile of the candidate for a particular job vacancy using the data from LinkedIn which is the social network for business. This methodology uses natural language processing techniques and association rules. The need for natural language processing arises when the data are semi-structured and there is a need of keyword analysis. This paper describes the potentials of these methodologies by demonstrating them using the data from LinkedIn.
Keywords- Data Mining, Association Rule, Natural Language Processing, LinkedIn, TF-IDF

1. INTRODUCTION

Data mining has become an extremely comprehensive area and its methods have been applied in different areas in order to generate knowledge. Among these areas, one that has stood out is the data mining in social networks, since the generation of their data is irregular and is covered in the most varied types and themes. Nevertheless, social networks are distinguished by the diversity of content and the number of users, which impact and are impacted by cultural and social media due to the wide sharing of information and knowledge by other users from different locations. Among the existing social networks, stand out Facebook, LinkedIn, Research Gate, Glass Door, Hirist, among others. In particular, LinkedIn, which is of interest to this work, is characterized by being a social network of businesses, in which professionals can present their aptitude through an online curriculum so that others can endorse their specialties, giving credibility to the user. LinkedIn is currently the largest professional social network on the web with more than 300 million registered users.

Motivated by the volume of data inherent in the LinkedIn social network and by the investigative possibilities through data mining methods and methodologies, this paper aims at mining job vacancies advertised on LinkedIn. As a goal, it is desired to identify requirements that are of greater influence in the process of candidate selection. It should be noted that data mining activities are not trivial [5] and become more complex when based on semi-structured data, such as those provided by LinkedIn. Such data are provided in the textual structure, thus, text mining methods and natural language processes were required to allow the extraction and analysis of information.

In order to select more relevant requirements, which are therefore the most requested by different companies, key words were extracted from the advertised job vacancies. To do this, an algorithm was implemented that calculates the TF-IDF (2), which establishes a weight for each word in the document and, based on metric results, allows inferences about its relevance [7]. In addition, the Apriori algorithm was also used to analyze the requirements through their relationships. As a result of the keyword analysis, it was possible to identify requirements that stand out from others because they are more requested in job postings. The association rules allowed the user to assist in decision-making processes, since relationships between keywords consent, for example, to identify compound requirements, i.e. requirements that are requested together, which in the case of LinkedIn, are shown to be useful in activities of suggestion of contiguous competences. This paper is divided as follows: Section II presents correlated works that are comparatively relevant to this study; Section III describes the methodology with the theoretical basis for its application and use; already in Section IV is presented a case study that illustrates potential of the methodology in data mining processes; and finally, Section V discusses the conclusions and future work.

2. RELATED WORK

Several works are related through research resulting from data mining in social networks. However, these works are not directed to a single social network and its intention, for many times, is not to characterize information relevant to users. For the most part, related jobs focus on the behavioral characterization of their individuals, while our work focused on mining through LinkedIn job openings. Thus, in this section, the work comparatively relevant to this proposal is analyzed, while at the same time the approach adopted to solve the problem is justified in a detailed way. Work [3] presents a study of mining possibilities using the data obtained through the LinkedIn platform.

[1] Assistant. Professor, Department of MCA, AIMIT, St Aloysius College(Autonomous), Beeri, Mangalore
[2] Student. MCA, AIMIT, St Aloysius College(Autonomous), Beeri, Mangalore
[3] Student, MCA, AIMIT, St Aloysius College(Autonomous), Beeri, Mangalore

This paper presents a tool that allows analysis of the performance of research in social network data through advanced search clauses, which guarantees greater specificity when compared to the current tool of the respective social network. In addition, the proposal allows to work with raw data, so it does not require structuring information, tolerating searches on raw data obtained by scraping LinkedIn. As a result, it uses tools to analyze the information collected through diagrams and graphs. However, it does not discuss methods to apply the results obtained in data mining processes aimed at the evaluation of information, as is the case of the present research, using the data made available by the social network platform and processes it in order to mine job vacancies. However, the possibilities of characterizing the vacancies are limited, since the authors use this information to predict trends and not to identify the requirements that motivate them, this can be done through the results of the present research. Already in the work, [4] is proposed system of the recommendation of job vacancies for users of social networks LinkedIn and Facebook. The algorithm is based on a set of taxonomies, but does not use and does not compare its results with those obtained through the application of TF-IDF. Also, they do not introduce a clear methodology to characterize the proposed classification and recommendation method, as addressed by this work in Section III. Also, [10] presents a system of the recommendation of connections based on the characteristics of different users. Its purpose is to indicate possible acquaintances in social networks. As a case study, they make use of LinkedIn users by employing geographic location data and common skills. Nevertheless, they fail to use their recommendation system to indicate possible openings of interest based on the expertise of a given user.

Finally, in [8] an algorithm for association rules generation is discussed, with the aim of improving referral systems in social networks. However, it lacks demonstrations of its proposal in real scenarios, in order to allow sufficient results to justify its approaches. It is also observed that the study related to data mining and social referral systems are the subject of several studies already consolidated in the literature. However, there is still difficulty in processes related to the mining of characteristics arising from social networks. Among the works that cover these premises, those that focus on mining through keywords are minority. Thus, in the subsequent section, the proposed methodology is introduced, which covers the limitations of related jobs in the process of mining job data from LinkedIn.

3. PROPOSED METHODOLOGY

This methodology is capable of executing all data mining processes from LinkedIn, from data collection to analysis - providing support for decision-making activities. It was developed using processes that make use of the Python2.7 as well as R programming language. In addition, the methodology steps use the NLTK (Natural Language Toolkit) library for the purpose of providing methods for the natural language processing and also a file data set for data storage purposes. Its methodological flow is subdivided into four stages: (A) Data Collection, (B) Pre-Processing, (C) Data Mining and (D) Evaluation of Results. These processes were organized following the proposal of [5]. In order to illustrate the specific steps of the methodology, Figure 1 gives an overview of the proposed methodology. Specifically, in the (A) Data Collection stage, the LinkedIn Job Search API was used, it returns job post data published by companies on the social network. The API is based on the REST architecture style, which aims to communicate between client and server in a simplified and efficient way through HTTP (S) protocols. The connection is encrypted with the OAuth 2.0 protocol, which guarantees access to data only for authorized users. The results of the requests consist of XML or JSON. Once the data are obtained, we proceed to the step of (B) Pre-Processing, in which the data are prepared to be analyzed according to the premises to be validated. In the scope of this work, natural language processing was used, and the data returned from the API was organized for further processing. Consequently, the pre-processing was subdivided into the following steps (described in Figure 1):

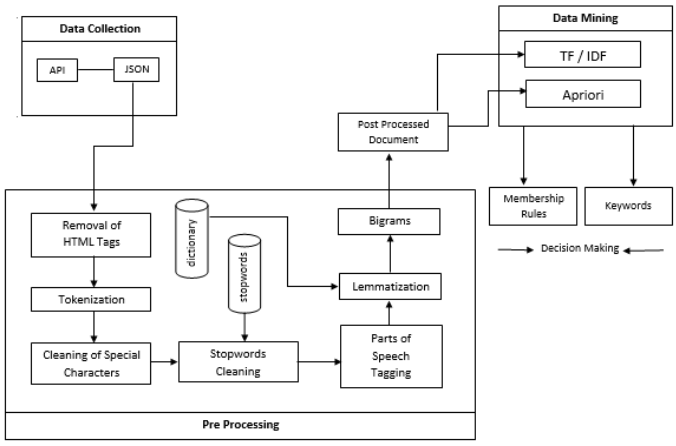


Figure 1. Different steps in Mining Methodology

3.1 Data Cleaning –

The cleaning process is present between several stages of the preprocessing activity; its intention is to remove the text elements that do not contribute to the discovery of knowledge. This process consists of eliminating HTML tags, removing

special characters and stop words - very frequent and meaningless words, besides the morphological analysis of the text by means of techniques capable of determining the grammatical classes of the elements of a text.

3.2 Tokenization –
The tokenization stage consists of separating the elements that make up the natural language discourse, this step is performed through regular expressions that are able to identify words while ignoring the punctuation marks.

3.3 Lemmatization –
This step consists in normalizing the words of the text in order to avoid variation in writing and formatting.

3.4 Bigram Identification –
The process of identifying bigrams consists in locating words that only have semantic meaning when accompanied by another, so it was considered a bigram the doubles that appear at least 10 times together.

3.5 Data Mining –
This step contains the extraction of keywords and the identification of association rules. The intention is to find the words that best describe the text and to find relationships between the most relevant words, in order to aid the decision-making processes. Keyword extraction consists of applying an algorithm that implements the Term Frequency-Inverse Document Frequency (TF-IDF) which, in turn, calculates the importance of a word in a collection of documents. This calculation consists of two steps, the first calculates the frequency of a term and then the inverse of the frequency of the same term in relation to the entire collection of documents. The intention of this process is to identify the most frequent, discarding those of lesser semantic importance. Demonstrate this calculation by Equation 3, which combines the Equations 1 and 2. In equations, the element D represents the collection of documents, and | D | denotes its number of occurrence; d ∈ D demonstrates a single set, which is a single job vacancy notice, while t ∈ d represents a set consisting all of the advertisements. Finally, $t_{f,d}$ illustrates the gross frequency of a term that is part of one advertisement d.

Term frequency:

$$tf\ (t,d) = 0.5 + 0.5\times \frac{f_{t,d}}{\max\{f_{t',d}: t' \in d\}} \quad (1)$$

Inverse document frequency:

$$idf(t,D) = \log \frac{|D|}{|\{d \in D: t \in d\}|} \quad (2)$$

Term frequency – Inverse document frequency:

$$tfidf(t,d,D) = tf\ (t,d)\times idf(t,D) \quad (3)$$

The TF-IDF values were used in the keyword extraction process, in which, through a vector space model, each element of the set of texts was organized into a matrix, in which each line represented a word and each word had its associated TF-IDF value. Therefore, the most used words had lower values, while the less used, higher values. In order to identify the relevance of a word, the maximum value of 0.5 was empirically established. For the extraction of association rules, this methodology uses the Apriori algorithm following the assumptions established by [1] and [6]. The first step of this algorithm is to calculate the frequency of each item generating sets of frequent items of unit size. The following steps are subdivided into two steps. First, the sets generated in the previous step are used to generate the new candidates for frequent items, subsequently, the support value (percentage of documents in the dataset containing the identified rule) of each candidate is calculated, at the same time as eliminating items that are less than a minimum. In the second stage, the association rules are discovered, the procedure is to find the measure of trust (ratio between the support and the percentage of documents that contain the rule) in each set of frequent items, and if it is greater than a minimum trust the rule is established. In this work, the minimum support is considered as 3.5 and the confidence is 4.5. The next section presents a case study that illustrates the potential of steps in the data mining process in LinkedIn data.

4. CASE STUDY

To carry out this case study, approximately 4370 vacancies related to Information Technology Software Industry were collected during September and October 2017.

4.1 Filtering and Classification –
At the collection stage, data were filtered by top Cities of India. For example, Bengaluru, Ahmedabad, Chennai, Ernakulum etc. The filtering provided by the LinkedIn API fails in the process of categorizing data, returning vacancies in regions other than those listed above. To solve this problem, each of them had to be validated before proceeding to the analysis phase. The intention of the validation is to verify if the data collected corresponds to the data of interest to the domain under investigation.

4.2 Data Collection –

Figure 2 shows a graph that describes the number of vacancies in the area of Information Technology, according to the search filters previously discussed. Vacancies are not limited to those collected, however, because of the difficulty in acquiring relevant data, the focus was only on this data set. Nonetheless, it should be noted that this methodology is suitable for anytime, it is highly scalable in terms of volume and compatible with several categories of LinkedIn data.

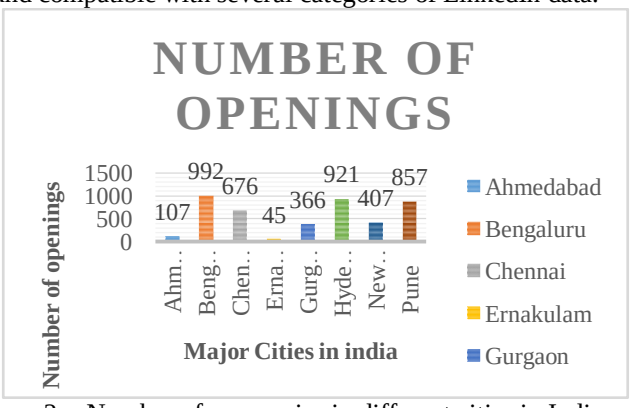


Figure 2. Number of vaccancies in different cities in India

4.3 Pre Processing –

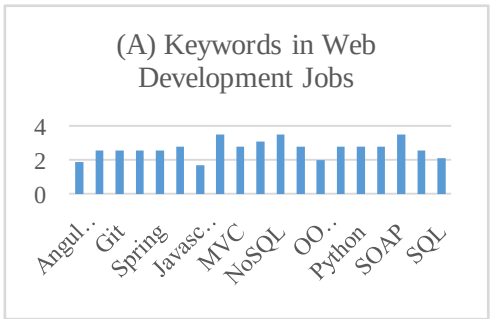
Each of the vacancies collected went through pre-processing processes. In fact, this step is the most laborious of data mining. The resulting data were reduced (in number of characters) to 92% on average. These processes refer to those described in Section III. In The following is a synthetic example in which the raw data and its respective post-processed data are highlighted.

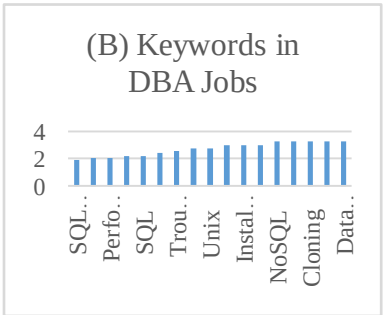
<h2 class="jobs-description-content__title jobs-box__title">Job description</h2> Mandatory Skill 2 UI Flex / Spring Mandatory Skill 3 Java Java script Mandatory Skill 4 Unix/Scripting Unix/perl/shell script Skill 5 Oracle/PL-SQL Oracle PL/SQL Qualifications UI Developer with strong analytical and technical ability. 3-5 years of experience in Java/J2EE, Spring framework, MyBatis, Web services, Flex/AngularJS
Working knowledge of PL/SQL is a good to have</li
Bachelor's degree (in Science, Computers, Information Technology or Engineering) </ul

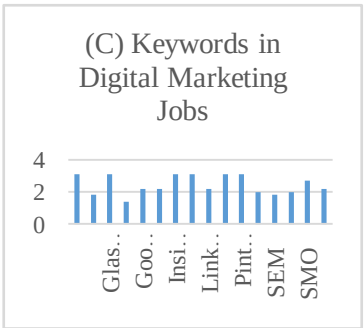
Post-Process Data: "UI, Flex, Spring, Java, Javascript,Unix/Scripting,Unix/Perl,Shell script, Oracle, PL/SQL, Java/J2ee, Spring, MyBatis,Web services,Flex/Angular JS "

4.4 TDF-IDF Algorithm –

Figures A to E indicate the results obtained by applying the TF-IDF algorithm. Due to the space limitation of this work, only 4 cases of analysis were considered, the others are analogous to those presented. Thus, the following keywords are considered as identified in the job categories of (A) Web Application Development; (B)DBA Jobs; (C) vacancies in Digital Marketing; (D) Data Analytics and (E) vacancies in Mobile Application Development; as shown below. The horizontal axis of each graph indicates its value of TF-IDF (following Equation 3), while the vertical axis presents the keywords of major importance for the analysis scenario. In addition, note that not all words meet the relevance requirements, which as previously discussed correspond to the minimum significance of 0.5 - note that the significance is inversely proportional to the relevance of a word, so the smaller the value of the more significant TFIDF becomes the term analyzed







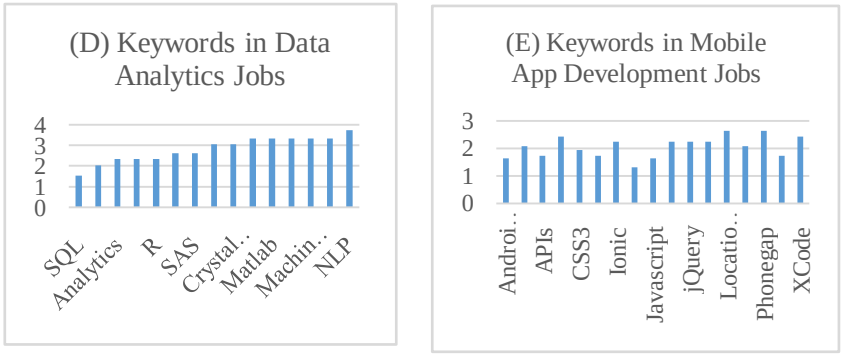


Figure 3: results obtained by applying the TF-IDF algorithm

4.5 Apriori Algorithm –

From the keywords identified as relevant by Equation 3 of the TF-IDF, the Apriori algorithm was applied to identify the respective association rules. The results of this step are shown in Table 1, which discusses the support and trust values, followed by their association rule. The support value indicates the percentage of documents in which the rule is valid and the trust value indicates the relationship between the support value and the percentage of documents that contains the first term of the rule but does not contain the second. For example, considering the "Data Analytics → "SQL" rule, we conclude that 45% of the vacancies involving "Data Analytics" requires the knowledge of "SQL", the trust of this rule is 75%, that is, all times where word "Data Analytics" appeared also the word "SQL" was present. The interpretation of these relationships allows, for example, activities of recommendation in the form of suggestion of competences to the users. In the case exemplified above, if a user has knowledge of "Data Mining" among his skills, but at the same time does not have knowledge about "SQL", this competence can be recommended to the user, who can exploit it in order to fortify your resume and increase your chances of pursuing a profession through the LinkedIn platform.

Table 1: Analyzing the Results of Apriori Algorithm

Association	Support	Confidence
SEM ==> SEO	80	100
Security ==> Backup and Restore	53	75
Oracle 11g/12g ==> PL/SQL	53	42
Security ==> Data Guard	53	100
SEO ==> Google Analytics	50	60
Google AdWords ==> SEO	50	75
Data Visualization ==> R	50	100
iOS==> Objective C	51	40
AngularJS ==> .NET	49	100
AngularJS ==> CSS3, HTML5	49	80
Data Analytics ==> SQL	45	75
HTML5 ==> CSS3	40	69
HTML5 ==> JavaScript	39	80
JQuery ==> AJAX	38	63
PL/SQL ==> SQL	36	100
REST ==> Java	35	41
JQuery ==> JavaScript	35	45

5. CONCLUSION AND FUTURE WORK

In this article, we present a methodology that applies data mining techniques to job vacancies advertised on LinkedIn. The data generation of this network is intermittent, justifying its analysis to identify patterns. The proposed methodological approach demonstrates results in all stages of data mining, from data collection to data analysis. In order to do so, it was mined vacancies in the area of Information Technology, through the analysis of the relevance of keywords using the TF-IDF and through association rules generated by the Apriori algorithm. The results of these processes are in the form of discussions and analysis of a case study. Through TF-IDF, it was possible to identify the relevance of the terms used in the dissemination of vacancies. While, with the application of the association rules generation method, it was possible to identify relations between terms present in the documents. It would be noticeable the diversity of vacancies and requirements related to the area under

analysis, which is the reason for the generation of exacerbated deregulated content, justifying the need to analyze such information. Finally, as future work, it is expected to test the processes developed in other social networks other than LinkedIn. In addition, new results will be generated from the analysis of other professional areas, all the cities, and even other countries. We highlight here, areas such as healthcare, business administration, finance, sales, human resources and others.

6. REFERENCES

[1] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th int. conf. very large databases, VLDB. Vol. 1215. 1994.

[2] Berry, Michael W., and Jacob Kogan, eds. Text mining: applications and theory. John Wiley & Sons, 2010.

[3] Bradbury, Danny. "Data mining with LinkedIn." Computer Fraud & Security 2011.10 (2011): 5-8.

[4] Diaby, Mamadou, and Emmanuel Viennet. "Taxonomy-based job recommender systems on Facebook and LinkedIn profiles." Research Challenges in Information Science (RCIS), 2014 IEEE Eighth International Conference on. IEEE, 2014.

[5] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data." Communications of the ACM 39.11 (1996): 27-34.

[6] Hipp, Jochen, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. "Algorithms for association rule mining—a general survey and comparison." ACM sigkdd explorations newsletter 2.1 (2000): 58-64.

[7] Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. "A brief survey of text mining." Ldv Forum. Vol. 20. No. 1. 2005.

[8] Lin, Weiyang, Sergio A. Alvarez, and Carolina Ruiz. "Efficient adaptive-support association rule mining for recommender systems." Data mining and knowledge discovery 6.1 (2002): 83-105.

[9] Tajbakhsh, Mir Saman, and Vahid Solouk. "Semantic geolocation friend recommendation system; LinkedIn user case." Information and Knowledge Technology (IKT), 2014 6th Conference on. IEEE, 2014.

[10] Ye, Yanbin, and Chia-Chu Chiang. "A parallel apriori algorithm for frequent itemsets mining." Software Engineering Research, Management and Applications, 2006. Fourth International Conference on. IEEE, 2006.