EXTRACTING INFORMATION FROM IT JOB ADVERTISEMENTS USING TEXT MINING TECHNIQUES

Bojan Ilijoski[309]
Zaneta Popeska[310]

https://doi.org/10.31410/itema.2018.888

Abstract: *Finding a suitable job requires a lot of time in searching the job advertisements in order to find just that job openings that fit certain skills. Usually the job advertisements are written like free text and they don't follow some specific structure or pattern and finding relevant information inside them is pretty difficult. Also if we want to analyze the job market so that we can uncover the needs of the industry probably we should go through the advertisements one by one and analyze them. This process can be facilitated by using some text mining techniques for extracting specific information. This will help jobseekers to find the employment that suits those best, job requirements they can cover and also help them to discover the job market needs so they can be profiled in that direction and acquire the necessary skills. Also, if we analyze a longer period of time, we can see how the needs of the labor market have changed over time, how the required competencies have changed, and maybe we can predict in which direction future requirements will be made. By using text mining methods for extracting specific information from the texts, job advertisements, we obtain just the major job requirements and skills. Also from the same texts we can extract additional information like experience needed for certain job or minimum wage for some job position. We are using jobs advertisements collected in period of 5 years in the field of IT industry. For fast growing industries, this kind of analysis is very important in order to follow the progress and needs of this industry. Our findings, in addition to showing us the trends in the labor market, can also be used by universities and other educational and training institutions for creating curriculums that the industry needs.*

Keywords: *Text mining, job advertisement, job analysis, labor market analysis*

1. INTRODUCTION

The job market can be big and messy. A lot of job advertisements are generated daily especially in the fast growing industry like IT. How important these analyzes are can be shown the large number of papers written on this subject and the time period they are analyzing [1] – [4]. Analyzing of the job market can be very important from several aspects. One of them is to find a way to give a right direction to the education, so the education can follow the industry's latest trends, and the students to gain the right knowledge base, so they can easier fit in the industry after. It also raises the question whether the theory, technical or business skills should be developed and considered as more important. The job advertisements can also be used for talent recruitment. The big brand companies are using their status and

[309] Faculty of Computer Science and Engineering, University "Ss. Cyril and Methodius", ul.Rudzer Boshkovikj 16, P.O. 393, 1000 Skopje, Macedonia

[310] Faculty of Computer Science and Engineering, University "Ss. Cyril and Methodius", ul.Rudzer Boshkovikj 16, P.O. 393, 1000 Skopje, Macedonia

facilities to make up their job advertisements to invoke the young talents to work for them. Also the job advertisements can provide information about economy and social status in the region by analyzing salaries for specific job position.

In order to extract some valuable information from the job advertisements we use some simple text mining techniques which can help us to process the jobs advertisements faster and more accurate. The main idea is by using those techniques to extract the most wanted jobs and the most important skills for them. For this purpose, we analyzed job advertisements in the IT industry and we have obtained the most demanded positions and technical skills in the past few years. This has allowed us to see the trends in the IT industry, what is in demand and in what direction and with which speed it is developing. In the next section we give some of the other articles that are trying to solve the same or the similar problem on job advertisements data set. In the third section we are talking about the dataset that we used and the methods that we created, and in the last section we give the conclusion and ideas for the future work.

2. RELATED WORK

There are many other articles that target this problem. All of them have different methods or approaches in order to extract some relevant information for the job advertisements. We give a quick review of some of them that are related to our idea.

In “Employers’ extractions: A probabilistic text mining model” [5] the authors are analyzing more than 20,000 job advertisements from various web sites. They are analyzing the sentences by finding some of the main keywords and then by using the Bayesian theorem they are trying to extract the job qualifications. After, then by using the LDA techniques, one of the most known technique for topic modeling, to identify the groups of skills, which lead to general job positions. In the article „Linguistic information extraction for job ads (SIRE project)“ [6] the authors manually tagged 200 jobs advertisements so they can try to relate them with the corresponding ontological features of the job and they obtain set of empirical labels. Then with techniques similar to LSA and MWE they are creating v-lexicon from 1081 IPs jobs advertisements. The “Changing Trends in LIS Job Advertisements” [7] is speaking about extracting information from job advertisements in Australia. They are using context analysis which includes counting how often words, phrases or themes appear individually or in combinations and categorization dictionaries in order to find the most needed skills in the job advertisements. They are making cooperation between their study and the similar study from the 2004. The conclusion is that changing work practices have led to job definitions which no longer rigidly demarcate role functions and now a wider range of skills is required. “Text clustering based on centrality measures: an application on job advertisements” [8] articles speak about clustering the job advertisements. The authors are using similarity and adjacent matrix to find the overlapping between job advertisements. Then by using mixed approaches for clustering they are identifying the job advertisements clusters. They analyzed 1650 job advertisements and they are getting five clusters labeled with the educational profile like “Graduates in marketing, social sciences and humanities”, “Engineering management”, “Informatics”, “computer engineering”, “mathematics and statistics”, “Economics” and “No qualifications”. The author of “Skills and Vacancy Analysis with Data Mining Techniques” [9] is analyzing 4846 IT vacancies, which in first step are into ten exclusive classes. After preprocess and usage of several text mining techniques, the author concludes that the kind and the Naïve Bayes algorithm show the best performance. Similar to the previous article in the “Job Opportunity Finding by Text Classification” [10] the authors are using k-NN, Naïve Bayes, decision tree, neural network, SVM, and Linear Least Squares Fit to classify the chins jobs in several predefined classes. The job market is also analyzed by using some date mining teaching techniques in “Data Mining Approach to Monitoring the Requirements of the Job Market: A Case Study” [11]. The authors are using O*NET database data science and related techniques. They are using standard preprocessing techniques like removing common expressions and words (e.g. stop words) and stemming. After these steps they are creating the TF-IDF matrix and LSI model for identifying the most demanded occupations in the job market.

3. THE DATA AND THE METHODS

For this propose we created web scraper to collect the data from Macedonian job advertisements web sites. The collected data is from January 2013 to July 2017 and more than 3500 job advertisements are collected or in details - 590 in 2013, 654 in 2014, 818 in 2015, 969 in 2016 and 580 for July 2017. We were scraping just the IT jobs so the upper trend in those numbers is expected. All job advertisements are already divided in many categories but, because those categories were too detailed we merge them in ten major job positions (fields, classes). They are Programmer/Software developer, Sales/Marketing, Customer Service, QA/Tester, Intern, Designer, Network engineer/System administrator, Data bases, Education and Other. The idea is to get the basic rough information about job market requirements. At the figures 1 – 5 are shown the percentages of the job advertisements classes by the year. So the trend in the job positions is retained. In all those years the Programmer/Software developer is the most needed job position with around half of the jobs (except the 2013 where it is the same with the Other). On the second place is Other, with 10% - 20 % of the jobs (except the 2013). We should consider that in the other part there are jobs like Office manager, Service worker, Account manager, Informatitian, Business analysis, Brand manager and even a Driver. All those job categories are categorized in the IT sometimes like a mistake of the job advertisement enterer or because the company that advertises is an IT company. Also all jobs in the IT part but without specific category are also putted in the Other class. That’s why in the 2013 they are that much jobs in the Other and in the next years the number is drastically smaller. On the third place is always Sales/Marketing which again is not strictly IT job, but the positions are related with IT or are from companies that are classified as IT companies. On the next places with similar percentage between them and over the years are Customer Service, QA / Tester, Inter, Designer, Network engineer / System administrator.

Figure 1: Job advertisements in 2013

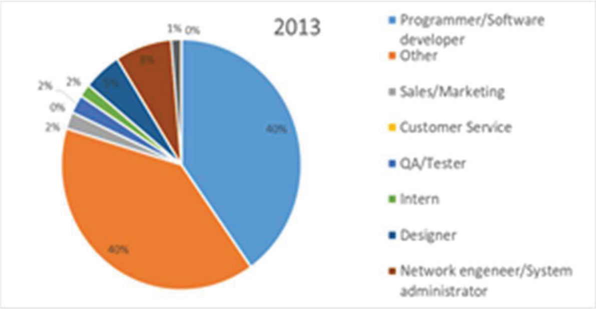


Figure 2: Job advertisements in 2014

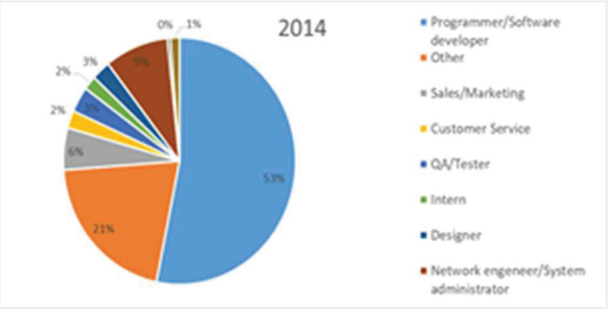


Figure 3: Job advertisements in 2015

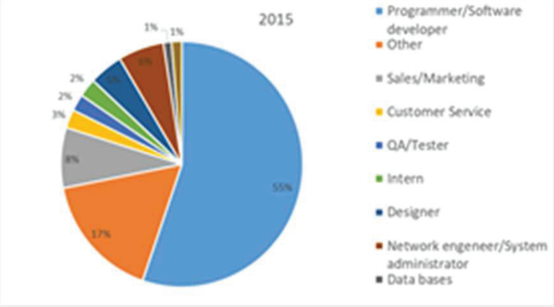


Figure 4: Job advertisements in 2016

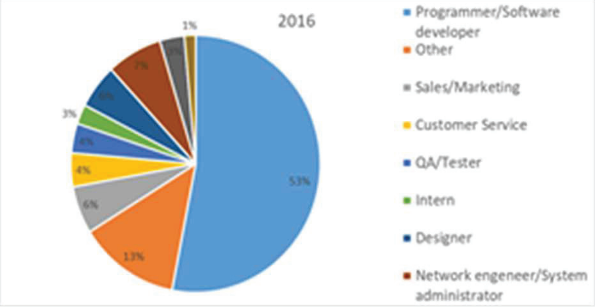
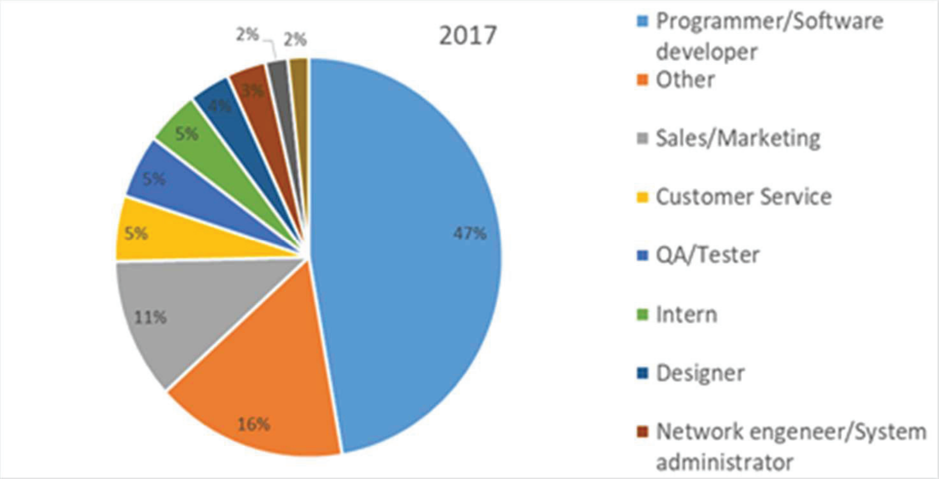


Figure 5: Job advertisements in 2017

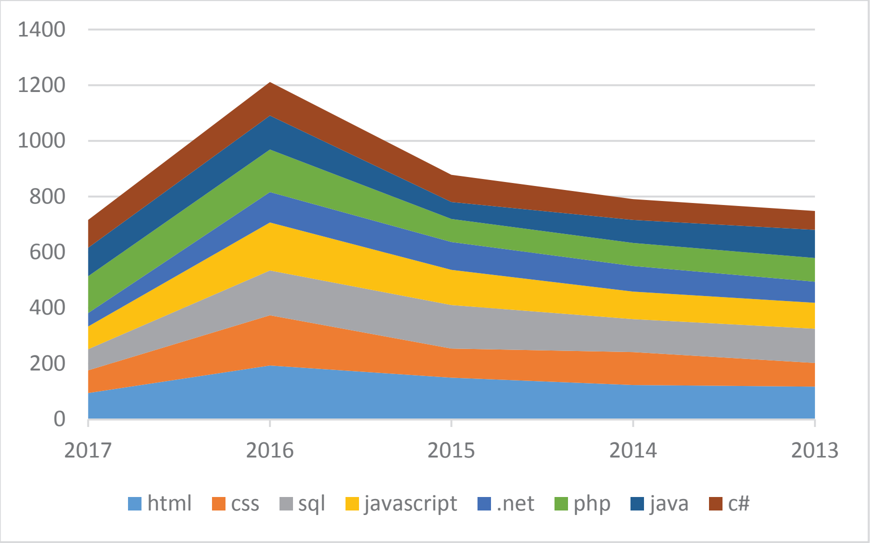


Because almost half of the jobs in those 5 years are classified as Programmer/Software engineer we extract the certain skills that are required for this job. In the Table 1 is shown the number of jobs that required certain skill (technology) per year. We extracted just the most commonly required skills in the jobs advertisements.

Table 1: Number of jobs that required certain skill/technology

	2017	**2016**	**2015**	**2014**	**2013**
html	94	193	150	123	117
css	82	181	104	119	86
sql	76	161	156	118	122
javascript	81	172	127	99	93
.net	49	109	100	92	76
php	132	153	83	83	86
java	102	122	61	82	101
c#	100	121	97	75	67

Figure 6: Number of required skills in jobs



From the Table 1 and Figure 6 we can see that the eight most demanding technologies have some stable trend over the years and almost all of them have increasing trend relatively related to the growth of the number of jobs. We should keep on mind that we have data from just a first half of 2017.

4. CONCLUSION AND FUTURE WORK

By analyzing job advertisements we can extract very important information. In our case with simple text mining techniques we scrap the IT job advertisements and extract the main categories from them. As we can see in the results section the Programmer/Software engineer is the most sought-after job from employers. This information can be used in creating curriculum from schools in order to develop most suitably educated cadres and also for the student to choose their education in the direction that they can find work easily. Also this information can be used from some superior structures so they can see all the shortcomings in the industry and will know in which direction to put their policies. From the most demanding technologies part we can conclude which technologies are more demanded and in the rising position. We can also see the period when one technology is alive and in how the all those technologies are changing and developing.

There is a lot of space for this work to be extended. The non-technical skills, soft skills, should be also included in the research because they are also very important part for the employers, and can be bust for job seekers in order to know what other skills they should develop. Also the salary part can be included, if it is available, so we can get clearer picture what position and what skills are the most valuable. The combination of the skills should also be considered. Do some skills always go together and what skills are correlated and cannot function without each other? By using more advanced text mining techniques maybe some non-trivial information can be extracted, like new categorization job advertisements patters etc. Also, some predictions about job need in the future can be provided.

The main concussion is that the work done is very valuable and gives us very important information about IT labor market, and what are the trends. But, there is much that can be done in the field, so the more information can be extracted and greater concussions can be made.

ACKNOWLEDGEMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss.Cyril and Methodius: University"

REFERENCES

[1] Peter, A. T., McKeen, J. D., Gallupe, R. B. (1995) *The Evolution of IS Job Skills: A Content Analysis of IS Job Advertisements from 1970 to 1990*, Management Information Systems Research Center, University of Minnesota, MIS Quarterly Vol. 19, No. 1 (Mar., 1995), pp. 1-27.

[2] Bennett, R., (2002) *Employers' Demands for Personal Transferable Skills in Graduates: a content analysis of 1000 job advertisements and an associated empirical study*, Journal of Vocational Education & Training Volume 54, 2002 - Issue 4, pp. 457-476.

[3] Elving W. J. L., Westhoff J. J. C., Meeusen K., Schoonderbeek J. W., (2013) *The war for talent? The relevance of employer branding in job advertisements for becoming an employer of choice*, Journal of Brand Management, Volume 20, Issue 5, Palgrave Macmillan UK, pp. 355-373.

[4] Kim J., Warga E., Moen W. E. (2013) *Competencies Required for Digital Curation: An Analysis of Job Advertisements*, The International Journal of Digital Curation Volume 8, Issue 1, pp. 66-83.

[5] Gao L., Eldin N. (2014) *Employers' extractions: A probabilistic text mining model*, Creative Construction Conference, Procedia Engineering 85, pp. 175-182.

[6] Loth R., Battistelli D., Chaumartin F. R., De Mazancourt H., Minel J. L., Vinckx A. (2010) *Linguistic information extraction for job ads (SIRE project)*, 9th international conference on Adaptivity, Personalization and Fusion of Heterogeneous Information, Paris, France pp. 300-303.

[7] Wise S., Henninger M., Kennan M. A. (2011) *Changing Trends in LIS Job Advertisements*, Australian Academic and Research Libraries 43(4), pp. 268-295.

[8] Iezzi D. F., Mastrangelo M., Sarlo S. (2007) *Text clustering based on centrality measures: an application on job advertisements*, 11es Journées Internationales d'analyse statistique des données textuelles, Liegi, Belgium, pp. 515–524.

[9] Wowczko I. A. (2015) *Essential of Strategic Management*, Informatics, pp. 31-49.

[10] Zhang S., Li H., Zhang S. (2012) *Job Opportunity Finding by Text Classification*, International Workshop on Information and Electronics Engineering (IWIEE), pp. 1528–1532.

[11] Karakatsanisa I., AlKhader W., MacCrory F., Alibasic A., Omar M.A., Aung Z., Woon W.L. (2017) *Data Mining Approach to Monitoring the Requirements of the Job Market: A Case Study*, Information Systems, Volume 65, pp. 1-6.