Searching, Examining, and Exploiting In-demand Technical (SEE IT) Skills using Web Data Mining

Taeghyun Kang
Department of Computer Science
University of Central Missouri
Warrensburg, MO, USA
tkang@ucmo.edu

Hyungbae Park
Department of Computer Science
University of Central Missouri
Warrensburg, MO, USA
park@ucmo.edu

Sunae Shin
Department of Computer Science
University of Central Missouri
Warrensburg, MO, USA
sshin@ucmo.edu

***Abstract*—There has always been a mismatch between academic education and industry demands in various industry fields. Especially, in the computer related jobs such as software developers or engineers, employers and job seekers are experiencing a large curriculum or skill gap compared to other industry fields due to its fast-changing nature. In order to minimize the gap, identifying the current job skill trends and offering/learning relevant skills are critically important for both educators and students in the fast-paced job market. However, due to the lack of relative information about the trends of in demand technical skills in certain regions or around the nation, it is difficult for educators to keep up with these changes and determine what needs to be conveyed to students so they can be ready for the job on day one. In this paper, in order to address these challenges, we focused on the analysis of in demand technical skills and their trends in software developer and engineer jobs around the nation. We collected and analyzed over 120,000 software developer or engineer job postings and summarized demanding and required skills from different regions throughout the nation.**

***Index Terms*—Web Data Mining, Job Trend Analysis, Software Developer, Software Engineer, Programming Languages**

I. Introduction

The academic program's job placement rate is one of important performance indicators that can be used to measure the success of a program. Programs, which demonstrate a strong job placement rate, frequently use this information for advertisement. It is critical for educators to convey current knowledge that is currently demanded or required by employers. However, due to the fast-changing nature in the requirements of computer related jobs, it is challenging that we minimize the gap between academic education and industry requirements. This is not the matter of the quality of the academic program or students being lazy and not studying hard enough to meet the job requirements. The rapid changes of technologies are a nature of industry caused by factors such as satisfying the customers' demands and achieving a company's revenue goal efficiently. As an educator, we need to keep asking ourselves, "Do our graduates possess the skills employers need?" and "How do we keep up with fast-changing industry requirements?" In order to clearly answer the questions, it is inevitable to make a systematic way that continuously monitors and analyzes the trends and changes in job markets. There are various job search engines available such as Glassdoor, Indeed, Handshake, etc. Even though they provide advanced search filter so the job seekers can easily identify the jobs they are looking for, search features of these websites have limited capability and do not show the trends of jobs or specific skills. Many academic programs have their industry advisory board meeting to minimize the gap, but it is still difficult to see the overall trends of required skills between states and most in-demand skills in each state. Job seekers and students also proactively find available resources to identify job trends or requirements so that they can be prepared for a job and find their job effectively and efficiently. If failed, they may need to re-educate themselves.

In this paper, we focus on making a systematic tool to monitor and analyze in demand technical skills and their trends in software developer and engineer jobs around the nation rather than addressing programming languages expected to be offered in specific regions. We collected and analyzed over 120,000 software developer or engineer job postings around the nation over the five month period from October 2019 to February 2020 and summarized demanding and required skills from different regions throughout the nation. The results show the popularity of integrated development environments, project management tools, and programming languages and their correlations. The remainder of the paper is structured as follows: Section II describes the motivation of the paper and reviews background knowledge and literature references. Section III describes the proposed tool for job skills trend analysis and the data sets processed by the tool. Section IV shows the results of our analysis and Section V concludes the paper.

II. Related Work

Web mining is the technique that collects and processes freely available data from web postings and web documents and discovers and extracts useful insight, knowledge, and information. Web mining can be categorized into three broad areas [14], [17]–[19] such as web usage mining, web content mining, and web structure mining. The enormous amount of web data can be counterproductive without proper processing procedures or automated and systematic methodologies. It will take a long time to collect and digest data and this tedious and un-automated process can mislead the viewers. There are various important techniques in order to make

DOI reference number: 10.18293/SEKE2020-058

data mining effective such as generalization, tracking patterns, classification, characterization, association, outlier detection, clustering, regression, and prediction [8], [9]. Many research work have used these web data mining techniques to find and extract hidden information from the ever-growing number of web postings and documents.

Milad Eftekhar et al. [15] proposed two models such as intrinsic burst model and social burst model, in order to find and extract knowledge and information from burst behaviors and neighbors' influences when identifying bursts on social network sites. They used two graphs called action graph and holistic graph to characterize and identify users' bursty behavior.

Nicola Barbieri et al. [16] proposed the Community–Cascade Network(CCN) model that utilizes information propagation and community formation in a social network can be explained according to the level of active involvement and the degree of passive involvement, which ultimately guide a user behavior within the network. They validated the proposed CCN model by applying it to real-world social networks.

Wilden et al. [10] presented a framework that suggests the effectiveness of a brand signal to potential employees. The employee-based brand equity influenced by employer brand clarity, consistency, brand investments and the credibility of brand signals. Furthermore, the paper present that the brand investment influences both attractiveness of a prospective employer and employee based brand equity.

Xu et al. [11] investigated the prediction of job change occasion based on career mobility and daily activity patterns at the individual level. In order to model the job change motivations and correlations between professional and daily life, the work experiences and check-in records of individuals are collected. They found that the job change occasions are predictable and shown on the experiment based on the large real-world data set.

The talent exchange prediction method is developed in [12] for predicting the possible companies for the potential employees. The proposed talent circle detection model extracts talent circles that includes the organizations with similar talent patterns. In addition, the semantic meaning is offered for detected circles are labeled with job description. With the proposed method, the organizations are able to seek the right talent during recruitment and job seekers can find appropriate jobs.

H. Li et al. [13] proposed a model that focuses on predicting the turnover and career progression of talents. The survival analysis approach shows survival status at a sequence of time intervals for turnover behaviors of employees. Moreover, the prediction of the relative occupational level is framed for modeling career progression.

Among various applications mentioned above, the applications related to job search and trends [10]–[13] are our interest in this paper. This analysis will help identify and reduce the gap between academia education and industry demands. L. Buth et al. [1] analyzed the needs of industry by interviewing employers and the situation of a university responsible for educating students. They identified the gaps in 1) applied knowledge and problem solving skills, 2) communication skills, and 3) self-discipline and positive work attitude. Bracey [22] pointed particularly to the absence of teaching "soft skills" in academic degree programs. She mentioned that majority of employers are demanding soft skills as a pre-condition for their employment. McGill [24] particularly surveyed the hiring needs of game industry and compared them against game development curriculum at post-secondary institutions. Hynninen et al. [23] conducted the similar survey in the field of software testing and quality assurance. They identified the differences between what is considered important skills by the industry and those taught in academia. Cheng et al. [25] presented an analysis of job transition network for recruitment. The authors suggested a real-time system for mining job–related patterns collected from social media sources.

The references listed above indicate the need from field professionals to understand the employable skills in this dynamic job market. This is critical for employers and students because they seek qualified employees and these job opportunities, respectively. For employers, it will improve the chances of hiring the right person and for students it will improve the chances of getting hired.

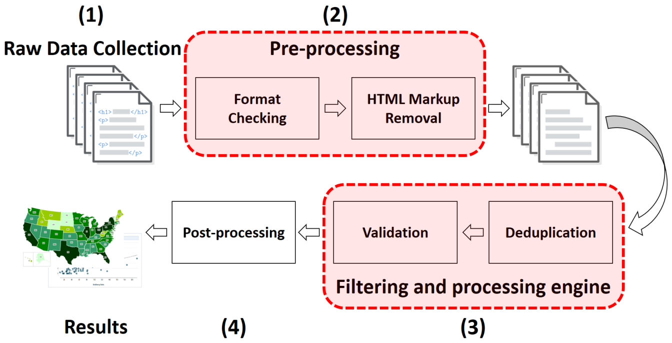

Fig. 1. The overview of data processing flow through multiple phases

III. Data Mining Tool and Data Sets

We developed a home-grown web data mining tool that crawls job posting web pages from job search websites (i.e., indeed), collects raw data of job postings, and extracts necessary information from each job posting in order to achieve the objectives of our analysis. In order to improve the efficiency of our tool and the accuracy of the data sets, we conducted black box and white box testings on our tool. Through this continuous improvement process, we were able to get more accurate data sets within a shorter amount of time.

The series of data handling processes can be divided into several phases. In the raw data collection phase (1) in Figure 1, we developed our own crawler that is customized for different job search websites. Each job search website has a distinct structure and it uses different web page management methods.

In order to efficiently discover and collect job postings from the web, we needed to customize our crawler for different job search websites and collected more than 120,000 job postings from the web. This may seem that we are having an enormous amount of opportunities. However the enormity of the data can be counterproductive in many ways. If we have to check out each job posting individually, then it would take a long time to tabulates and extract what educators, students, and job seekers need to focus and it may give misleading information to the viewers. In worst case, they may need to start over learning new programming languages. Therefore the collected raw data sets will be refined throughout the following three phases.

The next pre-processing phase (2) in Figure 1 utilizes the Beautiful Soup Python package [7] in order to efficiently extract necessary information from each job posting. The Beautiful Soup Python package makes it easy to parse HTML and XML documents and generates a parse tree. This parse tree can be used to extract specific data from HTML.

In the Filtering and processing phase (3) in Figure 1, we check the duplication of job postings and remove redundant job postings from each company. We noticed that there are many duplicated job postings for the same position from the same company. For each job posting, a unique value is assigned in the tag called “vjk”. This is a unique identifier for each unique job description. We utilized the “vjk” ids and job descriptions to filter out unnecessary redundancies from the data sets. In order to achieve a highly accurate and quality collection of the data, the collected data was deduplicated and validated.

In the last phase (4) in Figure 1, we extract only information that corresponds to our analysis. After the collected raw data sets are processed through the four phases described in Figure 1, we were able to identify total 2,171 unique job descriptions out of over 120,000 job postings.

IV. Job Trend Analysis and Results

As described in the previous section, we have identified total 2,171 unique job descriptions out of over 120,000 job postings. This shows that there are on average 55 job postings for the same position with the exactly same job descriptions from the same company. Various reasons can cause the enormous amount of duplications; 1) a position has not been filled for a while, 2) a company wants to expose their job descriptions more frequently than others, 3) a company needs more employees for the same position with the same skill sets, etc. In this section, we will focus on those uniquely specified jobs for our analysis in order to minimize noise that may be caused by enormous duplications. The following subsections describe job requirement trends in several categories such as integrated development environments, software management tools, programming languages, etc.

A. Integrated Development Environments (IDEs)

IDEs have been used for decades and increase software developers’ productivity. Therefore, this is definitely one of the required skills for any programmer positions around the world. However, the commonly used or required IDEs have not been specified in most of the job descriptions. The IDEs are specified by only a few companies around the nation. This is because each team even within a company uses multiple or different IDEs depending on programming languages they use to develop software or applications or to complete given tasks. For example, IntelliJ would be used for a project written in Java, Android studio for Android, RubyMine for Ruby, and etc. Therefore, it would be difficult for employers to specify any IDEs used in their job descriptions. This suggests that job seekers may want to expose themselves to multiple IDEs for different programming languages. Programmers need to be flexible and should be able to use any IDEs in their work environments. Based on the popularity of programming languages, the IDEs that support Java are placed top ranks (see Figure 4). Xcode is placed the fourth place and the popularity of Xcode is due to the popularity of Apple’s iPhone around the world. We can see the strong correlation between the popularity of both IDEs and programming languages.

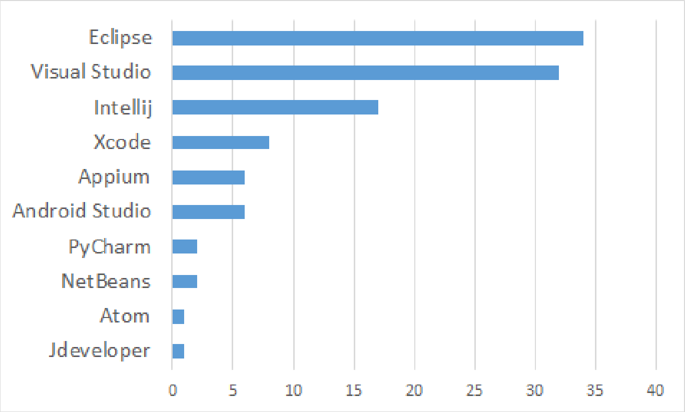

Fig. 2. Popularity of integrated development environments

B. Project Management Tools

We also conducted an analysis to identify which software development models or project management tools are commonly used in industry. Interestingly, the vast majority of companies is not specifying or asking a specific software development model or a project management tool in their job descriptions. However, companies in certain areas such as Washington, California, Arizona, Colorado, Missouri, Georgia, and New York showed a strong trend that these specifications are listed in their job description.

As shown in Figure 3, the Agile development is being specified most frequently compared to other agile project management processes such as Scrum, Lean, Kanban, etc. Traditional project management models such as the spiral or evolutionary models have not been specified by any of the job postings but only the waterfall model has been specified along with the agile process as a required background. These job descriptions required understanding of both agile and waterfall developments. This may indicate that the traditional software

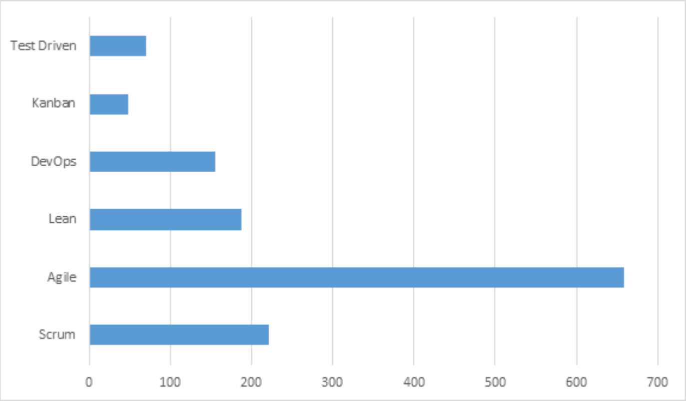

Fig. 3. Project management tools

development process has been converged to the most broadly used model among those traditional software development models. It also indicates that the trend of software development cycle has been shifted from the traditional project management models to the agile project management models. This does not mean that the traditional software development models are not being utilized in industry. There are many companies that use the combination of traditional software development and agile process.

C. Programming Languages

As we see in Figure 4, Java is the most popular programming languages around the nation. This is evident by the top ranked IDEs in Figure 2. All the top 4 IDEs in the list support Java. Compared to the popularity of Xcode shown in Figure 2, 'Swift' ranked lower in the category of popular programming languages. This is because Xcode support not just Swift, but also other programming languages such as C, C++, Objective-C, Objective-C++, Java, Python, and Ruby.

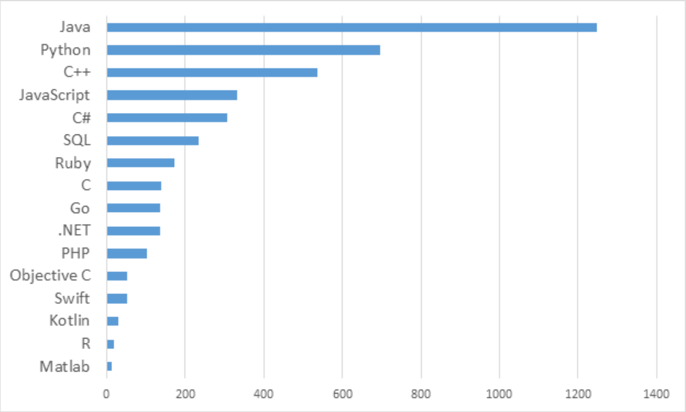

Fig. 4. Popularity of programming languages

D. Soft Skills

Teamwork, written and spoken communication skills, critical thinking, leadership, and work ethics are critical soft skills required and demanded by employers. Especially, some large corporations already have well-organized intensive programs for their new employees to teach hard skills and they even more emphasize the importance of soft skills as soft skills are hard to teach in a short period of time. It was evident by many surveys such as [22]–[24]. However, the vast majority of job descriptions may mislead job seekers, students, and educators as they are not specifying these critical skill sets. Even though the vast majority of companies wants to hire people who has excellent soft skills, they do not actively require or specify these soft skills in the job description in reality. It is understandable as soft skills are not only hard to teach but also hard to evaluate. But a clear job description with soft skills would help job seekers, students, and educators be aware of the importance of soft skills and prepare for it.

E. New Technologies

Along with the traditional programming languages and IDEs, we also conducted analysis for relatively new technologies such as blockchain, and TensorFlow. We have found 29 and 23 job postings specifying blockchain technologies and TensorFlow knowledge and background, respectively. These numbers are relatively lower than other traditional and popular programming languages. The adaptation of these new technologies will be collected and analyzed in our future work.

V. Conclusion

Offering and learning appropriate programming languages and demanding skills are critically important for both educators and students in the fast-paced job market. However, due to the lack of a systematic approach that analyzes job skills demands and trends, it is difficult to keep up with these fast changes. In this paper, we developed a job trend analysis tool, analyzed over 120,000 job postings, and summarized demanding and required skills from different regions throughout the nation.

As a continued effort toward this work, we will expand our analysis to other fast-paced and emerging disciplines such as cybersecurity and etc. This will also help both educators and students to identify in demand technical skills in the industry and frequently exploited vulnerabilities. In addition, we will analyze other countries' in demand technical skills to investigate the worldwide trend for a certain job. In addition, we are currently working on integrating machine learning features (e.g., TensorFlow [20]) so the tool can automatically generate reports that show the trends of job and programming languages without involving or requiring a manual analysis.

References

[1] L. Buth, V. Bhakar, N. Sihag, G. Posselt, S. Bohme, K.S. Sangwan, and C. Herrmann, "Bridging the qualification gap between academia and industry in India," Proceedings of the 7th conference on Learning Factories (CLF), Volume 9, pp.275–282, Elsevier, 2017.

[2] M.M. McGill, "Defining the expectation gap: a comparison of industry needs and existing game development curriculum," Proceedings of the 4th International Conference on Foundations of Digital Games, pp.129–136, April 2009.

[3] A.H. Harris, T.H. Greer, S.A. Morris and W.J. Clark, "Information systems job market late 1970'2-early 2010's," Journal of Computer Information Systems, pp.72–79, 2012.

[4] C. McLean, “A foot in the door: IT job-search strategies,” Certification Magazine, Volume 8(4), pp.38–40, 2006.

[5] C. Litecky, A. Aken, A. Ahmad, and H.J. Nelson, “Mining for computing jobs.” IEEE Software, Volume 27(1), pp.78–85, 2010.

[6] S. Zhong, “Information intelligent system based on web data mining,” Proceedings of the International Symposium on Electronic Commerce and Security, pp.514–517, 2008.

[7] https://www.crummy.com/software/BeautifulSoup/#Download

[8] P. Berkhin, “Survey of Clustering Data Mining Techniques,” Grouping Multidimensional Data, Springer, 2006.

[9] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, “Data mining techniques and applications – A decade review from 2000 to 2011,” Expert Systems with Applications, Volume 39, Issue 12, pp.11303–11311, Elsevier, 2012.

[10] R. Wilden, S. Gudergan, and I. Lings, “Employer branding: strategic implications for staff recruitment,” Journal of Marketing Management, 26(1-2), pp.56–73, 2010.

[11] H. Xu, Z. Yu, H. Xiong, B. Guo, and H. Zhu, “Learning career mobility and human activity patterns for job change analysis,” IEEE International Conference on Data Mining, pp.1057–1062, 2015.

[12] H. Xu, Z. Yu, J. Yang, H. Xiong, and H. Zhu, “Talent circle detection in job transition networks,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.655–664, 2016.

[13] H. Li, Y. Ge, H. Zhu, H. Xiong, and H. Zhao, “Prospecting the career development of talents: A survival analysis perspective,” Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.917–925, 2017.

[14] N. R. Satish, “A Study on Applications, Approaches and Issues of Web Content Mining,” International Journal of Trend in Research and Development, Volume 4(6), ISSN: 2394-9333, 2017.

[15] Milad Eftekhar, Nick Koudas, and Yashar Ganjali, “Bursty Subgraphs in Social Networks,” Proceedings of the 6th ACM international conference on Web search and data mining (WSDM), pp.213–222, Rome, Italy, 2013.

[16] Nicola Barbieri, Francesco Bonchi, Giuseppe Manco, “Cascade-based Community Detection,” Proceedings of the 6th ACM international conference on Web search and data mining (WSDM), pp.33–42, Rome, Italy, 2013.

[17] Ujwala Manoj Patil, J.B. Patil, “Web Data Mining Trends and Techniques,” International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp.962–965, 2012.

[18] B. Liu, “Web Data Mining Exploring Hyperlinks, Contents, and Usages Data,” 2nd Edition, Springer, 2007.

[19] Jai Prakash, Bankim Patel, Atul Patel, “Web Mining: Opinion and Feedback Analysis for Educational Institutions,” IJCA, Volume 84(6), 2013.

[20] “TensorFlow: An end-to-end open source machine learning platform,” url: https://www.tensorflow.org/

[21] D. Smith and A. Ali, “Analyzing Computer Programming Job Trend Using Web Data Mining,” Issues in Informing Science and Information Technology, Volume 11, pp.203–214, 2014.

[22] P. Bracey, “Analyzing Internet Job Advertisements to Compare IT Employer Demands versus Undergraduate IT Program Curriculum in Texas,” Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, pp.37–42, ISBN 978-1-880094-90-7, 2011.

[23] T. Hynninen, J. Kasurinen, A. Knutas, and O. Taipale, “Guidelines for software testing education objectives from industry practices with a constructive alignment approach,” pp.278–283. 10.1145/3197091.3197108, 2018.

[24] M. M. McGill, “Defining the expectation gap: a comparison of industry needs and existing game development curriculum,” Proceedings of the 4th International Conference on Foundations of Digital GamesApril (FDG), pp.129-–136, 2009.

[25] Y. Cheng, Y. Xie, Z. Chen, A. Agrawal, A. Choudhary, and S. Guo, “Jobminer: A real-time system for mining job-related patterns from social media,” Proceedings of ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), 2013