ANALYSIS AND PREDICTION OF COMPUTER INDUSTRY TALENT DEMAND BASED ON MACHINE LEARNING

Yu Han and Haiyi Zhang  
Acadia University, Computer Science, Acadia University, Wolfville, NS, B4P 2R6 Canada

ABSTRACT
From the perspective of enterprise and social demand, this research analyzes the characteristics of talent demand in Canadian computer industries by using big data thinking, text analysis and machine learning. Firstly, through crawler technology, we obtained the recruitment information on mainstream recruitment websites in Canada and developed a standardized collection process. After processing the data, we used IBM Watson Studio to split the text and constructed the skill keyword dictionary to extract the talent demand. In addition, the time series algorithm was used to predict the changes of the computer skill keyword over time and visualize the results. After analyzing the experiment results, it is found that computer science related jobs are changing over time and require new skills to adapt. Finally, based on the analysis results, suggestions and plans for cultivating talents related to computer science are proposed from the perspectives of enterprises, governments and campuses. The governments should pay attention to policy making, campuses and enterprises should strengthen cooperation and create a new talent training model.

KEYWORDS
Computer Industry, Talent Demand, Text Mining, Machine Learning, Time Series

1. INTRODUCTION
The number of computer-related industries has rapidly increased since computer science emerged as a specialization and associated technology has rapidly developed. At the same time, the demand for talent to fill positions in these new industries has significantly increased. According to Computing Technology Industry Association (CompTIA) reported that 282,000 computer jobs were created since 2011, and the group expects some computing occupations to grow by double-digit percentages between 2019 and 2027 (Computing Technology Industry Association, 2020). Rapidly changing knowledge and skill requirements further aggravates the relationship between supply and demand of trained computer talent in Canada.

In the past decade, research on the characteristics of talent needed by industry and text mining based on Big Data have increasingly attracted attention from researchers. Therefore, taking the perspective of business and organizational enterprise requirements, my research aims to make research on the topic of computer industry talent skills requirements more efficient and by using Big Data and text mining. The intent of this research is to inform the knowledge base of the computer industry so that the industry has a way of tracking changes in talent demand characteristics. Promoting talent training and matching the supply with the demand for computer science talent can ultimately promote the sustainable and healthy development of the computer science industry.

This research used text mining of the web to broadly explore the theory behind unstructured data research. I incorporated online recruitment information as my data source. Then, this research used web crawler technology to collect my data, which extensively saved labour and time and improved efficiency. This research used web text mining technology to construct a multi-dimensional and expandable talent demand skills dictionary from which I extracted features in the recruitment textual data. This research used another text mining algorithm to analyze the invisible knowledge mode of the textual data, which kept my research more systematic. I introduced a time dimension to my data. Using text mining and time series analysis methods allowed my research to comprehensively study the talent demand characteristics of the Canadian computer industry by allowing me to correlate the certain industry jobs with various skills to discover the strength of the relationship between the two and to study how skills keywords changed over time so that I could predict future needs and level of demand.
2. BACKGROUND STUDIES

2.1 Web Text Mining

Text mining is a comprehensive analysis method that intersects multiple fields such as machine learning, data mining, information retrieval, natural language processing, and knowledge management (Feldman and Dagan, 1995). As in the principle of data mining, text mining identifies and explores laws and rules to extract information from data sources sufficient to solve the problem. Web text mining is the process of discovering an implicit pattern, \( P \), from a collection, \( C \), of a large number of text documents that originate on the web. If \( C \) is the input and \( P \) is the output, then web text mining maps \( \sigma \) from input to output: \( C \rightarrow P \) (Feldman et al., 1998).

The process of web text mining is generally divided into seven steps: text collection, text preprocessing, feature extraction, text representation, text mining, quality evaluation, and visualization of mining results.

2.2 Time Series

Exponential smoothing is a time series forecasting algorithm, generally split into single exponential smoothing, double exponential smoothing, and triple exponential smoothing. The forecast produced by the exponential smoothing method is a weighted average of prior observations. The nearer the observation time, the higher the correlation weight (Bekhradnia, 2007). The exponential smoothing method is sometimes called the ETS model, which refers to the explicit modelling of errors, trends, and seasonality (Hyndman and Athanasopoulos, 2018).

Equation 1 describes the process.

\[
S_t = ay_t + (1-a)S_{t-1}
\]

Where \( S_t \) represents the predicted value at time, \( t \) (or the value of the exponential). \( y_t \) represents the actual value at \( t \), and \( S_{t-1} \) represents the predicted value at \( t-1 \) (or the value of the exponential). \( a \) is a smoothing coefficient (Hyndman and Athanasopoulos, 2018). The initial value is the predicted value at the first time step.

When the number of items in the original series is greater than 15, the observed value at the first time step or the observed value at the previous time step is selected as the initial value. When the number of items is less than 15, the average of the observed values at the first three time steps is selected as the initial value.

In performing exponential smoothing, the key is the value of \( a \). When the time series data are relatively stable, \( a \) should be a smaller value, generally between 0.05 and 0.1. However, when the time series data fluctuate, and the long-term changes are not large, \( a \) slightly larger value for \( a \) should be chosen, between 0.2 and 0.5. When the time series data are increasing or decreasing, \( a \) should be larger still, between 0.6 and 1.0.

3. THEROY

3.1 Data Collection

For my research, I regularly collected the required data, twice per month, based on defined time nodes using web crawler technology. First, the web crawler selected mainstream recruitment URLs as sub-URLs to form URL categories. Then the software crawled web pages for interpretation and analysis. When crawling, the software continuously obtained new sub-URLs to join the queue and finally crawled all sub-URLs under specific conditions (Amudha and Phil, 2017). The task of the web crawler was to crawl the recruitment information of Canadian computer-related job postings on the three recruitment websites and crawl out my required fields to provide text information regarding the characteristics of talent demand of the industry.

3.2 Data Preprocessing

The collected recruitment information belonged to the unstructured text dataset, including text and symbols. I used the N-gram algorithm to process the recruitment information. I extracted core skills data from the
sampled data corresponding to a skill dimension, and I created a skills dictionary as a function of time. The combining text segmentation and N-gram segmentation, a the general search-based morphological analysis method combined with a non-dictionary-based N-gram segmentation, improves the search quality (Rijmenam, 2013).

After preprocessing the recruitment information, I retained the following analysis fields: recruitment position, work location, crawl date, work experience, academic requirements, salary, and job description. To identify the name of the recruiting post, I added a new “job category” field. Because each organization had a different name for its job posting, to facilitate uniformity and standardization, I classified specific posts from a data analysis perspective. The specific classifications are Technology, Product, Design, Operating and Marketing.

3.3 Skill Dictionary

I used relevant software tools extracted the skills keywords. Skills keywords of the same job category were grouped into one category. Semantic skills keywords were also classified into one category. This method only performed feature extraction of skills keywords with the goal of identifying existing skills keywords while simultaneously performing real-time analysis of the data.

The topic classification method of the skills keywords should learn from existing relevant skills keywords and added newly appearing skills keywords. The job skills in computer-related industries are divided into three dimensions: technical skills, business skills, and comprehensive skills (Vassakis, Petrakis & Kopanakis, 2018) and eight first-level skills indicators and thirty second-level skills indicators are constructed, as listed in Table 1.

Table 1. The skill dimensions and indicators of skill dictionary

<table>
<thead>
<tr>
<th>The dimension</th>
<th>First-level skills indicators</th>
<th>Second-level skills indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical skills</td>
<td>Basic computer knowledge, Website construction and maintenance</td>
<td>Operating system knowledge, computer protocol knowledge, database application, website testing,</td>
</tr>
<tr>
<td></td>
<td>Search engine optimization</td>
<td>visual perception, graphic design, website architecture/operation, web page production, video</td>
</tr>
<tr>
<td></td>
<td></td>
<td>clip, search engine application, website ranking</td>
</tr>
<tr>
<td>Business skills</td>
<td>The network marketing, Marketing management</td>
<td>Social networking, customer acquisition and retention, online sales platform, e-commerce model,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sales skills, website topics and activities, data analysis</td>
</tr>
<tr>
<td>Comprehensive skills</td>
<td>Professional quality, Interpersonal communication, Office automation</td>
<td>Work attitude, executive ability, psychological quality, learning ability, thinking ability,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>language ability, team ability, management ability, office software, copy editing and writing</td>
</tr>
</tbody>
</table>

4. EMPIRICAL STUDIES

4.1 Analysis of the Overall Need for Talent in the Computer Industry

I statistically analyzed the collected recruitment information for computer science-related positions in terms of overall demand, demand for various positions, academic requirements, and salary levels. I used various graphical methods to display the results of all aspects of my analysis.
I present the statistical trends found in the total amount of recruitment information in Figure 1(a). The period ranged from February to August 2021, spanning three financial quarters. The figure shows that during March to April, the demand for jobs slightly declined. When summer arrived, the market for computer jobs peaked compared to February through May. The market high endured over the next few months.

To clearly analyze my data, I divided my almost 12 months of data into four quarters. January to March is the first quarter; April to June is the second quarter; July to September is the third quarter; and October to December is the fourth quarter. I chose the median month of each quarter to represent its quarter. I classified recruitment positions into job categories, and I grouped the job categories into five categories: product, technology, marketing, design, and operation.

Figure 1(b) presents the statistics of the recruitment information data for the first three quarters of 2021. In terms of job categories, in the first quarter, technical jobs had the most significant market demand. In the second and third quarters, the number of job types changed significantly. The number of operation and design job postings gradually and steadily increased, while the number of product and marketing job postings slightly decreased. Commodity circulation channels in the computer industry are increasing so operations are critical to any computer-related enterprise and demand in this talent pool is relatively extraordinary. The demand for talent in marketing and product in computer-related industry weakened, especially in marketing.

Because business enterprises and organizations use different salary calculation methods when posting recruitment information, it was impossible for me to make a unified analysis. I adjusted all the salary postings to determine an annual salary for each posting. Figure 1(c) compares the salaries for the computer industry postings. The salary range with the most significant proportion of available jobs during the three quarters was the 40k to 80k per year salary range. The proportion of available jobs in the second and third quarters for this salary range progressively dropped, but they still surpassed 60% for those two quarters. The lowest paying jobs were in the less than 40k per year range and the fewest available jobs in that range were in the first quarter. An increase in talent demand indicates that the computer industry has increased its ability to pay competitive salaries, which, in turn, attracts more talent to the industry.

Figure 1(d) shows that the overall change in level of education was not significant over the three quarters for each educational requirement. The educational requirements for available positions mainly required a Bachelor’s degree. The annual statistics showed that a Bachelor’s degree was required for more than 40% of available jobs, followed by the Master’s degree and college diploma, both required by approximately 10% of available jobs. The percentage of jobs requiring Doctoral degrees or only high school diplomas was not very high, and these educational requirements were polarized.
4.2 The Talent Skill Demand of Computer Industry

I divided the skills requirements into technical dimensions, business dimensions, and comprehensive dimensions. Using the skills dimensions of my dictionary, I determined eight first-level indicators, 30 second-level indicators, corresponding text information, and word frequency statistics which I listed in Table 2. In the table, “Q1”, “Q2” and “Q3” represent the three quarters; “freq” is the short form of “frequency”, the number of times that keywords appear in the text; and “Prop” is the short form of “Proportion”, which represents the percentage of the total number of analyzed texts.

From an overall perspective, the top three first-level indicators in the first quarter were Website construction and maintenance (40.33%), Basic computer knowledge (24.66%), and Interpersonal communication (22.38%), followed by Office automation and Marketing management. The ranking of the first-level indicators in the second quarter did not change, and the overall ranking was the same. In the third quarter, the ranking of the top three indicators changed, but Website construction and maintenance still ranked first. However, there were apparent changes in Interpersonal communication, which accounted for 5.66% more than Basic computer knowledge in the total text proportion.

Table 2. Skill dimension frequency statistics

<table>
<thead>
<tr>
<th>1st Level Skills indicators</th>
<th>2nd Level Skills indicators</th>
<th>Q1 freq</th>
<th>Prop</th>
<th>Q2 freq</th>
<th>Prop</th>
<th>Q3 freq</th>
<th>Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic computer knowledge</td>
<td>The operating system</td>
<td>1337</td>
<td>24.66%</td>
<td>1822</td>
<td>34.17%</td>
<td>1622</td>
<td>25.83%</td>
</tr>
<tr>
<td>Website construction and maintenance</td>
<td>programming language</td>
<td>2180</td>
<td>40.33%</td>
<td>2953</td>
<td>55.38%</td>
<td>2824</td>
<td>44.98%</td>
</tr>
<tr>
<td>Interpersonal communication</td>
<td>Language ability</td>
<td>1234</td>
<td>22.83%</td>
<td>1757</td>
<td>32.95%</td>
<td>1977</td>
<td>31.49%</td>
</tr>
</tbody>
</table>

The second-level indicators show that the top three indices in the first quarter of the class varied. Website construction and maintenance focused on the mastery of programming languages and the use of databases, accounting for about 15% of the total text. The rankings show that programming ability was no longer a major skill requirement for Website construction and maintenance, and more attention was paid to the level of the website itself. As for Basic knowledge of computers, more attention was being paid to Operating system skills. For skills in Interpersonal communication, language ability was most required.

Table 3. Skills indicators and skills keywords

<table>
<thead>
<tr>
<th>1st Level Skills indicators</th>
<th>The top skills keywords in different quarter</th>
<th>1st Level Skills indicators</th>
<th>The top skills keywords in different quarter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic computer knowledge</td>
<td>Linux, Linux, Linux</td>
<td>Sales</td>
<td>Walmart, Amazon, Facebook</td>
</tr>
<tr>
<td>Website construction and maintenance</td>
<td>JavaScript, C++, Web</td>
<td>Marketing management</td>
<td>Data Analysis, Activities performed, Activities performed</td>
</tr>
<tr>
<td>Search engine optimization</td>
<td>Google, Google, Code optimization</td>
<td>Interpersonal communication</td>
<td>Communication, Experiences, Experiences</td>
</tr>
<tr>
<td>Office automation</td>
<td>Office, Office, Word</td>
<td>Professional quality</td>
<td>Learning ability</td>
</tr>
</tbody>
</table>

According to the Basic computer knowledge index, knowledge of Linux, Windows, iOS, Android, TCP/IP and other computer and mobile operating systems and computer protocols should still be mastered. Table 3 shows that JavaScript, Java, and C++ were the skills mainly required in the first quarter, while C++ and Web technology ranked first in the second and third quarters. For other skill indicators, there was no significant difference in the rankings of specific skills keywords in the three quarters. Google, code optimization frequently appeared under Search engine optimization. Office automation focused on the operation of office software and the writing of schedule text. Sales focused on the familiarity and understanding of major social platforms and online shopping platforms. Marketing management focused on
data analysis and activities performed. In terms of Interpersonal communication, organizations wanted more communication skills and experience. In terms of Professional quality, organizations focused on individuals’ learning ability.

The higher the keywords of a certain (type) skill are, the more basic the skill demand is and the more frequently it is used. In Figure 2, skill keywords for five job categories in different quarters are shown, with three colors representing three quarters. Technical class for technical requirements on data, for SQL has been a high concern. The product category has high requirements for data analysis. Design class requires a full understanding of product requirements and skilled use of Photoshop. Operation requires innovative ability of product solutions and post-implementation analysis. Marketing candidates are required to have a positive attitude towards work and customers.

Relevance in my study referred to the degree of correlation between documents consistent with the search criteria and the number of documents in the set of all documents that contained the specified skills dimensions or keywords. I computed the correlation using Equation 2.

Equation 2: \( C = \frac{P_n/P_N=(n_s/n)/(N_s/N_s)}{N_s/N} \). Where \( C \) represents the correlation value. \( n \) is the frequency of a certain type of job post appearing in the recruitment information. \( N \) is the quantity of the total recruitment information data. \( P_n \) is the frequency of a specific type of skills keyword, \( S \), appearing in certain types of recruitment information data. \( P_N \) is the frequency of a certain type of skills keyword appearing in the total recruitment information data. \( n_s \) is the frequency of a certain type of skills keyword, \( S \), appearing in certain types of recruitment information data. \( N_s \) is the frequency of \( S \) in the total recruitment information data.

In addition, after data processing and analysis, I found that the correlation between each position category and the first-level skills index was generally consistent. Using the first quarter as an example, Table 4 shows the correlation between job categories and first-level skills in that quarter of 2021. Among them, Basic computer knowledge, Website Construction and maintenance, Sales and Search Engine optimization are most related to technical positions. Marketing Management, Professional Quality and Office Automation are more focused on operations. Interpersonal communication is most related to product positions.

Table 4. Relevance between position categories and skill indicators

<table>
<thead>
<tr>
<th>1st Level Skill indicators</th>
<th>Relevance</th>
<th>1st Level Skill indicators</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic computer knowledge</td>
<td>Technical</td>
<td>3.6</td>
<td>Personal communication</td>
</tr>
<tr>
<td>Website construction and maintenance</td>
<td>Technical</td>
<td>3.2</td>
<td>Professional quality</td>
</tr>
<tr>
<td>Sales</td>
<td>Technical</td>
<td>2.9</td>
<td>Search engine optimization</td>
</tr>
<tr>
<td>Marketing Management</td>
<td>Operation</td>
<td>3.2</td>
<td>Office automation</td>
</tr>
</tbody>
</table>
4.3 The Trends of Computer Industry Skills Demand

Because the amount of data gathered at each time step was essentially the same, I used the frequency as the observed value to facilitate analysis and calculation. I selected Programming languages, Database, Web development as related indicators using the frequency of the core skills of keywords as observations, including Java, JavaScript, C++, C#, Python, SQL, and CSS.

I used a test algorithm to observe the trend of the skill word itself. Finally, I computed the smoothing coefficient for each skills keyword after running a test algorithm for which I set the smoothing coefficient of all the skills keywords to be 0.8.

Figures 3(a) and (b) reveals a significant downward trend in spring for the skills keywords, Java and JavaScript. The figure shows a good fit of the predicted values to the observed values so I am confident that the algorithm can be used to predict the skills trend for the fourth quarter. Figures 3(c) and (d), show that the observed values of C++ and C# in April and June noticeably dropped. Moreover, the predicted value of C# sharply declined from the first quarter to the third quarter before stabilizing in the fourth quarter. The predicted results of skills keywords SQL and Python are shown in Figures 3(e) and (f). Both of these keywords had an overall upward trend, and the predicted values of the fourth quarter declined compared with the observed values of the third quarter. The figure also shows that the number of jobs requiring SQL and Python skills in the Canadian job market increased since the start of 2021. Because website development remains a popular profession, there is still high demand for this skill. Figure 3(g) for CSS shows that CSS had the same overall trend as the other keywords for which the observed values and predicted values increased. In the fourth quarter, the graph shows that CSS was predicted to decrease.
4.4 Suggestions for Training Industry Talent

From the perspective of businesses and other enterprise growth, I suggest that businesses and organizations develop their human resources continuously by making talent reserves, planning human resource needs in advance. Businesses should offer employees relevant business training with regards to the industry and general business background knowledge as it applies to operations of the organization. In addition, Business enterprises can also cooperate with colleges and universities to actively help orient training.

Regional governments can focus on policy. Regional governments can provide support in terms of capital, preferential terms, and talent introduction by formulating relevant policies. In addition, colleges and universities in these regions could create related professional and research institutions with regards to the demand for talent of computer-related industry.

Colleges and universities should pay more attention to the impact of such technological transformation on talent training and adjust their curriculum accordingly. I suggest that academic institutions address these higher level requirements. Practical experiences can be offered through classroom simulations and more cooperative training programs between institutions and businesses and organizations.

5. CONCLUSION AND FUTURE WORK

The purpose of my research was to analyze the characteristics of talent demand in the Canadian computer science industry and to forecast the future trends. I extracted keywords related to required skills from numerous recruitment information posts on three recruitment websites to construct a talent skills dictionary, and I analyzed the degree of correlation between skills and related job categories. Lastly, using a machine learning algorithm, I used timeseries analysis to analyze and forecast job demand trends with regards to my skills inventory of keywords. Combining my research results, the methods and technology of Big Data with emphasis on industry personnel training associated with business and organizational needs and the external environment factors, I confidently suggest improvements to talent training for the computer industry.

The following are some thoughts for future work:

There is a particular deviation between the actual results and expectations. In the future, I suggest in-depth research at longer time scales for the recruitment information data to further examine the need for computer science talent.

My time series analysis method is relatively simple and had few data variables, so the degree of confidence for the curve fitting was relatively poor. In the future, I hope to add more data and make more accurate predictions by using more complex time series methods, such as double exponential smoothing or triple exponential smoothing.

REFERENCES