A Text Mining Approach to Analyze Trends in the IT Industry Job Market through Text Extraction from Images

Sadeepa Kuruppu
Division of Information Technology
Institute of Technology
University of Moratuwa
Diyagama, Homagama, Sri Lanka
sandeepak@itum.mrt.ac.lk

Kalpana Galappaththi
Division of Information Technology
Institute of Technology
University of Moratuwa
Diyagama, Homagama, Sri Lanka
kgalappaththi@itum.mrt.ac.lk

***Abstract* —Mining the details given in job vacancies, helps to understand trends in the job market. In online job vacancy repositories, details of vacancies are stored as images and text information embedded on them. The text on the images is extracted by using a text extraction process, first they need to be preprocessed and cleansed. After that, the preprocessed text is used in the process of text mining. Term Frequency-Inverse Document Frequency mining technique was used to discover keywords in the textual data while the Apriori algorithm was used to find the associations between job titles. As job titles and required qualifications for jobs are rapidly changing in the Information Technology industry, it is worth to understand current trends in the job market when applying for a job.**

Keywords — text mining, job trends, TF-IDF, Apriori

I. INTRODUCTION

Identifying the trends in the job market will be helpful for job seekers to match their qualifications, technical skills, and knowledge with job opportunities. Then job seekers will be able to upgrade their qualifications for jobs and technological skills required for jobs depending on the industry demands. Not only that, but also educators can update curriculum compliance with the current trends in the relevant job market. Identifying and following job trends takes special importance for the Information Technology (IT) profession as it is a rapidly changing industry. Therefore, it is important if there is a way those job seekers can identify what are the dynamics in the job market. Job holders, job seekers, and youth want to know which job prospects and career path look favorable [1-3].

At present, job advertisements are mainly published electronically online. In online job vacancy repositories, vacancies are stored as images and text information embedded in them. Image text is the text information embedded or written on an image. Text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources [4].

II. OBJECTIVES

This study aims to mine trends in the IT industry job market through text extraction from images that contain details of vacancies. This study will assist to minimize dissimilarities between skills demand and skills supply in the IT industry.

III. METHODOLOGY

The process of the proposed system is illustrated in Fig. 1, As the first stage, the texts are extracted from images. After that texts are preprocessed. The pre-processed textual data is used for text mining, to discover the trends in the job market.

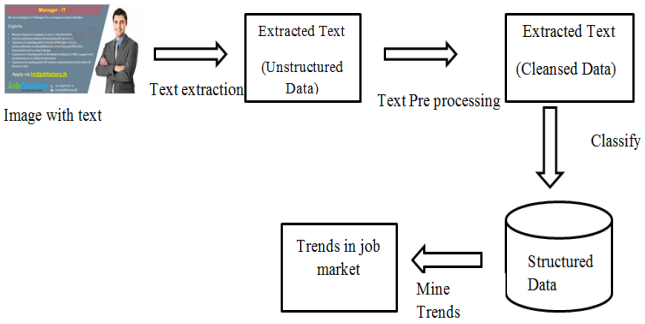

Fig. 1. Text extraction and mining process

According to the system introduced in this study, a data set with five hundred job vacancy images was used. The texts embedded in them were extracted by following Optical Character Recognition Algorithm (OCR). As the textual data embedded in images varied in length and structure, the extracted texts were considered as unstructured. They are not having a defined structure, these unstructured data are needed to pre-process to remove white spaces, special characters, and stop words [5-6]. Text preprocessing was done by following steps as given in Fig. 2; they are Tokenization, Stop word removal, and Stemming.

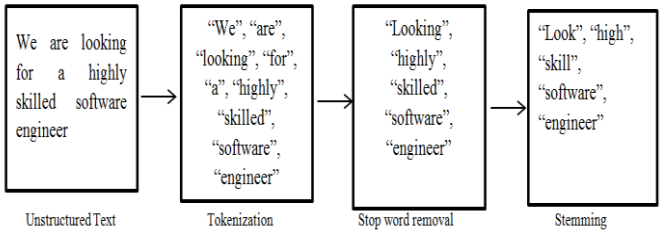

Fig. 2. Text pre-processing process

Tokenization is the process of separating a single sentence into words. This technique is used to explore the words in a sentence. Stop words do not contribute to the content or context of the textual document [6]. Also, stop words can be indicated as words that commonly repeating such as prepositions and conjunctions, therefore stop words need to be removed. Stemming is used to conflating variant forms of a word into a common representation [6]. During the process of stemming, prefixes and suffixes are removed from the word. Stemming is carried out to find the root or the base of the word. This root is known as the "lemma" in Natural Language Processing (NLP). After pre-processing and cleansing data, data is stored in the database and use for text mining purposes [7-9].

In this study, to mine trends in the IT industry job market two data mining descriptive functions were used. They are

mining of frequent patterns and Mining Associations [10-14]. To program the algorithm Python language was used.

Mining of frequent patterns was used to identify what are the most mentioned words in unstructured text data. This does not indicate only the frequency of how many times particular terms was appeared, but also consider its importance. In "Equation 1", W is a statistical measure which uses to refer to the weight that uses to evaluate the importance of a particular word in a document. "Equation 1" describes how weight is calculated in the Term Frequency-Inverse Document Frequency (TF-IDF) technique. The importance increase when a word appears in a document more frequently [15-17].

TF-IDF is a numerical statistic that shows the relevance of keywords to some specific documents [4]. Term Frequency (TF) measures how many times a term is present in a document.TF is the occurrence of any term in a document is divided by the total terms present in that document [4]. Inverse Document Frequency (IDF) assigns a lower weight to frequent words and assigns a greater weight for the words that are infrequent [4]. TF-IDF can be calculated by multiplying term frequency (TF) and inverse document frequency (IDF) [4].

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = Number of occurrences of i in j

df_i = Number of documents containing i

N = Total number of documents (1)

Association Rule Mining (ARM) is another text mining technique used to discover relationships among a large set of variables in a data set [18-21]. In this study, ARM was used to derive the associations with job titles. Mining associations help to identify what are the related job titles in the IT industry. It is used to identify frequent if-then associations. An association rule consists of two components. The antecedent (if) is the first component. Consequent (then) is the second component. The antecedent is the item found within the data. Consequent is the item found in combination with the antecedent [1] [22-24].

Apriori algorithm was used in this study, to mine the relationship between job titles, or what are the job titles that associate with each other. Apriori algorithm consists of three parts as support, confidence, and lift. Support is the default popularity of an item. It is calculated as given in "Equation 2"; for item B support

$$\text{Support (B)} = \frac{\text{Job titles contain (B)}}{\text{Total job titles}} \quad (2)$$

Confidence is the likelihood that job title B is also mentioned if job title A is mentioned. Confidence is calculated as given in "Equation 3"

$$\text{Confidence (A} \rightarrow \text{B)} = \frac{\text{Job titles contain both (A and B)}}{\text{Job titles contain (A)}} \quad (3)$$

Lift is the increment in the ratio of job title B with A .Lift is calculated as given in "Equation 4"

$$\text{Lift (A} \rightarrow \text{B)} = \frac{\text{(Confidence (A} \rightarrow \text{B))}}{\text{(Support (B))}} \quad (4)$$

IV. Results and Discussion

To identify the current dynamics in the job market, this study has proposed two approaches. TF-IDF algorithm was used to determine what the keywords in the data set were. The trained dataset consists of job titles which were extracted from five hundred job vacancies. The data set was collected using an online site which advertises IT industry job vacancies in Sri Lanka. Based on the trained dataset, Fig. 3, and Fig. 4, show the keywords extracted by processing the TF-IDF algorithm. The keywords shown in Fig. 3, are the keywords extracted in the technology attribute. The trending technologies that were mostly displayed in the advertisements are shown in Fig. 3. Whereas, Fig. 4, shows the keywords extracted from the job title attribute in the data set, which means Fig. 4, shows the most trending job title which extracted from the given dataset.

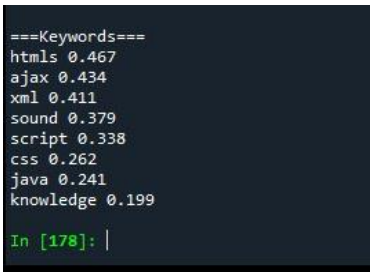

Fig. 3. The output of the TF-IDF technique – keywords in technologies

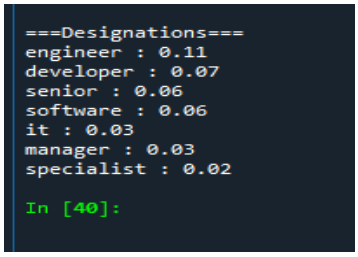

Fig. 4. The output of the TF-IDF technique – keywords in job title

To mine the associations between job titles, the Apriori algorithm was used. Fig. 5, shows antecedent, support, antecedent support, and consequent support values that exist between job titles. Support value used as an indicator to show how frequently the associated job titles appear together, in the dataset. An antecedent is an item that considers first, when creating an association rule. Consequent is the job title that found in combination with the antecedent. In Fig. 5, the first row represents the support value as 0.83, which means 83% occurrences are there which contain both "system administrator" and "business analyst" job titles out of the total dataset.

Fig. 6, shows the support, confidence, and lift values between job titles. This is useful to understand what the associated jobs in the IT field are. The support value of an association rule is defined as the percentage of records that contain both job titles, to the total number of records in the dataset. The lift value represents whether the association is positive or negative, also it shows the strength of the relationship which exists between the two job titles. If the lift value is greater than one, it indicates a positive relationship.

ISSN 2756-9160 / November 2020.
International Conference on Advances in Computing and Technology (ICACT–2020) Proceedings

If the lift value is less than one, it indicates a negative relationship, and if the lift value equals one, the job titles are independent and there exists no relationship between them [25]. Based on Fig. 6, it is clear that, the lift value that exists between all job titles shown here has a positive relationship. Confidence indicates the reliability of the derived association rule. Confidence shows the job titles which appear in an associate way in the used dataset.

antecedents	consequents	antecedent support	consequent support
(System administrator)	(Business Analyst)	0.083333	0.083333
(Business Analyst)	(System administrator)	0.083333	0.083333
(Product Manager)	(Financial Planning Analyst)	0.083333	0.083333
(Financial Planning Analyst)	(Product Manager)	0.083333	0.083333
(Lead Network Engineer)	(Senior Engineer)	0.083333	0.083333
(Senior Engineer)	(Lead Network Engineer)	0.083333	0.083333
(Senior technical specialist)	(Solutions support engineer)	0.083333	0.083333
(Solutions support engineer)	(Senior technical specialist)	0.083333	0.083333
(PHP software engineer)	(Web developer)	0.083333	0.125000
(Software engineer)	(Web developer)	0.083333	0.125000
(Web developer)	(PHP software engineer)	0.125000	0.083333
(Web developer)	(Software engineer)	0.125000	0.083333
(Senior software engineer)	(Web developer)	0.125000	0.125000
(Web developer)	(Senior software engineer)	0.125000	0.125000

Fig. 5. The output of the Apriori algorithm- association between job titles and their support value

antecedents	consequents	support	confidence	lift
(System administrator)	(Business Analyst)	0.083333	1.000000	12.000000
(Business Analyst)	(System administrator)	0.083333	1.000000	12.000000
(Product Manager)	(Financial Planning Analyst)	0.083333	1.000000	12.000000
(Financial Planning Analyst)	(Product Manager)	0.083333	1.000000	12.000000
(Lead Network Engineer)	(Senior Engineer)	0.083333	1.000000	12.000000
(Senior Engineer)	(Lead Network Engineer)	0.083333	1.000000	12.000000
(Senior technical specialist)	(Solutions support engineer)	0.083333	1.000000	12.000000
(Solutions support engineer)	(Senior technical specialist)	0.083333	1.000000	12.000000
(PHP software engineer)	(Web developer)	0.083333	1.000000	8.000000
(Software engineer)	(Web developer)	0.083333	1.000000	8.000000
(Web developer)	(PHP software engineer)	0.083333	0.666667	8.000000
(Web developer)	(Software engineer)	0.083333	0.666667	8.000000
(Senior software engineer)	(Web developer)	0.083333	0.666667	5.333333
(Web developer)	(Senior software engineer)	0.083333	0.666667	5.333333

Fig. 6. The output of the Apriori algorithm-assocition between job titles and their confidence

V. CONCLUSION

This study was conducted to mine the trends in the IT industry job market. The data were collected from online job vacancy repositories. First, texts were extracted from images and subjected to the preprocessing process. After that, the cleansed data were used to mine knowledge by using two text mining technologies, which are the TF-IDF technique and the Apriori algorithm. By using the TF-IDF techniques it was clear that what the key job titles were. With the help of Apriori algorithm the associations exist between job titles. This study was based to discover what were the key job titles in the IT industry and what are the associations between them, for further, this study can continue to mine what are the qualifications, experiences and technical knowledge associate with each job titles.

REFERENCES

[1] T Smith, D., & Ali, A. (2014). Analyzing computer programming job trend using web data mining. *Issues in Informing Science and Information Technology*, *11*, 203-214.

[2] Sibarani, E. M., Scerri, S., Morales, C., Auer, S., & Collarana, D. (2017). Ontology-guided Job Market Demand Analysis. *Proceedings of the 13th International Conference on Semantic Systems - Semantics2017*.

[3] Wowczko, I. (2015). Skills and Vacancy Analysis with Data Mining Techniques.*Informatics*,*2*(4),31–49.

[4] Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, *181*(1), 25-29.

[5] Nayak, A. S., & Kanive, A. P. (2016). Survey on Pre-Processing Techniques for Text Mining. *International Journal Of Engineering*.

[6] Kadhim, A. (2018). An evaluation of pre-processing techniques for text classification. *International Journal of Computer Science and Information Security*.

[7] Thilagavathi, K., & Shanmugapriya, V. (2014). A survey on text mining techniques. *International Journal of Study in Computer Applications and Robotics*, 41-49.

[8] Munková, D., Munk, M., & Vozár, M. (2013). Data Pre-processing Evaluation for Text Mining: Transaction/Sequence Model.

[9] Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2005). Overview of Text Mining. *Text Mining*, 1-13.

[10] Rajman, M., & Besançon, R. (1998). Text Mining: Natural Language techniques and Text Mining applications. *Data Mining and Reverse Engineering*, 50-64.

[11] Singh, S. (2018). Natural language processing information extraction.

[12] Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 76-82.

[13] Talib, R., Kashif, M., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, *7*(11).

[14] Natei, K. N., Viradiya, J., & Sasikumar, S. (2018). Extraction text from image document and displaying its related information. *K.N.Natei Jouranl of Engineering Study and Publications*, 27-33.

[15] Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. 2016 International Conference on Electrical, Electronics, and Optimization Techniques

[16] Yu, N. (2018). A Visualized Pattern Discovery Model for Text Mining Based on TF-IDF Weight Method. *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics*

[17] Christian, H., Agus, M. P., & Suhartono, D. (2016). Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, *7*(4), 285.

[18] Zhou, L., & Yau, S. (2010). Association Rule and Quantitative Association Rule Mining among Infrequent Items. *Rare Association Rule Mining and Knowledge Discovery*

[19] Association Rule Mining II. (n.d.). *Principles of Data Mining*.

[20] Sakurai, S. (2008). Rule Discovery from Textual Data. *Emerging Technologies of Text Mining*, 120-138.

[21] Malik, U. (n.d.). Association Rule Mining via Apriori Algorithm in Python.RetrievedOctober10,2020,fromhttps://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/

[22] 10 Apriori. (n.d.). Retrieved October 10, 2020, from https://docs.oracle.com/cd/E18283_01/datamine.112/e16808/algo_apriori.htm

[23] Shi, Y., & Zhou, Y. (2010). An Improved Apriori Algorithm. *2010 IEEE International Conference on Granular Computing*.

[24] Singh, S. K., & Kumar, P. (2016). I2Apriori: An improved apriori algorithm based on infrequent count. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*.

[25] Mahmood, S., Shahbaz, M., & Guergachi, A. (2014). Negative and Positive Association Rules Mining from Text Using Frequent and Infrequent Itemsets. *The Scientific World Journal*, *2014*, 1-11.