# Constructing Recommendation about Skills Combinations Frequently Sought in IT Industries Based on Apriori Algorithm

Latifah
*Information System Dept.*
*STMIK Jakarta STI&K*
Jakarta, Indonesia
latifahbahrudinsuryobroto@gmail.com

Tubagus Mohammad Akhriza
(*Corresponding Author*)
*Pradnya Paramita School of*
*Informatics Management & Computer*
Malang, Indonesia
akhriza@stimata.ac.id

Laras Dewi Adistia
*Information System Dept.*
*STMIK Jakarta STI&K*
Jakarta, Indonesia
LDAdistia@gmail.com

***Abstract*—To adapt the IT curriculum to the requirements of the IT industry skills, several methods have been proposed. Among them is the method of mining job advertisement data to find skills that are being sought by the industry. However, so far no significant research has focused on providing recommendations on skills that need to be taken along with other popular skills to fill the job vacancies offered. Traditional recommendation methods cannot be applied because information related to user or industry ratings on a skill is not available in advertisements. This article proposes an alternative solution to this need by developing recommendation techniques based on skill association rules, where the rules are mined using Apriori algorithm. The recommendation results were confirmed to curriculum managers in several universities, and had obtained quite good recall and precision, namely 70% and 76% respectively. The proposed recommendation system is also able to find skill combinations that are prominent in job advertisements.**

***Keywords*—*association rule, recommendation system, skillset***

## I. INTRODUCTION

The growth in the number of research and development in the field of information technology (IT) occurs rapidly and massively. Likewise, the implementation of this technology in the industry has led to the need for expertise in IT-based industries that also evolves rapidly. Some technologies such as Big data and wearable user interfaces require five to ten years to spread worldwide in the industry, while others such as location intelligence only need less than two years [1]. Even earlier, an article described that the evolution of the need for IT expertise in the industry will not stop, because new branches will be formed in this field and certainly require new experts as well [2]. The speed of this change cannot be counterbalanced by changes in IT curriculum in university, especially undergraduate programs; and this situation has created a gap between curriculum and industry needs [3].

To adjust the IT curriculum with the IT industry skills requirements, some methods have been proposed to find the skills needed in the industry. Among them are those who utilize data mining techniques to find the clusters of the most requested job titles in the industry along with the skills needed to fill these job positions [4]; Some other efforts are aimed at finding skills related to information systems using the content analysis approach [5–7]. Another method using frequent patterns mining is also proposed to find the skill combinations that are most often requested by the industry [3]. All of the methods mentioned are using IT job advertisements as a source of information about skills that are needed in the industry.

However, so far no significant research has focused on providing recommendations on new skills that need to be taken along with other popular skills to fill the job vacancies offered. Borrowing traditional recommendation system (RS), such as book [8,9] or film recommendations [10,11], to recommend an item (book or film) to a user, the system processes data about other users' ratings on the item, or what is also called collaborative filtering technique. If there is no rating data, then alternatively, the data attributes or types of content of items that are liked by a user in the past can be used to recommend new items to the respective user at present time, or also called the content-based filtering technique. The combination of these two approaches is called the hybrid method [12–16]. To the contrary, recommendations regarding skills that should be studied along with certain skills have their own characteristics. The three aforementioned RS approaches cannot be applied because there is no information in the job advertisements that supports the determination of skill recommendations, such as industry's or people's favorite rating data for a skill. This article provides answers to these needs through the development of skill recommendation techniques based on the model, namely skills association rule (AR) that are mined from job advertisement datasets using the Apriori algorithm.

The experiments produce quite interesting findings where some of the recommended skills may be skills that are still less popular, but pair with popular skills. Recall and precision are used to measure the performance of recommendations produced, compared to the recommendations of curriculum managers from universities affiliated with author. As a result, 76% precision was obtained which showed the level of approval of the managers, whereas 70% recall indicated the fact that the managers were not aware of the existence of new skills in job advertisements, so they did not recommend this skill.

## II. The Proposed Method

### A. *The Recommendation System Framework*

All processes needed to develop recommendations for skills are arranged in a framework as shown in Fig. 1. Process 1 is the data pre- processing stage, where job advertisements downloaded from online job search engines are prepared to be processed in Process 2. Data that is ready for processing is stored in the job skills dataset (JDS). Process 1 includes extracting only the names of the skills mentioned in the adverts that are usually represented by the name of the software, professional certification or terms in the IT field. These names need to be uniform because for example, people write MySQL and My SQL for the same purpose. More details of this naming process can be seen in the literature [3]. In Process 2, JDS is processed by the Apriori algorithm to produce skills association rules, which are then stored in the rule base (RB). Process 3 ranks rules in the RB based on rules dominance in the dataset, probability that rules appear in the dataset and dependency measurement between new skills recommended to be paired with popular skills.

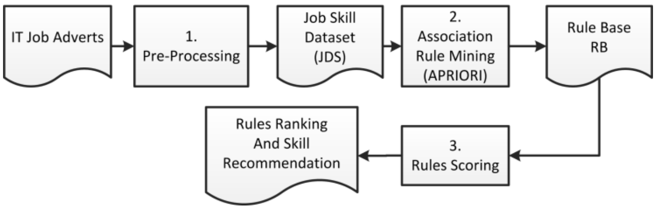

Fig. 1. Proposed Recommender System Framework

The skills association rule mining algorithm is implemented using the Python language and accepts JDS in the form of text files as input and also produces text-based RB. Each line in JDS contains a record number (001, 002, so on) and a set of unique skills that represents a job title that is opened in job advertisements (Fig. 2), while each line in RB is a rule that is generally in the following form (Fig. 3)

```
antecedent->consequent (support, confidence, lift)
```

Thus, extracted data from job advertisements is processed in such a way as to produce a text file in which each record is formatted such as shown in Fig. 2. Three interestingness measurement values in parentheses of each rule in Fig. 3 are explained in the next subsection.

```
001 mysql php ajax linux
002 word office windows
003 ceh linux security
004 mongodb java perl mysql
```

Fig. 2. Example of records in JDS

```
mysql -> php (0.1, 0.8, 1.3)
security -> ceh (0.05, 0.9, 1.4)
mongodb mysql -> java (0.05, 0.6, 1.7)
```

Fig. 3. Example of Rules in RB

### B. *Skills Association Rule Mining*

Mining AR was first proposed to find patterns of customer behaviour in a market when buying items [17–19]. Extending formal definition of AR, given a JDS D with N records and set I containing all skills (represented by skill code), namely all unique skill names in the D; X and Y are combinations of skills or skillset ($\in I$) so that $X, Y \;\; D$. The rule R is generally defined as a form of implication if X then Y, or R: X ® Y where X is called antecedent and Y is the consequent of R. Rules are arranged by associating the skillsets X and Y with a number of main interestingness measures, namely support, confidence, and lift. As explained in [17,18,20], support of X is explained in (1) which computes the number of records *t* in D containing X, relative to N. If it is not divided by N, then it is called absolute support. Confidence is the probability that if the skill X is required then so is skill Y, written as in (2). Lift shows a correlation between X and Y. If Lift > 1 then X and Y are interdependent and XY is a good rule to predict that X and Y will always appear together. If Lift = 1 then X and Y are mutually independent, and if Lift <1 then X and Y are negatively correlated, that is, the appearance of X does not promote the appearance of Y [20].

$$Sup(X) = \frac{|\{t \in D\}; X \; t|}{N} \qquad (1)$$

$$Conf(XY) \text{ or } Conf(XY) = \frac{Sup(XY)}{Sup(X)} \qquad (2)$$

$$Lift(XY) = \frac{Sup(XY)}{Sup(X)Sup(Y)} = \frac{Conf(XY)}{Sup(Y)} \qquad (3)$$

In this work, the activity of exploiting skills AR from D is divided into two sub-activities: 1) Find all skillsets which most often appear, or called as frequent skillset, that is all skillsets X in D where Sup (X) ≥ minsup and 2) find all the rules R: X⇒Y in D, so Sup(XY) ≥ minsup, Conf(XY) ≥ minconf, and Lift(XY) > 1 where X∩Y = ∅.and |Y| = 1

Both *minsup* and *minconf* are the minimum support and minimum confidence thresholds specified by the data analyst when using the AR mining algorithm, i.e. the Apriori algorithm. The fundamental approach of Apriori algorithm is to find all possible combinations of skills whose support satisfies *minsup*. First, this algorithm looks for all 1-frequent skillset (FS) (skillset that contains one skill), then combines 1-FS to get 2-FS, and so on until all *k*-FSs are formed, and then look for all possible rules X®Y that can be formed from all FS found [19]. But in this study, part Y is set to consist of only 1-FS, or |Y| = 1, and Y is defined as the recommended skill that can be studied with X in order to fill a certain job position. In the experiment, different *minsup* values are used to see the number of rules formed.

### C. *Ranking the Rules*

In this work, a formula to rank the rule R generated is proposed such as shown in (4)

$$Score(R) = Sup(R) * Conf(R) * Lift(R) * Length(R) \qquad (4)$$

If only Sup(R) is applied to rank the rules then the number of rules that should be considered is very large and may obscure the existence of an interesting rule. Confidence and lift are used to help filter rules that have more value i.e. skillsets that have the probability to emerge and the skillsets that are most positively correlated. Therefore, these two measures are added to the rule score. Length(R) or the number of skills in R is included in the score to put long skillsets at the top rankings. Long skillset is useful for generating skill combinations consisting of popular and unpopular skills. With this method, unpopular skills will emerge [3]. It is important to note that the XY combination will not be frequent if either X or Y is not a frequent skillset. (i.e. Sup(X) or Sup(Y) < *minsup*); This is called the monotonic nature of frequent itemset [17]. One other consequence of this monotonic property is that shorter rules usually have greater support; and vice versa. That is, in the experiment *minsup* is not set too large so that popular and unpopular skill combinations can be found.

TABLE I. SOME TOP RANK RULES

| No | X | Y | Supp(X) | Supp(Y) | Conf | Lift | Score |
|---|---|---|---|---|---|---|---|
| 1 | crystal, naturallyspeaking, | *dragon* | 0.001 | 0.002 | 1 | 525 | 2.211 |
| 2 | chromebox, mpls, ceh | *nat* | 0.001 | 0.004 | 1 | 285 | 1.257 |
| 3 | mongodb, digital | *anarchy* | 0.021 | 0.034 | 1 | 29 | 1.225 |
| 4 | jquery, mobile, avantlink, joyent | *postgresql* | 0.001 | 0.005 | 1 | 191 | 1.202 |
| 5 | allegorithmic, citysearch, avantlink, mongodb | *air* | 0.003 | 0.001 | 0.5 | 363 | 0.965 |

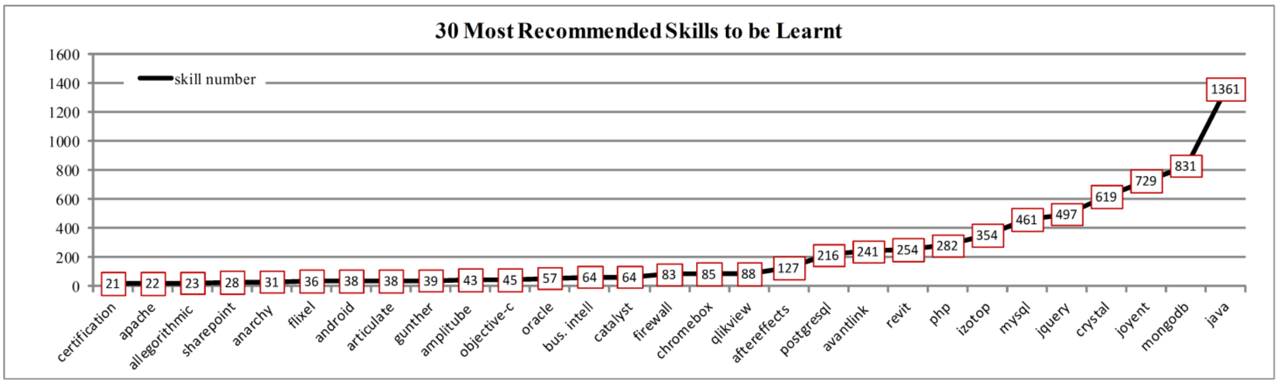

Fig. 4. Top 30 most recommended skills to be learnt

### D. *Recommendation Evaluation*

The proposed recommendation system is tested using JDS taken from IT related job advertisements for one year (2017–2018) from leading international job search engines namely Monster.com and Jobsdb. Jobs locations were in Indonesia, Malaysia, Singapore, Hong Kong, the UK, the US, and Australia. JDS contains 19,950 records and 852 skills. Evaluation of RS performance is done by calculating the recall and precision of recommendations produced by the proposed method against recommendations proposed by experts. The curriculum managers from the affiliate campus of the author are involved as expert. Recall and precision are calculated by (5):

$$Recall = \frac{TP}{TP+FN} \; ; \; Precision = \frac{TP}{TP+FP} \qquad (5)$$

Here, true positive (TP) is a true occurrence that one skillet (consequent) recommended to teach along with other skillset (antecedent) by RS is also recommended at least by one expert, false positive (FP) is a recommendation from RS not recommended by experts, while false negative (FN) happened when expert recommendations are not recommended by RS [12].

## III. RESULTS AND DISCUSSION

Number of rules produced using *minsup* = 0.01, 0.005, 0.001 and *minconf* = 0.5 respectively is 110, 343 and 10,634 rules. As shown, the smaller the *minsup*, the greater the number of frequent skillsets produced and so does the number of rules and this is the effect of the monotonous nature of frequent patterns explained earlier. In the following discussion regarding the recommended skills, rules are generated using the smallest *minsup* value i.e. 0.001 over 19,950 records or equal with 19.95 in absolute support. This means that there are at least 19 job openings that should contain a skillset, so then the skillset is said to be frequent. Analyst can change *minsup* value accordingly, if desired.

The results of ranking rules place several long rules in the top ranking position (Table I), and all skills in column Y are those recommended to be studied together with corresponding skillset in column X, but only a few skills that are considered interesting in discussion are included in column X. As also seen, some shorter rules also go into the top ranking because of the large value of support and Lift.

Scores are calculated using (4), but support of rules is not listed in the table, instead support for each X and Y is given. Nonetheless, according to the monotonous nature of frequent itemset it can be known, for example in the first row of Table I, support X = 0.001 and support Y = 0.002, thus support (XY) cannot possibly be greater than either of these two values. All rules have 100% confidence, except the last rule is only 50% but support X is higher than Y. This last phenomenon is displayed in order to show that a long skillset may have higher support than shorter ones. Here it means that the $\{air\}$ skill is not yet very popular, but is sticking to, or can be said to be "requested", along with a series of skills i.e. {allegorithmic, citysearch, avantlink, mongodb} that are more popular today.

On the other side, top 30 most recommended skills to be learnt based on the rules' consequent are given in Fig. 4. One skill that is quite dominant in job advertisements collected is MongoDB which has 59% support or means that around 11,170 job advertisements require this skill. This NoSQL-based database system began to emerge its existence along with the entry of the Big data era to the industry. Direct inspection into the rule base found that in some rules, some skills are seen paired with MongoDB as consequent, for example, with skills that are also well known such as PHP, CodeIgniter, Catalyst and MySQL with confidence of around 60% - 80%. This fact is an indication that these skill's combinations are sought after in the industry and can be used as a reference for the preparation of the college curriculum. Additionally, some rules have consequent parts with smaller support than antecedent while the lift value is quite high (Table I). This can be interpreted that the skill $\{air\}$ although is still unpopular, is recommended to be studied with a skillset $\{allegorithmic, citysearch, avantlink, mongodb\}$. The last position is occupied by certification which includes CEH (Certified Ethical Hacker), CISSP (Certified Information Systems Security Professional) and MOS (Microsoft Office Specialist), an indication that professional certification is also sought by the industry so that their curriculum can be considered for adoption into the university curriculum.

Here there are some skills that might look unfamiliar, such as a $\{dragon\}$ that appears together with {*crystal*, *naturallyspeaking*}. Searching on the Internet found that *Dragon Naturallyspeaking* is a speech recognition software package developed by Dragon Systems of Newton and there are as many about 40 job vacancies that need skills in this software. Similarly, $\{air\}$ is required with {*allegorithmic*, *citysearch*, *avantlink*, *mongodb*} and this rule can be found in digital designer jobs. Substance Air and Allegorithmic are products of Adobe, a known leading multimedia software company. Developer Anarchy can be found in software development job. NAT occurs with {*Chromebox*, MPLS and CEH} and as known, all these individual skill names are closely related to security of network.

The results of the evaluation of recommendations performance are given here. Two members of the curriculum management team from the affiliated campus of the authors are involved as experts to give recommendations regarding skills (consequent) that can be studied together with a skillset (antecedent) of 50 rules chosen. Expert recommended skills are compared with the skills recommended by the rules produced by the proposed recommendation system. The calculation results get TP = 38, TN = 12, and FN = 16, thus $Recall = \frac{38}{38+16} = 0.704$ and $Precision = \frac{38}{38+12} = 0.76$.

The high recall result is obtained from the small number of skills recommended by experts to be paired with the given skillset. Based on their evidence, many skills they did not understand beforehand, although when giving recommendations they were given time to find information on the Internet related to the skillset given. On the other hand, a high value of precision is obtained from the number of skills (consequent) in 50 rules that are in accordance with the recommendations of experts, i.e. 38 skills thus resulting in precision = 76%. These results illustrate that the rule produced can help universities management in curriculum renewal activities by considering adding new emerging skills in the industries.

The results of recall and precision depend subjectively on the experience and knowledge of the experts on the development of IT that is happening right now, so the results may differ from other experts. However, this can happen given that the need for IT skills in the industry continues to evolve in the future.

## IV. Conclusion

In addition to having good recall and precision, the proposed recommendation method can help curriculum management to determine new skills that they did not know before, but are the most sought after in the industry today.

## References

[1] I. Bojanova, "The digital revolution: What's on the horizon", IT Prof., vol. 16, pp. 8-12, 2014.

[2] J. Liu, "Computing as an evolving discipline: 10 observations", IEEE Computer (Long. Beach. Calif), 2007

[3] T.M. Akhriza, Y. Ma, and J. Li, "Revealing the Gap Between Skills of Students and the Evolving Skills Required by the Industry of Information and Communication Technology", Int. J. Softw. Eng. Knowl. Eng., vol. 27, pp. 675–98, 2017.

[4] C. Litecky, A. Aken, A. Ahmad, and H.J. Nelson, "Mining for Computing Jobs", IEEE Softw., vol. 27, pp. 78–85, 2010.

[5] I. Wowczko, "Skills and Vacancy Analysis with Data Mining Techniques," Informatics, vol. 2, pp. 31–49, 2015.

[6] M. A. Kennan, P. Willard, D. Cecez-Kecmanovic and C.S. Wilson C, "A Content Analysis of Australian IS Early Career Job Advertisements," Australas. J. Inf. Syst., vol. 15, pp. 169–90, 2009.

[7] H. E. Longenecker, D. Feinstein and J.D. Clark, "Information Systems Curricula: A Fifty-Year Journey", Information Systems Educators Conf., vol 29, pp. 1-26, 2012.

[8] A. Simović, "A Big Data smart library recommender system for an educational institution," Libr. Hi Tech, vol. 36, pp. 498–523, 2018.

[9] W. Ji, S. Liu, Y. Song and J. Qi, "Research of Intelligent recommendation system based on the user and association rules mining for books", 2016, pp. 294–9.

[10] P. Symeonidis, A. Nanopoulos and Y. Manolopoulos, "MoviExplain: A Recommender System with Explanations", Proceedings of the 3rd ACM conference on Recommender systems, 2009.

[11] R. Katarya and O. P. Verma, "An effective collaborative movie recommender system with cuckoo search", Egypt. Informatics J., 2017.

[12] J. Bobadilla, F. Ortega, A. Hernando and A. Gutiérrez A, "Recommender systems survey", Knowledge-Based Syst., vol. 46, pp. 109–32, 2013.

[13] H. Alharthi, D. Inkpen and S. Szpakowicz, “A survey of book recommender systems,” J. Intell. Inf. Syst., vol. 51, pp. 139–60, 2017.

[14] R. Burke. “Hybrid Web Recommender Systems”, The Adaptive Web, Berlin: Springer, 2007, pp 377–408.

[15] J. Lu, D. Wu, M. Mao, W. Wang and G. Zhang, “Recommender system application developments: A survey,” Decis. Support Syst., vol. 74, pp. 12–32, 2015.

[16] S. Sivapalan, A. Sadeghian, H. Rahnama and A.M. Madni, “Recommender systems in e-commerce”, World Automation Congress Proceedings, 2014.

[17] S.K. Solanki and J.T Patel, “A survey on association rule mining”, International Conference on Advanced Computing and Communication Technologies, 2015.

[18] R. Agrawal, T. Imieliński T and A. Swami, “Mining association rules between sets of items in large databases”, ACM SIGMOD Rec., 1993.

[19] R. Agrawal and R. Srikant R, “Fast Algorithms for Mining Association Rules”, the 20th Int. Conf. Very Large Data Bases, 1994.

[20] C. Ju, F. Bao, C. Xu and X. Fu, “A Novel Method of Interestingness Measures for Association Rules Mining Based on Profit Discret”. Dyn. Nat. Soc., pp. 1-10, 2015.