ICIC Express Letters
Volume **15**, Number **5**, May **2021**

ICIC International ©2021 ISSN 1881-803X
pp. 439–447

THE APPLICATION OF WEB-BASED MEASUREMENT OF STUDY PROGRAMS AND OCCUPATIONAL PROFESSIONS

Lely Prananingrum[1,*], Adang Suhendra[1], Imam Suryansyah[2]
Lily Wulandari[1], Lintang Yuniar Banowosari[1]
and Muhammad Haviansyah[2]

[1]Doctoral Program in Information Technology
[2]Information Systems Department, School of Information Systems
Gunadarma University
Jl. Margonda Raya, No. 100, Kota Depok 16424, Indonesia
{ asuhendra; suryacemerlang7; lilywulan; mpo.lintang; haviansyah09 }@gmail.com
*Corresponding author: lelyprana@gmail.com

Received September 2020; accepted December 2020

Abstract. *This present time is still often encountered many people who have a college degree, but still do not have a job. This happens because of several reasons, and one of them is the lack of a level of relevance between the courses taught in the lectures and the competencies needed to get a job in the working world. Thus, this study develops a web-based application that is capable of measuring the suitability level of a study program based on the courses taught at the academy or university with the competencies needed to get a profession. To get a professional mapping based on competency, the National Occupation Map is used as an official document that contains a mapping of jobs with the competencies needed to achieve that position or profession. Furthermore, to get details regarding what aspects are used in the Indonesian National Job Competence, so that Standard documents containing aspects of work attitudes, skills, and knowledge competence must be adapted to the aspects of attitudes, skills and competence knowledge to be adapted to the aspects of knowledge, attitudes and skills to be achieved in courses taken from the Semester Learning Plan document. The method that is used to check the level of compatibility between the majors and occupation is the LDA (Latent Dirichlet Allocation) and cosine similarity methods. This study produces a website-based application which is useful for calculating the relevance of the study program and occupation maps.*

Keywords: LDA (Latent Dirichlet Allocation), Cosine similarity, Occupation map, Study program

1. **Introduction.** Education is an important part of national development, because education, especially for young people, will bring the life of the nation to a more advanced level with the development of the level of knowledge in society. One of the targets of education is to create graduates who are able to work in working world and business. However, in fact, according to data from Badan Pusat Statistik (BPS) in February 2019 the number of unemployment according to the highest education obtained at university was 83.3% people and the number is increasing from 78.8% in the period of February 2018 [1].

The post-education unemployment problem is most likely due to one of the obstacles, namely the relevance of the education curriculum to the job market. Therefore, the government of the Republic of Indonesia has worked to solve the problem of the relevance of these competencies by compiling a national occupation map and also the Indonesian National Work Competency Standards. This occupational map is a facility that can be used in the world of education to carry out self-examination whether the curriculum taught

DOI: 10.24507/icicel.15.05.439

in each educational agency or institution has met the professional competency targets in accordance with their fields. To make a match between the curriculum and competencies, the relevance method is used which will measure the extent of the relevance of the curriculum to competencies. As the types of competency documents on occupational maps and also curricula in educational institutions are in the form of text documents, the relevance method used is a method that can make inter-text relevance. One of the methods that can be used to measure the relevance between texts is *Latent Dirichlet Allocation* (LDA) + *Cosine Similarity*, where LDA is a machine learning method, which is useful for modeling topics from a collection of documents into a vector, and cosine similarity is an algorithm that measures the compatibility between document vectors containing topics, assuming that the documents have relevance based on the same topics in the document.

Recently, several studies to calculate the relevance of study programs to professions using a machine learning approach usually only focus on using the method by matching words in each document [2], so it has a weakness when there are words that have the same meaning but are considered different and vice versa. In this study, the matching approach used the topic so that inaccuracies caused by differences in meaning between words can be reduced.

The finding of this study will be of broad use to the community, especially college graduates, to be able to find out the level of compatibility of the competencies they have learned at academy or university with the professional needs they may be involved in. Furthermore, the results of this research can also be useful for higher education institutions in order to evaluate the competencies taught at the institutions in order to suit the needs of the world of work.

This paper is structured as follows: Section 2 presents related work, Section 3 presents the study about latent Dirichlet allocation, Section 4 presents the study about cosine similarity, Section 5 presents the proposed methodology for measurement of study programs and occupational professions, Section 6 presents assessment and result of the proposed method and finally Section 7 presents conclusion.

2. **Related Research.** Celikyilmaz et al. conducted research using the LDA method to characterize the similarities between the candidate's questions and answers by giving a ranking score [3]. This study found that extracting hidden concepts (called topics in LDA) improved the results of the question answer classification model. Furthermore, Suyanwar developed a semantic web application using the TF-IDF (Term Frequency – Inverse Document Frequency) method and the Jaccard coefficient to measure the relevance of courses to occupation and with results depending on the document content of each body of knowledge, so there are several courses that are wrongly related to the body of knowledge, and the result is that it affects the percentage of relevance of the course to the chosen occupation [2].

3. **Latent Dirichlet Allocation.** Latent Dirichlet Allocation (LDA) is a generative probabilistic model for discrete data collections such as corpus text. LDA is a three-level hierarchical Bayesian models, where each item collection is modeled as finite mixture over the underlying set of topics. Each topic is in turn modeled as an infinite mixture of probabilities of the underlying topics. In the context of text modeling, topic probability provides an explicit representation of a document. The relation between the document and the topic is based on the Dirichlet distribution and the relationship between the topic and the word is based on the polynomial distribution. The generative process of LDA can be seen in Figure 1.

Figure 1 is what known as a plate diagram of an LDA model, where

1) α is the per-document topic distributions,

2) β is the per-topic word distribution,

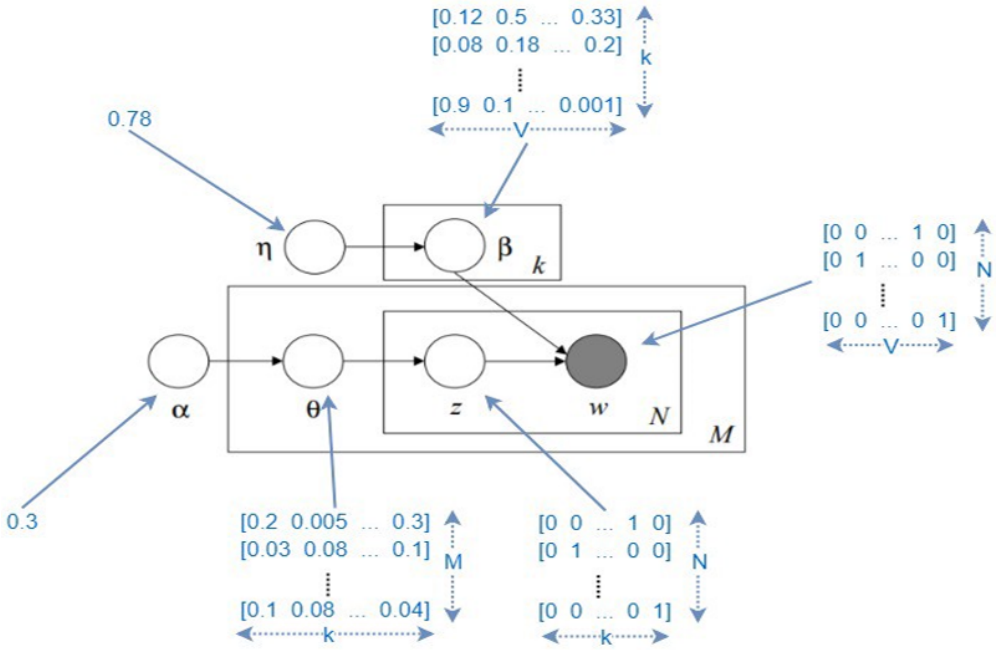


Figure 1. Graphical representation of the LDA model

3) θ is the topic distribution for document m,
4) φ is the word distribution for topic k,
5) z is the topic for the n-th word in document m, and
6) w is the specific word

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \tag{1}$$

Based on Equation (1), the generative process in modeling topics using LDA is as follows:

1) For hidden topic i, calculate f polynomial distribution of the words constructing the topic based on the Dirichlet distribution.
2) Get the number of N words in the document according to the Poisson distribution.
3) Calculate the j topic probability distribution for each text.
4) For each constructor word of each document, each document does
   a) Randomly select hidden topic z from those topics with a probability distribution of j.
   b) Select the word constructor randomly from the polynomial of the topic distribution.

4. **Cosine Similarity.**

In document-query cases, a document can be represented as a term vector that the vector's dimensions refer to the terms available in the document [5]. Cosine similarity is defined by the equation of the cosine angle between two vectors equal to a multiplication of dot product of two vectors, divided by the multiplication between vector length. In this algorithm the value that needs to be sought is the cosine angle between two vectors with the approach that two vectors are said to be similar if the two vectors are parallel or have a cosine value of 0 and two vectors are said to be different if the two vectors are perpendicular (orthogonal) or have a cosine value of 1. The cosine similarity formula can be seen in Equations (2)-(4).

$$similarity = \cos(\theta) \tag{2}$$

$$similarity = \frac{A \cdot B}{|A||B|} \tag{3}$$

$$similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i}\sqrt{\sum_{i=1}^{n} B_i}} \tag{4}$$

It can be seen that A and B are vectors of the results of modeling the topic of a document with LDA, which contains a collection of topics and the level of confidence a document contains the topic.

Vectors A and B are multiplied by the dot product which is then divided by the multiplication between the lengths of vectors A and B which results in the cosine values of vectors A and B.

5. **Methodology.** The methodology used in this study is as follows:

1) Data requirement analysis
   At this stage, an analysis of data requirements is carried out and also the data collection required.
2) Data pre-processing
   The data that has been collected is pre-processed so that it can be used in the next stage.
3) Model training LDA
   Machine learning model training is conducted by entering the data set into the model.
4) Topic prediction
   Topic predictions are carried out using LDA from the selected documents to produce a topic probability vector of a document.
5) Relevance calculation
   At this stage, the cosine similarity method is carried out to calculate the relevance between documents by calculating the similarity of two topic probability vectors to a document generated from the LDA.

5.1. **Requirement analysis.** Data collection is carried out by collecting data from the National Agency for Professional Certification in the form of an Occupational Map document which contains professions in the information technology field, along with the competencies needed to achieve that profession, then the Indonesian National Work Competency Standards document from the Ministry of Manpower which contains competency units which contain aspects of ability, knowledge and work attitude, and the last is the Semester Learning Plan document which is a learning planning document that is prepared as a guide for students in carrying out lecture activities for one semester to achieve predetermined learning outcomes. From the learning outcomes in the Semester Learning Plan, each subject is broken down into three aspects, that is ability, knowledge, and work attitude to match the format on competencies in Indonesian National Work Competency Standards.

5.2. **System design.** System design is designing a good system, which contains the operational steps in the data processing process and procedures to support system operation. The general description of the system is in Figure 2.

The system is divided into several parts, that is

1) Frontend: Frontend is a sub system that will communicate directly with the user as a graphical interface that will receive direct input from the user and display the output directly to the user.
2) Web Services: The Web Services sub system in this study will be made with an object-oriented programming scheme; therefore, a class diagram design is needed to define the class structure of the system to be built.
3) ML Microsystem: ML Microsystem is a subsystem that will run the machine learning method (LDA and cosine similarity). As a micro-system, this sub system will run if

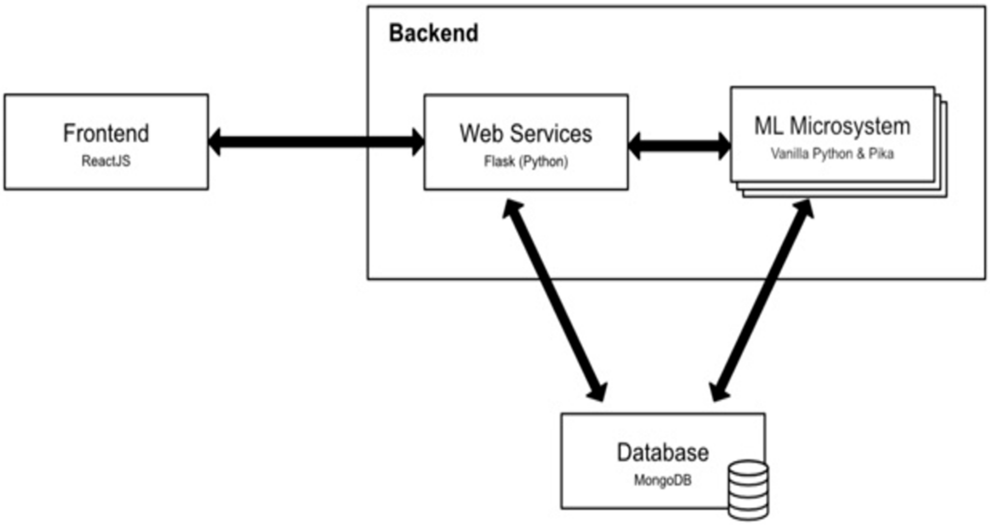


Figure 2. System overview

there is a trigger in the form of an order to run the process. This section will receive messages that contain orders and data to be processed in this section. The orders that can be received by this sub-system include “train” and “testing”.

4) Database: In making the application in this study, a useful database is needed to store useful data in the application. As for designing the database in the form of a structure from the collection in the database. There are 4 collections in the database that will be used in this application, namely “Majors”, “Occupations”, “Courses”, and “Competencies”.

5.3. **Model training.** Model training must coincide with data messages containing document IDs that have just been manipulated and will run data training sub-routines, which includes data collection, preprocessing, and training, and the stages are as follows:

- Data collection. The document text data that has just been manipulated will be retrieved from the database by performing data searches based on the document ID sent together with the “train” command.
- Preprocessing. After the data is taken from the database, the text preprocessing stage is carried out which includes 3 stages, namely
  a) Tokenization. Tokenization is the process of dividing text which can be a sentence, paragraph or document into tokens.
  b) Filtering. Filtering is the process of filtering the tokenization results of words that are not needed (stop words) for the weighting process, such as conjunctions or pronouns of people or subjects, so that the words that are stored are only words that have meaning.
  c) Stemming. Stemming serves to change words that have affixes into basic words. In this process, the Literature library is used, especially StemmerFactory() to remove additives from the tokenization and filtering results.

5.4. **Topic prediction.** The command “test” must coincide with a data message containing the ID of the document that you want to predict. The testing process is taking all tokens from the selected document from the database, after that the tokens are converted into bag of words based on a dictionary that has been stored in a file at the training stage, and then with a model that has been previously trained the data is tested to obtain predictions on the set of topics that are likely to be in the document and also the probability that these topics will be in the document. An example of the testing process using data from two competency documents of knowledge aspects, Analyzing Tools and Using Computer Devices, is as follows:

The input data:

- Knowledge aspect token of Analyzing Tools competency: [‘know’, ‘read’, ‘understand’, ‘notation’, ‘collect’, ‘need’, ‘device’, ‘soft’, ‘know’, ‘tools’, ‘benefit’, ‘collect’, ‘record’, ‘need’, ‘device’, ‘soft’].
- Knowledge aspect token of Using Computer Devices competency: [‘device’, ‘computer’, ‘point’, ‘use’, ‘computer’, ‘face’, ‘use’, ‘peripherals’, ‘program’, ‘application’].

The data after it is converted to bag of words:

- Bag of words of knowledge aspect document of Analyzing Tools competency: [$(1, 1)$, $(2, 2)$, $(5, 2)$, $(6, 1)$, $(7, 1)$, $(9, 1)$, $(10, 1)$, $(11, 1)$, $(14, 1)$, $(15, 2)$, $(16, 1)$].
- Bag of words of knowledge aspect document of Using Computer Devices competency: [$(0, 1)$, $(3, 2)$, $(4, 2)$, $(6, 1)$, $(8, 1)$, $(11, 2)$, $(12, 1)$, $(13, 1)$, $(17, 1)$].

The output results:

- Topic prediction from the knowledge aspect document of Analyzing Tools competency: [$(0, 0.01698199)$, $(1, 0.94932586)$, $(2, 0.016957296)$, $(3, 0.016734822)$].
- Topic prediction from the knowledge aspect document of Using Computer Devices competency: [$(0, 0.9419972)$, $(1, 0.019397814)$, $(2, 0.0193117)$, $(3, 0.0192933)$].

5.5. **Calculating relevances.** After obtaining the topic mapping vector in each document, to determine the similarity level of the two documents, the similarity calculation is carried out with cosine similarity. The cosine similarity calculation for competency document vectors Analyzing Tools as A and Using Computer Devices as B as in the previous example is illustrated in (5)-(9).

$$A = (0.0169, 0.9493, 0.0169, 0.0167) \tag{5}$$

$$B = (0.9149, 0.0193, 0.0193, 0.0192) \tag{6}$$

$$sim(A, B) = \frac{A \cdot B}{|A||B|} \tag{7}$$

$$sim(A, B) = \frac{0.0349}{0.9497 + 0.9424} \tag{8}$$

$$sim(A, B) = 0.01844 \tag{9}$$

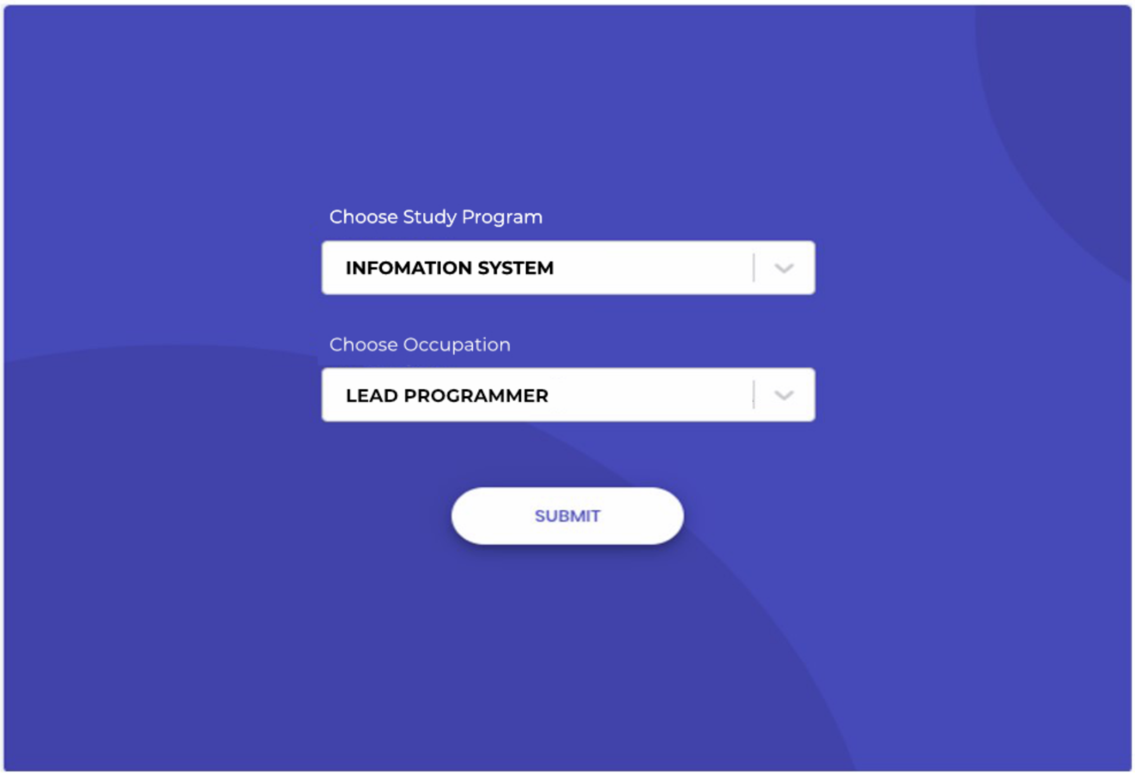
6. **Assessment and Result.** The result of this study is a web application which is useful for checking the relevance of study programs and occupations. Therefore, website application pages in this study are

- User input display

  In this user input display, there are two input forms in the form of options for filling in the study program and occupation for which relevance matching will be carried out. The user input display can be seen in Figure 3.
- Result display

  This page contains the percentage of matches selected and also a list of competencies in occupations with the status met or not in Figure 4.

Accuracy testing will measure how accurate the LDA and cosine similarity methods are in measuring the suitability between the study program and the selected occupation. At this stage the suitability of the Information Systems and Occupational Lead Program study program is tested using a system that has implemented the LDA and cosine similarity methods compared to the relevance of the competencies in the chosen occupation using the human brain. The result can be seen in Table 1.



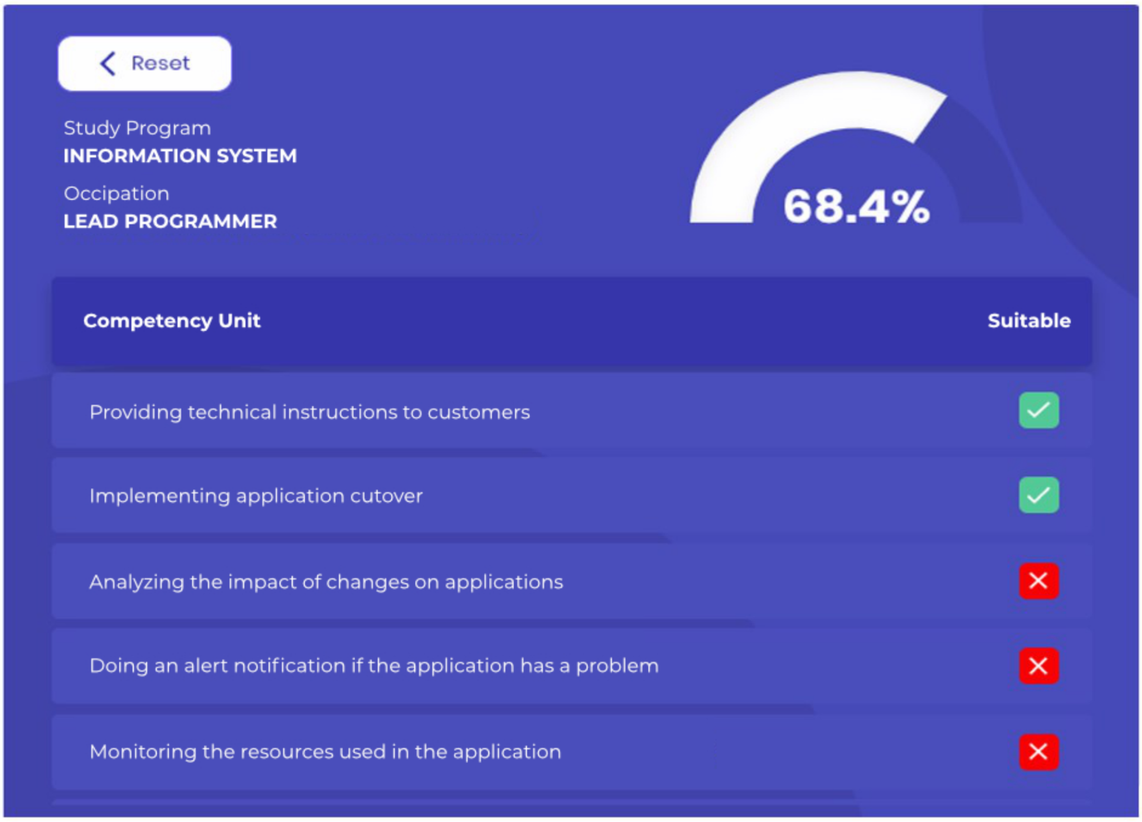
Choose Study Program

INFOMATION SYSTEM

Choose Occupation

LEAD PROGRAMMER

Figure 3. User input display



Study Program

INFORMATION SYSTEM

Occipation

LEAD PROGRAMMER

Competency Unit	Suitable
Providing technical instructions to customers	✓
Implementing application cutover	✓
Analyzing the impact of changes on applications	✗
Doing an alert notification if the application has a problem	✗
Monitoring the resources used in the application	✗

Figure 4. Result display

Table 1. Accuracy testing result

No	Competencies	System	Manual
1	Analyzing the scalability of the software	Suitable	Suitable
2	Designing user experience	Suitable	Suitable
3	Designing application architecture	Suitable	Suitable
4	Implementing network programming	Not suitable	Suitable
5	Implementing real-time programming	Suitable	Suitable
6	Implementing parallel programming	Suitable	Suitable
7	Implementing multimedia programming	Suitable	Suitable
8	Doing a code review	Not suitable	Not suitable
9	Doing static program code testing	Suitable	Not suitable
10	Performing user acceptance test	Suitable	Not suitable
11	Providing technical instructions to customers	Not suitable	Not suitable
12	Implementing application cutover	Suitable	Suitable
13	Analyzing the impact of changes on applications	Suitable	Suitable
14	Doing an alert notification if the application has a problem	Suitable	Suitable
15	Monitoring the resources used in the application	Suitable	Suitable
16	Performing software updates	Suitable	Suitable
17	Managing information security risks	Not suitable	Suitable
18	Specifying hardware architecture	Suitable	Suitable
Number of competencies		18	18
Number of suitable results		15	15
Calculation		$15/18 * 100\%$	$14/18 * 100\%$
Percentage		83.3%	77.8%

From the test results in Table 1, the confusion matrix is used to calculate the accuracy of this application. The confusion matrix is done by creating a table as in Table 2.

$$Accuracy = \frac{TP + TN}{total} \tag{10}$$

$$Accuracy = \frac{11 + 1}{18} \tag{11}$$

$$Accuracy = 0.67 \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Precision = \frac{11}{11 + 3} \tag{14}$$

$$Precision = 0.79 \tag{15}$$

Table 2. Confusion matrix

$n = 18$	System: Not Match	System: Match	
Manual: Not Match	$TN = 1$	$FP = 3$	4
Manual: Match	$FN = 3$	$TP = 11$	14
	4	14	

7. **Conclusions.** From the accuracy test, the percentage of compatibility between the Information Systems study program and the occupation of the Lead Programmer using an application that implements the LDA and cosine similarity methods is 83.3% and if it is done manually matching with the human brain, the percentage is 77.8%. This means that not all of the competencies required for the job of a Lead Programmer are relevant to the courses taught in the Information Systems study program. After that the level of accuracy obtained from the confusion matrix calculation is 0.67 and the precision level from this implementation is 0.79. Imperfect application accuracy and precision levels could be due to incorrect selection of the number of topics, alpha parameters, and beta parameters which can affect the accuracy and precision of the model. The LDA (Latent Dirichlet Allocation) method also depends on how many texts and building topics of a document, so that the selection of the number of topics becomes an obstacle that hinders research.

REFERENCES

[1] B. P. Statistics, *Open Unemployment According to the Highest Education Completed 1986-2019*, http://www.bps.go.id/, Accessed on 06 June 2020.

[2] Suyanwar, *Implementation of the TF-IDF Algorithm and the Jaccard Coefficient to Measure Eye Relevance Lecture with Semantic Web-Based Occupation*, https://library.gunadarma.ac.id/repository/implementasi-algoritma-tfidf-dan-koefisien-jaccardguna-mengukur-relevansi-mata-kuliah-dengan-okupasi-berbasis-web-semantik-jurnal, 2018.

[3] A. Celikyilmaz, D. Hakkani-Tur and G. Tur, LDA based similarity modeling for question answering, *Proc. of the NAACL HLT 2010 Workshop on Semantic Search*, 2010.

[4] B. M. Achmad, *Occpational Map, SKKNI and KKNI in the Field of Information and Communication Technology (Peta Okupasi, Skkni Dan Kkni Bidang Teknologi Informasi Dan Komunikasi)*, Aptikom Riau One Day Seminar, SATURDAY, Pekanbaru, DOI: 10.13140/RG.2.2.21154.27842, 2018.

[5] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, vol.24, no.5, pp.513-523, 1988.

[6] A. Nugroho, *Software Engineering Using UML and Java (Rekayasa Perangkat Lunak Menggunakan UML Dan Java)*, Andi, 2009.

[7] D. M. Blei, Probabilistic topic models, *Proc. of the 17th ACM SIGKDD International Conference Tutorials*, vol.55, no.4, pp.77-84, 2012.

[8] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol.3, no.1, pp.993-1022, 2003.

[9] S. K. Ketenagakerjaan, *About SKKNI*, https://skkni.kemnaker.go.id/tentang-skkni/, Accessed on 15 June 2020.

[10] S. Sanisah, Higher education and open unemployment: A dilemma. educational lantern (Pendidikan tinggi dan pengangguran terbuka: Sebuah dilema. Lentera Pendidikan), *Journal Education and Teacher Training*, vol.13, no.2, pp.147-159, 2010.