Developing alumni job prediction models based on term occurrences and word class analysis of search engine page results

Hapnes Toba, Evelyn A. Wijaya, Maresha C. Wijanto & Oscar Karnalim

Maranatha Christian University
Bandung, Indonesia

ABSTRACT: In this study, the authors propose several prediction models for alumni, which are based on term occurrences, word class analysis, such as nouns and verbs, and the clusters aggregation method. During the data collection process, a clustering-based method to disambiguate alumni names from the search engine page results was implemented. Further, several job prediction classifiers based on word occurrences, verbs-nouns word classes were developed and a cluster-based aggregation method. The classifiers were evaluated by using a real alumni tracer study, and tested in a cross-validation (5-fold) and hold-out combinations. The experimental results revealed that in the engineering and economics fields, there are more variations of alumni professions. Based on the experimental data, the approach seems to be useful for finding people in general, especially, for *ordinary* alumni names. As future work, it would be interesting to explore various social media, beyond LinkedIn. Moreover, friendship relations in social media, such as Facebook and Twitter, might also be beneficial for further alumni relational analysis.

INTRODUCTION

One of the objectives of alumni tracer study is to find the most recent job of alumni in their professional world after graduation. In the authors' experience, it is difficult to do this by means of surveys, in either manual or computer-based systems. The alumni tend to ignore this kind of surveys. At the other extreme, the development of information technology offers the possibility for everyone to exist in the virtual world via social media, such as LinkedIn, Facebook and Twitter. The authors postulate that if one could find the correct alumni name in search engines, then, it would be easier to locate their social pages, and find some useful keywords to determine their professional jobs, especially, in LinkedIn.

In this article, the authors report on the job classification task, as an extension of their previous work [1], i.e. the alumni name disambiguation task, which is based on an unsupervised clustering method [2]. Based on the cluster relevance, one could have a number of cluster candidates, which contain the most relevant information about an alumnus/alumna. The job classification task was conducted by using Naïve Bayes classification approach based on job terms. A job term is a term, which is related to a certain job, e.g. *coding* is a job term for a *programmer*.

The classification method is preferred ahead of clustering, since all major-related jobs have been analysed and listed before with the help of alumni associations. Naïve Bayes was selected due to its simple implementation and high probability of accuracy [3]. In several cases, this algorithm is also more powerful than other complex learning algorithms [4][5]. The authors also report on their prototype system, which may be connected to LinkedIn and can automatically extract important information about alumni.

MODEL DEVELOPMENT AND EXPERIMENT

After the most relevant search engine result pages (SERPs) cluster has been selected based on name co-occurrences in SERPs [1], the recent alumni profession is determined by using the Naïve Bayes classification method. The general formula of Naïve Bayes classification method can be seen in Formula (1):

$$P(job_x \mid term_y) = \frac{P(job_x)P(term_y \mid job_x)}{P(term_y)} \quad \ldots \tag{1}$$

This job prediction activity is implemented to fulfil the university's needs to recognise how many alumni are working in a relevant field based on their major. The job prediction models in the experiments were determined in several scenarios:

1. By calculating the term occurrences of a job target in a profession field (see Table 1). A profession field is considered as related to a profession, if it contains (one or more) specific job terms (e.g. programmer, doctor and lecturer). In this case, a set of 134 jobs was selected, based on the job categories listed in the Indonesian Wikipedia. Since the focus is research on Maranatha Christian University (MCU) alumni, the above selected jobs were mapped into nine profession fields, based on the faculties (which are medical, economics, information technology, engineering, art and design, literature, psychology, law and dentistry). This list was, then, combined with some other relevant jobs obtained from the most common higher education nomenclature in Indonesia. The complete job categories can be seen in Table 2;
2. By determining the occurrences of nouns and verbs word classes in alumni SERPs clusters;
3. By determining an aggregation method. The aggregation is done by calculate the highest probability score of the top-3 clusters, which constructed using the *Red*-UPND algorithm [1].

Table 1: Job prediction targets.

Professional field	Job targets (profession models)
Medical (including: dentistry)	Surgeons, nutritionists, pharmacists, assistant pharmacists, midwives, doctors, paramedics, nurses, dentist and psychiatrists
Economics	Accountants, public accountants, auditors, real estate brokers, economists and brokers
Information technology	Hackers, system administrators and programmers
Engineering	Engineers, technicians and architects
Literature	Curators, novelists, translators, authors, screenwriters and poets
Art and design	Designers, font designers, illustrators, painters, sculptors, tailors, graphic designers, (travel) guides and artists
Law	Politicians, dictators, judges, prosecutors and lawyers
Psychology	Psychologists, human resource managements and counsellors
Education	Teachers, professors, school counsellors and librarians
Theology	Counsellors, marriage advisors, cleric, clerics, pastors and priests
Entertainment	Actor, go-go dancers and masters of ceremony
Security	Police, special police train and the guard
Aerostation	Pilots, co-pilots and flight attendants

The second and third scenarios were designed to evaluate the significance of each profession model, so that one can capture whether special terms effect a particular profession model. The evaluation was conducted by exploiting 119 alumni data from MCU alumni tracer study, from 2009 to 2013. The experiments and the evaluation were organised between May 2015 and April 2016.

EVALUATION AND DISCUSSION

Model Evaluation

As the first step of the evaluation, the authors evaluated the quality of the first model scenario described in the previous section. This evaluation was conducted in two approaches, which are: the *k*-fold cross validation and the hold-out testing between training and testing data. Hold-out was conducted based on three proportions, i.e. 80:20, 70:30 and 60:40. Table 2 shows cross validation results for the job prediction experiments.

Table 2: Cross validation accuracy (in percentage).

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	SD
74.3%	80.1%	84.4%	80.6%	64.3%	76.7%	7.8%

The highest accuracy was found at Fold-3 of the cross validation, which is 84.4%. Second was Fold-4 of the cross validation with 80.6% accuracy. The remaining *k*-folds yielded lower results, which were 80.1% (Fold-2), 74.3% (Fold-1), and 64.3% (Fold-5). The overall accuracy average of the cross validation is 76.7% with a 7.8% standard deviation. Based on the data, the authors considered the cross-validation evaluation process achieved its stability at 80% accuracy.

Further investigation shows that the errors in Fold-5 was mainly caused by unseen terms in the engineering profession fields models, which were related to the information technology and engineering faculties. Specific terms, such as manager, marketing or trainer are related to another profession field, i.e. economics, in the models. This result suggests that the engineering alumni are not only contributing to the real engineering job, but also have broader job roles in society.

Table 3 shows the experiment results conducted by the using hold-out mechanisms. The findings in Table 3 indicate that the proposed classification models have covered the prediction needs for each profession. Job prediction in the 70-30 composition yielded the lowest accuracy, though its training data is not the lowest size. After further analysis,

these phenomena were found to be caused by the unbalanced composition of the testing dataset from the economics and engineering majors (can be seen in highlighted cells in Table 4). These findings also strengthen the results in Table 2 that in the engineering and economics fields, there are more variations of alumni professions, which are not covered in the models.

Table 3: Hold-out evaluation accuracy (in percentage).

80:20	70:30	60:40
90.91	85.71	89.19

Table 4: The comparison of the numbers of testing data between 70-30 and 60-40 in hold-out evaluation.

Faculty	70:30	60:40
Medical	9	9
Engineering	6	15
Law	1	1
Economics	5	8
Psychology	4	4
Art and design	4	4
Literature	2	2

The authors investigated these variations by calculating the influences of matching specific words, all words models, verbs and nouns word classes' models, and the cluster aggregation mechanism. The result in Table 5 gives the appropriate classification methods for alumni job prediction based on each faculty.

Table 5: Classification methods for alumni job prediction.

Faculty	1st most suitable method	2nd most suitable method	3rd most suitable method
Medical	Nouns, e.g. hospital	Verbs, e.g. diagnose	Specific words, e.g. physician
Engineering	Specific words, e.g. electricity	Nouns, e.g. technician	-
Law	Specific words, e.g. lawyer	-	-
Economics	All words, e.g. marketing	Aggregations, e.g. financial	-
Psychology	Specific words, e.g. psychologist	-	-
Arts and design	Specific words, e.g. designer	-	-
Letters	Verbs, e.g. teach	-	-
Information technology	Specific words, e.g. programmer	All words, e.g. marketing	Nouns, e.g. computer
Dentistry	Nouns, e.g. clinic	Verbs, e.g. diagnose	Specific words, e.g. dentist

The highlighted cells in Table 5 also reveal that there are some faculties in which the alumni have a broader contribution to society, beyond their original study fields. From the data, these phenomena occur mostly in the fields of engineering and information technology. It is of course interesting for future engineering curriculum development to strengthen soft skills and entrepreneurship subjects during the study period.

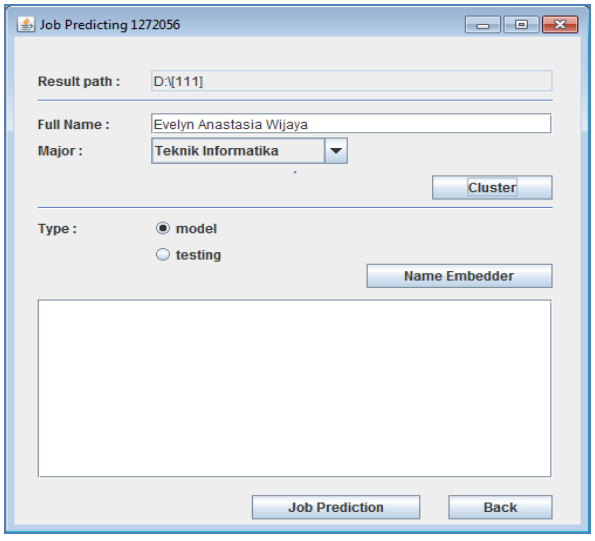

Figure 1: Back-end user interface.

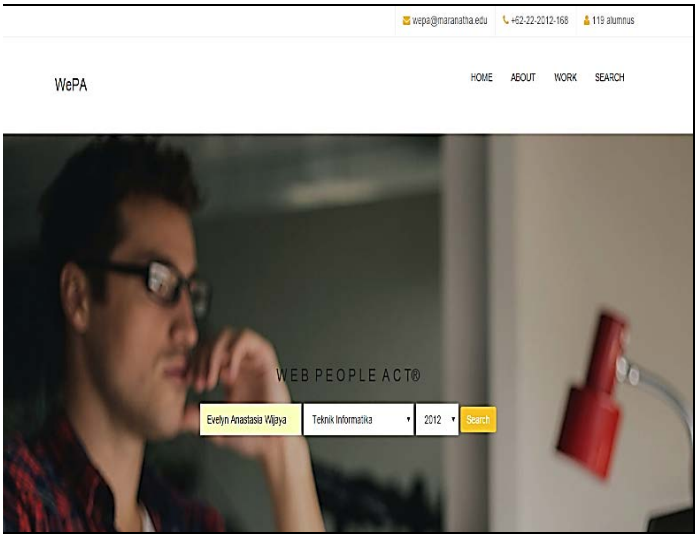

Figure 2: Front-end Web application.

Virtual Alumni Tracer Prototype

This section covers the proposed application, which consists of two parts: 1) desktop application (back-end component that handles administrator-based features); and 2) Web application (front-end component that interacts with users and alumni). Figure 1 shows back-end implementation of the prototype. Through this form, data from SERPs are clustered by using *Red*-UPND [1] and processed in several sub-components so that it yields the job prediction result.

Figure 2 presents the user interface of the main front-end system. Alumni information can be accessed with this page by clicking the search button. The prototype system will search for information on an alumnus/alumna and predict their job based on the SERPs. Information retrieved from this phase are subsequently shown in a Web page view as seen in Figure 3 and Figure 4, which may be connected to LinkedIn, as a means to validate the job classification model.

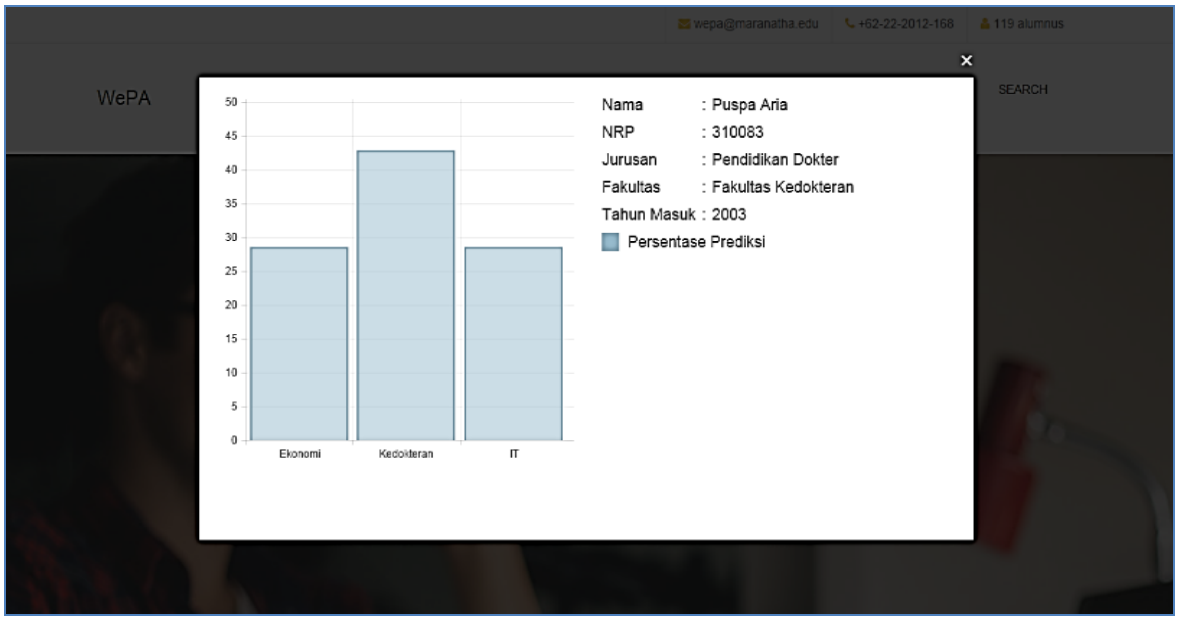

Figure 3: Scraped SERPs user interface and the alumni predicted profession.

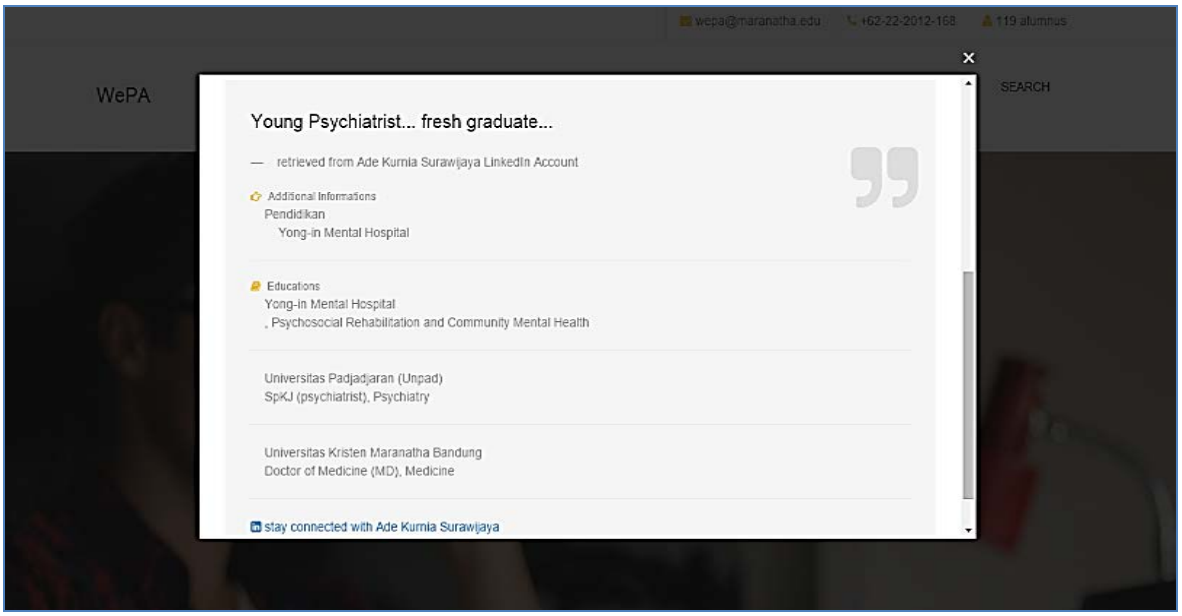

Figure 4: Alumni data from an alumni LinkedIn account.

CONCLUSION AND FUTURE WORK

Based on the research conducted, several conclusions can be drawn, as follows:

1) 134 alumni jobs have been successfully mapped into 14 more-general profession fields and treated as job prediction models.
2) The accuracy of the alumni job prediction is greatly affected by faculty name queries and by the number of data collections. More collections of data may yield more precise results.

3) The alumni job prediction incorporated in this system has proved to be effective since it yields 90.91% accuracy. Hold-out evaluation with the composition of 80% training data and 20% testing data yields the highest accuracy among hold training-test and cross validation.
4) This application predicts not only recent alumni jobs based on non-social media clusters, but it also incorporates a social media cluster (LinkedIn) to extract alumni information.

As consequences of these conclusions, several research aspects should be prepared in the near future, such as:

1) The SERPs collection size for the system should be enlarged, since a larger collection size is expected to yield more accurate results for predicting jobs and faculties.
2) Further evaluation should be considered for comparing automatic job prediction with manual-assigned jobs provided by social media, such as LinkedIn.
3) Enrich the Web portal with supplementary functionality that is beneficial for alumni promoting themselves. For example, providing a mechanism which allows alumni to post their resume and CV and facilitating job seekers to advertise their job vacancy in the portal.
4) Exploring other social media instead of LinkedIn for extracting alumni data, such as: Facebook and Twitter. Moreover, friendship relations in these social media may also be utilised for further alumni relationship analysis.

ACKNOWLEDGMENT

This research has been supported by a research grant provided by Maranatha Christian University.

REFERENCES

1. Toba, H., Wijaya, E.A, Wijanto, M.C. and Karnalim, O, Enhanced unsupervised person name disambiguation to support alumni tracer study. *Global J. of Engng. Educ.*, 19, **1** (2017) (Submitted).
2. Delgado, A.D., Martinez, R., Fresno, V. and Montalvo, S., A data driven approach for person name disambiguation in web search results. *Proc. Inter. Conf. on Computational Linguistics: Technical Papers*, Dublin, Ireland, 301-310 (2014).
3. Mitchell, T., *Generative and Discriminative Classifiers: Naïve Bayes and Logistic Regression in Machine Learning*. McGraw Hill (2015).
4. Croft, B., Metzler, D. and Strohman, T., *Search Engine: Information Retrieval in Practice*. Boston: Pearson Education Inc. (2010).
5. Bui, A.A. and Taira, R.K, *Medical Imaging Informatics*. London: Springer Science Business Media (2010).