Global Journal of Engineering Education

Enhanced unsupervised person name disambiguation to support alumni tracer study

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

Maranatha Christian University
Bandung, Indonesia

ABSTRACT: An alumni database is a valuable information source for the development of a university. However, alumni databases tend to be incomplete. It is always possible for phone numbers and home or e-mail addresses to change. In this study, the authors propose an information collection strategy by gathering information spread across the Internet through search engines. The research is focused on the evaluation of efficiency factors during the name disambiguation process. The authors suggest a combination of reduction (*Red*-UPND) and supervised queries strategies, which improve the efficiency of the disambiguation process to around 67% compared to the baseline unsupervised person name disambiguation (UPND). The experiment results show that the approach fits the case university's requirements to support an alumni tracer study and to find people automatically, especially for people with *ordinary* names.

Keywords: Text mining, information extraction, clustering, person name disambiguation, virtual alumni tracer

INTRODUCTION

An alumni database is required to support further developments at a university, such as during the curriculum reconstruction process or for promotional needs. Unfortunately, up-to-date data are difficult to obtain. This might be because of the unmaintainable relationship between a university and its alumni after the graduation. Although most of the alumni provide their personal data upon graduation, like telephone numbers or electronic media contacts, these data might change in the future. Besides that, most of the alumni have a very low incentive to update their data directly with the university. Manual survey methods can be used to update an alumni database, but this method is quite difficult, since most of the locations of the alumni are unknown. Even if the alumni could be reached, only a small percentage would be willing to complete it, as the survey may contain a lot of crucial personal information [1].

Considering the difficulties of distributing hardcopy surveys and the options provided by technology advances (especially the Internet), some universities focus their surveys on electronic versions. The process could be done semi-automatically, with alumni not needing to fill out surveys physically, and also it could be prepared anytime and anywhere [2][3]. However, changes in variations of the surveys are still unable to increase alumni interest in filling out surveys.

Along with the development of Internet technology in recent times, most students have social media accounts, such as Facebook, Twitter, LinkedIn, personal blogs and other Web-based information. In contrast with alumni data in a university static database, alumni on-line data are frequently updated by their owners. In social media, most of the owners not only renew their profile data willingly, but also their contact information. In this study, the final objective is to develop an efficient method to disambiguate alumni names.

RELATED WORK AND RESEARCH CONTRIBUTION

Alumni data after the graduation day are quite difficult to obtain since there is almost no obligation for a post-relationship between the alumni and the university in general. Hardcopy surveys may be promising, but they are still inefficient in terms of time, and their coverage is limited. To tighten the bond between the university and its alumni, short messaging service (SMS) or social media approaches can also be utilised [2][3]. However, most efforts are considered to be unsuccessful since there are no specific advantages for alumni, which encourage them to update their

information on the Web. In this research, the low interest of alumni to update their information is handled by obtaining their information automatically based on search engine result pages (SERPs) [4]. Furthermore, an information extraction technique to excerpt information, from social media lists has been used. (As is well known, most Internet users update their social media accounts frequently). The link information from each SERP is accessed and its contents are converted into *n*-gram terms and will be used during the name disambiguation process.

Several recent pieces of research have discussed person name disambiguation. Based on Malin, name ambiguity can be caused by a typing error or exact-match name terms between two or more people [5]. In this research, the focus is on the latter one since a name found on a SERP may not represent an alumnus/alumna. Name ambiguity due to exact-match name terms usually occurs in paper authors [6], citations [7], names in e-mails [8], and names in Web pages [9-11].

Delgado et al proposed a mechanism to cluster names on SERPs without relying on a training dataset or certain threshold values. Their method is called the UPND, which yields more effective results than other supervised clustering algorithms in similar circumstances [9]. The main idea of UPND is to create independent clusters with the assumption that each cluster contains several SERPs that relate to a unique entity. Since the name disambiguation problem in this research is quite similar to the problem context of UPND, it has been selected as the baseline clustering algorithm for name disambiguation in this research, which has been further optimised.

RESEARCH DESIGN

Data Processing

The data extraction process from the search engine is managed by using a self-developed SERPs processor, which enables users to obtain SERPs from popular search engines, such as Google, Yahoo!, and Bing. In this research, Google has been selected as the target search engine due to its high relevancy, which is determined by using PageRank [12]. Furthermore, since the focus of SERPs is on the alumni relationship to a university, supervised query strategy is proposed. It is based on the alumni major (study programme) and the university name, in both Bahasa Indonesia and English. An example of supervised queries generated for *Evelyn Anastasia Wijaya* can be seen in Table 1.

These queries are generated based on three query terms, which are *Evelyn Anastasia Wijaya* as the alumnus/alumna name, *teknik informatika* (technical informatics) as her major, and *Universitas Kristen Maranatha* as the name of the university. Alumni and university name are represented as phrases (double quotation mark), whereas the study programme name is represented as natural query.

Table 1: Supervised query for searching.

Queries
"Evelyn Anastasia Wijaya"
"Evelyn Anastasia Wijaya" + *Teknik Informatika*
"Evelyn Anastasia Wijaya" + Technical Informatics
"Evelyn Anastasia Wijaya" + "*Universitas Kristen Maranatha*"
"Evelyn Anastasia Wijaya" + "Maranatha Christian University"
"Evelyn Anastasia Wijaya" + *Teknik Informatika* + "*Universitas Kristen Maranatha*"
"Evelyn Anastasia Wijaya" + Technical Informatics + "Maranatha Christian University"

To maintain the validity of the search results, specific SERPs only were selected, i.e. whether it is a social media site or its domain, it is categorised as a trusted domain. The social media site list was obtained from *alexa.com*, whereas trusted domains are defined by their suffixes. A domain is considered trusted, if its suffix represents education, organisation and government. Furthermore, the content of any Web page link in selected SERPs will be extracted and stored. Extraction is conducted in several steps, which are: removing html tags; tokenising text with non-alphanumeric as its delimiter; converting all tokens to lowercase; removing Indonesian stop words; word stemming and applying rank-concerned TF-IDF; weighting for each *n*-gram forms, where *n* is defined from 1 to 4, inclusively [4]. Rank-concerned TF-IDF weighting is the extension of standard TF-IDF weighting, which takes the page ranks in SERPs into consideration.

UPND Clustering

This phase is intended to disambiguate alumni names, since some alumni names given in the previous step might be ambiguous and, thus, not all SERPs are considered relevant. UPND generates several clusters and selects the most relevant cluster based on given query. Clustering is conducted using a UPND algorithm by which each page is represented as a document and terms are defined in *n*-dimensional vectors. This algorithm applies agglomerative clustering, which assumes each document is a cluster and each iteration will merge two or more clusters if each possible pair yields a similarity value higher than a certain threshold [9]. The complete UPND algorithm can be seen in Figure 1, where the threshold γ is measured by an *n*-gram difference between two Web pages in Equation (1).

$$\gamma\left(SERP_i^n, SERP_j^n\right) = \frac{\min(k,l) - shared(SERP_i^k, SERP_j^l)}{\max(k,l)} \quad \ldots \quad (1)$$

Selecting the most relevant cluster for an alumnus/alumna is based on a heuristic method, which measures the correlation between the numbers of Web pages in SERPs with the number of optimal clusters. Vectors for each document are determined from term weighting based on its respective terms. Weighting can be conducted with word counting or a relevance-involved method, such as the inverted document frequency (TF-IDF). The UPND algorithm has two tuneable parameters, i.e. the initial and maximum number of *n*-grams. For this research, the authors set the initial *n*-gram to 3 and maximum *n*-gram to 4 - the *k* and *l* parameters in Equation (1) - since these values yield the best combination for this algorithm [9].

Since UPND is quite inefficient in terms of execution time, due to its $O(n^2)$ complexity, the SERPs involved in UPND need to be reduced, with a basic assumption that the supervised queries might yield similar SERPs [4]. Those similar SERPs are removed from the collection, with its rank is assigned with the highest rank among them. By reducing the number of similar SERPs, one might expect that the UPND algorithm will performed better. Here, it has been called the *Red*-UPND algorithm. The details of both algorithms can be followed in Figure 1.

```
Algorithm 1 UPND(SERP, k, l)
Input: Set of web pages that shared a
  person name SERP = {SERP1, SERP2, ...,
  SERPN}, k, l2 ≥ 1 such that l ≥ k
Output: Set of person clusters Clust = {C1,
  C2, ..., Cl}
Procedure:
    for n = 1 to N do
      Clusti = {SERPi}
    end For
    Clust = {Clust1, Clust2, ... , ClustN}.
    for n = k to l do
      setNGrams (n, SERP).
      for i = 1 to n do
        for j = i + 1 to n do
          if Sim(SERP_i^n, SERP_j^n) ≥ γ(SERP_i^n,
          SERP_j^n) then
             Clusti = Clusti ∪ Clustj
             Clust = Clust \ {Clustj}
          end if
        end for
      end for
    end for
    return Clust
```

```
Algorithm 2 – Red-UPND(SERPL)
Input: SERPL a list of web links,
  i.e. the query results from the
  search engine
Output: UpndClust the UPND
clusters
Procedure:
for all links (Li, … , Lj) ∈ SERPL
do {
      Remove duplicate
      links(Li…Lj)
      Maintain the HURL, i.e.
      the highest unique rank
      link in (Li…Lj)
}
call UPND(HURL)->UpndClust
return UpndClust
end Procedure
```

Figure 1: The baseline UPND algorithm (left) and its enhancement the Red-UPND algorithm (right).

RESULTS, ANALYSIS AND DISCUSSION

In this section, the experiment results based on the approach used are reported. The evaluation was conducted by exploiting 119 pieces of alumni data from an MCU alumni tracer study, from 2009 to 2013. The experiments and the evaluation were organised during the period May 2015 - April 2016.

Clusters Impact for Completing Basic Information

Each query type of an alumnus/alumna is given as an input for the Internet search engine. It is expected that 100 search results (the first 10 SERPs) would be received for each query type. However, most queries yield lower numbers of search results due to their rare query terms. In test cases, the maximum number of search results for alumni data is only 99. This result strengthens the fact that most of MCU's alumni could be considered as *ordinary* people, since their data are not easily found on the Internet.

Table 2 shows the correlation between search results with the number of clusters based on the dataset. Intuitively, each query is expected to yield the lowest number of clusters since UPND will merge similar Web page contents into one cluster. Furthermore, the Pearson's correlation between the number of SERPs results and the number of clusters are also shown in Table 2. A high Pearson's correlation should represent more relatedness between the number of SERPs and clusters. Shaded-cells show how the university name affects the number of clusters. The university name in Bahasa Indonesia yields a lower number of clusters than the English one, although both of them yield similar SERPs accuracy, which are calculated according to the overlapped longest common subsequence (LCS) [13] of the alumni names (see also Table 3 in the subsequent section).

Table 2: Correlation between the number of SERPS and cluster formation for each query type.

Queries (119 alumni)	Average #SERPs	#Max page	Average #Clusters	#Max clusters	Pearson's correlation
Name + Ind Univ + Ind Major	5.27	68	1.81	26	0.87
Name + Ind Major	11.86	94	4.81	71	0.82
Name + Ind Univ	6.66	60	2.23	34	0.82
Name + Eng Major	15.85	97	5.34	56	0.73
Name + Eng Univ + Eng Major	7.07	94	1.92	22	0.69
Name + Eng Univ	7.42	71	2.82	65	0.67
Name only	21.37	99	4.29	37	0.27

Four query types have acceptable correlations between the number of SERPs and the number of clusters. These types are: *alumni name + Indonesian university name + Indonesian major name*, *alumni name + Indonesian university name*, *alumni name + English university name + English major name* and *alumni name + English university name*.

These results show that the combination of an alumnus/alumna name and the university name, regardless of the use of natural language (Indonesian or English), yields around two or three clusters for each alumnus/alumna. Besides that, combining the major name with an alumnus/alumna name, yields a higher number of clusters (four clusters or more). Since each cluster represents unique person data based on query, fewer clusters might be more advantageous when determining alumni-related clusters. The impact of the number of clusters for completing alumni basic information is presented in Table 3. This evaluation is intended to show the depth of information that can be reached, extracted from the search engine for forming each cluster. The accuracy is determined by matching the results with the (static) alumni database by calculating the percentage of overlapped LCS in specific fields. In Table 3, the evaluation result is sorted by e-mail accuracy, since in the case here, an e-mail address is the most beneficial piece of information for conducting a tracer study.

Alumni e-mail addresses might be valuable for constructing the relationship between a university and its alumni. It is quite interesting to see that the greatest accuracy for finding e-mail addresses can be achieved by using only the alumni name as a query. However, the overall information can only be successfully gained by combining alumni name, university name, and major name (regardless the use of natural language). Since the highest overall accuracy result from the combination of university and major name in English, it could also be derived, because most of the alumni data are stored in English Web pages.

Table 3: The impact of clusters, based on experimental dataset (sorted by LCS *e-mail match* accuracy).

Queries (119 alumni)	E-mail address	Name	High school	Province	Birth place	Date of birth	Office address	Home address	Average
Name only	30.78	84.59	18.64	14.66	17.17	20.78	20.54	14.75	26.12
Name + Eng Univ	29.27	59.52	23.52	25.68	34.52	24.03	16.68	16.57	29.21
Name + Eng Major	28.28	76.00	17.67	10.95	15.99	21.93	12.46	16.62	24.42
Name + Ind Univ + Ind Major	27.67	65.52	21.98	34.91	39.87	24.71	16.29	19.11	31.85
Name + Eng Univ + Eng Major	27.37	60.53	25.74	32.40	43.81	24.96	18.45	16.88	31.93
Name + Ind Univ	26.34	69.47	19.37	22.50	37.29	23.69	12.04	16.88	27.44
Name + Ind Major	24.45	74.23	20.47	21.86	22.25	21.70	13.23	17.18	26.04

Efficiency and Quality of the *Red*-UPND Algorithm

Table 4 shows the execution time differences between *Red*-UPND and UPND for constructing clusters, based on the test data for the 119 alumni. The *Red*-UPND algorithm improved significantly in terms of execution time. The original execution has been reduced to 95.25 hours, which is an advantage of 195.08 hours. Since 195.08 plus 95.25 equals 290.33, the efficiency rate is around 195.08/290.33 that is approximately 67.2%. Further investigation shows that the search results consist of many redundant Web pages.

Table 4: Execution time comparison between UPND and *Red-UPND*.

Type	Start date	End date	Execution time (in hours)
UPND	22 May 2015 (13:30)	03 Jun 2015 (15:50)	290.33
Red-UPND	09 Jul 2015 (16:10)	13 Jul 2015 (15:25)	95.25

Redundant Web pages might be caused by two aspects, which are:

1. Since the query is given as a combination of phrases of alumni name, university and the major name, there are many Web pages, which are considered as true positive in two or more sub-query phrases. True positive in two or more query phrases may yield a Web page considered to be two or more search results.
2. Many search results were linked to each other, especially for *more-popular* alumni in social media. Moreover, many social media pages split their view based on their server region (.id, .com, etc) despite the similarity of its contents.

When evaluated, based on the quality of the resulting clusters, the UPND and *Red*-UPND yield almost identical results. However, some queries yield more qualified clusters by utilising *Red*-UPND. These findings can be seen in Table 5, which shows the effectiveness differences between UPND and *Red*-UPND. Effectiveness is evaluated by using the partial match of LCS. Improved accuracy that is caused by utilising *Red-UPND* is shaded. It can be seen that the combination of alumni name, university and major name yields higher accuracy than the baseline (although it is not significant).

A significant difference (with $p = 0.05$) is only shown in the combination of alumni name and English university name wherein the *Red*-UPND declines about 2.5% for the overall alumni data. This finding indicates that several information sources might not be properly retrieved by the usage of *Red*-UPND for that particular query. In other words, there are many alumni that provide their information on English Internet Web pages. English Web pages do not always consist of *only-English* words. Most of them may contain another language with several parts presented in English. In general, effectiveness resulting from *Red*-UPND is quite similar to UPND. The best accuracy still results from the combination of the alumni name, university and major name regardless of natural language usage, i.e. the *Red*-UPND yields 32.57% accuracy for that combination, whereas UPND yields 32.66%.

Table 5: Effectiveness differences between UPND and *Red*-UPND.

Queries (119 alumni)	E-mail address	Name	High school	Province	Birth place	Date of birth	Office address	Home address	Avg.
Name + Eng Univ + Eng Major	11.72	(1.82)	1.95	3.05	(2.08)	1.26	0.72	0.52	0.73
Name + Ind Univ + Ind Major	9.23	2.93	(0.41)	0.36	2.02	0.37	0.52	0.27	0.72
Name + Eng Univ	0.62	(4.13)	0.27	(6.07)	(4.51)	(1.01)	(1.39)	(0.54)	(2.49)*
Name + Ind Major	(5.32)	0.85	0.19	0.23	(0.08)	0.22	0.43	0.03	0.23
Name + Eng Major	(7.67)	0.99	(1.06)	(0.69)	(2.11)	0.11	1.05	(0.19)	(0.19)
Name + Ind Univ	(11.52)	(1.40)	(2.25)	(6.32)	(1.63)	(0.41)	2.37	(0.91)	(1.77)
Name	(18.85)	0.91	0.35	(0.59)	(0.76)	(0.37)	(0.04)	(0.02)	(0.16)

Note: * Statistical significant with $p = 0.05$

CONCLUSIONS AND FUTURE WORK

Based on this research, two main conclusions can be drawn, as follows:

1. Relevant links for each alumnus/alumna could be extracted by providing a supervised query that involves the majors and university name. A natural major's query is expected to yield search results not only as majors, but also as job-related terms.
2. Alumni name ambiguity could mostly be handled with the *Red*-UPND algorithm. *Red-UPND* reduces the execution time of around 67.2% of the execution time with no significant precision loss against the standard UPND algorithm.

As consequences of the conclusions, several research aspects could be prepared in the near future, as follows:

1. Execution time for running UPND can be reduced with more enhanced approaches instead of only removing redundant pages. For example, by filtering particular terms, which are assumed to have a greater impact on the forming of the alumni clusters.
2. Exploring social media, such as LinkedIn (one of the reliable social media used during the experiments for this article) for extracting alumni data and job predicting tasks. Such social media, including Facebook and Twitter, might provide friendship relationships, which could be useful for further alumni analysis.

ACKNOWLEDGMENT

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

REFERENCES

1. Heidemann, L., Only successful graduates respond to tracer studies: a myth? Results from the German Cooperation Project. *Proc. Inter. Conf. on Human Capital and Employment in the European and Mediterranian Area*, Bologna, Italy, Working Papers 13 (2011).

2. Wibisono, A., Ulama, B.S.S. and Asmoro, W.A., Tracer study at Institut Teknologi Sepuluh November (ITS), promoting localization and multiple touch points to capture alumni. *Proc. Inter. Conf. on Experience with Link and Match in Higher Educ.: Result of Tracer Studies World Wide*, Bali, Indonesia (2012).
3. Noviyantono, E. and Aidil, Integration system of web based and SMS gateway for information system of tracer study. *Proc. Inter. Conf. on Engng. and Technol. Develop.*, Bandar Lampung, Indonesia, 86-92 (2012).
4. Croft, B., Metzler, D. and Strohman, T., *Search Engine: Information Retrieval in Practice*. Boston: Pearson Education Inc. (2010).
5. Malin, B., Unsupervised name disambiguation via social network similarity. *Proc. SIAM Inter. Conf. on Data Mining*, California, USA, 93-102 (2005).
6. Huang, J., Ertekin, S. and Giles, C.L., Efficient name disambiguation for large-scale databases. *Proc. European Conf. on Principles and Practice of Knowledge Discovery in Databases*, Berlin, Germany, 536-544 (2006).
7. Han, H., Zha, H. and Giles, C.L., Name disambiguation in author citations using a *k*-way spectral clustering method. *Proc. ACM/IEEE-CS Joint Conf. on Digital Libraries*, Denver, USA, 334-343 (2005).
8. Minkov, E., Cohen, W.W. and Ng, A.Y., Contextual search and name disambiguation in email using graphs. *Proc. SIGIR Conference on Research and Development in Information Retrieval*, Washington, USA, 27-34 (2006).
9. 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).
10. Mann, G.S. and Yarowsky, D., Unsupervised personal name disambiguation. *Proc. Natural Language Learning at HLT-NAACL*, Stroudsburg, PA, USA, 33-40 (2003).
11. Yoshida, M., Ikeda, M., Ono, S., Sato, I. and Nakagawa, H., Person name disambiguation by bootstrapping. *Proc. ACM SIGIR Conference on Research and Develop. in Infor. Retrieval*, Geneva, Switzerland, 10-17 (2010).
12. Brin, S. and Page, L., Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 56, **18**, 3825-3833 (2012).
13. Yalniz, I. and Manmatha, R., An efficient framework for searching text in noisy document images. *Proc. Inter. Workshop on Document Analysis Systems*, Queenslands, Australia, 48-52 (2012).

BIOGRAPHIES

Hapnes Toba graduated in 2002 with a Master of Science from Delft University of Technology in the Netherlands, and completed his doctoral degree at Universitas Indonesia in 2015. He has been working as a faculty member in the Faculty of Information Technology at Maranatha Christian University since 2003. His speciality is in the field of information retrieval, text processing and data mining. Further information can be found on Research Gate.

Evelyn Anastasia Wijaya is an alumna of the Faculty of Information Technology at Maranatha Christian University. She graduated in 2016, and is now working for a nation-wide company in Indonesia.

Maresha Caroline Wijanto is an alumna of the Faculty of Information Technology at Maranatha Christian University and graduated at Bandung Institute of Technology (ITB) with her Master's degree in computer science. She joined the Faculty of Information Technology at Maranatha Christian University in 2010. Her specialty is in the field of natural language processing and datamining.

Oscar Karnalim graduated with a Bachelor of Engineering degree from Parahyangan Catholic University in 2011, and completed his Master degree at Bandung Institute of Technology (ITB) in 2014. He is interested in the software engineering domain, especially in source code analysis. Several other topics also attract his attention, such as algorithm visualisation, reverse engineering, and information retrieval. Further information can be found on Research Gate.