Gathering Alumni Information from a Web Social Network

Gabriel Resende Gonçalves; Anderson A. Ferreira; Guilherme Tavares de Assis
Departamento de Computação
Universidade Federal de Ouro Preto
Ouro Preto, MG, Brazil
gbrll.rg@gmail.com, {ferreira, gtassis}@iceb.ufop.br

Andrea Iabrudi Tavares
Jasper Design Automation
Mountain View, CA, USA
andrea.iabrudi@gmail.com

***Abstract*—An undergraduate program must prepare its students for the major needs of the labor market. One of the main ways to identify what are the demands to be met is creating a manner to manage information of its alumni. This consists of gathering data from program's alumni and finding out what are their main areas of employment on the labor market or which are their main fields of research in the academy. Usually, this data is obtained through available forms on the Web or forwarded by mail or email; however, these methods, in addition to being laborious, do not present good feedback from the alumni. Thus, this work proposes a novel method to help teaching staffs of undergraduate programs to gather information on the desired population of alumni, semi-automatically, on the Web. Overall, by using a few alumni pages as an initial set of sample pages, the proposed method was capable of gathering information concerning a number of alumni twice as bigger than adopted conventional methods.**

***Keywords*-alumni information management; web social network; search engine; focused crawling; linkedin.**

I. Introduction

Nowadays, in academic environments, one of the major concerns of teaching staffs of undergraduate programs is analyzing how their students adapt to the professional life after their graduation. This concern exists, essentially, as a means to know if the curricular grating and the program syllabus of a given undergraduate program, passed onto its students, are managing to meet the needs of the current labor market.

Within this context, the definition of an efficient method to obtain professional data from alumni of undergraduate programs becomes necessary. According to Lousada and Martins [1], the traditional methods employed to gather alumni information do not have good collaboration rate, which lead to interrupt of gathering such a information. The problem of managing alumni information is recurrent in several universities. Thus, this work aims to propose a tool capable of semi-automatically retrieving professional data of alumni from the Web to help teaching staffs of undergraduate programs. The tool has methods for gathering relevant data about alumni from a social networks using a given available search machine.

Social networks are oriented either towards entertainment of its users or spreading professional data about their users. For the first purpose, there exist, for instance, MySpace[1], Twitter[2] and Facebook[3]. For professional purposes there exist, for instance, LinkedIn[4] and Bayt[5] [2].

Our tool uses a search engine for getting from LinkedIn a initial set of candidate pages. Next, the tool selects from the set of candidate pages the ones similar to the alumni population from a given under graduate program. To calculate the similarity, we use some pages gathered from LinkedIn as examples. Instead of providing data about the undergraduate program and institution of a alumni list, our tool receives as input a few LinkedIn alumni pages from this list. Furthermore, in order to perform effectively, our tool also receives as input the list of alumni's name from an undergraduate program.

In sum, the main contribution of this work is the definition of a new method performed by our tool that is capable of gathering information regarding alumni of a given undergraduate program. Moreover, this work presents an experimental comparative evaluation between our method and a traditional classifier in the retrieval of alumni pages from the web.

The rest of this paper is organized as follows. Section II presents works directly related to the goal of this work. In Section III, we describe our proposed method for gathering semi-automatically information on alumni of a given undergraduate program. In Section IV, we evaluate our method. Finally, in Section V, we present conclusion and perspectives for future work.

II. Related Work

Our proposed method, in this work, aims to perform a focused crawling of information on alumni on the web and extracts such a information. Based on that, the works related to this paper are divided in the following subsections: alumni information management and web pages focused crawling.

[1] http://www.myspacecom
[2] http://www.twitter.com
[3] http://www.facebook.com
[4] http://www.linkedin.com
[5] http://www.bayt.com

978-1-4799-6953-1/14 $31.00 © 2014 IEEE
DOI 10.1109/LAWeb.2014.17

A. Alumni Information Management

Taking into account that promoting management of alumni information is an important task to undergraduate programs of higher education Brazilian institutions, some studies were conducted to evaluate the benefits that such management may bring to the programs. Thus, several alumni monitoring programs have already been created in Brazil.

In [3], the authors conducted studies on the difficulties encountered in the management of alumni from educational institutions in Brazil. The authors introduce the concept of alumni within the Brazilian sphere. On that basis, alumni were categorized in graduates, students transferred to other schools, students that were dismissed from the institution and students that quit their programs. It was concluded that the monitoring concerning alumni is a means of assessing education outcomes of an institution so that improvements in its teaching methodology can be made. Furthermore, the work also highlights that, in order for the alumni management to be accurate, those must be considered according to the four previously mentioned categories.

In [4], the authors perform studies on benefits and potentialities that might be explored by an institution, when promoting alumni monitoring management. It is noteworthy that the challenges concerning the alumni information management does not happen solely in Brazil, but in a number of countries. As a conclusion, the authors mention that the alumni management provides better effectiveness of the institutional actions promoted by higher education institutions.

In the Federal University of Santa Catarina, a system was created to perform alumni tracking. The method contacts the alumni using a online portal where alumni enter and answer some questions [5]. Up to the moment of publication of the results, only about 6.8% of all alumni from the institution had contributed to the portal. On the other hand, regarding the alumni monitoring program from the environmental engineering program of the University of São Paulo, the institution achieved 79.3% of responses to emails that were sent to the alumni up to the publishing moment [6]; the explanation to such high response rate obtained by the referred program lies on the fact that the undergraduate program is not old, showing a number of only 140 alumni.

In the University of São Paulo, a system called Egressos-USP renders the management of information on the institution's alumni [7]. The data gathering is done through an online survey divided in four pieces. In the first piece, a alumnus must answer about its professional situation. In the second piece, the alumnus answers questions concerning the university structure. In the third piece, there exist questions about personal opinion on the labor market. In the fourth piece, the alumnus answer questions about ”life goals” proposed in [8]. This system has obtained the collaboration of about 5% of the university alumni population, graduated in the last 10 years.

In the Department of Computer Science at Federal University of Viçosa, a survey was made on the professional profile of alumni of their undergraduate program[9]. Upon the survey, the computer science program had 357 alumni. The survey obtained the collaboration of 94 alumni, which represents about 26.33% of the total population of these alumni. In this case, the alumni population is relatively small, which made it viable for the research to be conducted through standard means: data was collected through an available form at the department website.

Overall, the traditional methods for obtaining alumni information are only effective for those programs that present a small number of alumni. In case of a larger alumni population, traditional methods may show inaccurate results, since the taken sample might be too small. Furthermore, such methods take weeks or even months to achieve the desired collection. In this work, alumni information is found through means of data publicly available on social networks from the alumni themselves. Since the process is done in a semi-automatic manner, it does not depend on the collaboration of the alumni. Moreover, the proposed method ensures to gather alumni information from a given program within a reduced period of time, depending on the limitations presented in Section III.

B. Web Pages Focused Crawling

Focused crawler aims to crawl web pages that are considered relevant to a specific user interest. There exist several works concerning focused crawling, involving proposed heuristics to such end [10], [11] and classification schemes [12].

Particularly, in [13], the authors propose a heuristic for focused crawling based on genre and content of the information contained on the web pages. Several experiments were conducted in order to demonstrate the effectiveness and efficiency of the proposed heuristic; some experiments were based on the information crawling from a few disciplines of the Computer Science program. The results showed that the proposed heuristic can reach a F1 value [14] greater than 0.92. In [15], the authors improved the heuristic efficiency by considering the link context of web pages.

In [12], the authors show a comparison between focused crawlers based on the SVM, Neural Networks and Naive Bayes classifiers. Among all three crawlers, the one based on SVM was the most effective one. Although experiments show that Naive Bayes is the worst choice within the three classifiers, this is the one that presents the lower cost for generating the classification model.

Unlike the mentioned studies, the proposed method in the present work does not make use of classifiers in order to determine the relevance of a page and does not correspond to a heuristic that can be applied in any context, depending on the user needs on information. The proposed method focuses on retrieving only information of alumni from a social network, thus becoming more efficient and effective within such context.

III. Proposed Method

As previously mentioned, in this section, we describe the proposed method for the construction of our tool aiming gathering from the Web, semi-automatically, information on alumni of a given undergraduate program. Figure 1 presents the functioning architecture of the proposed method.

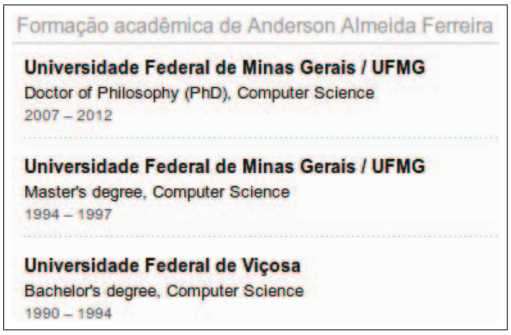
Notice that the method encompasses three main modules and two repositories. The first module, called *Searcher*, aims at searching, from a social network on the web, candidate pages to belonging to alumni from an undergraduate program, through a given available search engine. This module receives, as input, a list of the desired alumni's names. The second module, called *Filter*, aims at filtering, among the candidate pages retrieved by the first module, the ones that are in fact of alumni from an undergraduate program. The third module, named *Extraction*, aims at extracting, from the pages filtered by the second module, the data of the alumni (the alumni data). These data may be academic, professional or personal, depending on what is available within its content. The first repository, called *Pages Repository*, stores the pages from the initial set of samples, being yet incremented as the *Filter* module determines relevant pages of alumni. The second repository, named *Final Database*, corresponds to a database where the data on each alumnus is stored.

This section is organized as it follows. In subsections III-A, III-B and III-C, the *Searcher*, *Filter* and *Extraction* modules, we describe each module in detail.

A. Searcher

The first task, which must be performed by the tool based on the proposed method, is searching for candidate pages for the *Filter* module, which means determining what pages might belong to the alumni set of an undergraduate program. Those pages are retrieved from a social network, through public pages available on the web. In order to do so, the module receives as input the list of alumni's names from an undergraduate program.

We use LinkedIn as the social network for searching professional data on alumni. LinkedIn recently became a powerful professional contact network within the labor market. Nowadays, the network holds over 277 million users around the world, being Brazil the third greater country in number of records on the site, with a little more than 16 million users[6]. That represents around 8% of the total country population and 20% of all Brazilians with Internet access [16]. On average, 35% of LinkedIn users access the website daily and other 32% access it at least once a week. The social network obtains, on average, 2 new members per second [17]; in other words, LinkedIn grows at a fast pace with, approximately, 172.700 new users a day.



Formação acadêmica de Anderson Almeida Ferreira

Universidade Federal de Minas Gerais / UFMG
Doctor of Philosophy (PhD), Computer Science
2007 – 2012

Universidade Federal de Minas Gerais / UFMG
Master's degree, Computer Science
1994 – 1997

Universidade Federal de Viçosa
Bachelor's degree, Computer Science
1990 – 1994

Figure 2: Example of Academic Data from a LinkedIn Page.

Another important LinkedIn feature is the fact that the network offers, in a reduced format, its user pages publicly on the web. That causes different existing search engines index a large number of its pages. In addition, the indexed reduced pages contains several data such as name, professional address and academic degrees.

Figure 2 shows academic data of a LinkedIn user. Notice that, there exist some data that can be irrelevant to the *Filter* module, as for instance the graduation date. However, data concerning the undergraduate program, program degree and institution are relevant and used by the method's *Filter* module.

LinkedIn has an Application Programming Interface (API) in order to search data from its users. This API were not used in this module because of searching limitations. First of all, the API must undergo authentication with a registered user profile on the network; after authenticating, it only allows the user to automatically visit and retrieve data from other users that are separated from such a user by few degrees of separation. Moreover, another limitation of the API is the number of available attributes for consult: the social network is more careful in exposing its data through API that enable extraction; therefore, LinkedIn provide more attributes on network's public pages available on the Web.

Thus, our tool becomes more effective to extract the desired information on users of the social network from the public pages available on the web, instead of consulting the social network itself. There exist several methods used to extract information from pages of social networks [18]. However, in order to retrieve them, one can use the API of a given available search machine. In this work, the API from Google[7], called Custom Search Engine (CSE), was used. Other works have already adopted such a API to perform crawling on the web [19]. Google's search engine holds a massive repository of indexed social network

[6] http://press.linkedin.com/about (as of Apr. 2014)

[7] http://www.google.com

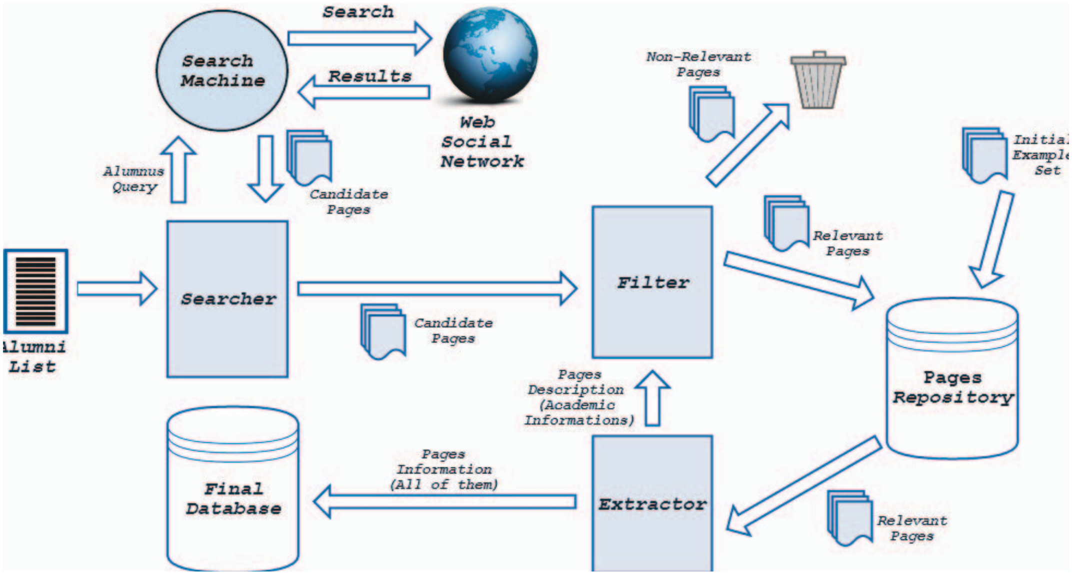


Figure 1: Functioning architecture of the proposed method

pages and that makes it a good tool when performing the assignment proposed on the *Searcher* module. On average, an undergraduate program features hundreds of alumni and, since the module searches for each alumnus, one must search via CSE the same number of alumni, for each undergraduate program. CSE API is limited regarding its usage, offering 100 daily free searches with 100 results per query.

Using the CSE, the *Searcher* module provides to the search engine a combination of the first, middle and last names of a given alumnus. This causes the proposed tool to be robust enough to retrieve user pages who do not use their full names on social networks. However, if we generate a large number of combinations, more results are retrieved by the queries, which might lead to large amount of candidate pages and a bottleneck in the runtime of the module. For instance, to search for the alumnus called ”Gabriel Resende Gonçalves”, the query includes the names ”Gabriel Resende”, ”Gabriel Gonçalves” and ”Gabriel Resende Gonçalves”.

B. Filter

The *Filter* module, the main module of the proposed method, aims at determining the significance of candidate pages, provided by the *Searcher* module, filtering the pages that in indeed belong to the alumni population of the desired program by the user. In order to do so, the module receives, as input, a initial set of sample pages as well as the candidate pages returned by the previous module. *Pages Repository* is started with such initial set of sample pages (small set of alumni pages initially obtained) so that the *Filter* module is able to start the classification, and carry on being incremented as the module recognizes a new page as relevant one, which means it is related to an alumnus from an undergraduate program.

In this module, we calculate the similarity among pages using only their academic data (i.e., data about undergraduate program, attended institution and degree). Thus, the method does not attempt to determine whether a certain candidate page indeed concerns the sought alumnus, but if the page belongs to the population described on the set of sample pages.

There are several classifiers based on supervised machine learning [20] that obtain fine results to page or text classification [21]. Most of them can be used in order to determine whether a candidate page belongs or not to the population set. One of the possible classifiers to be adopted is the *Naive Bayes*, which consists of a method that calculates the probabilities of page instances belonging to the predefine classes on the training sets; those probabilities are, usually, calculate through the Bag of Words model, based on word occurrence counting. Naive Bayes assumes that the terms of each classifying instance is independent of one another, and that it bad for databases which do not exhibit such a feature [22]. Another classifier which could also be used is the SVM that attempts to find the hyper plane that best separates the instances of the training set into smaller subsets. However, since the set of examples is incremental in this work, the cost for generating a new hyper plane, every time an instance is classified, might be high, which makes this an unusable method in terms of time. Nonetheless, another technique that may be adopted to classify candidate pages is the application of a heuristic function that relates a given candidate page to the training set. This function might be, for instance, the cosine similarity function [14], which determines the similarity between two vectors based on their relative opening angle in the vector space; in this case, each vector would be defined on the page terms. For this technique, it is necessary to determine a similarity threshold for checking the similarity among the pages.

Beyond using classifiers and heuristic functions, we may evaluate a candidate page by means of its similarity regarding a set of predefined terms. In this case, the terms must be defined by a specialist and represent different information about a given undergraduate program such as the program and institution names. However, such specification may be a laborious task to a specialist, since these types of information may show many variations and, in order the tool presents a good result, most of them must be predicted. Another difficulty is the fact that the specified terms are not unitary, since an ambiguous term cannot have the same meaning as an unique term. For instance, in the case of the ”Computer Science” program, the unique terms ”Science” and ”Computer” do not have the same meaning of ”Computer Science”. Thus, the values that represent the significance of the terms must be empirically determined, and that also consists of a laborious task.

Hence, for our *Filter* module, we propose a new strategy to select the candidate pages returned by the *Searcher* module. The strategy does not depend on the user for determining the terms, instead, an initial set of positive page examples must be manually defined. The terms are extract from these examples. Moreover, in order to improve the *Filter* module, our proposed strategy requires the specification of a γ value that is the minimum percentage of pages in which a certain term must figure to take in account on the similarity calculation. For instance, if one defines the γ value as 0.15, the term must figure in 15% of the pages in order to be considered.

Usually, a classifier would apply a similarity function in order to relate all terms from a given candidate page with the terms from the page examples. In this work, the terms are separated in three groups: undergraduate program name, institution and program degree. The heuristic function is applied separately for each of the three types of terms and the final result is given by an average of the three obtained similarities.

The threshold used to determine whether a candidate page concerns an alumnus of a given program consists in the lower value obtained by applying the similarity function to determine the resemblance of each page of the set of sample pages with the other ones since, in the such set, all pages belong to the desired alumni.

The selected similarity function to build the relation between candidate pages given by the *Searcher* module and the set of page examples was *Cosine Similarity* [15]. Preliminary experiments were conducted using other functions, such as Jaccard Distance function; however, the results were not satisfactory.

In order to calculate the *Cosine Similarity*, it is necessary to create a n-dimensional vector for each term type (undergraduate program, institution and degree), where n is the number of terms with frequency over the pre established γ value. After defining the vector, calculation is made according to Equation 1.

$$Cosine = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

where A is the vector of terms from the page examples and B is the vector of terms from the evaluated candidate page.

C. Extraction

The *Extraction* module is responsible for extracting, from the HTML pages returned by the *Filter* module, relevant information of alumni from an undergraduate program. In this work, since the page structures from a same social network (LinkedIn) are single, i.e., all pages feature the same pattern, our strategy for extracting the data is based on regular expression [23] manually defined. Table I shows the information along with their respective regular expressions.

This module extracts academic, professional and/or personal information from the filtered pages by the *Filter* module. In this work, according to the architecture shown in Figure 1, this information have two purposes: the first is the storage on the *Final Database* repository, in order to further analyzing the obtained results (in this case, all information is stored); the second is defining and enhancing the description of the set of sample pages used by the *Filter* module (in this case, only academic information is required).

IV. Experiments

In this section, we describe the experiments and discuss the results, in order to validate the proposed method for constructing the tool that gathers on the Web, semi-automatically, professional data on the alumni of an undergraduate program. The section is organized as follows. In Subsection IV-A, we present experiment setup: alumni lists, baseline and evaluation metrics. In Subsection IV-B, we present and discuss the obtained results.

A. Experimental Setup

In this subsection, we describe the experimental setup and inputs of our method. The experiments for validating our proposed method of semi-automatic gathering professional data on alumni of an undergraduate program (see Figure 1) consisted on the conduction, for each alumni list, of 10 distinct runs and, for each run, a random set of 15 initial page examples are used as training examples.

Table I: Regular Expressions Used by *Extraction* Module

Information	Regular Expressions
Given Name	<span class=”givenname”>(.*?)</span>
Family Name	<span class=”familyname”>(.*?)</span>
Jobs	<li>\n(.+)\n<span.*at.>.+</span>\n(?:<a.+><.+summary”>)?(.+)
Currently Job Location	<.*locality.*>\n*(.*?),.*\n
Programs	<span.*major.*>\n(.*)</span>
Program’s Degrees	<span.*degree.*>(.*)</span>
Schools	<div.*education.*>\n<h3.*summary fn org.*>\n(.+)\n</h3>

Alumni Lists

As previously mentioned, our proposed method receive, as input, a list of alumni’s names from an undergraduate program. We perform experiments with five alumni lists, available on the web, concerning the following undergraduate programs: Computer Science of the Federal University of Minas Gerais (UFMG) [8], Metallurgical Engineering of the Federal University of Ouro Preto (UFOP) [9], Chemistry of the University of So Paulo (USP) [10], Computer Science of the USP [11], and Computer Science of the Catholic Pontifical University of Paran (PUC-PR) [12]. Table II shows the number of alumni available in each list.

Table II: Population size of each alumni list.

Alumni list	Population size
UFMG	1,542
UFOP	1,579
USP - Comp. Sci.	1,259
USP - Chemistry	900
PUC-PR	812

Baseline

In this work, for evaluating our proposed method that semi-automatically gathers alumni information of an undergraduate program, we compare it with the *Naive Bayes* classifier. Naive Bayes generates a classification model with low cost, since it only counts the term occurrences according to the *Bag of Words* model. Naive Bayes considers the program, institution and degree terms mutually independent. Furthermore, others classifiers such as *SVM* and *Neural Network* need for training set with, at least, two training sets. Our proposed method uses only a set of pages with true examples.

Evaluation Metrics

Typically, works related to the information retrieval area are evaluated by precision, recall and F-mean metrics [14]. However, for this work, precision and the number of retrieved relevant pages were adopted. Precision can be calculated as demonstrated in Equation 2. The reason why the recall value is not used was the impossibility of obtaining the exact number of alumni pages available on a social network.

$$Precision = \frac{RelevantPages \cap RetrievedPages}{RetrievedPages} \qquad (2)$$

B. *Experimental Evaluation*

γ Determination

As previously mentioned, in order to improve the process of filtering candidate pages performed by the *Filter* module, the strategy to such filtering requires the specification of a γ value, which corresponds to the minimum percentage of pages in which a given term must feature. Therefore, initially, several tests were conducted in order to determine the best γ value to be used in our experimental evaluation. The results are the average value obtained for each γ value on 10 runs.

Figure 3 shows, for different γ values, the precisions and the number of relevant pages obtained by our method varying the γ value from 0 to 1. Notice that a great value for γ was 0.2 since, for this value, a great number of relevant pages was obtained, maintaining precision at a satisfactory level. Thus, we use such γ value in our experimental evaluation, i.e., only the terms, that features at least 20% of the set of page examples, are considered.

Experimental Results

Table III shows, for each alumni list, the most frequent terms encountered on the performed experiments. This table shows the terms concern, for each alumni list, the test that obtained the greater precision between all 10 runs.

Table IV shows the results gathered when performing experiments considering, respectively, our proposed method and the baseline. This table contains the average precision and the average number of found relevant pages, for each alumni list, considering a 99% confidence interval.

Notice that the baseline has retrieved a greater number of pages; however, the precision obtained is much inferior when in comparison with our proposed method. Considering all alumni lists, our proposed method obtained the average precision from 0.83 to 0.91, while the best result for the baseline was 0.65 on average precision. Unlike our proposed

[8] http://dcc.ufmg.br/dcc/index.php?option=com_content&view=article&id=274&Itemid=8; (as of Apr. 2014)

[9] http://www.em.ufop.br/exalunos; (as of Apr. 2014)

[10] http://www.iqsc.usp.br/acad1/egressos/app/graduacao/lista/index; (as of Apr. 2014)

[11] http://www.ime.usp.br/c̃gmac/ex-alunos/res.html; (as of Apr. 2014)

[12] http://www.pucpr.br/graduacao/cienciacomputacao/egressos_curso.php; (as of Apr. 2014)

Table III: Most frequently terms for γ = 0.2 using cosine similarity function.

Alumni list	Program	School	Degree
UFMG	”ciencia da computacao” ”computer”	”ufmg”, ”universidade federal de minas gerais”	”ma”,”bachelor” ”bacharel”, ”bs”
UFOP	”engenharia”,”metalurgia”	”universidade federal” de ouro preto”	”engenheiro”
USP - Comp. Sci.	”ciencia da computacao” ”computer”	”usp” ”universidade”	”bachelor” ”master”
USP - Chemistry	”quimica”	”usp”, ”universidade	”bacharel”, ”doutor”
PUC-PR	”bacharel”	”ciencia da computacao”	”pontificia universidade” catolica do parana”, ”puc”

Table IV: Number of Pages Retrieved and Precision Results For Proposed Method and Baseline.

	Pages Retrieved		Precision	
Alumni list	Proposed Method	Baseline	Proposed Method	Baseline
UFMG	127 ± 6.04	381 ± 85.01	0.919 ± 0.01	0.368 ± 0.08
UFOP	85 ± 3.86	242 ± 64.21	0.839 ± 0.01	0.554 ± 0.06
USP - Comp. Sci.	155 ± 23.13	237 ± 89.97	0.86 ± 0.02	0.658 ± 0.14
USP - Chemistry	70 ± 6.297	129 ± 78.087	0.85 ± 0.02	0.593 ± 0.21
PUC-PR	34 ± 6.29	70 ± 57.28	0.893 ± 0.09	0.508 ± 0.13

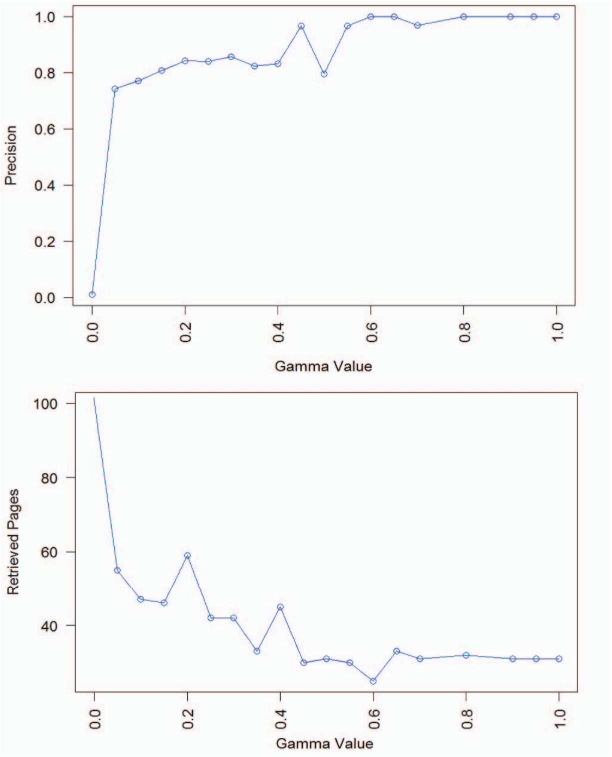


Figure 3: Results of γ Variation on Precision (Top) and Number of Retrieved Pages (Bottom).

method, the baseline results were too sensitive towards the initial set of training pages; therefore, the experiments showed significant variation on the results. That can be observed through means of the confidence interval on Table IV.

Estimating the coverage of LinkedIn

As we may not determine the recall value on a domain where the exact number of alumni pages is unknown, we perform another experiment for estimating the percentage of true alumni pages on LinkedIn. We manually search for alumni pages that are not found by our method in our experimental evaluation. For each alumni list, 50 names was randomly chosen. The Table V shows the percentage of pages that are found in this search. We can notice that, except for the experiment of Metallurgical Engineering from UFOP, a few alumni, who have pages on LinkedIn, were not found by our method. For Metallurgical Engineering from UFOP, the diversity in the course name filled in the pages leads to a poor performance.

Table V: Percentage of alumni pages found in the manual Search.

Alumni list	Percentage of pages
UFMG	0
UFOP	0.22
USP - Comp. Sci.	0.14
USP - Chemistry	0.10
PUC-PR	0.12

Descriptive analysis of the results

After gathering information on alumni of the 5 undergraduate programs considered in the performed experiments, we may inserted the data into the final repository (*Final Database*) and analyze their information. The repository was populated based on the test that showed greater precision for each one of the undergraduate programs.

Figure 4 shows the alumni distribution per year of graduation regarding the alumni of the 5 undergraduate programs adopted. Distributions regarding alumni from all programs were similar. One can observe the high concentration of alumni in the years from 2000 to 2010. That can be explained by the fact that such alumni, newly graduated, are looking for new jobs. Hence, many seek to promote their professional resume on the Web and LinkedIn is a great social network for such purpose. Another relevant factor is the fact that LinkedIn network was created in the year of 2002; moreover, people who joined the labor market before 2003 could, at that time, not have interest and/or ease to engage in social networks.

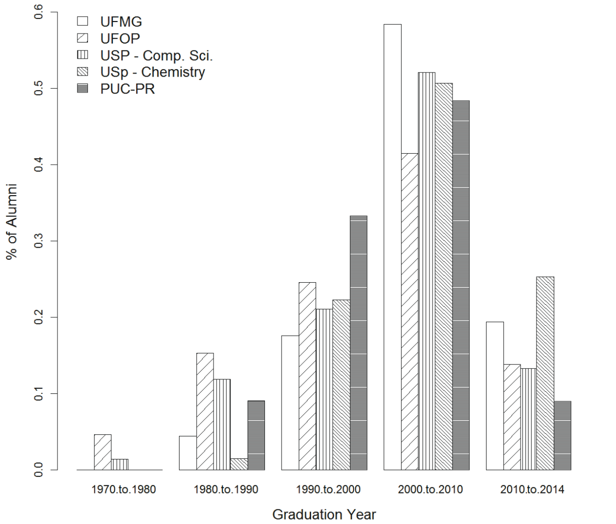


Figure 4: Graduation year graphic of all experiments.

For exemplification purposes, Figures 5 and 6 show location of the current employment of the alumni of all 5 undergraduate programs considered in the performed experiments. Images were generated with the aid of the maps API from Google Maps[13]. We can notice, based on Figures 5 and 6, a high concentration of alumni in the city of origin of their program and, with the exception of alumni in Computer Science from USP, all other ones are located, entirely, in the American continent.

V. Conclusion

Our proposed method for semi-automatic gathering professional data on alumni of an undergraduate program makes use of a list of alumni names and an initial set of relevant pages. At first, the method automatically performs searches for pages within a social network using a search machine, which are candidates to belonging to the alumni. Subsequently, the method uses a proposed strategy, based on similarity metrics, in order to determine the relevance of the candidate pages and, thus, retrieve only the ones that concern the alumni of the desired program.

Once prospective tool users, made from the proposed method, belong to teaching staffs of undergraduate programs, acquiring the necessary list of alumni names is not a problem. Obtaining the initial set of sample pages might come to be the greater complicating factor in the method's functioning since, in any social network, manually acquiring some pages may be a laborious task, even in small numbers.

In our experimental evaluation, considering alumni lists from five different undergraduate programs, our proposed method was capable of gathering information on them, within the social network LinkedIn, with satisfactory precision. The experiments show that the proposed method was able to find a great number of alumni pages. For undergraduate programs with over 1.000 alumni, the method was able to find, on average, 7.5% of alumni. Particularly, for alumni in Computer Science from USP, the method was able to find 12.2%. By using conventional methods for gathering information about alumni of undergraduate programs with a high number of alumni, the feedback rate is around 6% of the alumni population [5]. Moreover, conventional methods usually take days or even months to obtain such results, while our proposed method performs in few minutes, depending on limitations of the search engine API.

As future work, we intend to improve our method for gathering alumni pages from the web without providing an initial set of page examples. Another important work is the proposal of a strategy that enables the non necessity of a list of alumni names, in order to not restrict the use of the tool to only the users who have access to the this list of a certain undergraduate program. Furthermore, we intend to experiment the method on other purposes, such as retrieving relevant pages on a specific topic or individuals from a given group that is not from the academic sphere.

Acknowledgment

This research is partially funded by Fapemig (Foundation for Research Support of the State of Minas Gerais). Furthermore, this research was carried out on the GAID/UFOP Laboratory.

References

[1] A. C. Z. Lousada and G. d. A. Martins, "Alumni as a source of information to management accounting courses (in portuguese)," *Journal of Accounting and Finance*, vol. 16, no. 37, pp. 73–84, 2005.

[2] N. B. Ellison *et al.*, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[3] M. D. Pena, "Alumni monitoring: Conceptual analysis and its application in brazilian educational context (in portuguese)," *Technological Education of Belo Horizonte*, vol. 5, no. 2, pp. 25–30, 2000.

[4] L. S. Michelan, C. A. Harger, G. Ehrhardt, and R. P. O. Moé, "Alumni management in higher education institutions: Possibilities and potential (in portuguese)," *IX International Colloquium on University Management in South America*, 2011.

[13] http://maps.google.com

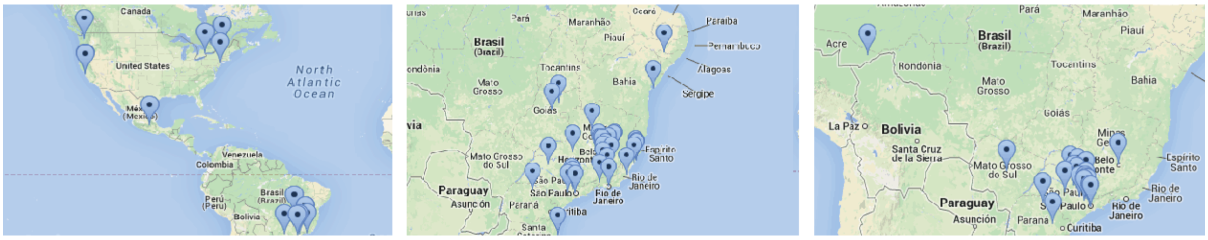


Figure 5: Alumni localization. UFMG on left, UFOP-Met. Eng. on center and USP-Chemistry on right.

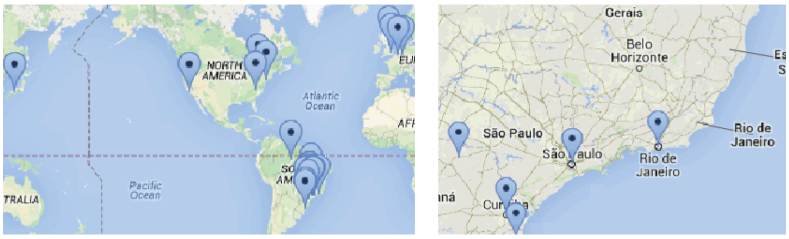


Figure 6: Alumni localization. Usp-Comp. Sci. on left and PUC-PR on right.

[5] J. M. Silva, R. d. S. Nunes, and A. d. L. Jacobsen, "The alumni monitoring program of the Federal University of Santa Catarina: The profile of students definition in the period 1970-2011 (in portuguese)," *IX International Colloquium on University Management in South America*, 2011.

[6] R. P. Morgado, C. G. Geroto, and A. C. G. Ramalho, "Course evaluation and professional status of the environmental management program ESALQ/USP (in portuguese)," *Electronic Journal of Master in Environmental Education*, vol. 27, 2013.

[7] U. O. Media, "Research reveals alumni profile of USP courses (in portuguese)," Retrieved April 12, 2014, from http://www.usp.br/imprensa/?p=31718, 2013.

[8] E. L. Deci and R. M. Ryan, *Self-Determination*. Wiley Online Library, 2010.

[9] J. L. Imbrizi, F. G; Filfo, "Alumni research - conclusive review (in portuguese)," Retrieved April 12, 2014, from http://www.dpi.ufv.br/arquivos/diversos/pesqegressos.pdf, 2003.

[10] G. Pant, K. Tsioutsiouliklis, J. Johnson, and C. L. Giles, "Panorama: Extending Digital Libraries with Topical Crawlers," in *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, Tuscon, AZ, USA, 2004, pp. 142–150.

[11] P. Srinivasan, F. Menczer, and G. Pant, "A General Evaluation Framework for Topical Crawlers," *Information Retrieval*, vol. 8, no. 3, pp. 417–447, 2005.

[12] G. Pant and P. Srinivasan, "Learning to Crawl: Comparing Classification Schemes," *ACM Transactions on Information Systems*, vol. 23, no. 4, pp. 430–462, 2005.

[13] G. T. De Assis, A. H. Laender, M. A. Gonçalves, and A. S. Da Silva, "A genre-aware approach to focused crawling," *World Wide Web*, vol. 12, no. 3, pp. 285–319, 2009.

[14] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[15] V. Mangaravite, G. T. Assis, and A. A. Ferreira, "Improving the efficiency of a genre-aware approach to focused crawling based on link context," in *Proceedings of the Eighth Latin American Web Congress (LA-WEB)*. IEEE, 2012, pp. 17–23.

[16] IBGE-Brazil, "Internet access and possession of mobile cell phone(in portuguese)," Retrieved April 19, 2014, from http://loja.ibge.gov.br/pnad-2011-sintese-dos-indicadores.html.

[17] L. Rao, "Linkedin now adding two new members every second," Retrieved April 19, 2014, from http://techcrunch.com/2011/08/04/linkedin-now-adding-two-new-members-every-second.

[18] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: a survey," *arXiv preprint arXiv:1207.0246*, 2013.

[19] M. Allauddin and F. Azam, "Service crawling using google custom search api," *International Journal of Computer Applications*, vol. 34, no. 7, 2011.

[20] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques." *Informatica (03505596)*, vol. 31, no. 3, 2007.

[21] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Machine learning: ECML-98*. Springer, 1998, pp. 4–15.

[22] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[23] V. Alfred, "Algorithms for finding patterns in strings," *Handbook of Theoretical Computer Science: Algorithms and complexity*, vol. 1, p. 255, 1990.