Machine Learned Job Recommendation

Ioannis Paparrizos
EPFL
Lausanne, Switzerland
ioannis.paparrizos@epfl.ch

B. Barla Cambazoglu
Yahoo! Research
Barcelona, Spain
barla@yahoo-inc.com

Aristides Gionis
Yahoo! Research
Barcelona, Spain
gionis@yahoo-inc.com

ABSTRACT

We address the problem of recommending suitable jobs to people who are seeking a new job. We formulate this recommendation problem as a supervised machine learning problem. Our technique exploits all past job transitions as well as the data associated with employees and institutions to predict an employee's next job transition. We train a machine learning model using a large number of job transitions extracted from the publicly available employee profiles in the Web. Experiments show that job transitions can be accurately predicted, significantly improving over a baseline that always predicts the most frequent institution in the data.

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General

General Terms

Design, Experimentation, Performance

Keywords

Job recommendation, job transition, employee, institution

1. INTRODUCTION

The highly competitive and dynamic nature of the job market as well as personal preferences and goals lead individuals to change their jobs at some point in their lives. Moving to a new job, however, is not an easy decision, which may depend on many factors, such as salary, job description, and geographical location. Making successful job transitions is essential for a successful professional career.

In this work, we build an automated system that can recommend jobs to people based on their past job histories in order to facilitate the process of selecting a new job. We believe that such a system can successfully exploit the job transitions performed by other employees. That is, we propose recommending jobs to people based on inference from the job transition patterns observed in the past. These patterns may involve features extracted from the business profiles of employees (e.g., years of experience, educational degree, job title), the profiles of institutions[1] (e.g., industry, type, size), and the job transitions themselves (e.g., frequency of transitions between jobs, average time spent in a job).

The framework we propose is based on supervised machine learning. Given an employee's past job history, the objective of the learning model is to accurately predict the next institution that the employee will move to. The predicted institution can then be recommended to the employee as the next step in his/her career.

To evaluate our framework, we use a large sample of job transitions extracted from the publicly available employee profiles in the Web. From this sample, we extract a number of features that we use to train and test our machine learning model. The results of our experiments demonstrate that the transition of an employee to an institution can be quite accurately predicted, significantly improving over a baseline predictor that always predicts the most frequent institution in the data. Our results indicate that the most important feature in predicting a job transition is the current institution of the employee.

The rest of the paper is organized as follows. Section 2 gives an overview of the dataset we use. We describe the technical details of our work in Sections 3 and 4. Section 5 contains the details of our experimental setup. In Section 6, we report the performance results. We survey the related work in Section 7. Finally, the paper is concluded in Section 8, including pointers to future work.

2. DATASET

We use a large sample of job transitions and associated meta-data extracted from the publicly available employee profiles (about five million) in the Web. The profiles contain information about the employees' professional and educational experiences. Each profile is composed of three sections. The first section contains personal information about the employee, such as the first name, last name, and geolocation. The second section contains information about the current and past professional positions held by the employee. This section includes company names, positions, company descriptions, job start dates, and job finish dates. The company description field may further contain information about the company (e.g., the number of employees and industry). The third section contains information about the current

[1]In our context, the term institution refers to private or public companies as well as educational or research institutions.

RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.

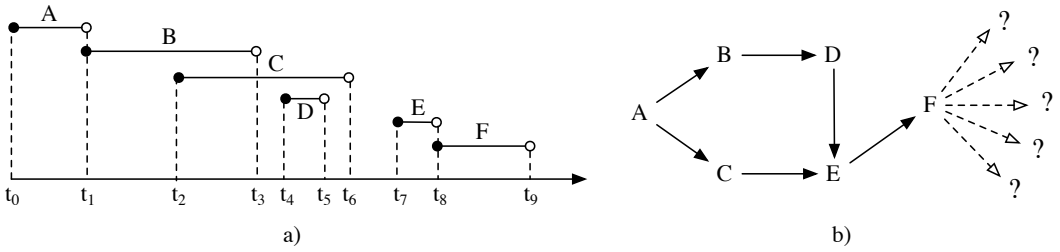


Figure 1: a) Past job transitions of an employee and b) the corresponding job transition graph.

Table 1: Dataset statistics

Description	Value
Number of profiles	5,298,912
Number of unique company affiliations	1,278,240
Number of unique university affiliations	195,849
Average number of company affiliations	3.04
Average number of university affiliations	1.27

and past educational experiences of the employee, such as university names, degrees, fields of education, start and finish dates. Table 1 shows various global statistics computed over the entire dataset. The number of unique company and university affiliations are reported after some cleansing and ignoring those that occur only once in the data.

3. JOB TRANSITION GRAPH

A job transition refers to the movement of an employee from one job to another. Although our dataset contains a lot of information about employees, there is no direct information regarding job transitions. In principle, transitions can be constructed using the start and end dates of employment. However, in practice, there are two issues that complicate the formation of job transitions. First, a person may be employed by many institutions at the same time. Second, there may be periods during which a person is not affiliated with any institution. Therefore, we define a set of rules to extract job transitions based on employment dates.

We represent transitions of an employee as a directed graph $(\mathcal{I}, \mathcal{T})$, which we refer to as a job transition graph. The set of nodes $\mathcal{I}$ represents institutions, and the set of edges $\mathcal{T}$ represents transitions between institutions. A directed edge $(u \rightarrow v) \in \mathcal{T}$ denotes a transition from institution $u \in \mathcal{I}$ to institution $v \in \mathcal{I}$ for a specific employee. The transition graph enables effective representation of transitions as well as efficient computation of some features that are used by our prediction model.

We explain the technique we used to create the job transition graph over a simple example that involves a single employee. Fig. 1(a) shows the time intervals a person was employed in six different institutions, namely A, B, C, D, E, and F. The entire employment interval for the person is $[t_0, t_9]$. During the period $[t_2, t_6]$, the person was affiliated with more than one institution while he/she was not employed during the period $[t_6, t_7]$. For each institution X that has employed the person, we have a start time $s(X)$ and an end time $e(X)$. Hence, in our example, $s(A) = t_0$ and $e(A) = t_1$.

Our definition of the job transition graph assumes that a transition $(u \rightarrow v)$ takes places if and only if the following two conditions are both met:

- The end time $e(u)$ of an employee at institution u should be before the start time $s(v)$ of the employee at institution v, i.e., $e(u) \leq s(v)$.
- There is no institution w such that $s(w) > e(u)$ and $e(w) < s(v)$.

The two conditions given above imply that any transition to a new institution can take place only from the most recent previous institution(s).

In Fig. 1(b), we show the job transition graph that corresponds to the job history shown in Fig. 1(a). Note that, in the graph, the transition to institution E is represented by both $(C \rightarrow E)$ and $(D \rightarrow E)$. The transition $(F \rightarrow E)$ is omitted from the graph because the first condition is not satisfied (i.e., $t_f(F) > t_s(E)$). Transitions $(A \rightarrow E)$ and $(B \rightarrow E)$ are omitted because the second condition is not satisfied (i.e., $s(D) > e(A)$ and $s(D) > e(B)$).

For each employee profile, we use the following two-phase algorithm to create a part of the job transition graph, respecting the two conditions mentioned above.

Phase 1. (creation of transitions): Take the node with the maximum start date (referred to as the max node) and put it in the subgraph. Iterate over all other nodes (referred to as node) and compare their end date with the start date of the max node. If the end time is before the start time of the max node, add a directed edge from the node to the max node. Otherwise, check if the edge from the max node to the node is needed and if so place the edge. Remove the max node of the list. Get the next max node and repeat the same procedure.

Phase 2. (removal of redundant transitions): Take the node with the minimum start date (referred to as the min node). Iterate over all other nodes and check if there is a link from the min node to the node. If there is, remove it from the graph and check if there is any other shortest path from the min node to the node. If there is such a path, terminate the process. Otherwise, put the edge back in the graph.

4. PREDICTION MODEL

The prediction problem we address in this work is the following. Given an individual who is currently employed in an institution, we want to predict the next institution where the individual will be employed. If the accuracy of such predictions is sufficiently high, we may use our model to recommend institutions to employees who are seeking jobs.

We build such a predictive model by supervised machine learning. An instance in our learning model corresponds to

Table 2: Features used by the prediction model

Feature type	Feature	Range
institution	`company title`	String
	`industry`	String
	`company type`	{public, private}
	`number of employees`	$\mathbb{Z}$
employee	`number of jobs`	$\mathbb{Z}$
	`position title`	String
	`best position title`	String
	`years of experience`	$\mathbb{Z}$
	`number of universities`	$\mathbb{Z}$
	`educational degree`	String

Table 3: Setups used in the experiments

Setup	Data sample	Set sizes	
		Train set	Test set
`I`	Top 100 universities + top 100 companies	65,622	28,124
`II`	Top 100 companies	52,142	22,346
`III`	Top 25 companies	45,891	19,668

a person who is employed in an institution. Each instance is represented by a set of features extracted from the data. In our learning model, the class labels that we try to predict correspond to the institutions in the data. Table 2 demonstrates the features used in our prediction model.

5. EXPERIMENTAL SETUP

To train and evaluate our machine learning model, we used the Weka machine-learning toolkit.[2] We experimented with several machine-learning algorithms, but we present here the results only for the decision table/naïve Bayes hybrid classifier (DTNB) [6], which achieves the highest accuracy. A brief description of the DTNB algorithm is as follows. The set of attributes is split into two subsets by evaluating a gain function. One subset of attributes is used to build a decision table and the other to build a naïve Bayes model. Consequently, the algorithm employs a forward-selection search process, where at each step, a set of selected attributes are modeled by naïve Bayes and the rest by the decision table. At each step, the algorithm considers dropping entirely a feature from the model.

We repeat our experiments for three different setups, each using a different sample from our data. In all setups, the predicted class is an institution among the most frequent 25 companies in the full data. However, in each setup, we pick the instances where the current employee position is limited to a specific set of companies or universities. In setup `I`, the current position is either in one of the most frequent 100 universities or in one of the most frequent 100 companies. In setup `II`, the current position is in one of the most frequent 100 companies. In setup `III`, the current position is in one of the most frequent 25 companies, i.e., those that also form the set of class labels. Train and test set sizes for these three setups are shown in Table 3.

[2] `http://www.cs.waikato.ac.nz/ml/weka/`

Table 4: Prediction accuracy

Setup	Accuracy (%)		
	Baseline	DTNB	Difference
`I`	15.21	66.78	51.57
`II`	15.40	78.26	62.86
`III`	15.97	86.09	70.12

Figure 2: Feature importance.

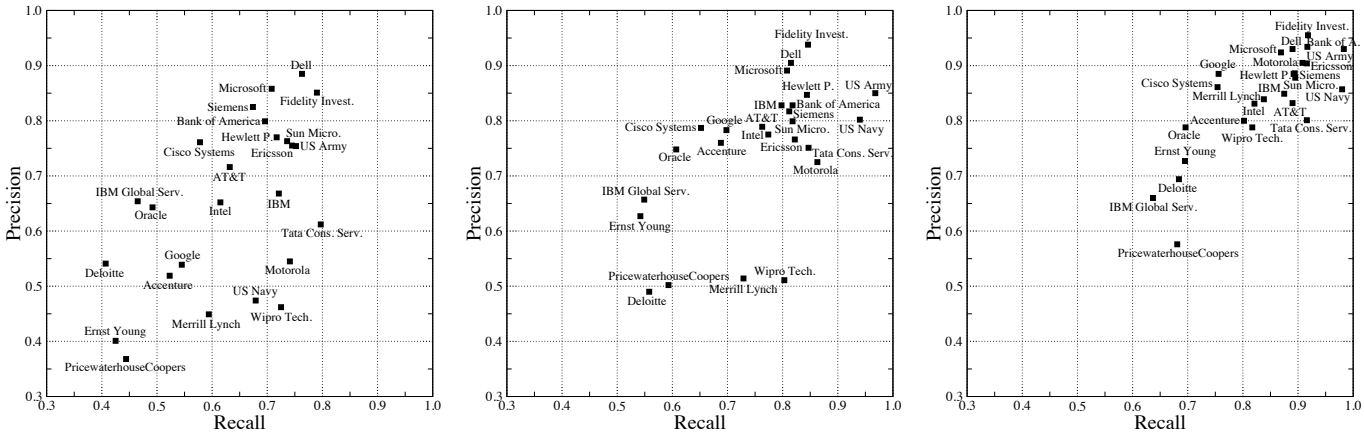
6. EXPERIMENTAL RESULTS

Table 4 shows the accuracy values achieved by our prediction model for the three different setups: 66.8%, 78.2%, and 86.0%, for setups `I`, `II`, and `III`, respectively. For each setup, we consider a simple baseline that predicts always the majority class in the train data. In all cases, the accuracy of the baseline predictor is around 15%.

Fig. 2 shows the relative importance values for features (the values are computed by the χ^2 statistics). We see that company name and the industry are relatively more important features. This may indicate that a large number of people follow similar career paths.

In Fig. 3, we present scatter plots that show the precision and recall values for every predicted class, i.e., the most frequent 25 companies in our data. The results confirm the intuition that `I` forms a more difficult prediction problem than `II`, which in turn is a more difficult case than `III`. Setup `I` is the most difficult case because it contains transitions from universities to companies. The vast majority of those transitions correspond to the very first jobs of people after graduation and are more difficult to predict. On the other hand, it is easier to predict a subsequent transition of a person once this person has already established an employment history in a company.

During our study, we performed a large number of experiments with many parameters and configurations, excluded due to the lack of space. For example, we tried to include information about the position that somebody would get in the next institution (to suggest institutions for a target job position) and the results were only slightly better. We also performed experiments with setups similar to `II` and `III`, but limiting the predicted institution to the current industry of the employee. Our experiments showed that we can achieve comparable accuracies.



(a) Setup I (b) Setup II (c) Setup III

Figure 3: Precision and recall for different setups.

7. RELATED WORK

The study of methods for job search and factors that determine the success of individuals in a labor market is an extensive topic of research in the area of labor economics. In this area, the analysis of social networks has played an important role in analyzing how people are searching for jobs. For instance, in his seminal work, Granovetter showed that weak ties are superior to strong ties for providing support in getting a job [5]. This evidence has led to a number of theoretical models in economics that explore the importance of social networks in labor markets (e.g., Calvó-Armengol [1], Calvó-Armengol and Zenou [2], Galeotti and Merlino [4]).

Another well investigated scenario in assigning jobs to individuals is via the theory of matching [3, 7, 9]. In this scenario, many applicants rank a set of available job positions in terms of preferences, and similarly the employers in those positions rank the applicants. Then, an optimal matching is sought, in various technical notions of optimality. This setting is obviously very different than the problem we study in this paper.

More similar to our work, the problem of recommending jobs to individuals (as well as the dual problem of recommending applicants for job profiles) was studied by Malinowski et al. [8], who propose to learn a probabilistic model that estimates the probability that an applicant likes a job. This approach differs from our work in that it requires more information. It is assumed that not only applicant profiles are available, but also job opening profiles, and the goal is to match applicant to job opening profiles. In our work, we do not assume that we know the job opening profiles, instead we recommend job positions to applicants based only on the previous job history of a number of other employees. Another difference is that through the machine learning approach and the prediction methodology we follow, we offer a quantitative way of evaluation.

Finally, a somewhat related concept is that of churn prediction [10]. Here, the challenge is to predict when a customer (and not an employee) will leave a company and possibly sign up with a competitor company. The churn prediction scenario is orthogonal to the problem we focus on, where we try to predict not when someone will leave, but what would be the next move.

8. CONCLUSIONS AND FUTURE WORK

We studied the problem of recommending jobs to people who are seeking a new job. We developed a supervised learning model that let us predict the next job transition of a person and recommend jobs to people. We trained a prediction model using a large number of job transitions obtained from the Web, and we found that a relatively high accuracy can be obtained relative to a simple baseline predictor.

As a future work, we plan to extend our work with more features and compare it against stronger baselines. We also intend to study the influence of social aspects. In particular, we aim to build models that exploit the job transition patterns of the people in the social circle of an employee (e.g., friends and friends of friends).

9. REFERENCES

[1] A. Calvó-Armengol. Job contact networks. *Journal of Economic Theory*, 115, 2004.

[2] A. Calvó-Armengol and Y. Zenou. Job matching, social network and word-of-mouth communication. Technical Report 771, IZA, 2003.

[3] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *American Mathematical Monthly*, 69(1), 1962.

[4] A. Galeotti and L. P. Merlino. Endogenous job contact networks, 2010.

[5] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6), 1973.

[6] M. Hall and E. Frank. Combining naive Bayes and decision tables. In *FLAIRS*, 2008.

[7] R. Irving. Matching medical students to pairs of hospitals: a new variation on a well-known theme. In *ESA*, 1998.

[8] J. Malinowski, T. Keim, O. Wendt, and T. Weitzel. Matching people and jobs: a bilateral recommendation approach. In *HICSS*, 2006.

[9] A. E. Roth. The economics of matching: stability and incentives. *Mathematics of Operations Research*, 1982.

[10] C. Wei and I.-T. Chiu. Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, 23, 2002.