Modeling Career Path Trajectories

David Mimno and Andrew McCallum
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA
{mimno,mccallum}@cs.umass.edu

January 19, 2008

Abstract

Understanding the structure and dynamics of the job market is important both from the local perspective of individual job hunters and from the global perspective of economists and policy makers. In this paper, we explore such questions by analyzing the text of a corpus of resumes and their job transitions. We first demonstrate the use of a statistical topic model to discover the latent skills that make up each job description and to map the cooccurrence patters of various skills. Although these topical features alone are good at discovering the structure of the job market, they are relatively poor at predicting job transitions. We next present a topical sequence model trained on topic features that has improved ability to predict subsequent job titles.

1 Introduction

Understanding the structure and dynamics of the job market is important from a variety of perspectives. Individuals are clearly very much concerned with establishing good career paths. Knowing what skills and combinations of skills are valued in various positions is very valuable. Understanding how to plan a career and seek positions that will lead to desirable career outcomes is another vital capacity. Institutions and policy makers should also understand patterns and trends in the job market in order to set policy and focus training resources where they can be most effective.

This paper presents the problem of career path modeling: predicting subsequent positions given previous work experience. We present a topical

sequence model that constructs a low-dimensional representation of the job market and learns a hidden state model that predicts job transitions.

Our first goal is to provide a tool that can analyze the static structure of jobs within businesses. Specifically, we are interested in learning about the duties, responsibilities, and technical skills that make up jobs. We are also interested in learning in what ways people interact within organizations. Our second goal is to examine career paths over time. Having a model of the probability of various career path transitions could, for example, support a career counseling application that would find job opportunities that maximize the probability of some stated career goal.

The data for this study consists of 9722 resumes. Each resume contains some number of records describing previous work experience. There are 54,549 such records in total, containing 2,383,402 words after the removal of stopwords.

Each job description is labeled with a job title. These titles are useful, but extremely noisy. The first problem is that there is little standardization in job titles, so they tend to be very sparse, exhibiting the common "long tail" phenomenon. Of the 28,828 distinct job titles in the corpus after lower casing and removing non-letter characters, only 536 or 1.9% appear 10 or more times. In contrast, 85% of distinct job titles appear only once. Additionally, only 33.5% of job descriptions have one of the frequent titles that appear 10 or more times. The second problem is that frequent job titles are often vague. For example, the most common job title is Consultant, a title that can cover a wide range of duties.

We propose a topical sequence model, which learns hidden states that comprise career path trajectories based on a low dimensional representation of the components of job description text produced by an off-the-shelf topic model. We evaluate several models for predicting subsequent job titles. The topical sequence model produces better predictive likelihood than topics alone and the previous job title. In addition, the topical sequence model discovers coherent hidden states that have meaningful topic transition probabilities. The hidden states provide an alternative to the extremely sparse job titles.

2 Topical Components of Resumes

The topical sequence model depends on the dimensionality reduction provided by a statistical topic model [1, 2]. These models are capable of learning underlying hidden topical components in the presence of polysemy (words

with multiple meanings) and synonymy (multiple words with the same meaning). The topic model effectively breaks the language of job descriptions down into distinct components. The low dimensional topic representation allows us to learn a sequence model for job transitions more efficiently, but it is also interesting and useful in its own right.

Each job description in a work experience record within a resume contains words that describe duties, responsibilities, and accomplishments. Jobs are often combinations of distinct sets of such skills and responsibilities. Moreover, within an organization it is common for interactions between employees to focus on specific topics and functions. We apply a statistical topic model to the problem of discovering both the clusters of duties performed by employees and the interactions between employees.

We train a latent Dirichlet allocation model [1] on each job description using 200 hidden topics. Examples of the most probable words for several topics are shown in Table 1. The job titles with the highest average number of words in this topic are shown underneath each list of topical words. The first topic in this table, *responsible, maintaining, included*, is one of the most common topics. The job titles associated with this topic are the most common titles in the corpus. This topic represents the language common to most job descriptions. The next four topics are more interesting. All four prominently include language about management, but they distinguish between several aspects of management. The first relates to technology project management. The second and third topics are more similar, both relating to staff and training, but the second topic appears to involve more direct interaction (*supervised, maintained*) while the third involves higher-level manangement (*operations, planning*). The job titles associated with these topics (Office Manager vs. Operations Manager) reinforce this impression. The fourth management topic includes words about high level strategic planning.

The next two topics distinguish between financial topics, separating accounting from trading and investment. Many of the topics in this corpus relate to technology and specific skill sets. The next five topics distinguish between software development in general, Oracle and UNIX development, Microsoft based development, Java web application development, and web design. Except for the web topic, the job titles associated with these topics are largely the same. The final four topics demonstrate some of the range of topics discovered in the corpus of resumes.

The topic model discovers distinct clusters of skills and responsibilities. We are also interested in the cooccurrence patterns between topics. For each pair of topics t_1 and t_2, we calculate the mutual information between the event that at least one word in a document is assigned to topic t_1 and the

Table 1: Examples from 200 topics discovered by the statistical topic model.

---

responsible maintaining included creating responsibilities managing include duties developing
(Project Manager, Administrative Assistant, Consultant)

management systems project implementation technology system design services development
(Project Manager, Consultant, Senior Consultant, Senior Software Engineer)
managed supervised staff trained employees department maintained assisted daily training
(Office Manager, Assistant Manager, Manager)
management responsible staff development operations including training planning personnel
(Operations Manager, General Manager, and Project Manager)
business management development team process strategic support developed strategy processes
(Manager, Consultant, Senior Consultant, President)

financial analysis reporting budget monthly reports accounting management cost annual
(Financial Analyst, Senior Financial Analyst, Controller, Senior Accountant)
financial clients trading investment client funds fund stock mutual securities
(Financial Advisor, Consultant, Registered Representative)

software developed system time code designed development real interface application
(Software Engineer, Senior Software Engineer, Consultant)
oracle database sql unix server system application data support shell
(Consultant, Programmer Analyst, Software Engineer)
server sql net visual web asp application applications microsoft development
(Consultant, Programmer Analyst, Software Engineer)
java application server xml web oracle jsp ee system websphere
(Software Engineer, Senior Software Engineer, Project Manager)
web site design sites internet html content development online website
(Web Developer, Web Designer, Consultant)

customer store customers cash service inventory sales merchandise stock register
(Sales Associate, Cashier, Assistant Manager)
office duties filing entry data answering phones mail phone typing
(Administrative Assistant, Receptionist, Office Manager)
news articles editor wrote magazine editorial newspaper edited copy publication
(Editor, Editorial Assistant, Associate Editor)
students school taught teaching student classes computer high education courses
(Teacher, Substitute Teacher, Instructor)

event that at least one word in that document is assigned to topic t_2. A graph showing all pairs of topics with mutual information above a threshold is shown in Figure 1. In order to make the graph more clear, we have removed several management-related topics that have strong connections to large numbers of other topics.

The most dominant feature of the graph is a pair of highly connected subgraphs, corresponding to systems administration and tech support on the left and software development on the right. The topics that connect these two clusters represent technical infrastrastructure: *server, database, system*. One side supports these systems, the other side uses them for development. To the right of the development cluster are topics for sales and marketing. These two subgraphs are largely disconnected. One of the few connections to the development subgraph is between the *business sales development services marketing company strategic software product developed* topic and the *requirements business system user project process design application analysis management* topic. This pattern shows that job descriptions frequently include *java* and *database* along with *requirements*, and they frequently include *requirements* along with *sales* and *marketing*, but do not frequently include *database* along with *sales*. The other prominent connection between the technical and business clusters is between *web site design* and *marketing*. Other clusters include design, engineering and production at the bottom right, administrative support on the right, and retail, call center, and food service topics at the top. Two law related topics form a disconnected subgraph at the bottom left.

The structure of this graph could potentially be useful to job seekers. For example, a person who has a particular skill set could identify other skills that would be useful in moving to interesting or desirable careers.

3 Topical Sequence Model

The topic model described above is effective at discovering a low-dimensional set of hidden components that make up job descriptions. Although a topic-based analysis of job descriptions is useful in discovering the structure of the job market, we are also interested in modeling career paths and job transitions.

One simple approach to modeling career paths is to look at the probability of transitions between job titles. This approach is hindered by both the sparsity and the lack of specificity of job titles, as discussed previously.

We present another approach based on learning a set of hidden states.

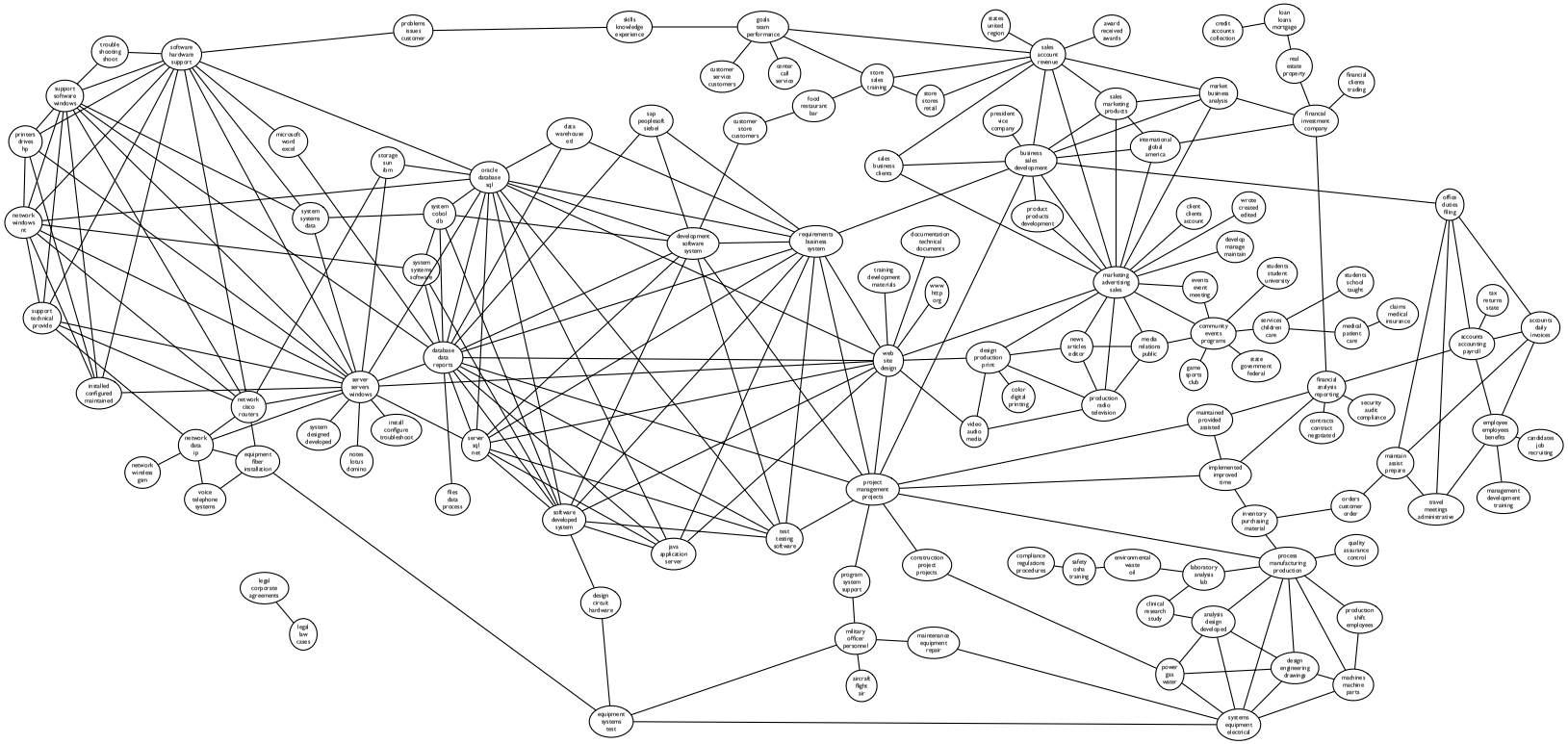

Figure 1: The network of topics. Links represent topic-topic mutual information scores above a threshold. Zoom in to view in online versions. A camera-ready version will include alternatives.

This method is based on a generative model for a sequence of job descriptions in a single resume. In the topical sequence model, the observations at each time step are drawn from one of a finite number of mixtures of multinomial distributions, one for each state. The probability of choosing each mixture of multinomials at a given time step depends on the previous mixture's distribution over states.

Each state has a multinomial over topics, drawn from a symmetric Dirichlet prior. We generate a resume as follows.

1. Sample a multinomial over states π from Dirichlet(δ).
2. For each state s, sample a multinomial over states ξ_s from Dirichlet(γ).
3. For each state s, sample a multinomial over topics θ_s from Dirichlet(α).
4. For each topic t, sample a multinomial over words ϕ_t from Dirichlet(β).
5. For each resume r,
   (a) Sample a state s_0 from π.
   (b) For each subsequent time step t, sample a state s_t from ξ_{t-1}.
   (c) For each time step t,
      i. For each word i in description r_t,
         A. Sample z_i from θ_{s_t}.
         B. Sample w_i from ϕ_{z_i}.

In the first timestep we begin by sampling a hidden state from a multinomial over states. We then generate the words comprising a job description for that time step according to the standard LDA generative model. We then generate the next state from the state transition distributions ξ_s given the current state.

Previous work combining topic models with HMMs, the HMM-LDA model of Griffiths et al. [3], models each word as having a hidden state determining whether it is drawn from a document-specific topic distribution or a state-specific multinomial. While the current model has strong ties to the HMM-LDA model, it differs in that we are concerned with the sequence of *documents*, treating the sequences of words within the document as i.i.d. given the document, whereas HMM-LDA is concerned with the sequence of *words*, treating the documents as i.i.d. The main point of this work, however, is to explore the application of topic sequence models on career path trajectory data.

We train the sequence model with Gibbs sampling. Because we found learning the nested hidden variables in the model to be unstable, we start with the converged LDA topic model described in the previous section. Using topics trained without access to sequence data is substantially more efficient to train, and allows us to directly compare the ability of topics alone to predict subsequent job titles to the ability of the same topics and a topical sequence model to perform the same task. We intend, however, to revisit jointly trained topical sequence models in future work.

The sampling distribution for a given state depends on the probability of the topics in the job description given the state and the transition probabilities from the previous state to the current state and from the current state to the next state. These are all Dirichlet-multinomial distributions, in which the multinomial parameters π, ξ, and θ can all be integrated out analytically. The resulting predictive distributions can be easily calculated by multiplying a factor for every individual count that is added.

Examples of the most probable topics and job titles for selected states are shown in Table 2.

Figure 2 shows the resulting graph of states and state transitions. Each state is labeled with the single most common job title assigned to that state. As shown in the topic mutual information graph, there is a distinction between technical job states toward the top right and other aspects of business, mostly to the bottom left. There are also two types of management states, one for each cluster.

Table 3 shows the probability of being in a given state after between one and three job transitions. Starting as a software engineer, the probability of staying a software engineer after one transition is quite high. After two and three transitions, however, the probability of moving into a technology management position increases until it is the single most likely career option.

4 Evaluation

We evaluate the predictive ability of several models. The task is to predict the next job title based on the previous job record. We present four models for this task. In every case, we divide the resumes into testing and training sets with 10-fold cross-validation. Because of the sparsity of job titles, in order to have sufficient training data we only consider job titles that appear 10 or more times in the corpus.

- The first model (TITLE) predicts the next job title based on the current job title. We train the model by counting the number of times

Table 2: Most frequent topics and job titles for hidden states

Topic %	Topic	Title %	Title
16.9	oracle database sql unix	4.8	software engineer
16.2	server sql net visual	4.6	consultant
6.9	database data reports system	3.5	programmer analyst
6.3	java application server xml	2.4	web developer
4.2	software developed system time	2.2	senior software engineer
3.6	management systems project implementation	1.9	programmer
34.1	design production print advertising	8.1	graphic designer
12.0	color digital printing print	3.7	art director
2.4	work worked time job	2.4	graphic artist
2.1	store stores retail sales	1.8	designer
1.5	skills knowledge experience ability	1.8	creative director
1.5	responsible maintaining included creating	1.4	production manager
23.6	laboratory analysis lab testing	5.9	research assistant
18.2	clinical research study cell	4.0	research associate
3.6	analysis design developed data	2.6	chemist
3.5	process manufacturing production product	1.6	lab technician
2.8	maintained provided assisted performed	1.1	laboratory technician
2.2	provide ensure develop work	1.0	analytical chemist
29.1	server servers windows network	4.7	network administrator
15.7	support software windows users	3.1	systems administrator
6.8	software hardware support computer	2.7	network engineer
5.8	network windows nt novell	2.2	systems engineer
2.2	network cisco routers switches	1.9	system administrator
2.0	printers drives hp ibm	1.0	systems analyst
30.4	business sales development services	1.6	account executive
4.9	company due position franchise	1.6	president
4.0	international global america europe	1.3	ceo
3.3	state government federal agencies	1.3	partner
3.3	business clients services small	1.3	president ceo
2.9	american america latin mexico	1.0	project manager

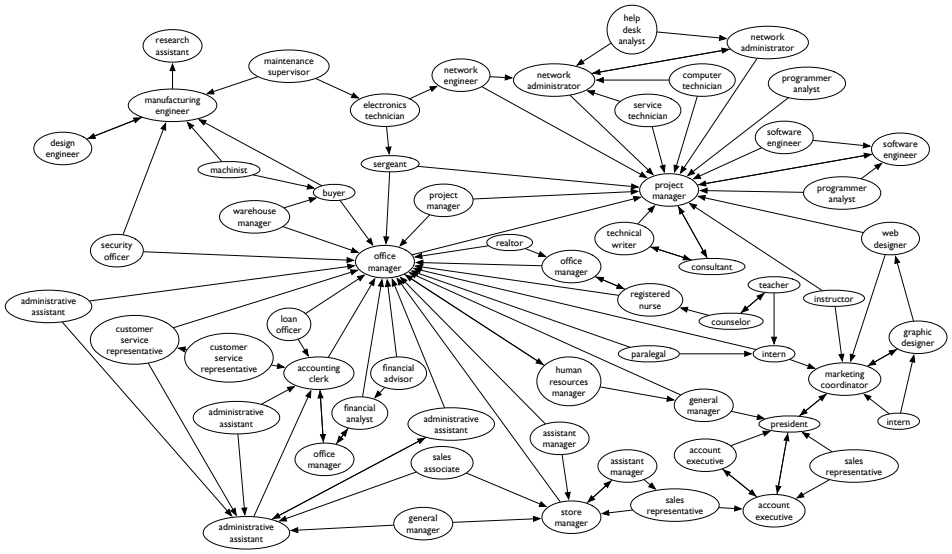

Figure 2: A graph of state transitions. The labels are the single most common title for jobs assigned to a given state; note that these can be extremely sparse. Self transitions are not shown. The states for business management (Office Manager) and technology management (Project Manager) can be reached from many other states.

Table 3: Transition probabilities for state *software engineer, consultant, programmer analyst*

Prob	Title after one transition	Prob	Title after two transitions	Prob	Title after three trans
0.375	software engineer	0.176	software engineer	0.134	**project manager**
0.177	**project manager**	0.166	**project manager**	0.098	software engineer
0.079	programmer analyst	0.059	programmer analyst	0.040	programmer analyst
0.048	web designer	0.039	consultant	0.034	consultant

each title t' follows a given title t, $N_t^{t'}$. $N_t^{\cdot}$ indicates the total number of times title t appears in the training data in a non-final position. In order to avoid zero probabilities we smooth the model using a Dirichlet prior ζ, which we set to 0.1. If t is an infrequent job title (ie $N_t^{\cdot} < 10$), we replace it with a single distinct symbol. The total number of frequent job titles is T. The predictive distribution is therefore

$$p_{TITLE}(t'|t) = \frac{N_t^{t'} + \zeta}{N_t^{\cdot} + \zeta T}. \quad (1)$$

- The second model (TOPIC) predicts the next job title based on the topics in the previous job description. We use a naive Bayes model, in which the class is the job title and the features $f(z)$ are indicators for whether a given topic comprised more than 10% of the words in the previous job description. Again, we use a Dirichlet prior to smooth the probability of a title given each topical feature. The prior probability of each job title is equal to the count of each job title in the training data with a Dirichlet prior. The predictive distribution is therefore

$$p_{TOPIC}(t|\mathbf{z}) \propto \frac{N_t + \eta}{N + \eta T} \sum_z \frac{N_t^{f(z)} + \zeta}{N_t^{\cdot} + \zeta T}. \quad (2)$$

- The third model (CURR-STATE) predicts the current job title given the current hidden state. We include this model only for the purposes of comparison to the PREV-STATE model, since it is the timestep that we are trying to predict. The predictive distribution for a title t given a state s is simply

$$p_{CURR-STATE}(t|s) = \frac{N_s^t + \zeta}{N_s^{\cdot} + \zeta T}. \quad (3)$$

- The fourth model (PREV-STATE) predicts the next job title given the previous hidden state. The predictive distribution for a title t given a previous state s is equal to the sum over all states s' of $p(s'|s)p(t|s')$,

$$p_{PREV-STATE}(t|s) = \sum_{s'} \frac{N_s^{s'} + \gamma}{N_s^{\cdot} + \gamma S} \frac{N_{s'}^t + \zeta}{N_{s'}^{\cdot} + \zeta T}. \quad (4)$$

The log likelihoods of testing instances averaged over 10-fold cross-validation is shown in Table 4. The CURR-STATE model has the highest log likelihood, but the PREV-STATE model is only slightly worse. Both hidden

state models are better at predicting the next job title than the previous job title alone. One interesting result is that the topical features by themselves are extremely poor at predicting the next job title, but the hidden state features trained from the same topical data provide very good predictive ability.

Table 4: Log likelihood of frequent titles in held-out resumes for the four models. The current hidden state model (an Oracle) performs best, but the previous hidden state model, which sums over all possible subsequent states, still beats the baseline of predicting the next title based on the current title. Topic occurrence features are fairly poor, performing well below the baseline. Despite this, however, the PREV-STATE model, a hidden state model trained on the same topics, performs better than either baseline model.

Model	log-likelihood
TITLE	-8526.618
TOPIC	-12280.248
CURR-STATE (Oracle)	-7362.468
PREV-STATE	-7662.267

5 Conclusions

We have presented the task of predicting job transitions given previous employment history. In addition, we have presented a new method, topical sequence modeling, that shows strong results in predicting job transitions. Job descriptions, and especially job titles, have many ways of describing similar skills and responsibilities. The topical sequence model is effective at providing an interpretable, low dimensional representation of this complex application.

6 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF Nano # DMI-0531171, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations

expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[2] W. Buntine. Applying discrete PCA in data analysis. In *UAI 2004*, 2004.

[3] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *NIPS 2004*, 2004.