Abstract

This study uses text mining techniques to analyze employment data posted over the internet. The objective is to identify knowledge areas, skills and expertise relevant to jobs in the construction industry. We utilized the fast growing online job search engines to understand the construction job market and employer expectations. Over 20,000 job advertisements were downloaded from various websites between Oct 14th 2012 and March 15th 2013. We developed a text mining method to identify derived job qualification information from the downloaded pages. The developed algorithm is capable to derive rules by automatically extracting statistically significant patterns present inside preselected qualifications. The selection rules can then be used to detect the presence of these qualification in new pages. Once the qualification are identified, we used the Latent Dirichlet Allocation (LDA) model to identify groups of skills that are required by employers. One of the major advantages of implementing LDA model is that it is an unsupervised approach and no training is needed. The algorithm was applied to a case study as an illustrative example.

Keywords: Construction Management Skills; Employer Expectation; Text Mining; Topic Model

1. Introduction

The construction industry is one of the nation’s largest industries. There is undoubtedly a demand for professionals with skills necessary to successfully manage today’s complex construction projects. Without capable workforce, companies will be unable to fulfill their potential growth.

Training is fundamental to develop the necessary skillsets particular to the construction industry. However, training must be demand-led (Fluitman et al.1989) by the market. While formal courses are offered to equip job seekers with derived knowledge and skills, people still need to recognize industries’ requirement. To effectively do so, it is advantageous to understand the job market requirements in greater detail (e.g., skills, qualifications and experiences). Moreover, the formal education programs must be constantly enhanced to address the needs of the industry and hence the job seekers. Extensive studies on identifying key competencies for construction graduates have been conducted in the past 30 years. For example, a previous research conducted by Lees (2002) stated that education should enhance employability by focusing on knowledge and understanding, developing skills, self-efficacy beliefs, and strategic thinking and reflection. This study also found that employers view team spirit as more important than numeracy and literacy. Archer and Davidson (2008) identified communication skills, teamwork skills, integrity, intellectual ability and confidence as important employability skills. Farooqui and Ahmed (2009) developed a questionnaire-based survey to identify required key skills from graduating construction management students, both undergraduate and graduate level. Rawlins and Marasini (2011) conducted a survey-based study of skill-gaps on construction management program in UK. Arain (2010) identified a range of necessary expertise for students to practice successfully in the construction industry. Ahn et al. (2012) examined U.S. construction industry perceptions regarding key competencies for construction graduates using a survey of recruiters for over 100 construction companies located in the eastern United States.

Most of the existing efforts to identify the key skill sets in the construction industry have been conducted by interviewing or surveying industry representatives. This approach is usually limited by the number of organizations who are willing to respond and by the variability introduced by the subjective nature of survey and interviews. This research introduces a new approach to address this issue by analyzing online job advertisements. Job advertisements posted by construction companies are usually found in various job search engines (e.g., Indeed.com and Monster.com). The fast growing online job search engines make it feasible to look for job skills information from massive amounts of readily available data. Job search engines provide millions of job openings with daily updates. Also, they provide detailed information on qualifications and special requirements for each job opening. Since the online job search services have become more and more popular, it is important to utilize this digital world material and analyze the advertisements of job openings. In this research, we collected more
than 20,000 construction related job openings in the past several months and analyzed this large sample to extract the employers’ expectations.

This paper is organized as follows. The first section discusses the methodology used to extract job qualification information from online advertisement pages. The second section discusses the results collected. The last section concludes this research and describes the future steps.

2. Methodology

To achieve the research objective, we developed a text mining method. This method consists of three main components:

1. Text preprocessing.
2. Qualification sentence identification.
3. Utilization of the LDA topic model in extracting the sought skill sets.

2.1. Text processing

The first step of the developed method is to prepare job advertisements in bulk. A computer program was developed to search Monster.com and Indeed.com between Oct 2012 and July 2013 for construction related jobs. More than 20,000 job advertisements webpages were downloaded during this time period. The downloaded webpages were preprocessed by removing the HTML tags and special character. This is because most of the job advertisements web pages are written in HTML, which uses open/closed tags to indicate webpage commands, i.e., “<” and “>”. By removing those HTML tags, the computational burden of the information extraction is reduced. We also used Natural Language Toolkit (NLTK) (Bird et al. 2009) to segment the text into individual sentences. At the end of this step, the downloaded job advertisements were processed into a large pool of sentence set denoted by $S = \{s_1, s_2, \ldots \}$.

2.2. Identify qualification sentences

Once the sentence segmentation was completed, we started to extract sentences that contained job qualification information from S. An online job posting webpage is usually composed of information such as job-title, location, employer name, job description, job qualification, advertisements, etc. For instance, the following ten sentences are taken from an online job post:

1. “XXX Inc. is comprised of talented professionals like you”.
2. “They are intuitive, goal oriented, and think progressively while working in a variety of roles to provide the highest level of client satisfaction”
3. “It is this mindset of success, coupled with our opportunities that will help develop and define your career”
4. “We need a Construction Estimator”
5. “Need to have at least 10 years experience”
6. “Willing to work long hard hours”
7. “There maybe some weekends”
8. “Must be familiar with reconstruction/restoration and some stucco”
9. “Have experience with excel, work and outlook”
10. “If interested please send us your resume”

In this example, only the sentences 5-9 contain information about job qualification. Therefore, it is important that the adopted method is able to differentiate sentences that contain job qualification information from those that do not. In this research, we utilized the naive Bayes classification method to help determine if a given sentence contains job qualification information or not. We consider each sentence $s$ to be represented as an $n$ - dimensional vector, $X$, of different feature values.

$$X = (x_1, x_2, \ldots, x_n)$$

where $x_i$ represents the value of the ith feature and $n$ is the number of features. In this research, the following features were used:

1. First word of the sentence
2. Last word of the sentence
3. First two words of the sentence
4. Last two words of the sentence
5. All consecutive pair of words
Following is a simple example to illustrate how the features of a sentence are generated.

“Must have excellent communication and written skills.”

For example, the corresponding features for the sentence above are:
1. “must”
2. “skills”
3. “must have”
4. “written skills”
5. “must have”, “have excellent”, “excellent communication”, “communication and”, “and written”, “written skills”

The intuition standing behind this is that if a sentence starts with (or ends with, or contains) the words “must have”, or “is preferred”, or “proficient”, it is very likely that this sentence contains job qualification information.

With the features defined, the Bayes classifier assigns each sentence to one of two classes $S_1$ (sentences containing job qualification information) and $S_2$ (sentences with other contents). A sentence $s$ is classified to the class for which it has the highest posterior probability conditioned on its features. That is, $s$ is classified to class $S_i$ if and only if,

$$p(s \in S_i | X) \geq p(s \in S_j | X), \text{ for } j \in \{0,1\} \text{ and } j \neq i$$

According to Bayes Theorem,

$$p(s \in S_i | X) = \frac{p(X|s \in S_i)p(s \in S_i)}{p(X)}$$

Since $p(X)$ is a constant factor which is equal for all classes, only the numerator $p(X|s \in S_i)p(s \in S_i)$ need to be calculated. Both of these two values, $p(X|x \in S_i)$ and $p(s \in S_i)$, can be calculated from the data set. However, it can be very challenging to compute $p(X|s \in S_i)$. If each component $x_i$ of $X$ can have one of $r$ values, there would be $r^n$ combinations to consider for each class. In order to simplify the calculation, the assumption of conditional independence was made. In other words, for each class, the attributes are assumed to be independent. The classifier resulting from this assumption is known as the Naive Bayes classifier. This assumption leads to

$$P(X|s \in S_i) = \prod_{k=1}^{n} p(x_k|s \in S_i)$$

To develop this classifier, we prepared a training set of 1,000 sentences and a testing set of 500 sentences. The training and testing sets were randomly selected from online job advertisements. The classifier had a precision rate of 0.83 and a recall rate of 0.90. Precision is the ratio of the number of sentences which was determined correctly by classifier to the number of sentences which classifiers classified to the $S_i$ class. Recall rate is the ratio of the number of sentences which was determined correctly by classifier to the number of sentences in $S_i$.

2.3. Topic model

Once the qualification sentences are extracted, we applied the Latent Dirichlet Allocation (LDA) topic model to identify the skills. This model is a popular method for modeling term frequency occurrences for documents in a given corpus. A topic basically consists of a set of words that co-occur frequently. A document is described as a mixture of different topics. In this study, the qualification sentences are treated as the documents and the skills are the topics to be discovered. The LDA model is a probabilistic topic model, which uses a generative probabilistic model for collections of discrete data (Blei et al. 2003). One of the major advantages of implementing LDA model is that it is an unsupervised approach and no training is necessary. In this paper, we implement the LDA model through the Gibbs sampling algorithm (Geman and Geman, 1984).

3. Case Study

In this case study, we applied the above methodology on more than 20,000 construction related job advertisements from online job search engines. Figure 1 shows the percentages of major job positions in the collected data. There are more than 7,000 project manager positions in the dataset. The rest are project engineers, estimators, superintendents and schedulers positions.
As discussed earlier in this paper, one of the advantages of implementing the LDA model is that it’s an unsupervised approach and no training is needed. The only input required is the number of topics K. In this study, we set K = 10 such that the top 10 skills are identified for each of the job type in Figure 1. Table 1 - 5 present the results of the LDA model. In addition, the representative examples for each topic are provided under each Table. As shown in Table 1 - 5, the major skills required by all job types are listed as follows.

1. Communication (10th topic in Table 1, 5th topic in Table 2, 4th topic in Table 3, 6th topic in Table 4, 10th topic in Table 5; including words like skill, strong, communication, written, excellent, organizational, good, verbal, interpersonal, et al.). Some examples of job qualification sentences are:
   a. “Ability to communicate effectively both verbally and in writing.”
   b. “Great interpersonal and communication skills (both verbal and written), the ability to interface and influence at all levels of an organization.”
   c. “Good verbal and written communication skills, as well as respect for customers and coworkers are a must.”

2. Degree (1st topic in Table 1, 1st topic in Table 2, 9th topic in Table 3, 1st topic in Table 4, 8th topic in Table 5; including words like construction, degree, engineering, management, preferred, related, bachelor, field, equivalent, civil, et al.). Examples are:
   a. “Bachelor’s degree in Construction Management or Engineering”
   b. “Bachelors or higher degree in an engineering or construction related discipline”
   c. “Four-year college degree in Engineering, Construction Science or related field or equivalent combinations technical training and or experience.”

3. Team work (4th topic in Table 1, 5th topic in Table 2, 6th topic in Table 3, 5th topic in Table 4, 5th topic in Table 5; including words like ability, effectively, team, client, manage, multiple, customer, communicate, level, service, et. al.). Examples are:
   a. “Strong management, leadership and teamwork skills.”
   b. “Must be pro-active, adaptable and able to motivate team members.”
   c. “Must be able to work as part of a team.”

4. Microsoft office software (2nd topic in Table 1, 8th topic in Table 2, 3rd topic in Table 3, 7th topic in Table 4, 3rd topic in Table 5; including words like computer, microsoft, word, project, office, excel, software, proficient, working, use, et. al.). Examples are:
   a. “Personal computer knowledge, including Microsoft Word and Excel.”
   b. “Proficient in Microsoft Excel & Word.”
   c. “Proficiency with all MS Office products (Word, Excel, PowerPoint, etc.”

There are also skills that are particularly demanded by a certain job type. For example, the ability to use Primavera and the knowledge of Critical Path are identified as schedulers’ skills in the 7th and 2nd topics of Table 2. Examples related to those topics include:
   • “Experience with Primavera is required”
   • “Intermediate level knowledge of Microsoft Office and P6 Primavera Scheduling Program”
   • “minimum of 5 years direct experience in applying scheduling principles to construction projects using the Primavera suite of software”
   • “Able to run and analyze the schedule Critical Path”
   • “5-10 years proficient knowledge in Primavera and Critical Path Methodology.”
   • “Five years experience with all aspects of Critical Path Method scheduling and its application, etc.”

For project engineer jobs, the professional engineers (pe) license requirement is identified in the 9th topic in Table 4.
Examples of job advertisements are:

- “A bachelor’s degree in Engineering is required and Professional Engineering License in good standing with at least one state is preferred.”
- “Professional Engineering (PE) license and/or PMP certification (preferred)”
- “An advanced Degree in Engineering, Engineering Management or MBA and a Professional Engineer license (PE) is preferred.”

The ability to read specification and drawings is required by the cost estimator position (6th topic in Table 1). Examples are:

- “Working knowledge of drawing systems and ability to coordinate specifications and drawings”
- “Experience in CAD and/or other digital drawing tools helpful”
- “Problem solving issues and questions related to drawing and specifications”

For project manager, the ability to understand contract is also identified in the 2nd topic of Table 5. Examples are:

- “Assists in conducting fee negotiations, preparing contract agreements and is sufficiently familiar with all agreements between the firm and the client to effectively manage the project in a professional and economic manner”
- “Knowledge of contract services and contract terms and conditions”
- “Ability to assess contract compliance and product/service quality”

Table 1. Cost Estimator

| Topic number | Topic words |
|--------------|-------------|
| 1 | construction, degree, engineering, management, preferred, related, bachelor, field, equivalent, civil |
| 2 | building, customer, service, client, both, new, relationship, professional, maintain, working |
| 3 | ability, work, environment, team, effectively, well, deadline, multiple, communicate, time |
| 4 | specification, plan, ability, read, drawing, interpret, project, requirement, document, understand |
| 5 | experience, estimating, construction, project, 5, minimum, industry, estimator, candidate, commercial |
| 6 | cost, project, estimate, knowledge, bid, system, control, estimating, design, process |
| 7 | problem, solving, company, complex, concrete, situation, development, deal, proven, provide |
| 8 | skill, strong, communication, written, excellent, organizational, good, verbal, interpersonal, computer |

Table 2. Scheduler

| Topic number | Topic words |
|--------------|-------------|
| 1 | experience, construction, project, degree, engineering, scheduling, management, related, bachelor, 5 |
| 2 | schedule, able, analysis, cost, critical, environment, resource, path, impact, program |
| 3 | ability, deadline, demonstrated, multiple, others, drawing, environment, fast-paced, job, obtain |
| 4 | skill, communication, strong, written, team, excellent, demonstrated, verbal, client, effectively |
| 5 | financial, process, provides, area, development, well, risk, reporting, business, requirement |
| 6 | primavera, knowledge, p6, working, software, scheduling, p3, using, preferred, candidate |
| 7 | microsoft, excel, proficient, office, word, computer, proficiency, application, powerpoint, suite |
| 8 | work, production, good, able, rsquo, environment, safety, job, vessel, plant |
| 9 | demonstrated, ability, travel, complex, function, environment, contractor, maintain, condition, solution |
### Table 3. Superintendent

| Topic number | Topic words                                                                 |
|--------------|-----------------------------------------------------------------------------|
| 1            | project, plan, specification, construction, site, schedule, safety, job, work, quality |
| 2            | ability, able, read, travel, willing, interpret, understand, perform, job, report |
| 3            | computer, microsoft, word, project, office, excel, software, proficient, working, use |
| 4            | skill, strong, communication, written, organizational, excellent, leadership, good, interpersonal, verbal |
| 5            | ability, able, area, license, work, up, driver, position, certification, valid |
| 6            | work, environment, ability, team, multiple, manage, company, deadline, proven, task |
| 7            | ability, effectively, problem, customer, communicate, subcontractor, employee, maintain, solving, client |
| 8            | knowledge, construction, working, control, equipment, maintenance, building, system, cost, electrical |
| 9            | degree, preferred, management, construction, field, related, engineering, experience, bachelor, college |
| 10           | experience, construction, project, 5, minimum, superintendent, 10, candidate, commercial, supervisory |

### Table 4. Project Engineer

| Topic number | Topic words                                                                 |
|--------------|-----------------------------------------------------------------------------|
| 1            | engineering, degree, bachelor, related, mechanical, civil, field, b, construction, electrical |
| 2            | design, experience, system, process, equipment, knowledge, operation, power, building, control |
| 3            | ability, work, able, environment, travel, multiple, manage, time, independently, task |
| 4            | project, construction, specification, cost, technical, plan, requirement, drawing, schedule, support |
| 5            | ability, team, effectively, client, level, customer, communicate, well, service, position |
| 6            | skill, strong, communication, written, excellent, good, problem, verbal, organizational, interpersonal |
| 7            | knowledge, working, office, word, microsoft, software, excel, computer, proficient, autocad |
| 8            | experience, project, management, construction, minimum, 5, industry, 10, manufacturing, 3 |
| 9            | preferred, engineer, candidate, professional, within, license, e, environmental, pe, state |
| 10           | environment, business, safety, job, perform, area, demonstrated, able, position, apply |

### Table 5. Project Manager

| Topic number | Topic words                                                                 |
|--------------|-----------------------------------------------------------------------------|
| 1            | experience, project, construction, management, minimum, 5, manager, 10, commercial, managing |
| 2            | project, specification, plan, schedule, cost, contract, demonstrated, planning, management, development |
| 3            | microsoft, office, excel, word, software, computer, knowledge, proficient, project, program |
| 4            | knowledge, design, working, construction, system, building, process, facility, safety, operation |
| 5            | ability, effectively, team, client, manage, multiple, customer, communicate, level, service |
| 6            | experience, preferred, candidate, professional, engineer, industry, certification, successful, should, within |
| 7            | work, environment, able, ability, travel, perform, job, time, position, individual |
| 8            | degree, engineering, bachelor, construction, management, related, field, civil, b, equivalent |
| 9            | ability, problem, technical, business, read, complex, solving, issue, interpret, understand |
| 10           | skill, strong, communication, written, excellent, good, organizational, verbal, interpersonal, oral |

### 4. Conclusion

In developing career pathways in construction industry, it will be important to give students skill sets which can be useful in their future career. In addition to the existing curriculum, it is also important to look at adding more industry needed skills. This research aims to provide insights into the skill sets through text mining the online information sources. In the case study, the proposed methodology is able to effectively extract skill sets from online job advertisements. The same methodology can be applied to job positions in other industries.

### References

Ahn, Y.H., Annie, R.P. and Kwon, H. (2012). Key competencies for us construction graduates: Industry perspective. Journal of Professional Issues in Engineering Education & Practice, 138(2):123–130.

Arain, F.M. (2010). Identifying competencies for baccalaureate level construction education. In Construction Research Congress.

Archer, W. and Davison, J. (2008). Graduate employability. The Council for Industry and Higher Education.

Blei, D.M. Ng, A.Y. and Jordan, M.I. (2003). Latent dirichlet allocation. The Journal of machine Learning research, 3:993–1022.
Farooqui, R.U. and Ahmed, S.M. (2009) Key skills for graduating construction management students—a comparative study of industry and academic perspectives. In Construction Research Congress, 2009.
Fluitman, F. (1989) OIT Vocational Training Branch, et al. Training for work in the informal sector. 1989.
Geman, S. and Geman, D. (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (6):721–741.
Lees, D. (2002). Graduate employability—literature review. LTSN Generic Centre.
Rawlins, J. and Marasini, R. (2011). Are the construction graduates on ciob accredited degree courses meeting the skills required by the industry? Management, 167.174.