Quantifying Skill Relevance to Job Titles

Wenjun Zhou*, Yun Zhu†, Faizan Javed†, Mahmudur Rahman‡, Janani Balaji† and Matt McNair†
*University of Tennessee, Knoxville, TN 37996 wzhou4@utk.edu
†CareerBuilder, LLC., Atlanta, GA {yun.zhu,faizan.javed,janani.balaji,matt.mcnair}@careerbuilder.com
‡Computer & Info. Sci., IUPUI, Indianapolis, IN 46202 mmrahman@iupui.edu

***Abstract*—Eliminating or reducing skill gaps in the job market is critical to putting people back to work, reducing the unemployment rate, and increasing the labor market participation rate. A key element in closing the skills gap is accurately identifying the mismatch between the skills expected by employers and those possessed by job seekers. In this study, our goal was to profile job titles by effectively quantifying the relevance of skills. We started by using a naive, frequency-based skill ranking approach, which resulted in the most generic skills ranked on the top. We then adapted a number of alternative metrics and compared their performances on a number of job titles. The outcome of this study can support CareerCoach, an analytical solution CareerBuilder has piloted to provide insights and data dashboards to job seekers.**

***Keywords*-job market; career profiling; skills ranking;**

I. Introduction

Eliminating or reducing the skills gap is critical to putting people back to work, reducing the unemployment rate, and increasing the labor market participation rate. A key element in closing the skills gap is accurately identifying the mismatch between the skills expected by employers and those possessed by job seekers. In particular, with those skill gaps identified, it will be possible to recommend new skills or educational products to job seekers to help make them more competitive in the job market and adaptive to the changing job requirements.

Identifying the value of skills has been of long standing interest in business and economics. On a macro level, skill requirements are important for industries and the economy. For example, a study of information systems (IS) jobs over a twenty-year period (1970-1990) [1] reported that the requirements on technical skills and business skills evolved with different trends in the industry. For example, a *System Analyst* was expected to have increasingly have more technical skills compared to business skills over the 20-year time horizon. Such evolution in skill requirements have strong implications for educators and industrial practitioners. With recent acceleration of industrial developments, we are seeing more rapid changes in job needs, including new positions and emerging skills. On a micro level, skill requirements are important for job seekers and employers. For example, the same skills may indicate different levels of utility for different workers in the labor market [2]. The value of employee skills is found to contribute significantly to the growth of an organization [3], [4].

In recent years, as online recruiting and big data applications become more and more common, many models are proposed in the information retrieval, data mining, and machine learning fields to identify relevant skills in an automatic fashion. A large body of work focuses on skill *extraction*. By processing textual information from the World Wide Web (e.g., Wikipedia), user profiles (such as LinkedIn and organizational internal networks like IBM), job postings, resumes, and other information, it is possible to extract key words that relate to people or positions. CareerBuilder has built an in-house skill term extraction system, called SKILL [5], that can extract skill keywords from both job ads and job seekers' resumes. Even though such extracted skills are each attached with a confidence score, the scores represent the likelihood of each skill, but not the relevance of the skill to the job title. Learning skill *relevance* to job titles or job seekers remains a challenging problem. Endorsement, for example, is a feature on LinkedIn that integrates labels contributed by people's connections. This feature relies heavily on the quality of human input, and thus, may suffer from errors due to heterogeneity in users' efficacy.

In this study, we focus on the problem of quantifying skill relevance to job titles, relying on just the job posts and skill keywords information. This study is motivated by a number of use cases at CareerBuilder (see Section II-A). Our initial attempts based on skill requirement frequency resulted in many generic skills being ranked on top. For example, *Business Administration* is commonly a required skill for an *Accountant*, but it is not quite informative with respect to how important or unique it is to the job title. As a result, we adjusted skill frequency using the term-frequency-inverse-document-frequency (TF-IDF) model. Moreover, we also adapted information theoretic metrics and measurements of variation to assess the (un)certainty of a skill to a title, to adjust for the frequency of very commonly required skills. The intuition is that if a skill is required by many job ads of the same title, it is likely an important skill. We consider global uniqueness and local variance as additional factors for score adjustments. This global-local score measures the extent of variation, which helps strengthen the level of certainty of each skill for each title.

The rest of this paper is organized as follows. In Section II, we provide preliminary details, such as terminology

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

and problem formulation. In Section III, we describe the metrics used to discount the frequency of common skills, and amplify more unique skills for different titles. In Section IV, we summarize experimental results. Related work is summarized in Section V, and we conclude in Section VI.

II. Preliminaries

The goal of this study is to develop a systematic approach to **assess the relevance of different skill terms for each job title**. This section will start by describing a motivating use case. Then, the terminology to be used throughout this paper will be clarified in the context of our system. Finally, a formal problem formulation will be provided.

A. CareerCoach: A Use Case

A number of products being developed at CareerBuilder rely on effectively evaluating the relevance of skills to job titles. The main use case of this study is career profiling, which has to do with identifying top skills for a given job title. The first goal is to identify the **core skills**, which are the skills that are essential for a job title. Core skills are most frequently asked for by employers. To have a good chance of getting a job with this title, a candidate should have most, if not all, of these skills. Take the job title *Accountant* for example. Core skills for an accountant include *Accounting*, *Microsoft Excel*, *Financial Statements*, and *Auditing*. Once job applicants see lists of core skills under each job title, they may quickly identify their compatibility to that career, and identify the skill gap if any. Identifying the core skills provides the basis of the CareerCoach, an analytical solution that CareerBuilder has piloted to provide insights and data dashboards to job seekers.

Moving a step further, we could identify skills listed in each job seeker's resume[1], and contrast them with those in job ads. For example, it is possible to suggest **standout skills** to the job seekers. These are the skills that are likely to help job seekers stand out from others (i.e., getting noticed by employers), as they are frequently required by employers, but were less frequently listed on job seekers' resumes. Again, taking the job title *Accountant* for example, standout skills may include *Leadership*, *Staffing*, *Written Communications*, etc. A list of titles and their standout and core skills are listed at http://coach.careerbuilder.com/browse.

Moreover, CareerCoach aims at providing insights and advice to job seekers by automatically identifying the skill gaps. The identification of skill gaps can happen on two levels: the job market level, and a specific job seeker's level. The focus is to identify skills that are in demand and those in supply. Skills that are high in demand but low in supply are generally very valuable. Job seekers should revise their resumes or consider educational opportunities in order to add the skills. In particular, Figure 1 illustrates an example. In this visualization, sets of skill terms that correspond to the job title "Accountant" are shown. The large yellow circle on the left includes skill terms that are in supply. The large red circle on the right includes skill terms that are in demand. The circles within the large red circle, but not in the overlapping area, represent the current skill gap collectively in the job market.

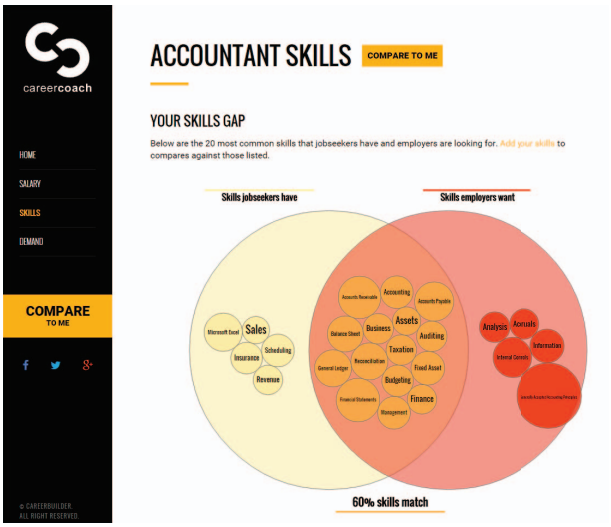


Figure 1. Skills in supply and in demand on the job market for an accountant position.

B. Terminology and System Overview

In the following, we first clarify the terminology used frequently throughout this paper.

- **Ad**: A job ad is a textual document that describes an open position. It has a title (described below) and a description. The description provides details about the position, including skill requirements (to be elaborated below).
- **Title**: Even though the original job ad can have any text in the job title field, we normalized them to a list of standardized job titles using Carotene, CareerBuilder's job title classification system [6]. For example, "Data Scientist - Atlanta, GA" would be normalized into "Data Scientist" [7]. The normalized titles correspond to O*NET coding standards[2].
- **Skill**: From the description of each job ad, we extracted skill terms using CareerBuilder's skill term extraction system [5]. Each skill is then represented as a normalized skill term (i.e., keyword or phrase).

The overall architecture of our framework is visualized in Figure 2. The raw job ads were persisted in the jobs database (i.e., Job DB). Each job ad then passed through our in-house

[1]Please note identifying skills from job seekers' resumes and checking the market gap are beyond the scope of this study. In this study, we will only focus on ranking skills from job ads.

[2]https://www.onetcodeconnector.org/

natural language processing (NLP) engines. Specifically, the job title was normalized using the Carotene system, which converts the raw title into a Carotene title, and the SKILL system extracted and normalized skill keywords from the job description. Therefore, after this NLP step, all job ads were processed into a standard format where each title was followed by a list of skill terms.

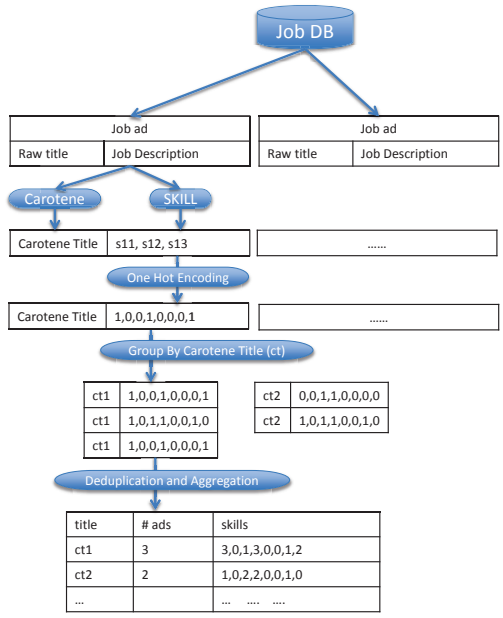
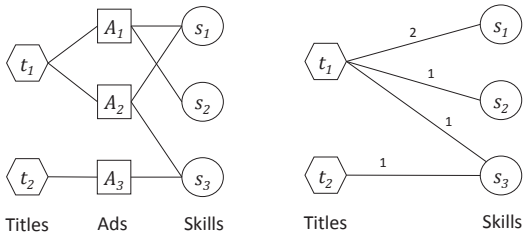


Figure 2. The framework.

We then converted the skill term lists into a binary format using one-hot encoding [8], and grouped job posts by normalized titles. In other words, under each title, there is one or more job ads, and each indicates whether each skill is required or not. Finally, we aggregate the skill counts by job title.



(a) Tripartite, Unweighted (b) Bipartite, Weighted

Figure 3. Concepts and their relation.

As illustrated by Figure 3(a), suppose that there were three job ads (henceforth referred to as ads), A_1, A_2, and A_3. After title normalization, we found that A_1 and A_2 shared the same title t_1, whereas A_3 had a title t_2. Note that each job ad has just one title, and there are many job ads under the same job title. Thus, the relationship between titles and ads is one-to-many. Moreover, after processing the job descriptions, we found that A_1 required skills s_1 and s_2, A_2 required skills s_1 and s_3, and A_3 required just skill and s_3. We can see that there are overlaps between sets of skills required by different ads, therefore, the relationship between ads and skills is many-to-many.

The relationship among titles, ads, and skills, as represented in Figure 3(a), forms an unweighted tripartite graph. The underlying assumption is that a skill is either required (i.e., an edge exists), or not (i.e., the edge does not exist), in an ad. This simplification allows us to consider the (co-)existence of skills in all ads with the same title. Since the primary goal of this study is focused on the relationship between titles and skills, we could simplify the tripartite graph into a weighted bipartite graph between titles and skills, as illustrated in Figure 3(b). The count on the edge that links title t_i and skill s_j represents the number of ads with title t_i that require skill s_j.

C. Problem Formulation

In this study, we were given a large vocabulary of all possible skills $S = \{s_1, s_2, \ldots, s_{|S|}\}$, and a large collection of job ads $\mathcal{A} = \{A_1, A_2, \ldots, A_{|\mathcal{A}|}\}$, where each job ad $A_i \subseteq S$ consists of a set of skills required by this job ad. Moreover, we extracted normalized job titles for each ad, so that multiple ads may be categorized under the same title. Suppose that $T = \{t_1, t_2, \ldots, t_{|T|}\}$ are the set of all unique job titles in our dataset. In the rest of the paper, unless otherwise specified, we will use s to index the skills, t to index the titles, and a to index ads. Moreover, we represent the number of ads under title t that require skill s as $n(s,t)$.

Our ultimate objective is to return a list of relevant skills for any given job title. Skills must be ranked based on the relevance to the job titles. Practically, we are going to learn a function that maps each pair of skill s and title t into a relevance score, which makes skill ranking easy. The very first idea is to use the frequency. In other words,

$$\text{relevance}(s,t) = n(s,t). \tag{1}$$

Then, for any given title t_0, we may sort skills by $n(s,t_0)$, since the more frequently a skill is required under that title, the more likely it is an essential skill for that title. However, this has resulted in the more generic skills to be ranked at the top. Sometimes such generic skills are not very informative indicators of the most useful skills for the title. This observation has motivated this study.

III. Quantifying Skill Relevance

In this section, we present the methodology. We will consider three levels of adjustments: TF-IDF based, global uniqueness, and local uniqueness.

A. TF-IDF Based Scores

As illustrated in our framework, the aggregated skill counts under each job title fit naturally into the document-term model that is well known in the information retrieval literature. Considering each job title as a document, and each skill as a word, we may compute the TF-IDF of each skill-title pair [9], [10], [11].

By definition, TF-IDF is the product of term frequency (TF) and inverse document frequency (IDF) [12]. We consider two variants of term frequency. Namely, the basic term frequency (Eqn. 2), and the logarithmic version (Eqn. 3):

$$\mathrm{TF}_{\mathrm{raw}}(s,t) = n(s,t), \quad (2)$$

$$\mathrm{TF}_{\log}(s,t) = 1+\log_2 n(s,t). \quad (3)$$

Moreover, we also consider two variants of inverse document frequency. Namely, the basic inverse document frequency (Eqn. 4), and the max version (Eqn. 5):

$$\mathrm{IDF}_{\mathrm{raw}}(s) = \log_2 \frac{|T|}{m_s}, \quad (4)$$

$$\mathrm{IDF}_{\mathrm{max}}(s) = \log_2 \left(1+\frac{\max_{\{s'\in S\}} m_{s'}}{m_s}\right), \quad (5)$$

where m_s is the number of job titles that require skill s, and $|T|$ is the number of all unique job titles.

The TF-IDF can then generally be calculated as

$$\text{TF-IDF}(s,t) = \mathrm{TF}(s,t)\times \mathrm{IDF}(s), \quad (6)$$

where TF can take any variant from Eqn. 2 or Eqn. 3, and IDF can take any variant from Eqn. 4 or Eqn. 5.

TF-IDF based methods are normally very good to penalize general words that appear in many documents. Using TF-IDF based method, we compute TF-IDF(s,t) for all valid (s,t) pairs in the dataset. We can then return a list of skills for any given title, sorted in decreasing order of TF-IDF.

B. Quantifying Skill Dispersion

Instead of using IDF, we have a few other possible alternatives to measure the uniqueness of a skill term for different titles. The basic idea is to leverage the dispersion of a skill term across different job titles. The intuition is the more titles that require a skill (i.e., the skill is "dispersed"), the less unique the skill is to any title. On the contrary, if a skill is required by just a few titles, it is quite unique to those titles.

Table I
MATRIX OF SKILL-TITLE RELATIVE FREQUENCIES

		Titles				
		1	2	…	$\lvert T\rvert$	Sum
Skills	1	p_{11}	p_{12}	$\cdots$	$p_{1\lvert T\rvert}$	1.00
	2	p_{21}	p_{22}	$\cdots$	$p_{2\lvert T\rvert}$	1.00
	⋮	⋮	⋮	⋱	⋮	⋮
	$\lvert S\rvert$	$p_{\lvert S\rvert 1}$	$p_{\lvert S\rvert 2}$	$\cdots$	$p_{\lvert S\rvert\lvert T\rvert}$	1.00

In particular, for any given skill s, we normalized its distribution across all job titles into percentages. In Table I, we have $p_{st} = \frac{n(s,t)}{\sum_t n(s,t)}$ as the relative frequency of skill s under job title t. We consider three variants of dispersion measures. Namely, the entropy-based (Eqn. 7), the dispersion of purity (DP) based (Eqn. 8), and the variation (VAR) based (Eqn. 9):

$$\mathrm{U}_{\mathrm{en}}(s) = \frac{1}{-\sum_t p_{st}\log(p_{st})}, \quad (7)$$

$$\mathrm{U}_{\mathrm{dp}}(s) = \max_t(p_{st}) - \frac{1}{|T|}, \quad (8)$$

$$\mathrm{U}_{\mathrm{var}}(s) = \frac{1}{|T|-1}\sum_t \left[p_{st}-\frac{1}{|T|}\right]^2. \quad (9)$$

The above U measures quantify the lack of dispersion (or, concentration) of skills across job titles. Each skill is weighted such that the more dispersion, the less the weight towards any given title. More specifically, we could replace the IDF component in the previous section with a dispersion metric listed here. As we can see, for all the above variants of U, skills with larger values are more unique across titles. In summary,

$$\text{TF-U}(s,t) = \mathrm{TF}(s,t)\times \mathrm{U}(s), \quad (10)$$

where TF can be any variant from Eqn. 2 or Eqn. 3, and U can be any variant from Eqn. 7 to Eqn. 9.

C. Skill (Un)certainty within Job Title

Methods summarized in the previous subsection have the underlying assumption that a skill required by many different titles is less unique. This *global* view has not considered the internal homogeneity within a job title. In this subsection, we leverage the *local* variation among job ads with the same title, and contrast it to the global variation.

In particular, for any given title, if a skill is required by almost all ads with the title, then we are almost certain that the skill is relevant to the title. On the contrary, if a skill is required by a relatively small fraction of ads with the title, we are also quite sure the skill is not relevant to the title. If around half of the ads require the skill, we are less certain whether the skill is relevant or not.

Table II
MATRIX OF SKILL-AD FRACTIONS

		Ads with Title t		
		Skill Required	Skill Not Required	Sum
Skills	1	q_{1t}	$1-q_{1t}$	1.00
	2	q_{2t}	$1-q_{2t}$	1.00
	⋮	⋮	⋮	⋮
	$\lvert S\rvert$	$q_{\lvert S\rvert t}$	$1-q_{\lvert S\rvert t}$	1.00

More specifically, for a given title t and each skill s, we may classify all ads with this title into two groups: those requiring skill s and those not requiring skill s. Then we can

create a matrix, like Table II, by normalizing the frequencies of the two classes into fractions. Based on this table, we can then calculate the same set of metrics as listed in the previous section (i.e., $\mathrm{U_{en}}(s,t)$, $\mathrm{U_{dp}}(s,t)$, $\mathrm{U_{var}}(s,t)$), plus a Gini-index based dispersion metric, calculated as follows:

$$\mathrm{U_g}(s,t) = q_{st}(1-q_{st}). \tag{11}$$

In summary, the weight of skill s for title t will be

$$\text{relevance}(s,t) = \mathrm{TF}(s,t) \cdot \frac{\mathrm{U^{global}}(s)}{\mathrm{U^{local}}(s,t)}, \tag{12}$$

where $\mathrm{U^{local}}(s,t)$ refers to any of the "local" certainty metrics calculated with the given title t, and $\mathrm{U^{global}}(s)$ refers to any of the "global" concentration metrics introduced in Section III-B.

Based on our observations, a trade-off needs to be considered between skill importance and skill uniqueness. Term frequency and its variants roughly represent the importance of a skill to a title. The intuition is that the more necessary a skill is, the more frequently it is mentioned. The IDF and the U (uncertainty, both at the global level and at the local level) scores can measure the uniqueness of a skill to a title. If a skill is associated with just a few titles, it is unique to those titles. If a skill is commonly required by many titles, it is not unique. Just like TF-IDF, when ranking skills for a title, we used a multiplication of the three factors, which requires the most relevant skills to be both important and unique to the title.

IV. Experiments

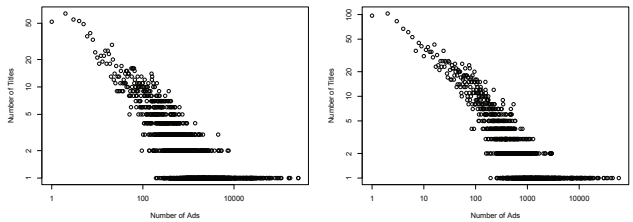
In this section, we present experimental results. We will first describe the data, the validation methodology, implementation details, and evaluation metrics. Then, we will discuss the results, including a summary of comparisons across ranking scores, performance on different job titles, and the deduplication effects.

A. Dataset and Preprocessing

Our experiments are based on a large-scale, real-world dataset. The initial dataset had 8,619,921 job ads classified into 4,983 unique, normalized job titles that involve 17,456 unique skill terms.

Deduplication was considered an important data preprocessing step in this study, because we are focused on skill-title relevance rather than the supply and demand of skills or positions. Since each skill is either required or not required in a job ad, the skill frequency is essentially the number of ads that require such a skill. Many duplicate job ads were created when employers frequently reposted or updated the opening's information. As a result, ads that are frequently reposted or ads with more openings will inflate the required skill frequencies. Therefore, we deduplicated the job ads by removing ads that have exactly the same normalized title and require exactly the same set of skills. After deduplication, we were left with 2,802,087 job ads.

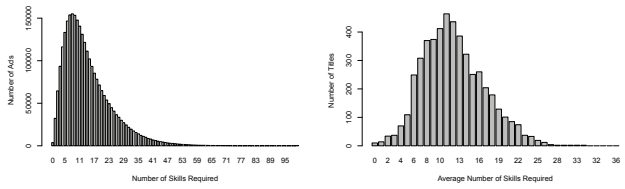
Figure 4 shows the distribution of the number of ads over various titles, before and after deduplication. Each subfigure uses the log-log scale, so that each data point indicates the number of ads per title (x-axis) and the corresponding number of job titles that have so many ads (y-axis) . We can see that both before and after deduplication, the scale-free property of the title-ad distribution is retained. In other words, there are a large number of titles that have just a few ads, and a small number of titles that have lots of ads.



(a) Before (b) After

Figure 4. Number of ads per title, before and after deduplication.

Figure 5 shows the distribution of the number of skills required by various titles. Figure 5(a) is the number of skills required per ad (with the frequency for "100 or more" aggregated into the last bar on the right). Figure 5(b) is the average number of skills required per ad across all ads with the same title. We can see that they both have a clear mode, yet skew slightly to the right. Requiring between 3 and 30 skills per ad is most common.



(a) Count By Ad (b) Average By Title

Figure 5. Number of skills required.

To evaluate the effectiveness of our approach, we collected ground truth skills for select job titles via expert ranking (see Section IV-B). Sixteen job titles were randomly selected, each out of a Standard Occupational Classification (SOC) major[3], and are listed in Table III. As we can see, our selection covers a variety of job titles, with a wide range in frequency (i.e., the number of ads under each title). In Table III, we provided the number of ads under each title before (i.e., "Raw") and after (i.e., "Dedup") deduplication. We can see that the pruning ratio (PR, i.e., percent of job ads removed) of the selected job titles varied from 39.01%

[3]http://www.bls.gov/soc/major_groups.htm

Table III
LIST OF JOB TITLES AND DEDUPLICATION EFFECTS

ID	Title	Counts			Overall (%)			TF_ONLY (%)			TF_LOG_DP_DP (%)		
		Dedup.	Raw	PR (%)	PK	MAP	NDCG	PK	MAP	NDCG	PK	MAP	NDCG
1	*Accountant*	30,102	52,920	43.12	0.00	0.00	+0.07	0.00	-0.29	-0.05	0.00	0.00	0.00
2	*Certified Nursing Assistant*	7,196	60,029	88.01	+13.33	+23.89	+6.27	+23.08	+39.99	+14.44	0.00	0.00	0.00
3	*Customer Service Representative*	58,509	258,969	77.41	+7.14	+1.92	+4.69	+7.69	+4.05	+6.88	+6.25	+1.04	+0.94
4	*Early Chdhd Spec Edu Teacher*	1,441	5,425	73.44	+44.44	+53.44	+28.32	+55.56	+156.39	+59.04	0.00	+2.80	+5.82
5	*Electrician*	1,291	4,051	68.13	+11.76	+36.91	+25.94	+30.77	+71.12	+40.06	0.00	0.00	-2.37
6	*Hair Stylist*	65	441	85.26	-16.67	-29.79	-3.49	-28.57	-47.23	-8.87	0.00	-13.65	-4.55
7	*Housekeeper*	2,459	14,229	82.72	+14.29	+41.58	+19.03	+55.56	+91.86	+37.89	0.00	+0.31	+0.07
8	*Maintenance Mechanic*	18,739	99,356	81.14	0.00	-0.55	+3.08	-6.67	-3.14	+4.68	+5.26	+5.54	+2.83
9	*Packer Inspector*	24	58	58.62	0.00	-2.44	+10.97	0.00	-2.08	+14.84	0.00	+3.80	+0.82
10	*Registered Nurse*	35,709	261,295	86.33	+30.77	+51.53	+17.93	+50.00	+78.51	+29.70	+5.56	+5.35	+2.78
11	*Research Scientist*	985	1,615	39.01	0.00	+6.13	+2.73	+7.69	+8.75	+3.36	0.00	0.00	+0.18
12	*Retail Merchandiser*	245	14,957	98.36	+36.36	+53.45	+21.65	+50.00	+74.89	+29.59	0.00	+5.62	+1.52
13	*Retail Sales Associate*	8,679	173,366	94.99	+23.08	+27.45	+16.60	+33.33	+29.17	+17.59	0.00	-0.46	+1.29
14	*Social Worker*	2,329	5,450	57.27	+5.26	+5.54	+2.02	0.00	+3.14	+1.69	0.00	0.00	+0.18
15	*Software Engineer*	41,376	87,140	52.52	0.00	0.00	+0.56	0.00	-2.25	-0.18	0.00	0.00	0.00
16	*Truck Driver*	5,535	164,306	96.63	+44.44	+46.66	+25.23	+42.86	+37.00	+25.23	+7.14	+6.77	+3.50

to 98.36%. Such a wide variation in pruning ratios might be related to two factors. First, the number of openings available under each title. In other words, the more openings or recruiting activities, the more similar ads (i.e., duplicates) were posted. Second, the heterogeneity in job skills required by a job title. In other words, if the job ads under the same title tended to require different sets of skills, the fewer similar ads would be found. Our evaluation dataset does cover job titles with different levels of pruning ratios.

B. Gathering Expert Ranking

For evaluation purposes, we created a validation dataset as follows. We started by generating a list of the top 50 skills by term frequency (before deduplication) for each of the 16 job titles. Then, we asked our in-house experts to label each of these skills into one of four levels:

- **Essential** skills correspond to those directly related to the job title, such as *Sales Management* to *Sales Manager*, and *Accounting* to *Accountant*.
- **Important** skills are those strongly related to the job. For example, *Microsoft Excel* and *Balance Sheets* are important skills for an *Accountant*.
- **OK** skills are somewhat related to the job. For example, *Financial Data Vendor* and *Business Support Systems* are OK skills for an *Accountant*.
- **Wrong** skills are those considered not relevant. For example, *Global Marketing* is a wrong skill for a *Legal Assistant*, and *Business Administration* is a wrong skill for a *Software Developer*.

Normally, the set of essential skills is not very big. We expect to see between 2 to 5 essential skills per title. The set of important skills makes up the majority of relevant skills. After all, given the way of the initial candidate generation process, we would expect many of the skills to be relevant. Accordingly, the "*OK*" group and the "*wrong*" group are usually rather small. Table IV and Table V provide select examples of hand labeled skills for a few titles.

Table IV
EXAMPLE SKILLS FOR TWO RETAIL RELATED TITLES

	Retail Sales Associate	*Retail Merchandiser*
Essential	• Retailing • Sales • Retail Sales	• Retailing • Sales • Merchandising
Important	• Customer Service • Cash Register • Merchandising	• Customer Service • Category Management • Stock Rotation

Table IV lists sample skills for two retail related job titles, the *Retail Sales Associate* and the *Retail Merchandiser*. Both job titles require essential skills of *Retailing* and *Sales*. While *Merchandising* is considered an essential skill for *Retail Merchandiser*, it is considered merely an important skill for *Retail Sales Associate*. These two job titles also share some important skills, such as *Customer Service*, and there are many other important skills that make them different. For example, while the *Cash Register* is an important skill for a *Retail Sales Associate*, it is not one for a *Retail Merchandiser*, who work more in the area of stock and inventory operations.

Table V
EXAMPLE SKILLS FOR TWO NURSING RELATED TITLES

	Registered Nurse	*Certified Nursing Assistant (CNA)*
Essential	• Health Care • Nursing • Registered Nurse	• Health Care • Nursing • Nursing Assistance • Certified Nursing Assistant
Important	• Hospitalization • Basic Life Support • Advanced Cardiovascular Life Support (ACLS) • Intensive Care Unit	• Hospitalization • Basic Life Support • Blood Pressure • Bed-Making • Assisted Living

Similarly, Table V lists sample skills for *Registered Nurse* and *Certified Nursing Assistant (CNA)*. These two nursing related job titles also share some of the essential and important skills (e.g., *Health Care*, *Basic Life Support*), and differ in others. We can also see that since the nursing related titles are quite different from retail related titles, the skills do not overlap across the two pairs of job titles.

C. Implementation

We ran the experiments on a 16-core Amazon EC2 instance (c4.4xlarge). All methods were implemented in Python. Specifically, we used the data analysis library Pandas [13] for data cleaning and transformation. Scientific computing libraries Numpy/Scipy [14] were employed for basic vector operations as well as some advanced statistics calculations, such as variance and entropy.

We implemented 23 metrics, as outlined in the first five columns in Table VI. Our baseline was M1, the approach based on raw frequency. We grouped the other metrics into four groups, depending on the uniqueness metric (i.e., IDF, entropy, DP, and VAR) used.

D. Evaluation Metrics

After we gather responses for a requested set of titles, we evaluated the performances of the methods considered. As all methods discussed above will return a ranked list of relevant skills for a title, we used the following evaluation metrics, which are commonly used for evaluating ranking performances:

- **Precision at k (P@K)** is calculated as
$$P@k = \frac{r}{k}, \tag{13}$$
where r is the number of relevant results out of the first k results on the list.
- **Mean average precision (MAP)** is calculated as
$$MAP(k) = \frac{1}{k}\sum_{i=1}^{k} P@i \tag{14}$$
- **Normalized discounted cumulative gain (NDCG)** is calculated as
$$nDCG(k) = \frac{DCG(k)}{IDCG(k)}, \tag{15}$$
where
$$DCG(k) = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \tag{16}$$
rel_i is the relevance score assigned to the i-th skill, and $IDCG(k)$ is the maximum possible (ideal) $DCG(k)$.

For the above metrics, we considered the precision of the top 25 skills found by each of our methods, in comparison with the expert labeled ground truth. This parameter was chosen because it was not only a reasonable number in practice (see Figure 5), but also close to the number of expert-labeled relevant skills (for most jobs).

E. A Summary of Results

A summary of performance is provided in Table VI and visualized in Figure 6. The columns PK, MAP, and NDCG were the performance scores (average taken across all job titles). We can see that they have different ranges (0.29–0.84, 0.17–0.78, 0.38–0.86, respectively), and were highly correlated ($\rho > 0.9$). The normalized average score (Norm. Avg.) is taken among the three different metrics after each was re-scaled by

$$newscore = \frac{oldscore - min(oldscore)}{max(oldscore) - min(oldscore)}.$$

We can see that except for M15, our uniqueness scores do provide enhancements over the simple, frequency-based baseline (i.e., M1). The classic TF-IDF measures (M2 and M3) performed fairly strong in comparison to M1; however, they were not successful when the logarithm was applied to the frequencies. The most successful metric turned out to be M16, which takes logarithm of the term frequency, and evaluates DP both globally and locally.

Since all three performance metrics are highly correlated with each other, we may look at one of them (e.g., NDCG) in detail. Focusing on NDCG@25, we visualized the distribution of the performance metric of each method over all datasets in Figure 6. The results were consistent with the overall summary, and provided more details about the distribution across job titles. We can see that compared to M1, M3 is better by having a higher average and lower variation. As our most successful approach, M16 performs better than others for many titles, but also has a couple of outliers with an NDCG score of around 0.6.

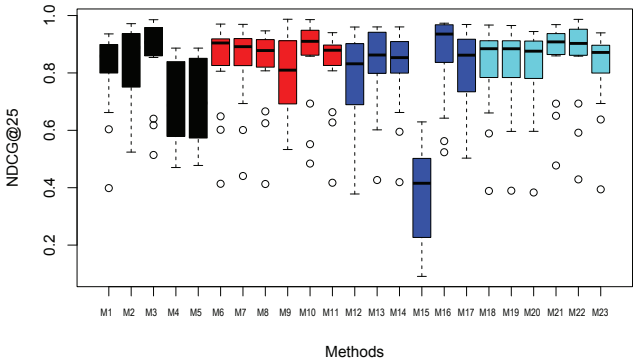


Figure 6. Performance comparison. Each bar represents a method, indexed on the x-axis and outlined in Table VI. Black = TF-IDF based methods (M1 to M5). Red = TF-Entropy based methods (M6 to M11). Dark blue = TF-DP based methods (M12 to M17). Light blue = TF-VAR based methods (M18 to M23).

Instead of looking at the mean performance, we might also summarize the number of times each method beats others. Figure 7 visualized the number of job titles on which each metric on the row outperforms (with a score not lower than) another on the columns. The color and size of circles are proportional to the number of job titles (out of a total of 16) when this happens. In other words, if we see a pattern in larger and darker circles in the columns, the corresponding method in the column is not that competitive, since it is dominated by many others. Similarly, if we see a pattern in larger and darker circles in the rows, the corresponding

Table VI
METHODS IMPLEMENTED AND OVERALL PERFORMANCE (ACROSS ALL TITLES)

ID	Method	Importance	Uniq.(global)	Uniq. (local)	P@K	MAP	NDCG	Norm. Avg.	Ranking
M1	TF_ONLY	TF_{raw}	(None)	(None)	0.7000	0.6153	0.8111	0.7896	20
M2	TF_RAW_IDF_RAW	TF_{raw}	IDF_{raw}	(None)	0.8063	0.7608	0.8371	0.9523	5
M3	TF_RAW_IDF_MAX	TF_{raw}	IDF_{max}	(None)	0.8094	0.7514	0.8611	0.9658	3
M4	TF_LOG_IDF_RAW	TF_{log}	IDF_{raw}	(None)	0.6656	0.5899	0.6908	0.6713	22
M5	TF_LOG_IDF_MAX	TF_{log}	IDF_{max}	(None)	0.6781	0.6033	0.7003	0.6928	21
M6	TF_RAW_ENTROPY_global	TF_{raw}	Entropy	(None)	0.7500	0.6705	0.8342	0.8664	12
M7	TF_RAW_ENTROPY_ENTROPY	TF_{raw}	Entropy	Entropy	0.7500	0.6675	0.8370	0.8668	11
M8	TF_RAW_ENTROPY_G	TF_{raw}	Entropy	G	0.7313	0.6495	0.8290	0.8399	14
M9	TF_LOG_ENTROPY_global	TF_{log}	Entropy	(None)	0.8063	0.7129	0.7873	0.8918	8
M10	TF_LOG_ENTROPY_ENTROPY	TF_{log}	Entropy	Entropy	0.8063	0.7454	0.8577	0.9582	4
M11	TF_LOG_ENTROPY_G	TF_{log}	Entropy	G	0.7375	0.6581	0.8247	0.8454	13
M12	TF_RAW_DP_global	TF_{raw}	DP	(None)	0.7375	0.6557	0.7819	0.8145	18
M13	TF_RAW_DP_DP	TF_{raw}	DP	DP	0.7688	0.6893	0.8283	0.8842	9
M14	TF_RAW_DP_G	TF_{raw}	DP	G	0.7688	0.6891	0.8173	0.8765	10
M15	TF_LOG_DP_global	TF_{log}	DP	(None)	0.2969	0.1690	0.3807	0.0000	23
M16	TF_LOG_DP_DP	TF_{log}	DP	DP	0.8375	0.7810	0.8622	1.0000	1
M17	TF_LOG_DP_G	TF_{log}	DP	G	0.8188	0.7577	0.8136	0.9421	7
M18	TF_RAW_VAR_global	TF_{raw}	VAR	(None)	0.7219	0.6431	0.8190	0.8237	16
M19	TF_RAW_VAR_VAR	TF_{raw}	VAR	VAR	0.7219	0.6418	0.8189	0.8229	17
M20	TF_RAW_VAR_G	TF_{raw}	VAR	G	0.7156	0.6322	0.8145	0.8108	19
M21	TF_LOG_VAR_global	TF_{log}	VAR	(None)	0.8219	0.7743	0.8612	0.9860	2
M22	TF_LOG_VAR_VAR	TF_{log}	VAR	VAR	0.8000	0.7383	0.8552	0.9488	6
M23	TF_LOG_VAR_G	TF_{log}	VAR	G	0.7313	0.6507	0.8187	0.8334	15

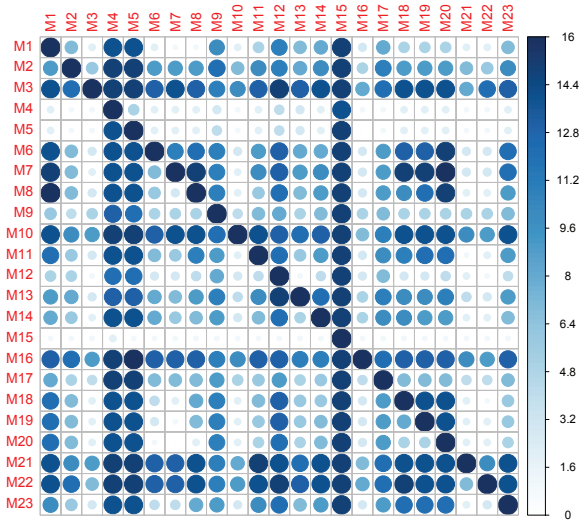


Figure 7. Pairwise comparison among methods listed in Table VI (based on NDCG, after deduplication). Cell (i, j) represents the number of job titles (out of all listed in Table III) on which Method i outperformed Method j.

method in the row is quite strong. We can see that M3, M10, M16, M21, M22 are among the stronger approaches, which is consistent with the overall summary.

F. De-Duplication Effects

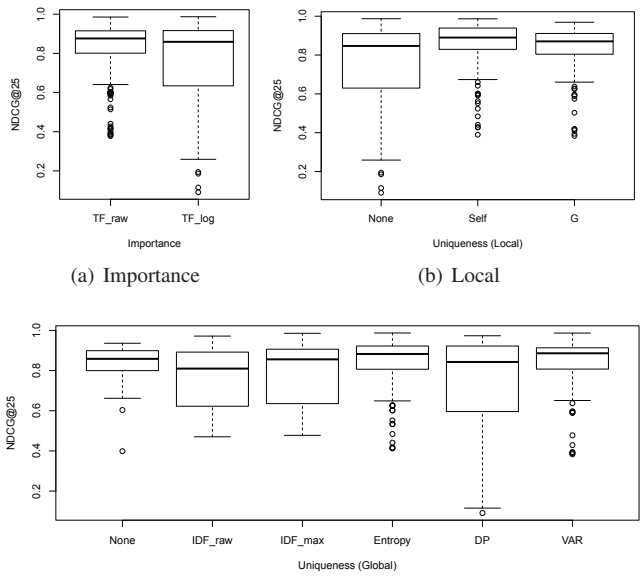
Table III shows the average improvement (in percentages) in performance metrics over all methods considered for each title. We can see that for each of the performance metrics, most titles showed an improvement after deduplication. In particular, *Certified Nursing Assistant (CNA)* and *Registered Nurse* are two similar titles with a high pruning ratio and much improvement from deduplication. We might infer that jobs under these titles are somewhat similar in skill requirements. The negative amounts for *Software Engineer* and *Maintenance Mechanic* are insignificant. The negative amounts for *Hair Stylist* likely indicate that the threshold we used to deduplicate was too high for this title, which was a less common job title with a smaller sample size. Deduplication made the sample size even smaller, and thus results were less stable. In contrast, for frequent titles, the extent of improvement still varies. For example, *Accountant* does not show much improvement after deduplication. Combined with the fact that the pruning ratio for this title was also mediocre, we might infer that this is a diverse title. In other words, ads under this title are quite heterogeneous in terms of skill requirements.

Comparing TF_ONLY (M1) and TF_LOG_DP_DP (M16), as shown in Table III, we can see that M16, which was the champion method, had many zeros or relatively small absolute values, meaning that deduplication did not help. This also implies that M16 is more robust to duplicate job postings.

G. Balancing Importance and Uniqueness

To study the effect of importance and uniqueness measures in isolation, we summarize the performance metrics by each component of the relevance score. Figure 8 compares the performance metrics by varying one component at a time. Figure 8(a) indicates that TF_RAW is generally better than TF_LOG, if importance is considered alone. Figure 8(c) indicates that at the global level, None, Entropy, or VAR are among the better choices, if uniqueness is the only factor to rank skill relevance. Figure 8(b) indicates that when a

uniqueness metric is used at the global level, it is best also used at the local level, or at least some adjustment of local uniqueness (e.g., using G).



(a) Importance (b) Local

(c) Global

Figure 8. Effect of single factors.

V. Related Work

The values of skills at the labor market have attracted extensive interest in business and management research. For example, career path development, the value of employee skills to organizations [15], [16], [4], utility of skills in an online labor market [2], and the evolution of skill requirements in an industry [1]. In particular, it has been reported that certain skills (e.g., analytics) play different roles in process driven and innovation driven companies [3].

A number of existing studies have attempted to address the problem of skill extraction from unstructured textual data. For example, a graph-based approach to skills extraction from text was proposed in [17]. The study exploits a list of skills obtained from LinkedIn, associates them with Wikipedia pages, and utilizes such associations to extract skill terms from any given query text. To infer whether a person possesses a skill, a factor graph model was used to integrate personal connections and skill associations [18]. SkillFinder is an algorithm for enhancing skill extraction from a skill-resume matching system [19]. WorkerRank is a system to infer workers' reputation in skills using employer implicit judgements [20]. Implicit ratings inferred from application, interviewing, and hiring decisions were used to model worker ratings. However, this requires hiring data from the employer's side. These studies have quite a different focus from ours, and have to leverage additional, external information.

Identifying skills and finding experts has been studied to create real-world systems. A number of studies on expert finding and team forming focused on learning partial and multiple labels, which can be noisy. As one of the largest professional social networking service sites, LinkedIn has done extensive research in skill extraction, such as the LinkedIn Endorsements[4] feature, and the skill recommender system [21]. Some related papers from IBM Research include identifying existing skills [22], and predicting and recommending skills [23] for employees of a big organization who are connected via online social tools.

CareerBuilder has created a suite of in-house data and products. Carotene, CareerBuilder's job title classification system [6], [7], can classify each job ad into a two-level job category taxonomy. Each job ad is converted into one of many normalized job titles. CareerBuilder's skill term extraction system [5] can extract skill keywords from both job ads and job seekers' resumes. Each skill is then represented as a normalized skill term, so that each job ad and each resume can be deemed as a document of words. CareerBuilder's employer name normalization system [24] can process employer names from both job postings and resumes from an entity resolution perspective.

VI. Conclusions and Future Work

In this study, we investigated the problem of quantifying skill relevance to job titles. The outcome of this study can support a number of intelligent solutions at CareerBuilder. Since frequency-based skill weighting resulted in the most generic skills ranked on the top, yet they were not the most relevant, we first considered TF-IDF adjustments. To further improve the TF-IDF metrics, we also considered measuring the concentration or dispersion of skills across job titles. While the TF-IDF measure only considers the skill-title relationship like document-term relations, we further considered the variation within a given job title, where the variation among job ads were considered. By collecting and comparing with expert ranked skills for a random set of job titles, our experiments showed that the (un)certainty measures did help improve skill rankings, especially when we used the DP for both global and local uniqueness measures. We also found that the performance of all such measure vary greatly among different titles, and deduplicating similar ads before computing relevance scores has consistently helped improve the performance.

For future work, we intend to explore the following options. Firstly, we shall consider weighted versions of the metrics, including the ad-title confidence scores and the skill-ad confidence scores. While we have such information available in the Carotene system, the current study has not leveraged these scores. Secondly, since the relationship among titles, ads, and skills can be represented with tripartite

[4]http://www.slideshare.net/pskomoroch/strata-endorsements-16939466

or bipartite graphs, graph learning techniques could be used to investigate skill relevance in a more flexible manner. Finally, designing a smart, dynamic learning system to do all the learning and evaluations automatically would be useful. For example, we could integrate the expert or crowd labeled skills into a semi-supervised ranking system. We could also devise a dynamic ensemble learning framework that uses the various metrics as input, and outputs the best results.

REFERENCES

[1] P. A. Todd, J. D. McKeen, and R. B. Gallupe, "The evolution of is job skills: a content analysis of IS job advertisements from 1970 to 1990," *MIS Quarterly*, pp. 1–27, 1995.

[2] M. Kokkodis and P. G. Ipeirotis, "The utility of skills in online labor markets," in *Proceedings of the 2014 International Conference on Information Systems (ICIS)*, 2014.

[3] L. Hitt, F. Jin, and L. Wu, "Data skills and value of social media: Evidence from large-sample firm value analysis," in *Proceedings of the 2015 International Conference on Information Systems (ICIS)*, 2015.

[4] L. Wu and L. M. Hitt, "How do data skills affect firm productivity: Evidence from process-driven vs. innovation-driven practices," The Wharton School Research Paper 86, February 9 2016, available at SSRN: http://ssrn.com/abstract=2744789.

[5] M. Zhao, F. Javed, F. Jacob, and M. McNair, "SKILL: A system for skill identification and normalization," in *Proceedings of the 28th Conference on Artificial Intelligence (AAAI 2014)*, 2014, pp. 4012–4018.

[6] F. Javed, M. McNair, F. Jacob, and M. Zhao, "Towards a job title classification system," in *WSDM 2014 Workshop on Web-scale Classification: Classifying Big Data from the Web (WSCBD)*, 2014.

[7] F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao, and T. S. Kang, "Carotene: A job title classification system for the online recruitment domain," in *Proceedings of the 1st IEEE International Conference on Big Data Computing Service and Applications (BigDataService)*, 2015, pp. 286–293.

[8] D. Harris and S. Harris, *Digital design and computer architecture*, 2nd ed. San Francisco, Calif.: Morgan Kaufmann, 2012, page 129.

[9] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[10] K. S. Jones, *Document Retrieval Systems*, 1988, ch. A Statistical Interpretation of Term Specificity and Its Application in Retrieval, pp. 132–142.

[11] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Transactions on Information Systems*, vol. 26, no. 3, pp. 13:1–13:37, 2008.

[12] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[13] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, 2010, pp. 51–56.

[14] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, http://www.scipy.org/ [Online; accessed 2016-05-26].

[15] S. Iranzo, F. Schivardi, and E. Tosetti, "Skill dispersion and firm productivity: An analysis with employer-employee matched data," *Journal of Labor Economics*, vol. 26, no. 2, pp. 247–285, April 2008.

[16] B. Mahy, F. Rycx, and G. Vermeylen, "Educational mismatch and firm productivity: Do skills, technology and uncertainty matter?" *De Economist*, vol. 163, no. 2, pp. 233–262, June 2015.

[17] I. Kivimäki, A. Panchenko, A. Dessy, D. Verdegem, P. Francq, C. Fairon, H. Bersini, and M. Saerens, "A graph-based approach to skill extraction from text," in *Proceedings of the TextGraphs-8 Workshop*, 2013, pp. 79–87.

[18] Z. Wang, S. Li, H. Shi, and G. Zhou, "Skill inference with personal and skill connections," in *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, 2014, pp. 520–529.

[19] T. R. Kalva, "Skill finder: Automated job-resume matching system," Master's thesis, Utah State University, 2013.

[20] M. Daltayanni, L. de Alfaro, and P. Papadimitriou, "WorkerRank: Using employer implicit judgements to infer worker reputation," in *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM)*, 2015, pp. 263–272.

[21] M. Bastian, M. Hayes, W. Vaughan, S. Shah, P. Skomoroch, H. Kim, S. Uryasev, and C. Lloyd, "Linkedin skills: large-scale topic extraction and inference," in *RecSys*, 2014, pp. 1–8.

[22] K. R. Varshney, J. Wang, A. Mojsilovic, D. Fang, and J. H. Bauer, "Predicting and recommending skills in the social enterprise," AAAI ICWSM Workshop on Social Computing for Workforce 2.0, AAAI Technical Report WS-13-02, 2013.

[23] K. R. Varshney, V. Chenthamarakshan, S. W. Fancher, J. Wang, D. Fang, and A. Mojsilović, "Predicting employee expertise for talent management in the enterprise," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014, pp. 1729–1738.

[24] Q. Liu, F. Javed, and M. McNair, "Companydepot: Employer name normalization in the online recruitment industry," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 521–530.