Predicting Employee Expertise for Talent Management in the Enterprise

Kush R. Varshney,[1] Vijil Chenthamarakshan,[1] Scott W. Fancher,[2]
Jun Wang,[1] Dongping Fang,[1] and Aleksandra Mojsilović[1]
[1]IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, New York
[2]IBM Corporate Headquarters, 1 New Orchard Road, Armonk, New York
{krvarshn,ecvijil,sfancher,wangjun,dfang,aleksand}@us.ibm.com

ABSTRACT

Strategic planning and talent management in large enterprises composed of knowledge workers requires complete, accurate, and up-to-date representation of the expertise of employees in a form that integrates with business processes. Like other similar organizations operating in dynamic environments, the IBM Corporation strives to maintain such current and correct information, specifically assessments of employees against job roles and skill sets from its expertise taxonomy. In this work, we deploy an analytics-driven solution that infers the expertise of employees through the mining of enterprise and social data that is not specifically generated and collected for expertise inference. We consider job role and specialty prediction and pose them as supervised classification problems. We evaluate a large number of feature sets, predictive models and postprocessing algorithms, and choose a combination for deployment. This expertise analytics system has been deployed for key employee population segments, yielding large reductions in manual effort and the ability to continually and consistently serve up-to-date and accurate data for several business functions. This expertise management system is in the process of being deployed throughout the corporation.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.2.8 [**Database Management**]: Database Applications—*Data Mining*

Keywords

expertise assessment; human resources; supervised classification; talent planning; workforce analytics

1. INTRODUCTION

The International Business Machines (IBM) Corporation is one of the largest employers of knowledge workers in the world. The global workforce is over 425,000 employees strong,

with greater and greater fractions selling or delivering information technology and business consulting services every day. Two key trends in the information technology industry include a quickening pace of technological innovation and an increased customer-centrism by all employees. In the face of these trends, it is important for IBM's strategic and tactical business decision-making to be informed by complete, precise, accurate, and up-to-date information on the expertise of its employees. What team—in terms of composition of employee skills—should serve a particular client, which emerging technology areas IBM has adequate skills and experience in to support (e.g., cloud computing, cognitive computing, social business), and other similar decisions are all predicated on such information [12, 18, 1].

The human capital, or talent and expertise, of its workforce is IBM's greatest asset; only through valid and trusted data on employee expertise can this human capital be utilized to its fullest potential. Such expertise information is captured by existing expertise management systems within the IBM enterprise, but can be enhanced using digital 'breadcrumbs' and data 'exhaust' from other enterprise systems of record that have 'footprints' of employees' work activities (e.g. project claims, sales pipeline, etc.). However these pieces of data often provide only a diffuse indication of expertise. The main contribution of this work is using predictive analytics to churn such nebulous data into clear indicators of employee expertise in a way that commingles with existing IBM business processes.

Human resources (HR) departments are transforming from support functions to strategy leadership in many ways. David Bernstein, Vice President of Big Data for HR at eQuest, recently said in discussing different levels of talent analytics (operational reporting, advanced reporting, advanced analytics, and predictive analytics) that[1] "predictive analytics is where HR develops predictive models and integrates with the organization's strategic planning. The majority of organizations, however, are not doing this, yet." The fraction of organizations engaged in predictive HR analytics is reported to be as low as 4% [2]. Recent predictive modeling initiatives put IBM's HR organization into the 4% category. However, like most other organizations, the newly adopted predictive HR analytics in IBM are mostly focused on recruitment and hiring, resource deployment and proactive talent retention [22, 16, 6, 21]. Predicting expertise is a much more open-ended problem than the others.

---

[1]http://www.bigdatarepublic.com/author.asp?section_id=3431

Extant expertise systems, such as LinkedIn's skill recommendation, are completely free-form in how skills are described. In order to operate the IBM business, however, we need a semi-structured approach which relies on IBM's existing expertise taxonomy to link to solutions and products that IBM sells and to tap into an entire ecosystem of processes and reporting tools built around the taxonomy. The existing expertise assessment system and taxonomy yield many benefits to the company, but have a few shortcomings in keeping data current and accurate as detailed in Section 2. Therefore, our goal is to start with the IBM expertise taxonomy and assessments, pull in other data sources informative of expertise, and create good predictions of employee proficiencies.

In our work, we specifically investigate the prediction of an employee's primary job role and job role specialty (which are the third and fourth levels of the five-level IBM expertise taxonomy) from enterprise systems of record and data from the internal corporate social networking site IBM Connections as features. Since employees have a single primary job role among a set of labels with not too exorbitant cardinality, we approach the predictive modeling problem as a classification problem. Since we have veracious data on a reasonable fraction of employees' job roles and job role specialties, we can further formulate the problem via supervised multi-category classification, using the veracious fraction of employees' data as a training set. We perform a comprehensive study of classification algorithms and feature sets, finding the Liblinear implementation of logistic regression using job title and HR information features to perform with best generalization accuracy [9]. Moreover, since employees tend to be grouped in the IBM organizational structure by job role and skill set, e.g. mostly strategy consultants under one manager and mostly information architects under a different manager, we attempt to improve classification accuracy taking these relationships into account, but do not find enough improvement to warrant inclusion in the deployment.

In previous work, we investigated the prediction of skills —the lowest level of the IBM expertise taxonomy—using matrix completion-based collaborative filtering with side information [24, 27], including in the cold-start regime [10]. This problem is different because unlike skills, which employees have many of, employees only have one primary job role and one primary job role specialty each. Therefore, collaborative filtering is not appropriate for job role and job role specialty prediction. Other approaches for predicting the expertise of employees within an enterprise such as IBM are focused on social media data only and importantly, are not meant to be integrated with existing business processes [13, 20, 4, 11, 5], whereas we consider other types of enterprise data as well and integrate with existing practices.

Predicting primary job roles and primary job specialties is mainly intended to improve processes related to and built upon the 'inventorying' of employees. This work represents a part of a larger effort to transform the way the IBM Corporation operates. A complimentary piece of work is enabling expertise search [23, 7, 8, 19, 14], i.e. the ability for employees to find other employees who are experts on some topic. Expertise search is driven by the need of employees to find the right person to answer a question, to bring to a client meeting, to fill out a team, and so on. Another part of the effort is the creation of enhanced profile pages known as digital business cards that better enable the collection of expertise-related information and provide enhanced user experience. These parts have been deployed within IBM through web-based and mobile-based applications.

The job role and job role specialty prediction algorithms we have developed are in various stages of deployment in IBM. Predicted job roles from our algorithm have been applied to the population of all salespeople worldwide in the company. This employee segment represents less than 10% of the employee population and is more active in maintaining accurate job roles than other employee populations. Administrators responsible for these employees used our outputs to accurately enter job role values for thousands of salespeople with either no job roles entered or an invalid or outdated job role entered. This initial single one-time deployment saved the company the equivalent of one employee's full-time effort for an entire year, not to mention all of the business benefits from current, correct data. Moreover, we are now in the process of rolling out the predictive models with a user-centered design to allow administrators in the entire company to continually maintain complete and accurate job roles and job role specialties with corresponding compounded effort savings.

The remainder of the paper is organized as follows. In Section 2, we provide a description of the expertise assessment system and processes that IBM currently uses. In Section 3, we formulate job role and job role specialty prediction as a supervised classification problem and describe the features available for making the prediction. Section 4 details our empirical study of various features and classification algorithms. We then describe the deployment of the approach within IBM in Section 5. The paper concludes in Section 6.

2. BACKGROUND AND SIGNIFICANCE

In this section, we focus on describing the problem in detail along with its significance. We first recapitulate the history of expertise assessment in IBM and show how this history motivates an evolution to predictive modeling, as developed in this work.

2.1 Expertise Taxonomy

IBM has a taxonomy of employee expertise with the following five coarse-to-fine levels: primary job category, secondary job category, job role, job role specialty, and skill. The taxonomy is a directed acyclic graph with parent-child relationships between values at different levels. We provide three examples of paths through the taxonomy with the five different levels separated by the greater than symbol: Sales > Industry Sales > Brand Client Representative > Brand Client Representative: BAO-Advanced Analytics & Optimization > Sell ILOG Optimization; Human Resources > Learning > Learning Consultant > Learning Consultant: Collaboration, Knowledge & Communities > Analyze Performance Improvement Needs; Research > Research Staff > Research Scientist > Research Scientist: Computational Biology > Develop Algorithms for Biological Data Analysis. An individual employee has a single primary job role and primary job role specialty, which are the values we predict in this work.

2.2 Existing Expertise Assessment System

IBM utilizes an internally-developed application to collect expertise assessments from employees. The system was deployed in the year 2000 and at that time, contained many

leading-edge functions, numerous approaches to customize for different parts of IBM's business organizations, and a befitting user interface for employees. The basic function of the application is:

1. collecting information from the employee on what area of the business he or she works in,

2. having the employee select a primary job role and job role specialty (what he or she is assigned to do), and

3. presenting each employee with a list of expertise skills to assess on a 5 point scale.

The application interfaces directly with the IBM expertise taxonomy, which contains the relationships of expertise skills to the job role classification hierarchy. For example, each job role has a set of predefined expertise skills that are associated with it; based upon the employee-provided job role, the application asks the employee to assess the skills related to that job role.

While it is simple in concept, as IBM organizations have desired to collect finer expertise information over time, the taxonomy has grown, the number of job role specialties has grown, and the business rules that determine which employees assess which job role specialties has become quite complex. In 2000, the number of job roles was around 200 and the number of specialties 1000; today, there are around 350 job roles and 7000 specialties. The goal is to use the information collected about employees to ensure that the expertise skills which they are asked to assess directly pertain to their success in the their job: 'the right skills to the right employees.' However, the business rules sometimes deliver more than 100 skills for employees to assess, which results in a suboptimal user experience and subsequent lack of compliance.

While the current expertise assessment application has served its purpose for many years, it was designed during a time when IBM's product portfolio changed perhaps once every three or four years. Updating the taxonomy requires adding new skills to reflect new product knowledge. Updating the application's expertise delivery business rules also requires some thought to determine who should assess these new product skills. Therefore in an environment where an organization could set up the tool once every two years and have employees assess against a relatively constant set of expertise skills, the work required to set up the tool was relatively small compared to the benefit of the resulting expertise information collected.

2.3 Motivation for Predictive Modeling

Today IBM is a much more dynamically-changing company with new products, solutions, and acquisitions emerging each quarter. Thus the time and work required to set up a structured assessment begins to limit the ability to have expertise information that is responsive to the rapidly-changing needs of the business. These rapidly-changing needs of the business are actually dictating that the company evolve to a new expertise assessment approach that is much more flexible and requires much less time and effort to set up. Any information collected directly from employees requires some degree of end user design, setup, and testing. Information inferred from normal work products bypasses much of this design, setup, and testing by its very nature.

Through the proposed work, IBM is moving toward a world in which predictive analytics based upon employees' digital footprints is almost constantly updating the current inventory of expertise across the company. The eventual goal is to move from the world of counting and taking inventory once a year based upon a taxonomic system that was put in place a year earlier, to a world of instantly updated inventories: updated as soon as new terminology finds its way to the folksonomy and as soon as an employee creates digital evidence of new expertise gained. All the benefits that supply chain systems have realized over the last twenty years, with just-in-time inventories, point-of-sale inventory updates, economic order quantities, and predictive inventory demand are becoming possible for human resources and expertise management.

3. MACHINE LEARNING FORMULATION

In this section, we discuss the decisions and tradeoffs in making design choices for the data mining solution to the job role and specialty prediction problem within the IBM enterprise. In Wagstaff's three stages of a machine learning research program, this section is devoted to the necessary preparation phase: phrasing the problem as a machine learning task, collecting data, and selecting or generating features [25].

3.1 Supervised Multi-Category Classification

Several different machine learning or information retrieval formulations are possible for predicting and recommending primary job roles and job role specialties. We first lay out the desiderata of the solution and then comment on the choices made.

As mentioned earlier, an employee is supposed to have one primary job role and one primary job role specialty. Each line of business (LOB) maintains a short list, or subset, of job roles and specialties in the taxonomy that its employees may validly have. A reasonable fraction of employees throughout IBM have valid job roles and specialties, but there exists a significant fraction with values outside their LOB's short list or with no value entered at all.

The task at hand is two-fold: first, filling in the job role and specialty of any employee whose value is blank or invalid according to their LOB; and second, identifying employees whose values, although valid, are nevertheless incorrect based on their current job responsibilities and correcting them. The users of the system are not intended to be individual employees, but managers, administrators, and planners. An analytics solution is not intended to be fully automatic, but to recommend predictions and corrections that the user may or may not accept for each employee. In this mode of operation, a list of top k predicted job roles and specialties along with confidence values in the predictions may be more useful than a single prediction without an indication of confidence (wth k not more than three or four from a human factors concern).

One approach to tackle this problem is through an information retrieval formulation. We can index various data sources that provide an indication of expertise. Then, performing searches using query terms related to each job role or specialty, we can see in which search result the employee is ranked highest. However, constructing appropriate query terms that well-differentiate often-similar job roles and specialties is difficult and this approach is an indirect way of

tackling the problem. The direct way of approaching the problem is via classification because the task is to label an individual from a fairly small set of labels. As discussed in Section 1, collaborative filtering (narrowly construed as predicting the preference of an individual by collecting preferences from many individuals) is not appropriate because each employee has only one primary job role and specialty.

Classification need not be based on statistical machine learning. Oftentimes, manually-specified logical classification rules developed by subject matter experts perform excellently. However, as discussed, such manual work is not sustainable in the dynamic business environment of today. For a machine learning approach, we need to have reliable training data. The reasonable fraction of employees with non-blank and valid labels is sufficient for the first task even though there is label noise (the whole reason for the second task), because this noise is not overwhelming.[2] For the second task, we can take a cross-validation approach to construct training and test sets in such a way that we can identify employees that do not fit the pattern when they are a part of the test set.

We do not consider semi-supervised learning because we want all unlabeled samples (the invalid- and blank-valued employees) to be a part of the test set in the first task. Moreover, we focus on classification rather than learning to rank because of the form of the training data and because we are typically only concerned with accurately obtaining the best one, two, three, or four labels for an employee, not the entire permutation of ranked job roles or specialties for an employee.

Thus overall for the skills analytics problem we are facing, the most direct and appropriate formulation is supervised multi-category classification. Moreover, due to the business structure, we learn separate classifiers for the different LOBs within IBM because there are different valid class labels and different feature distributions among the different LOBs. Multi-task learning could be possible in this setting to do joint training for different LOBs, but we choose not to pursue this direction because it introduces unnecessary complexity.

3.2 Data Sources

There are several potential data sources that give some indication (sometimes in a quite diffuse way) about an employee's expertise, specifically as manifested in his or her job role and job role specialty. Within the IBM context, we can divide the potential data sources into four categories. The first category is employees' descriptions of their job in their own words. For many years, one line shown underneath the employee's name in the corporate directory has been a free-text field entered and edited by the employee. This text field is known as the job title. A similar, but longer, text field on the recently deployed digital business card mentioned in Section 1 deemed 'what I'm known for' is also of this same category of data sources, but we do not use it in the work described herein because it is in its incipient stages of adoption.

The second data source category is standard HR information recorded about all employees as standard practice.

Example fields include the length of service; the work location; the names of the group, department, and business unit to which the employee belongs; the job category of the employee, which is the highest level of the expertise taxonomy, but is maintained in a separate manner than job role and job role specialty; the employee's pay grade; and whether he or she is on a commission-based compensation plan.

The third category of data sources is work products, which includes artifacts produced by employees as part of their regular job responsibilities. As such, in contrast to the other data source categories, this category is job-specific. For example, researchers produce publications and patents, software developers produce documentation and code, salespeople work on and record sales opportunities, delivery personnel bill claims for their services, and so on. These days, much of this activity is captured digitally in various systems, but accessing and working with such data is often difficult. In this paper, we present empirical results on the LOB containing salespeople and show results from sales opportunity data.

Activity on the internal corporate social networking site, IBM Connections, constitutes the fourth data source category. There are several different types of social media content, including profiles, status updates, blogs, wikis, communities, forums, bookmarks, and calendars. They are also all associated with tags. Certain social media content is private whereas other content is public. For this project, we may only access public content and tags. Private material is typically confidential precisely because it relates to ongoing work activities, and thus its exclusion presents a limitation. Public content is not always work-related: there are communities for sports leagues, ethnic affinity groups, etc. too.

3.3 Feature Engineering

For each employee, we collect content from the HR information management system, business data warehouse, personal profile, and internal enterprise social network. Fig. 1 demonstrates examples of the extracted raw data, including the job title information from the HR data warehouse, sales opportunity data from the business data warehouse, and the associated tags retrieved from the social network.

The key step in processing sales data is mapping sales opportunities to employees. Several roles are associated with opportunities, including opportunity owner, supporting team, contact person, etc. In this work, we take the text descriptions of all opportunities a salesperson owns and create bag-of-words features from this text. To create a bag-of-words representation, we first convert the text into lower case and eliminate stop words. Then the text is tokenized and numerical term frequencies are computed. We compute similar bag-of-words representations for job title and social tags. Furthermore, categorical HR fields such as reporting chain and work location are processed into binary features. In summary, we have derived multiple feature representations for the employee in the feature engineering stage; the predictive power of each feature set is studied in our experiments section.

4. CLASSIFICATION ALGORITHMS

This section is devoted to the second stage of a machine learning research program, described by Wagstaff as the "machine learning contribution" [25]. We discuss the choice and development of an algorithm, discuss the choice of a
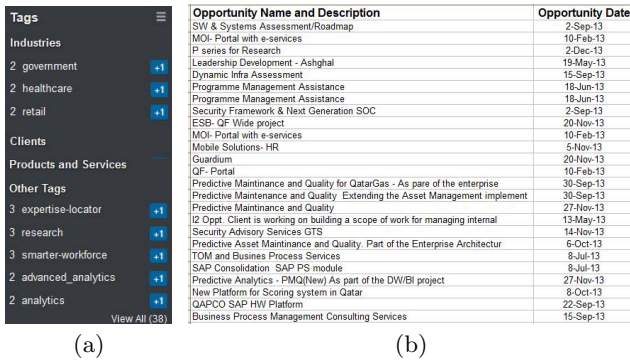
---

[2]We do not include any label noise correction, e.g. [17], to maintain implementational simplicity and because once the system has been deployed for some time, the label noise problem will fade away by itself.

Tags	≡
Industries	
2 government	+1
2 healthcare	+1
2 retail	+1
Clients	
Products and Services	
Other Tags	
3 expertise-locator	+1
3 research	+1
3 smarter-workforce	+1
2 advanced_analytics	+1
2 analytics	+1
	View All (38)

(a)

Opportunity Name and Description	Opportunity Date
SW & Systems Assessment/Roadmap	2-Sep-13
MOI- Portal with e-services	10-Feb-13
P series for Research	2-Dec-13
Leadership Development - Ashghal	19-May-13
Dynamic Infra Assessment	15-Sep-13
Programme Management Assistance	18-Jun-13
Programme Management Assistance	18-Jun-13
Security Framework & Next Generation SOC	2-Sep-13
ESB- QF Wide project	20-Nov-13
MOI- Portal with e-services	10-Feb-13
Mobile Solutions- HR	5-Nov-13
Guardium	20-Nov-13
QF- Portal	10-Feb-13
Predictive Maintenance and Quality for QatarGas - As pare of the enterprise	30-Sep-13
Predictive Maintenance and Quality Extending the Asset Management implement	30-Sep-13
Predictive Maintenance and Quality	27-Nov-13
I2 Oppt. Client is working on building a scope of work for managing internal	13-May-13
Security Advisory Services GTS	14-Nov-13
Predictive Asset Maintenance and Quality. Part of the Enterprise Architectur	6-Oct-13
TOM and Busines Process Services	8-Jul-13
SAP Consolidation SAP PS module	8-Jul-13
Predictive Analytics - PMQ(New) As part of the DW/BI project	27-Nov-13
New Platform for Scoring system in Qatar	8-Oct-13
QAPCO SAP HW Platform	22-Sep-13
Business Process Management Consulting Services	15-Sep-13

(b)

John Smith

Job Title
Workforce Diversity & Programs Professional: Workforce Analytics & Initiatives

What I'm Known For

"Deep experience in Enterprise Workforce Management, Expertise Data and Reporting, EA Tool, IBM Expertise, and Sales Eminence Initiative and Sales Eminence Top Sheet management system. Also lead the Sales Expertise Management Initiative (SEMI) community"

(c)

Figure 1: Examples of raw data for a single employee: (a) tags from IBM Connections, (b) sales opportunity data, and (c) job title.

metric, and conduct experiments to determine relative and absolute performance numbers.

4.1 Algorithms and Metrics

We would like to solve a multi-category classification problem with the features described in Section 3 to predict employee job role or job role specialty. Many different classification models may be applied to the problem; there is no particular a priori preference for any specific algorithm. We compare the performance of several models in the sequel. The performance metric of interest in this problem is classification accuracy. Although not the most sensible metric in many applications, it truly is the most sensible in job role prediction because the classes are not severely imbalanced and different types of misclassifications do not have different costs.

4.2 Empirical Study

In this empirical study, we focus on the job role classification problem rather than the job role specialty classification problem, but algorithmically, we treat both in the same way with the same features and evaluation criteria. The specific LOB that we focus on here is the set of all salespeople in the entire IBM Corporation worldwide. We have conducted similar empirical studies for other LOBs and populations, consistently finding similar results.

For this LOB, there are eleven valid job roles: Brand Client Representative (BCR), Client Representative (CR), Client Technical Architect (CTA), Client Technical Manager (CTM), Client Technical Specialist (CTS), Client Unit Executive (CUE), Industry Solution Representative (ISR), Mid-Market Client Representative (MCR), Solution Representative (SR), Solution Representative – Brand Specialist (SRB), and Solution Sales Manager (SSM). As may be guessed based on their names, several of these different job roles are fairly similar to each other. Due to this fact, it is not straightforward for employees to label themselves

quickly and accurately, not to say anything about a machine learning algorithm.

As discussed in Section 3, we take all employees with valid labels as the training set. This amounts to 36,709 employees, which is 89% of the salesperson population. The largest class, SRB, represents 0.2633 of all sellers and thus this value is the baseline classification accuracy. We compare four one-against-the-rest multi-category classification algorithms: linear logistic regression with ℓ_2 and ℓ_1 regularization, linear support vector machine, and naïve Bayes. The regularization parameters for the first three models are found by cross-validation.

In order to estimate generalization accuracy, we calculate fivefold cross-validation test error for the different classifiers. Cross-validation provides such an estimate because there is no reason to suspect any systemic bias in employees with blank or invalid class labels. We also compare the four different feature sets individually and in combination: job title, HR information, sales opportunities, and social tags. We take the 1000 terms with the highest term frequency from the text analytics for each of the data sources except for HR information. The HR information consists of 23 categorical features converted to binary and one numeric feature. The test accuracy results are given in Table 1.

It is clear in the results table that the naïve Bayes classifier performs the worst across all feature sets and the support vector machine is consistently a little bit worse than logistic regression. The two different regularization versions of logistic regression are nearly the same across all feature sets. The most interesting thing to notice is that the classification accuracy from only social tags is actually worse than the baseline accuracy. The performance from only sales opportunities is also quite poor. This poor performance can be attributed to the words in descriptions of sales opportunities seen in Fig. 1(b) being very specific and not well matched to differentiating often-similar sales roles. More so, public social tags give a different indication of a person's activities and skills than what is required for predicting job roles.

The best performance is achieved by using the HR information and the bag-of-words features of the job title, both individually and in combination. Adding the poor-performing social tags and sales opportunity features only introduces noise and degrades accuracy. Thus we go forward with the combination of HR and job title features for the deployment. Since both logistic regressions are essentially the same and the ℓ_2-norm is easier to optimize, we choose the ℓ_2-regularized logistic regression model. The 0.8016 accuracy is quite good for this challenging eleven-class problem. The logistic regression also provides posterior probabilities or confidence scores for each of the eleven classes for each sample.

We show the confusion matrix of the cross-validation test set in Table 2. We note that most mistakes are within two clusters of very similar job roles: technical sellers {CTA, CTM, CTS}, and brand solution reps {BCR, SR, SRB, SSM}, which is to be expected due to their similarities. The classifiers evaluated here all deal with employees as statistically independent samples; subject matter expertise suggests a possible way to improve accuracy based on relationships among employees, which we explore next.

Table 1: Fivefold Cross-Validation Test Accuracy

Feature Set	ℓ_2-Reg. Logistic Regress.	ℓ_1-Reg. Logistic Regress.	Support Vec. Mach.	Naïve Bayes
Job Title (A)	0.6746	0.6749	0.6695	0.6410
HR Info (B)	0.7661	0.7641	0.7604	0.6807
Social Tags (C)	0.2320	0.2396	0.2380	0.2573
Sales Opp (D)	0.3374	0.3404	0.3473	0.2306
(A) + (B)	0.8016	0.8031	0.7899	0.7330
(A) + (B) + (C)	0.7671	0.7703	0.7504	0.6118
(A) + (B) + (C) + (D)	0.7720	0.7733	0.7655	0.3952

Table 2: Confusion Matrix (rows are actuals, columns are predictions)

	BCR	CR	CTA	CTM	CTS	CUE	ISR	MCR	SR	SRB	SSM	Total
BCR	1084	88	3	1	11	4	3	5	70	913	119	2301
CR	33	4014	7	1	4	126	8	96	72	148	50	4559
CTA	5	12	2534	40	438	4	5	0	11	9	5	3063
CTM	0	1	55	955	89	4	0	0	5	10	45	1164
CTS	0	6	286	77	7358	2	1	0	7	57	4	7798
CUE	6	247	2	0	0	600	4	18	25	11	154	1067
ISR	17	46	12	0	2	9	51	3	31	60	6	237
MCR	54	202	0	0	1	13	2	294	16	91	6	679
SR	186	164	11	4	14	19	28	20	1891	892	119	3348
SRB	310	91	11	6	70	10	5	20	247	8572	322	9664
SSM	64	76	8	16	4	115	10	9	95	358	2074	2829
Total	1759	4947	2929	1100	7991	906	117	465	2470	11121	2904	36709

4.3 Org Chart-Assisted Postprocessing

IBM, like most large corporations, is a tree-structured organization with a management chain representing the path from an employee to the Chief Executive Officer. For ease of management, employees tend to be grouped by job function within the tree. We put forth a few methods to postprocess the predicted class probabilities output by the logistic regression and examine their effects on test accuracy.

We first construct groups of employees by taking sub-trees of the organizational structure at each level except for the highest three levels because they encompass the entire salesforce. Then, in the first postprocessing method, we consider all groups having all training set employees with the same class label and change the predicted class label of all test set employees to that common class label (if they are not already). This improves the accuracy from 0.8016 to 0.8022. As a further heuristic, if we only change those predicted test set class labels with confidence score less than 0.9, then the accuracy improves to 0.8039.

Another postprocessing method is provenance-assisted classification [26]. From the same sub-tree groupings, we construct a bipartite graph with one set of nodes being the groups and the other set of nodes being the employees. Then through an iterative expectation–maximization procedure similar to that found in error-correcting codes [15], we find a maximum likelihood estimate of the class probabilities under the enforcement of consistency in groupings based on the bipartite relationships. However, such an approach with all grouping nodes decreases the accuracy significantly to 0.5982. Only including grouping nodes that have all training set samples the same, as in the previous heuristic, also decreases the accuracy to 0.7999.

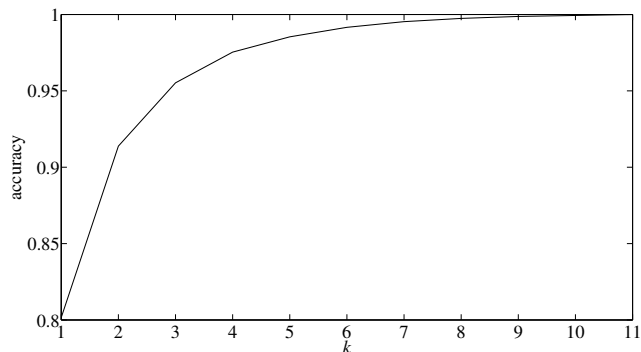Very limited improvement or even a worsening of performance occurs primarily because the HR information al-

Figure 2: Fivefold cross-validation accuracy as a function of number of predictions per employee.

ready includes categorical attributes indicating three levels of membership in the organizational structure and because while the grouping of employees in the tree is generally true, there are many exceptions which are identified correctly by the classifier but then smoothed out by the postprocessing.

4.4 Discussion

As discussed earlier, one use case of the classifier outputs is showing more than one predicted job role for an employee to a system user and allowing the user to choose the correct one. In this case, we can also measure the accuracy of whether the true class label appears in the k class labels with the highest confidence scores. This accuracy value is plotted in Fig. 2, showing that with two or three predictions per employee, we can achieve very good performance: 0.9138 and 0.9553 respectively. Most of the confusion is among similar

job roles, which can be alleviated by giving more than one prediction per person.

As also discussed earlier, even the labels that we use for training and cross-validation have some level of noise. We can get a rough indication of the level of noise by comparing the job role and the job role specialty values among the 36,709 salespeople in our population: if the job role specialty does not fall under the job role in the taxonomy, this mismatch indicates a possible error in the job role label. (Such a mismatch is only one possible manifestation of label noise; the job role label could be incorrect for a host of other reasons as well.) The employees with such a mismatch constitute 6.6% of the population. Among the mismatched population, the cross-validation test accuracy is 0.7207 compared to 0.8073 among the sellers with job role and job role specialty values that are in concordance. Thus in comparing 0.8073 to 0.8016, we see that label noise is not very significant, at least as manifested via job role and job role specialty discord.

Through the experimental results with different feature sets, classifiers, and predicted class probability postprocessing methods, we have seen that we can achieve approximately 80% accuracy using HR information and job title with the ℓ_2-regularized logistic regression without any postprocessing. These are the choices we go forward with in the deployment.

5. DEPLOYMENT AND IMPACT

In this section we describe the third stage of the machine learning research program: the deployment of the data mining system we have developed within the business processes of IBM. The section begins with the initial deployment among the population of all salespeople worldwide, and then discusses the deployment across the entire corporation that is currently in progress. Both of these parts discuss the business impact of the deployment. The section finishes with a discussion of lessons learned, from both the HR practioner's perspective and the data mining practitioner's perspective.

5.1 Worldwide Salespeople Deployment

We first deployed job role and speciality prediction to solve an issue with missing and incorrect job role information among IBM salespeople worldwide in the first part of 2014. The salesforce is a key segment of the IBM workforce because these employees garner the revenue for the company. Job role information is used in myriad salesforce planning and strategy functions [1].

In 2014, a new situation was encountered that required the primary job roles to be accurate and within a certain number of permitted values for 100% of the sellers. Even though sellers are asked to update job role information annually, approximately 4,000 of them fail to follow the directions or fail to provide the update. To get the employees to update this data field would have taken approximately 30 minutes per employee or manager, including contacting them, guiding them through the process, and then having them submit the update. Summed over the 4,000 employees, this is an effort of approximately 2,000 person-hours. Additional effort from corporate headquarters to create the instructions, deploy the communication, and assist with the roll out would have incurred approximately 500 person-hours.

Instead we used the result of the predictive analytics to recommend a primary job role and put that value into the primary job role field in the data. We then asked the managers to check the recommended job role and make any necessary changes. Since the predictions were accurate approximately 80% of the time, only a small number of managers had to physically make a change. Conservatively speaking, using predictive analytics saved 80% of the estimated 2,000 person-hours by employees and managers and most of the 500 person-hours by headquarters for setup and deployment, resulting in huge savings totaling approximately 2,100 person-hours, which is the equivalent of one person-year. A typical year of effort by an IBM salesperson results in revenue achievement in the range of $1 million; thus the time savings provided by the job role prediction can yield that much additional revenue for the company.

The salesforce is less than 10% of the total IBM workforce and is a population that is more compliant than many other LOBs. The less compliant an LOB is, the more time and effort savings the analytics provide. Therefore, deployment to the entire IBM workforce is expected to save more than 20 person-years of effort if used on an annual basis. However, the even greater impact is a fundamental transformation in the way the company operates. Through the analytics, job role data can be refreshed at a much higher frequency than annually to support many different existing business processes and enabling new business practices with their own monetary and effort savings or revenue achievement improvements. We next discuss the deployment to the entire company that is currently in progress.

5.2 Broad Deployment

We see many other opportunities to use the job role prediction for populations besides salespeople with similar circumstances. For example, IBM has some very large populations of employees in which the primary job role field is invalid almost all of the time unlike for sellers where it is valid most of the time. For these larger employee populations with very sparse data for primary job role, the savings will be even more significant. Going forward, we are enthusiastic about the ability to predict expertise from employees' digital breadcrumbs instead of constantly asking managers or employees to fill in the data. Invariably, there are compliance challenges with an annual update; keeping job roles and specialties accurate as people transfer, change organizations, or move to new jobs is almost impossible. As discussed above, the ability to predict the primary job role and specialty and keep each continuously updated will not only save countless person-hours but will yield more up-to-date and accurate information from which to make manpower planning decisions.

The deployment to the worldwide sales organization was performed in a one-time manner without any specific thought or development given to user interface and user experience design, or to system architecture. In preparing the deployment to the entire company, we are taking a design thinking approach [3]. This process has centered around constructing user stories for the job role and specialty prediction tasks. The initial outcome has been the identification of managers and other administrators as the primary system users rather than individual employees. Also, the form of the user experience has been set forth through an expertise inventory dashboard. Two initial wireframes of the user interface are
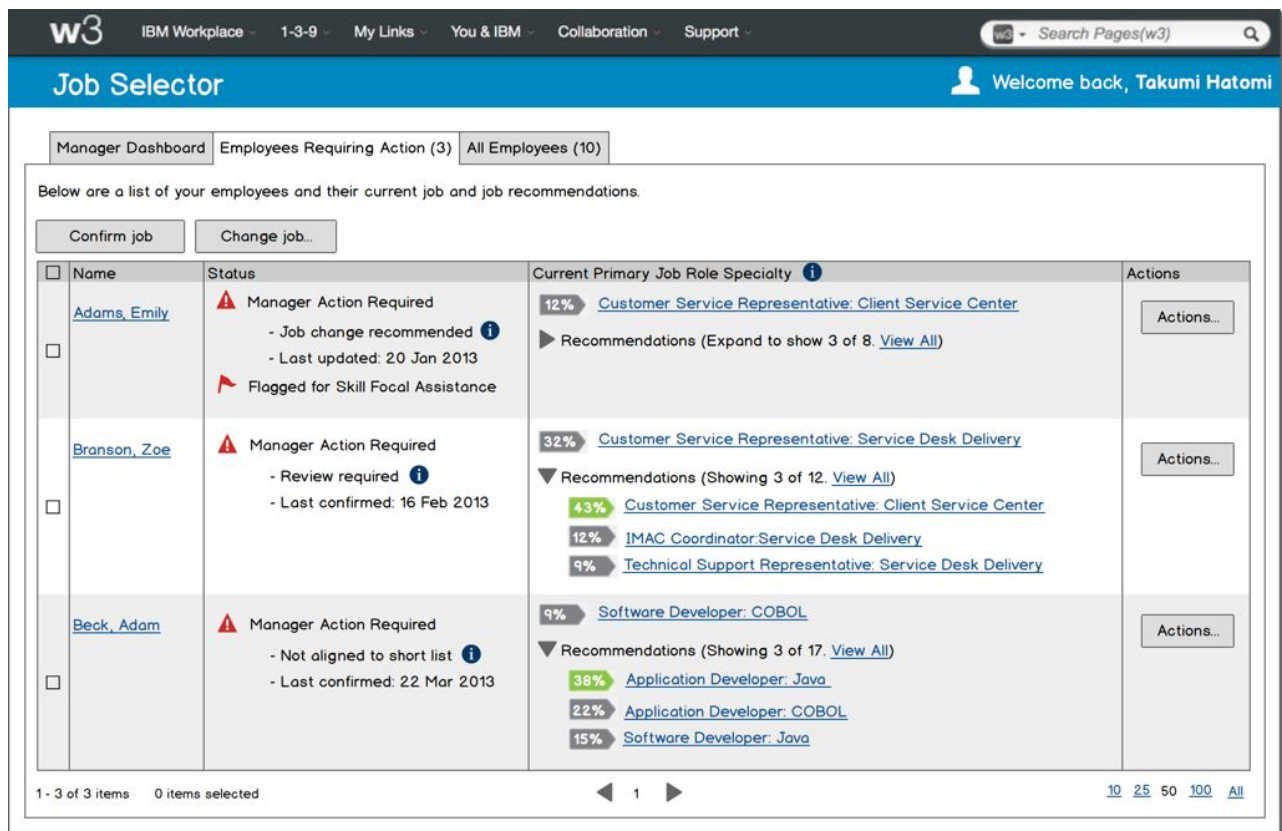
Figure 3: Wireframe of job role specialty prediction deployment.

shown in Fig. 3 and Fig. 4. The remaining work is in finalizing the user experience design and then developing the system architecture, which involves collecting and merging data from several siloed data warehouses, as well as putting together a flow of text analytics and supervised learning (through the Liblinear implementation of logistic regression) for cross-validation, training and testing. Finally, rolling out the system will require some level of education and evangelization within various business units throughout IBM.

5.3 Lessons Learned

In this section, we present lessons learned throughout the process of developing and deploying the predictive analytics from two perspectives. First, we give comments received from an HR practitioner involved in the project:

> "The prediction is only as good as the availability of data which helps the prediction. If you only have data that does not help inform the predictive analytics, you won't be able to achieve a very high accuracy level. So making data from other sources which helped us predict which specialty roles our sellers where working in was the key to success. And I think the more data that you can arm the predictive analytics with, the better the prediction will be. Determining whether different data types were useful in the prediction was a little trial and error and some sources that we thought ahead of time that would be good predictors turned out not to be. Also, adding

data to the prediction algorithms which are not related to the prediction will actually make the prediction less accurate as it adds noise to the algorithm. Also, it is very important to have a pretty large control group for which you know already know the predicted value that you are trying to achieve. This is necessary to be able to validate and determine the accuracy of the predicted values."

This process also involved quite a bit of learning for the machine learning practitioners as well. One of them reports:

> "Before starting this project, I was only vaguely familiar with the expertise taxonomy. It has so many intricacies that we had to learn about just to frame the problem properly. But by doing so and by being cognizant of all of the specific business rules that are in play, we can really make an impact. In the end, we selected a fairly basic classifier, which is less exciting than something more novel would have been, but it has many desirable properties including scalability and ease of use, which will make a big difference in the final deployment. Understanding the HR folks' perspective and figuring out how we can make a transformational journey together has been enlightening and enjoyable."
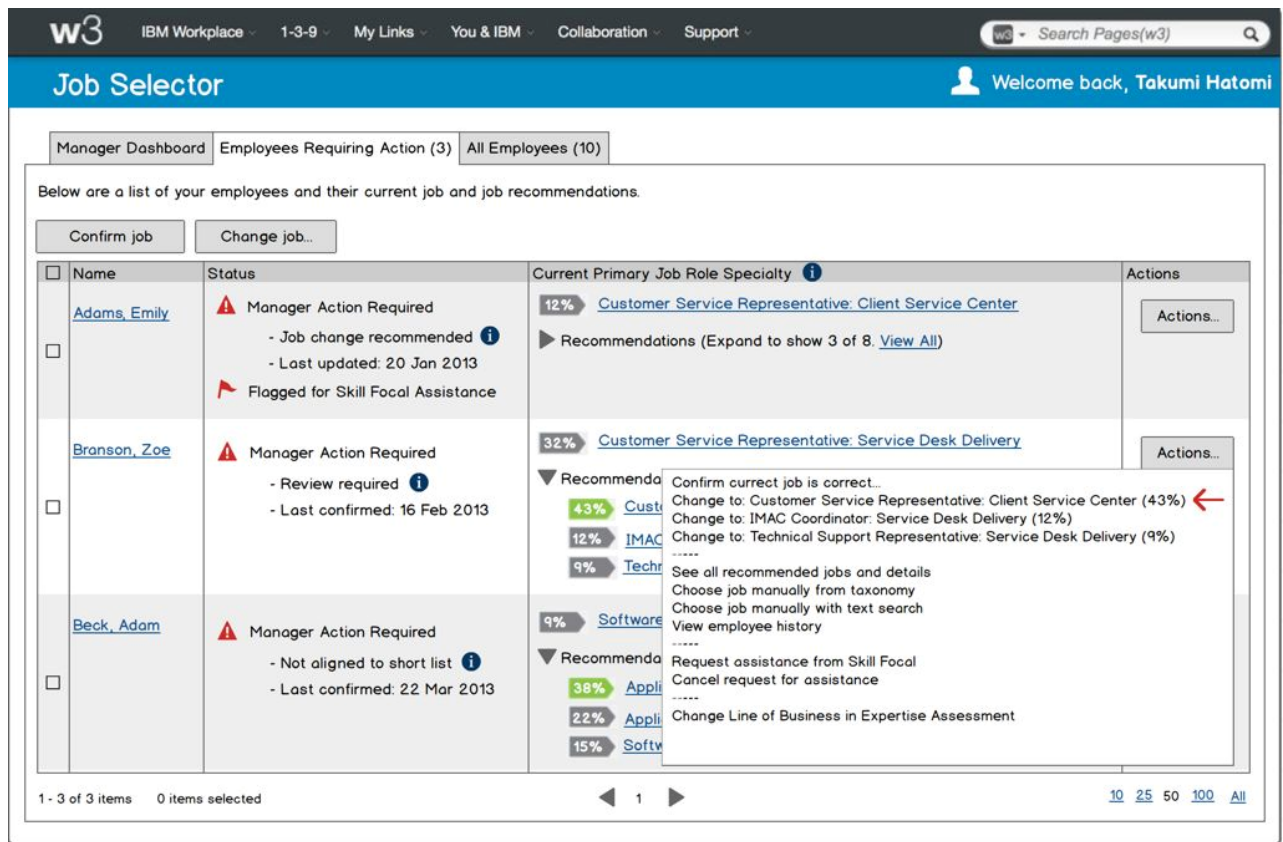
1736

Figure 4: Wireframe of job role specialty prediction deployment with user choosing to accept the top classification.

6. CONCLUSION

The talent and human capital of IBM is its most valuable resource that must be harnessed properly using trusted expertise information. In this work, we developed a classification methodology to predict the expertise of employees based on features derived from the digital footprints of employees with the label set coming from IBM's expertise taxonomy. We deployed this system initially to fill in missing and invalid job role records and to correct erroneous records in the worldwide population of 40,000 IBM salespeople. The reduction in manual effort obtained through using predictive analytics is estimated to be one entire person-year. We are now in the process of deploying the system for use by the entire corporation, which should result in approximately twenty person-years of savings in annual updates of job roles and specialties. The impact is even greater than the savings in manual effort, because all business processes that depend on complete, accurate, and updated expertise data benefit from the predictions. Additionally, because of the steep reduction in effort, it will now be possible to update expertise assessments much more frequently than once a year, which is a transformation required to compete in today's dynamic business environment.

In developing the approach, we evaluated features from four different data sources: job title, HR information, social tags, and work products; several classification algorithms; and a few organizational chart-assisted postprocessing methods. The final choice of Liblinear ℓ_2-regularized logistic re-

gression with bag-of-words features derived from the job title and 24 HR features yields cross-validation accuracy of 80% for a single prediction per employee and 96% accuracy if we allow three predictions per employee, which is more than enough for the application of interest. This classifier choice also allows us to present confidence values associated with predictions to the system user, which has emerged as an important feature from the user stories developed as part of the design thinking approach.

One immediate piece of future work is continued deployment and evangelization throughout the corporation. Other future work we have identified includes evaluation of additional work product data sources for different employee populations and expanding the response variable beyond job roles and job role specialties to industry specialization. Also, the simple tokenization and frequency-based word selection can be improved through advanced text analytics.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] M. Baier, J. E. Carballo, A. J. Chang, Y. Lu, A. Mojsilović, M. J. Richard, M. Singh, M. S. Squillante, and K. R. Varshney. Sales-force

performance analytics and optimization. *IBM J. Res. Dev.*, 56(6):8, Nov./Dec. 2012.

[2] J. Bersin, K. O'Leonard, and W. Wang-Audia. High-impact talent analytics: Building a world-class HR measurement and analytics function, Sept. 2013.

[3] T. Brown. Design thinking. *Harvard Bus. Rev.*, 86(6):84–92, June 2008.

[4] C. Chelmis, V. Sorathia, and V. K. Prasanna. Enterprise wisdom captured socially. In *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min.*, pages 1228–1235, Istanbul, Turkey, Aug. 2012.

[5] V. Chenthamarakshan, K. Dey, J. Hu, A. Mojsilovic, W. Riddle, and V. Sindhwani. Leveraging social networks for corporate staffing and expert recommendation. *IBM J. Res. Dev.*, 53(6):11, Nov. 2009.

[6] V. Chenthamarakshan, K. Dixit, M. Gattani, M. Goyal, P. Gupta, N. Kambhatla, R. Lotlikar, D. Majumdar, G. R. Parija, S. Roy, S. Soni, and K. Visweswariah. Effective decision support systems for workforce deployment. *IBM J. Res. Dev.*, 54(6):5, Nov./Dec. 2010.

[7] K. Ehrlich, C.-Y. Lin, and V. Griffiths-Fisher. Searching for experts in the enterprise: Combining text and social network analysis. In *Proc. ACM Conf. Supporting Group Work*, pages 117–126, Sanibel Island, FL, Nov. 2007.

[8] K. Ehrlich and N. S. Shami. Searching for expertise. In *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, pages 1093–1096, Florence, Italy, Apr. 2008.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[10] D. Fang, K. R. Varshney, J. Wang, K. N. Ramamurthy, A. Mojsilović, and J. H. Bauer. Quantifying and recommending expertise when new skills emerge. In *Proc. IEEE Int. Conf. Data Min. Workshops*, pages 672–679, Dallas, TX, Dec. 2013.

[11] I. Guy, U. Avraham, D. Carmel, S. Ur, M. Jacovi, and I. Ronen. Mining expertise and interests from social media. In *Proc. World Wide Web Conf.*, pages 515–525, Rio de Janiero, Brazil, May 2013.

[12] J. Hu, B. K. Ray, and M. Singh. Statistical methods for automated generation of service engagement staffing plans. *IBM J. Res. Dev.*, 51(3/4):281–293, May/Jul. 2007.

[13] A. John and D. Seligmann. Collaborative tagging and expertise in the enterprise. In *Proc. WWW Collabor. Web Tagging Workshop*, Edinburgh, UK, May 2006.

[14] C.-Y. Lin, L. Wu, Z. Wen, H. Tong, V. Griffiths-Fisher, L. Shi, and D. Lubensky. Social network analysis in enterprise. *Proc. IEEE*, 100(9):2759–2776, Sept. 2012.

[15] D. J. C. MacKay. Good error-correcting codes based on very sparse matrices. *IEEE Trans. Inf. Theory*, 45(2):399–431, Mar. 1999.

[16] S. Mehta, R. Pimplikar, A. Singh, L. R. Varshney, and K. Visweswariah. Efficient multifaceted screening of job applicants. In *Proc. Int. Conf. Extending Database Tech.*, pages 661–671, Genoa, Italy, Mar. 2013.

[17] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. In *Adv. Neural Inf. Process. Syst. 26*, pages 1196–1204, Lake Tahoe, NV, Dec. 2013.

[18] Y. Naveh, Y. Richter, Y. Altshuler, D. L. Gresh, and D. P. Connors. Workforce optimization: Identification and assignment of professional workers using constraint programming. *IBM J. Res. Dev.*, 51(3/4):263–279, May/Jul. 2007.

[19] A. Perer, I. Guy, E. Uziel, I. Ronen, and M. Jacovi. Unearthing people from the SaND: Relationship discovery with social media in the enterprise. In *Proc. Int. AAAI Conf. Weblogs Soc. Med.*, pages 582–585, Barcelona, Spain, July 2011.

[20] N. S. Shami, K. Ehrlich, G. Gay, and J. T. Hancock. Making sense of strangers' expertise from signals in digital artifacts. In *Proc. SIGCHI Conf. Hum. Fact. Comput. Syst.*, pages 69–78, Boston, MA, Apr. 2009.

[21] A. Singh, R. Catherine, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla. PROSPECT: A system for screening candidates for recruitment. In *ACM Conf. Inf. Knowl. Manage.*, pages 659–668, Toronto, Canada, Oct. 2010.

[22] M. Singh, K. R. Varshney, J. Wang, A. Mojsilović, A. R. Gill, P. I. Faur, and R. Ezry. An analytics approach for proactively combating voluntary attrition of employees. In *Proc. IEEE Int. Conf. Data Min. Workshops*, pages 317–323, Brussels, Belgium, Dec. 2012.

[23] L. Terveen and D. W. McDonald. Social matching: A framework and research agenda. *ACM Trans. Comput.-Hum. Int.*, 12(3):401–434, Sept. 2005.

[24] K. R. Varshney, J. Wang, A. Mojsilović, D. Fang, and J. H. Bauer. Predicting and recommending skills in the social enterprise. In *Proc. AAAI ICWSM Workshop Social Comput. Workforce 2.0*, pages 20–23, Cambridge, MA, July 2013.

[25] K. L. Wagstaff. Machine learning that matters. In *Proc. Int. Conf. Mach. Learn.*, pages 529–536, Edinburgh, United Kingdom, Jun.–Jul. 2012.

[26] D. Wang, M. T. Al Amin, T. Abdelzaher, D. Roth, C. Voss, L. Kaplan, S. Tratz, J. Laoudi, and D. Briesch. Provenance-assisted classification in social networks. *IEEE J. Sel. Topics Signal Process.*, 8(4), Aug. 2014.

[27] J. Wang, K. R. Varshney, A. Mojsilović, D. Fang, and J. H. Bauer. Expertise assessment with multi-cue semantic information. In *Proc. IEEE Int. Conf. Serv. Oper. Logist. Informat.*, pages 534–539, Dongguan, China, July 2013.