Information Retrieval, Fusion, Completion, and Clustering for Employee Expertise Estimation

Raya Horesh, Kush R. Varshney, and Jinfeng Yi
Solutions and Mathematical Sciences
IBM Thomas J. Watson Research Center
Yorktown Heights, NY, USA
Email: {rhoresh,krvarshn,jinfengy}@us.ibm.com

***Abstract*—Estimating the skills, talents, and expertise of employees is essential for human capital management in knowledge-based organizations across industries and sectors. In this paper, we describe an approach to infer the expertise of employees from their enterprise data and digital footprints. Using a novel big data workflow with components of information retrieval and search, data fusion, matrix completion, and ordinal regression clustering, we are able to automatically find evidence of expertise and determine appropriate evidence weights for different queries and data sources that we merge and present in a manner consumable by businesspeople. We illustrate the system on sample data from the IBM Corporation where it has been deployed.**

***Keywords*-human talent management; unsupervised learning; workforce analytics**

I. Introduction

Big data systems and techniques have found use in many different applications, including e-commerce, marketing, healthcare, energy, finance, agriculture, manufacturing, and public works. One key application for big data that is in its incipient stages of development and deployment is human talent management. Talent management is the science of using strategic human resource (HR) planning to improve business value; everything done to recruit, retain, develop, and reward people as well as make them productive forms a part of talent management. This application touches almost every industry sector and has, to now, been mostly informed by human instinct rather than big data. Transforming talent management through big data requires innovative technical solutions.

A. Context

Roger Martin of the Rotman School of Management recently stated that "over the past 50 years the U.S. economy has shifted decisively from financing the exploitation of natural resources to making the most of human talent." Creative non-routine positions were 16% of all jobs in 1960, but became 33% by 2010 [1]. This shift is not confined to the United States, but is an ongoing worldwide phenomenon: some of the largest worldwide employers today are knowledge-based enterprises whose most important asset is human capital. Smart creative knowledge workers are unique, each having individualized skills, competencies, and expertise that evolve and grow on a daily basis. Moreover, due to today's high pace of technological innovation, new skills are emerging faster than they can even be described [2].

In light of these trends, it is becoming increasingly important that organizations manage their talent well and individual employees manage their own skills and careers. In today's big data age, much information is captured about employees' skills within organizations, through: resumes, curricula vitae and certifications; explicit assessments of employee expertise; job position histories; digital artifacts from enterprise systems of record that have 'footprints' of employees' work activities, e.g., project claims, sales pipeline, software documentation, publications, etc.; and activity captured on internal corporate social media [3]. Outside of organizations, similar sorts of information are captured by public social networking sites such as LinkedIn [4]. This data along with advanced analytics algorithms for estimation, inference and optimization is starting to be used to inform talent management.

Big data and analytics (before the terms became en vogue) were first used to describe and inform decisions surrounding physical capital and financial capital, leading to the CFO becoming a strategic corporate role. Next, the big data revolution hit marketing, which continues into today. Now, the new area in which big data is playing a role is in describing and informing decisions on human capital.

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 [5] "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%. However, even among that small fraction of organizations, the newly adopted predictive HR analytics are mostly focused on recruitment and hiring, resource de-

ployment and proactive talent retention. Predicting expertise is a much more challenging and open-ended problem than the others, but has significant business value [6]–[8].

B. Inventories of Employees

Many tactical and strategic talent management activities are predicated on having up-to-date, complete, and accurate inventories of the skills and expertise of employees. One would think that companies already have good insight into the expertise of their employees, but that is in fact not the case. Most companies do not even have a language to describe the skills of their employees; the ones that do rely on hierarchical expertise taxonomies [9] that are difficult to create, maintain, and assess employees against [2].

Expertise is a fairly nebulous concept that can be defined at various granularities. Very fine level skills, such as *use Rational Unified Process for service-oriented modeling and architecture*, *analyze product to create proof package*, and *implement composite business services in WebSphere Business Services Fabric*, indicate specific competencies of employees. Coarser job roles, such as *delivery project executive for technical support services*, *internal auditor*, and *software architect for business intelligence*, indicate collections of still-specific skills that all employees with the given expertise have. Even coarser are broad general topics of expertise including industry sectors such as *financial services*, *chemicals and petroleum*, and *media and communication*, and technology areas such as *cloud computing*, *big data*, and *cybersecurity*, which do not indicate specific skills that employees must have, but delineate competencies within which employees have subsets.

In our previous work, we approached the fine-level skill estimation problem as one of collaborative filtering [6], [10], [11], and approached the medium-level job role estimation problem as one of multicategory classification from enterprise data [2]. The coarse-level problem of estimating broad industry or technology areas from enterprise data about employees, to the best of our knowledge, has not been studied in the literature. In [12], a related problem is studied, namely given an employee, return a profile of all the general expertise areas that that employee is competent in. Other approaches for expertise estimation at a coarse level are based on task-solving data rather than enterprise data not specifically collected for the reason of inferring skills [13].

C. Contributions

In this paper, we describe a system for estimating coarse-level general expertise areas that we have developed in collaboration with the global human resources (HR) organization of the IBM Corporation for deployment and use throughout the company. Starting with a given expertise area, the system performs information fusion on query results from a search and retrieval functionality that indexes numerous enterprise data sources. The system then further processes the fused results using low-rank matrix completion [14] and ordinal regression clustering [15] to produce a final list of employees labeled by their depth of expertise in that given expertise area. This is a novel workflow and system design that, as we detail in the remainder of the paper, addresses numerous issues encountered in this specific workforce analytics application.

D. Organization of Paper

The remainder of the paper is organized as follows. In Section II, we describe the problem of interest, highlighting the key domain-specific issues that arise. In Section III, we discuss the information retrieval and fusion aspects of the system. Section IV presents the matrix completion formulation and solution approach. Section V presents the ordinal regression clustering formulation and solution approach. We give an illustrative example on real-world expertise query data from IBM in Section VI and conclude in Section VII.

II. Problem Description and Approach

As described in the introduction, we would like to estimate how much expertise an employee has in a broad area such as *cloud computing* or *cybersecurity*. These are broad areas because they are really collections of many different competencies. Due to this breadth and the lack of strict definition, we feel that search-based information retrieval that brings back soft evidence for different queries should be the starting point for an estimation system. The queries should include keywords relevant to the topic and the search should be based on data sources about the employee captured digitally, including work products, employee assessments, curricula vitae, corporate social media activity, HR data, and so on. Even very strong experts in such expertise areas are not expected to have evidence of expertise on all of the queries because of their breadth. Given all of the evidence from all of the queries, the information contained therein must be fused together to obtain an overall score that can be easily consumed by the decision makers in the business. The fusion process should account for various biases so as not to incorrectly inflate or deflate the estimated expertise of individual employees.

At a high level, the information retrieval and fusion approach can be described as follows. A list of query terms related to the broad expertise area is generated. Search is conducted against each of those query terms to retrieve evidence by employee and data source. The various pieces of evidence are fused together, weighted by query, weighted by data source, and potentially weighted in other ways, to a single ordinal value (very deep, deep, moderate, some, limited) per employee indicating his or her depth of expertise in that broad area. The idea is illustrated in Figure 1.

The general idea can be understood in analogy with the blind men and the elephant. Just as blind men are myopic and only see a leaf, a snake, a spear, a wall, a tree trunk,

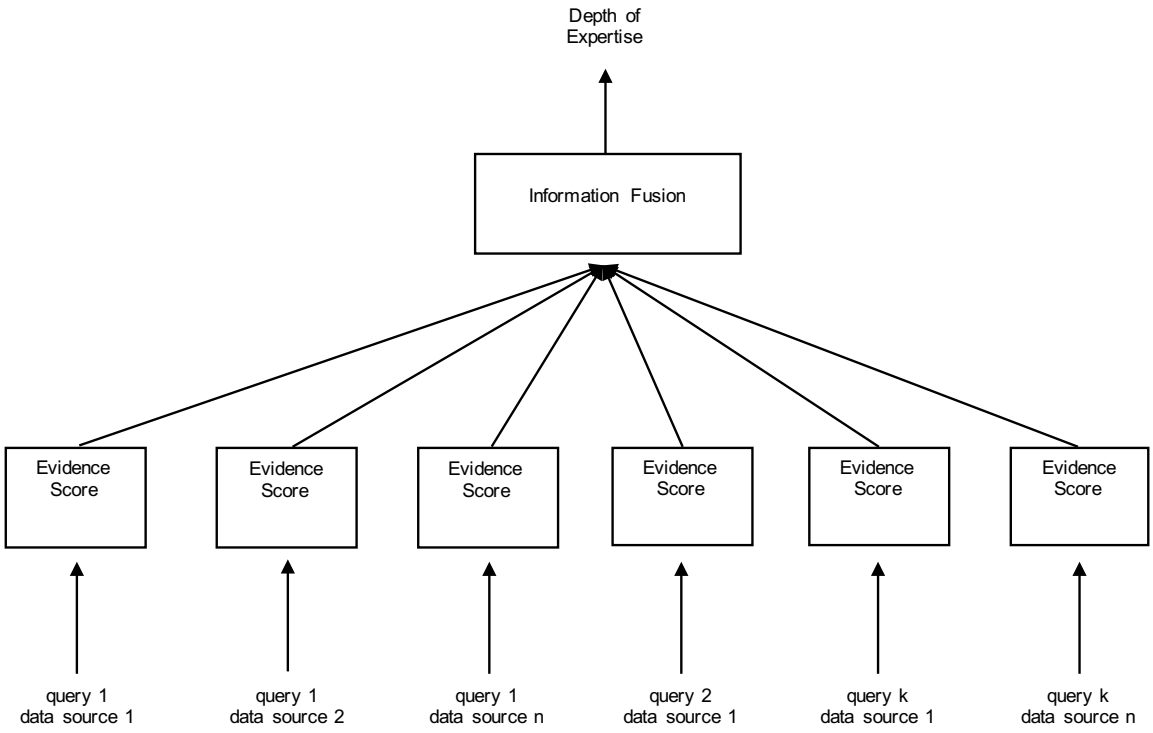

Figure 1. A high-level picture of the broad expertise estimation formulation.

or a rope when they touch an elephant, the various queries and data sources only 'see' one small part of the expertise of an employee and must be fused together to get an overall picture about the depth of expertise of an employee on a broad expertise area.

The first step in the process is determining thirty to fifty query terms related to the broad expertise area. There are two ways of doing so: eliciting the query terms from subject matter experts or generating the terms automatically using text analytics on an appropriate corpus. The second step in the process is information retrieval: executing searches against various indexed data sources using the query terms from the previous step. The final step is information fusion, which has several application-specific nuances to consider and for which we are operating in an unsupervised regime.

The nuances include:

1) Not all query terms are equivalent. Some keywords indicate greater depth of expertise than others and thus should be given greater weight in the fusion. For this, we use a heuristic that more esoteric keywords are given greater weight. This is discussed further in Section III.
2) Many of the data sources we utilize are not required to be complete for every employee. Simply treating missing data as zeroes is not correct because lack of data in a data source does not imply that an employee is less skilled. To address this issue, matrix completion technique is used to fill in estimates of evidence when an employee has no values or content in a data source. This technique is remarkably effective and has been successfully applied to many applications such as recommender systems [16], [17], multi-task learning [18], data clustering [19], [20], classification [21], [22], and crowdsourced learning [23], [24]. In Section IV, we will discuss how to solve the missing data problem by utilizing the matrix completion technique.
3) Certain employee demographics are less active in certain data sources (e.g. employees in marketing job roles are more active on corporate social media and technical services employees have more assessments. Therefore, simply scoring the quantity of documents in which evidence is found can over- and underscore different employees. To account for this nuance, we may normalize an employee's evidence counts by comparison to the average in the employee's peer group defined along demographic dimensions.
4) Similar to the first point, just as all keywords are not equivalent, not all data sources are equivalent and must be given different weight in the fusion. Moreover, the data source weights are dependent on the broad expertise area because different types of skills are reflected more or less in different types of data. We

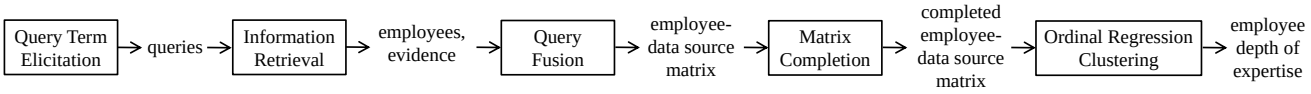

Figure 2. A block diagram of the overall expertise estimation procedure.

tackle the problem of learning data source weights in an unsupervised manner in conjunction with another domain-specific desideratum: for the business to consume and benchmark expertise inference, employees should be given labels from a small discrete set, e.g. {very deep, deep, moderate, some, limited}. As further discussed in Section V, we use multi-dimensional ordinal regression clustering to handle both problems. (A less desirable alternative in our earlier work was to use manually-determined data source weights (and query weights) to collapse evidence scores down to a single dimension and then cluster that single dimension to the discrete labels.)

Together these considerations lead us to the system depicted in the block diagram of Figure 2.

III. Information Retrieval and Query Fusion

As discussed earlier, 30 to 50 query terms are collected for a broad expertise area either by eliciting them from a subject matter expert or by using common text mining approaches for keyword extraction. For each keyword, a query is issued to an expertise discovery functionality described in [25]–[29]. The search is based on crawling and indexing a multitude of enterprise data sources, including internal social media but also many enterprise databases. The system and methodology is able to deal with the big data scale of the IBM Corporation effectively.

For each query, an evidence score per employee broken down by data source is returned. In the intended search use case, these evidence scores are not disaggregated across data sources and are used to rank employees for relevance to the query term. The evidence scores primarily indicate how often the keyword appears in documents from the data sources.

Scores resulting from the searches are returned for each of the query terms and must be merged together using some scheme. In our work, we take a linear combination of the scores with specific weights based on the following logic. The more people use a particular keyword, the more common it is. The fewer people use a particular keyword, the more esoteric it is. The use of more esoteric keywords indicates a greater depth of expertise. Therefore, we apply a hyperbolic weighting to queries that is one over the total number of employees returned by the search using that query. Since IBM has approximately $400,000$ employees, we are rarely in the situation where the total number of returned experts is small and thus the hyperbolic function is well-behaved in the regime in which we operate.

We apply the same query weight for all data sources resulting in an employee-data source matrix of evidence scores that has been merged from the 30–50 such matrices from each query. It is this matrix (suitably normalized) that is passed on to matrix completion.

IV. Inferring Unknown Evidence Scores by Efficient Matrix Completion

Since the employee-data source matrix is very sparse, i.e., typically more than 80% of its entries are unknown, our next challenge is to reliably recover the unknown entries based on the existing evidence scores. Let n and m be the number of employees and data sources, respectively. Let $\mathbf{A} \in \mathbb{R}^{n\times m}$ be the partially-observed employee-data source matrix. Then our goal is to reconstruct the full employee-data source matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ given $\mathbf{A}$. To this end, we need to make several reasonable assumptions about the relationship between $\mathbf{X}$ and $\mathbf{A}$.

Let $\Omega = \{(i,j) \in [n] \times [m]\}$ be the set of known entries in matrix $\mathbf{A}$. Then we define a matrix projection operator $\mathcal{P}_\Omega : \mathbb{R}^{n\times m} \mapsto \mathbb{R}^{n\times m}$ that takes a matrix $\mathbf{A}$ as the input and outputs a new matrix $\mathcal{P}_\Omega(\mathbf{A}) \in \mathbb{R}^{n\times m}$ as

$$[\mathcal{P}_\Omega(\mathbf{A})]_{ij} = \begin{cases} \mathbf{A}_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise,} \end{cases}$$

Since the known entries in matrix $\mathbf{A}$ are trustworthy scores generated by a set of query results, it is intuitive to assume that for every observed entry $(i,j) \in \Omega$, we have $\mathbf{X}_{ij} = \mathbf{A}_{ij}$, or alternatively,

$$\mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{A}). \tag{1}$$

The assumption specified in condition (1) is insufficient to recover the full similarity $\mathbf{X}$ as we can fill the unobserved entries (i.e., $(i,j) \notin \Omega$) in $\mathbf{X}$ with any values. An additional assumption is needed to make it possible to recover the full matrix from a partially observed one. To this end, we follow the setting of matrix completion [30] by assuming that the matrix $\mathbf{X}$ is of low-rank. This is a very natural assumption since usually only a few factors contribute to an employee's expertise level [31].

Combining the two assumptions together, the recovery problem can be cast into the following matrix completion problem

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{R}^{n\times m}} \quad & \|\mathbf{X}\|_* \\ \text{s.t.} \quad & \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{A}), \end{aligned} \tag{2}$$

where $\|\cdot\|_*$ stands for the nuclear norm, the convex surrogate of the non-convex rank function. According to the theory of matrix completion [30], [31], only a small number, i.e., $O[\max(n,m)\log\max(n,m)^2]$, of observed entries are needed to almost perfectly recover the low-rank matrix $\mathbf{X}$. This allows us to reliably infer the unknown entries from the existing evidence scores.

Despite the encouraging theoretical guarantees, we note that problem (2) is a semidefinite programming (SDP) problem, which is known to be computationally expensive in general. Although it can be solved by optimization methods such as interior-point algorithms [32] or accelerated proximal gradient [33], the methods suffer from high computational costs and their costs per iteration are no less than $O(n^2m^2)$ and $O(n^2m+nm^2+m^3)$, respectively. Given that IBM Corporation has approximately $400,000$ employees worldwide, it is intractable to recover the unknown entries by directly optimizing problem (2). In order to handle the data with hundreds of thousands of employees, we use a much more efficient algorithm to perform matrix completion. To this end, we add a smooth regularization term $\|\mathbf{X}\|_F^2$ to the objective function (2), namely

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{R}^{n\times m}} \quad & \lambda\|\mathbf{X}\|_* + \frac{1}{2}\|\mathbf{X}\|_F^2 \\ \text{s.t.} \quad & \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{A}). \end{aligned} \tag{3}$$

As shown by [14], when λ is large enough, the optimal solution of (3) converges to the matrix that nearly optimizes problem (2). Therefore, we optimize the problem (3) with a large λ instead of directly optimizing problem (2). The beauty of problem (3) is that its dual problem can be reformulated as efficiently computable linearized Bregman iterations. Starting with a $\mathbf{Y}_0 = 0^{n\times m}$, the algorithm iterates in the following way

$$\begin{cases} \mathbf{X}_t &= \mathcal{S}_\lambda(\mathbf{Y}_{t-1}) \\ \mathbf{Y}_t &= \mathbf{Y}_{t-1} + \eta_t\mathcal{P}_\Omega(\mathbf{A}-\mathbf{X}_t), \end{cases} \tag{4}$$

until a stopping criteria is reached. Here, η_t is the step size at the t-th iteration, and $\mathcal{S}_\lambda(\cdot)$ is the singular value shrinkage operator[1] [14], which can be efficiently computed without performing a time-demanding singular value decomposition [34]. Since the linearized Bregman iterations (4) can be efficiently computed, and the algorithm converges fast in practice, we are able to reconstruct the employee-data source matrix in a relatively fast way.

V. Ordinal Regression Clustering

Once we have applied query weights to merge together results from different queries and performed matrix completion to account for missing data, the remaining steps are to determine data source weights and to determine a final expertise depth value for an employee. These two items are combined into one algorithm: multivariate ordinal regression clustering. Unlike k-means clustering [35], ordinal regression clustering finds clusters that are bands in the space of data sources which have a direct translation into the levels of very deep, deep, moderate, some, and limited. Clusters resulting from k-means in any dimension but one have no natural ordering. The cluster boundaries from ordinal regression clustering are hyperplanes, all parallel to each other. The slope vector of the hyperplanes that is found is precisely the set of data source weights we are seeking and the intercepts of the hyperplanes the clip levels of the different depths of expertise. Figure 3 and Figure 4 contrast the form of output in the multidimensional cases of k-means and ordinal regression clustering.

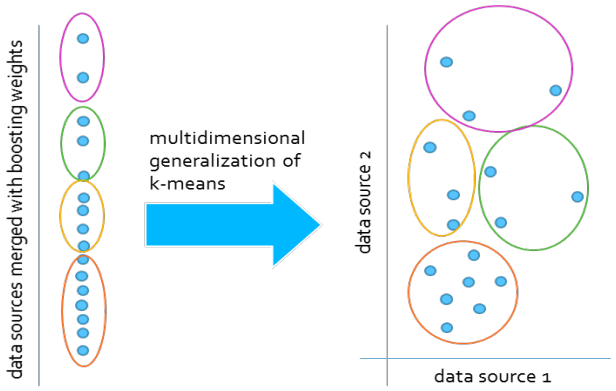

Figure 3. Illustration showing that the generalization of k-means clustering from one dimension to higher dimensions results in clusters that are blobs without any ordering.

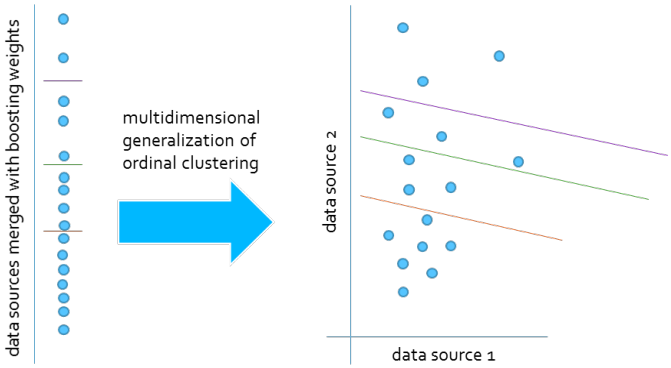

Figure 4. Illustration showing that the generalization of ordinal regression clustering from one dimension (in which it has the same form of output as k-means) to higher dimensions results in clusters that are bands separated by parallel hyperplanes that have a well-defined ordering.

The formulation for ordinal regression clustering, as proposed in [15] uses the linear support vector machine (SVM) algorithm as a component and tries to find the set of hyperplanes in the multi-dimensional space of data sources that best separate the employees into bands, defined using

[1]If $\mathbf{Y}$ has the singular value decomposition $\mathbf{U}\Sigma\mathbf{V}^T$, then $\mathcal{S}_\lambda(\mathbf{Y}) = \mathbf{U}\max(0,\Sigma-\lambda\mathbf{I})\mathbf{V}^T$.

the concept of margin.

Formally, let the completed data source vectors for the employees be $\mathbf{x}_j \in \mathbb{R}^m$, $i = 1, \ldots, n$ and the number of desired clusters be k (5 in our case). Let the vector of data source weights, equivalently the common slope of the hyperplanes, be $\mathbf{w} \in \mathbf{R}^m$. Let the clip levels, equivalently the intercepts of the hyperplanes, be $\mathbf{b} \in \mathbf{R}^{k+1}$; note that there are two more hyperplanes than the $k-1$ needed to separate k clusters with intercepts b_0 a large negative number and b_{k+1} a large positive number to bracket all of the data points. A given point $\mathbf{x}_i$ is in cluster l if $b_{l-1} \leq \mathbf{w}^\top \mathbf{x}_i < b_l$.

The clustering formulation is the following SVM-like optimization problem and can be minimized iteratively with standard SVM solvers as described in [15]:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_{i=1}^{n} \xi_i \qquad (5) \\ \text{s.t.} \quad & \max_{l=1,\ldots,k} \left(\frac{b_l - b_{l-1}}{2} - \left| \mathbf{w}^\top \mathbf{x}_i - \frac{b_l + b_{l-1}}{2} \right| \right) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

Existing solvers for the linear SVM and new ones to be developed in the future make the big data regime tractable [36].

The basic version of the algorithm is unconstrained in terms of the data source weights it produces. For semantic business reasons, it may be necessary to introduce constraints such as non-negativity and specific partial orderings on the weights. These are incorporated as projections within the iterations of the minimization algorithm.

VI. Empirical Results

We illustrate the workings of the full system of components: retrieval, fusion, completion, and clustering on one representative example expertise area that the IBM Corporation has an interest in characterizing employee expertise on: *big data and analytics*. Due to the proprietary nature of the data, we are only able to show results from a small subpopulation of IBM employees whose identity we cannot reveal. This restriction limits the size of the data set that we can show in terms of number of employees, but the distribution of the values is meaningful and does allow us to showcase the main points of the approach's operation.

Table I gives the set of queries that were elicited from subject matter experts within the company. It is interesting to note that several of the query terms are specific IBM products, e.g. Cognos, CPLEX, and Netezza; this characteristic of queries is common across different expertise areas because the businesspeople who consume the analysis desire an IBM-biased perspective. It is also interesting to note that some terms are specific and some are more general. The union of the number of employees from the subpopulation returned by the queries is 782. The inverse weighting based on the number of employees returned by query is obtained and also shown in Table I. One of the more esoteric terms on the list *front office analytics*, obtains the largest weight.

Table I
Queries for the Big Data and Analytics Expertise Area

Query	Weight
Advanced Analytics and Optimization	0.018
Analyst Notebook	0.017
Analytics Decision Making	0.017
BAO	0.047
Big Data	0.017
BigInsights	0.017
Business Intelligence	0.046
Cognos	0.021
Content Analytics	0.022
Coremetrics	0.016
CPLEX	0.017
Cross-Channel Selling and Marketing	0.016
Data Explorer	0.016
Data Management	0.021
Data Scientist	0.016
Data Warehousing	0.024
Database Administration	0.016
DBA	0.016
Descriptive Analytics	0.016
Digital Marketing Optimization	0.016
ECM	0.018
Enterprise Content Management	0.017
Entity Analytics	0.016
Financial Performance Management	0.017
Forensic Analysis	0.017
Fraud Analytics	0.017
Front Office Analytics	0.070
Hadoop	0.016
HBase	0.017
i2	0.016
IBM Research First-of-a-Kind (FOAK)	0.017
ILOG	0.016
Information Integration	0.018
Information Management	0.017
Marketing Performance Optimization	0.017
Mathematical Modeling	0.016
Mathematical Optimization	0.016
Netezza	0.017
Performance Management	0.019
Predictive Analytics	0.020
Pricing Promotion and Assortment Optimization	0.017
Regulatory Compliance and Risk Management	0.016
Risk Analytics	0.023
Sentiment Analytics	0.016
Social Media Analytics	0.017
Spend Analytics	0.016
SPSS	0.020
Steams	0.017
Tealeaf	0.016
Unica	0.016
Web and Digital Analytics	0.016

The information retrieval from which we obtain the employees is based on 31 data sources. In addition to the identity of the employees, the information retrieval also brings back evidence from the data sources that is fused using the query weights to obtain a 782×31 employee-data source matrix. The matrix has only $4,245$ known entries and is thus 82.5% missing. The maximum value of the matrix entries is 204. It is shown in the left panel of Figure 5.

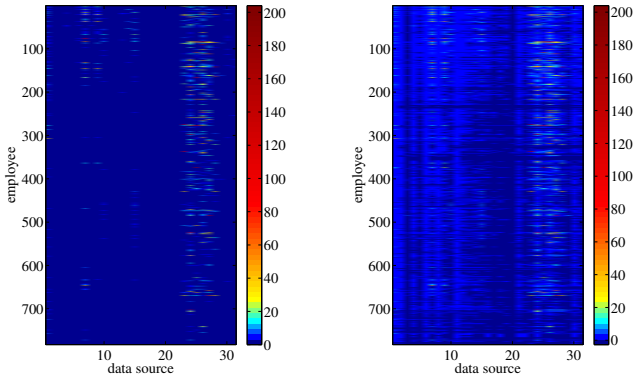

Figure 5. The example employee-data source matrix for the big data and analytics expertise area after query fusion (left) and after further matrix completion (right).

Such a sparsity level is consistent with the type of matrices constructed from larger populations of employees that we encounter in practice.

We complete the missing values as discussed in Section IV with the result being shown in the right panel of Figure 5. The relative error for the matrix completion, which is computed using

$$\frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}\frac{|\mathbf{X}_{ij}-\mathbf{A}_{ij}|}{\mathbf{A}_{ij}}\times 100\%,$$

is only 0.0099%. In addition to computing the accuracy of matrix completion algorithm on the observed entries, we also perform 10-fold cross validation, where the observed entries are randomly partitioned into 10 equal size subsamples. Each time, a single subsample is considered as testing group and is left out and the algorithm uses the other 9 groups for model construction. The root mean square error (RMSE) and the mean absolute error (MAE) are computed on each held out subsample. We repeat this 10 times, one time for each subsample. From this test, we have obtained an average RMSE of 8.68 and an average MAE of 2.46. The error numbers, when compared to the maximum value of the matrix, 204, are quite small, especially when noting the fraction of missing values. Such accuracies show that the assumptions made by our matrix completion approach are valid.

Finally, we apply ordinal regression clustering with a non-negativity constraint to the completed matrix with $k = 5$ clusters labeled as *limited*, *some*, *moderate*, *deep*, and *very deep*. The data source weights w, i.e. the slope of the hyperplanes, of the solution are presented in Table II. The non-extremal clip levels of the solution b are 17.803, 24.053, 36.841, and 50.057. The solution yields 216 employees in the *limited* cluster, 165 in *some*, 214 in *moderate*, 74 in *deep*, and 113 in *very deep*. (This particular subpopulation of employees has a larger fraction of deep experts than we typically find in other subpopulations. This result is not surprising for this subpopulation when its identity is known.) The relative weights of the data sources should be understood relative to the values that the data sources take on. Some of the most important data sources that result from the procedure are related to job role and job role skill set (JRSS), the expertise locator and answer network (ELAN), and career framework, which are all known by the business to be important descriptors of employee expertise. Among data sources from the internal corporate social media, only *what I'm known for*, *featured work*, and *community* have large weights. Others, such as *bookmark*, *wiki*, and *status update* have small weight, indicating that the internal social media content is not such a strong reflection of expertise except when used specifically to showcase one's best work or to indicate membership in groups. Two of the data sources have zero weight in the solution (and may have in fact received negative weights were it not for the non-negativity constraint), which illustrates that more evidence is not always indicative of more expertise, especially in noisy data sources like *tags* that in the IBM internal social media can be edited by anyone, not just the employee.

Table II
ORDINAL REGRESSION CLUSTERING SOLUTION

Data Source	Max	Weight
Tags	34	0.000
What I'm Known For	8	1.974
Distinguished Engineer/Fellow	1	0.000
ELAN Tags	5	2.317
Job Category	2	2.869
ELAN Expertise Name	8	0.101
ELAN Expertise Area	88	0.400
ELAN Business-Nominated Name	8	3.132
ELAN Business-Nominated Area	60	1.306
ELAN Business-Nominated Taxonomy	14	0.932
Job Responsibility	4	0.200
Primary JRSS	6	4.130
Expertise Assessment Primary JRSS	6	0.061
Expertise Assessment Secondary JRSS	12	0.287
ELAN Expertise Taxonomy	22	1.239
Primary Job Role	1	4.188
Secondary Job Role	1	0.889
Company	1	1.378
Industries	2	1.527
Sector Names	1	0.000
Work Location	3	2.183
Featured Work	2	4.014
Bookmark	151	0.154
Wiki	94	0.309
Status Update	204	0.135
Forum	147	0.172
File	120	0.272
Activity	129	0.040
Calendar	8	0.622
Career Framework	9	1.296
Community	4	4.092

VII. CONCLUSION

The advantages of the division of labor, specialization, and collective intelligence are accelerated when organiza-

tions and even society-at-large has a proper inventory of the expertise of all individuals because information and communication technologies can then be used to allocate human capital. In recent years, the nature of work has become much more knowledge-based and specialized, while also requiring integration of many disparate competencies distributed around the world [37]. Such changes to business and industry bring forth a clear and present need for expertise inference from big data about people, which is validated by the large investment that the IBM Corporation has made in expertise analytics.

This paper gives an account of one of the big data projects conducted under that Expertise-at-IBM initiative. We developed a system to infer coarse-level, loosely defined expertise areas from big enterprise data having the characteristics of volume, velocity and variety, along with different levels of value and veracity. Our proposed approach, which is now deployed within the IBM Corporation, is a unique and novel combination of data processing and machine learning capabilities that allow for accurate and close to real-time expertise inference.

We start with an information retrieval component that handles large volumes, velocities, and varieties of data by intelligent indexing and search, and then build an information fusion component atop query results that are output from the information retrieval. The main contributions described in this paper are the various ways to determine veracity and assign value to the evidence returned by the queries. We develop weighting schemes for different queries based on the cardinalities of search results, develop an approach to complete non-existent data based on matrix completion, and develop a method for giving weight to various data sources of different values and veracities while also making the inference more consumable to decision makers through ordinal regression clustering (a very new paradigm in the machine learning literature).

We illustrate the functioning of the matrix completion and ordinal regression clustering on a small real-world sample data set containing actual query results from the IBM Corporation that have been merged using the hyperbolic weighting scheme described in Section III. The reason for the example in the paper is not to illustrate the scalability of the approach, but to show how the data distributions that are typical in the application interact with the proposed machine learning algorithms. Due to the proprietary nature of the data and privacy regulations concerning it, we are unable to show the full-scale data sets on which the deployed system is currently operating.

One direction of future work is understanding how to value generalists, specialists, versatilists [38] in the broad expertise areas, i.e. those individuals who have great breadth across the expertise area but little depth in any of the subareas, individuals who have great depth in one of the subareas but little breadth across the area, and individuals somewhere between the two extremes. With some supervision, this problem can be approached by learning weights of linear combinations of order statistics [39].

Another direction for future work is calibration: the current output depths of expertise (very deep, deep, moderate, some, limited) are only valid within the data set, i.e. only within the IBM Corporation. The labels have no external calibration. To be able to say whether a very deep expert within IBM is also a very deep expert universally, we would need to gather and integrate an expertise data set that samples from the larger population of interest, which would certainly have different data sources than the internal data.

ACKNOWLEDGMENT

The authors thank Inbal Ronen, Udo Litschauer, Jun Wang, and Aleksandra Mojsilović for technical discussions as well as Scott W. Fancher, Paul Mastrangelo, and their team for subject matter discussions.

REFERENCES

[1] R. L. Martin, "The rise (and likely fall) of the talent economy," *Harvard Bus. Rev.*, Oct. 2014.

[2] 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 *KDD*, 2014, pp. 1729–1738.

[3] A. Mojsilović and K. R. Varshney, "Assessing expertise in the enterprise: The recommender point of view," in *Proc. ACM Conf. Recommender Syst.*, Vienna, Austria, Sep. 2015, p. 231.

[4] Y. Xu, Z. Li, A. Gupta, A. Bugdayci, and A. Bhasin, "Modeling professional similarity by mining professional career trajectories," in *KDD*, 2014, pp. 1945–1954.

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

[6] 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*, Cambridge, MA, Jul. 2013, pp. 20–23.

[7] D. Wei, K. R. Varshney, and M. Wagman, "Optigrow: People analytics for job transfers," in *Proc. IEEE Int. Congress Big Data*, New York, NY, Jun.–Jul. 2015, pp. 535–542.

[8] K. N. Ramamurthy, M. Singh, M. Davis, J. A. Kevern, U. Klein, and M. Peran, "Identifying employees for re-skilling using an analytics-based approach," in *Proc. IEEE Int. Conf. Data Min. Workshops*, Atlantic City, NJ, Nov. 2015, pp. 345–354.

[9] D. R. Ilgen and J. R. Hollenbeck, "The structure of work: Job design and roles," in *Handbook of Industrial and Organizational Psychology*, M. D. Dunnette and L. M. Hough, Eds. Palo Alto, CA: Consulting Psychologists Press, 1991, pp. 165–207.

[10] 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.*, Dongguan, China, Jul. 2013, pp. 534–539.

[11] 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*, Dallas, TX, Dec. 2013, pp. 672–679.

[12] K. Balog and M. De Rijke, "Determining expert profiles (with an application to expert finding)," in *Proc. Int. Joint Conf. Artif. Intell.*, Bangalore, India, Jan. 2007, pp. 2657–2662.

[13] F. Han, S. Tan, H. Sun, M. Srivatsa, D. Cai, and X. Yan, "Distributed representations of expertise," in *Proc. SIAM Int. Conf. Data Min.*, Miami, FL, May 2016.

[14] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Opt.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[15] Y. Xiao, B. Liu, and Z. Hao, "A maximum margin approach for semisupervised ordinal regression clustering," *IEEE Trans. Neural Netw.*, vol. 27, no. 5, pp. 1003–1019, May 2016.

[16] Y. Koren, R. Bell, C. Volinsky *et al.*, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[17] J. Yi, R. Jin, S. Jain, and A. K. Jain, "Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach," in *Proc. AAAI Conf. Human Comput. Crowdsourcing*, Palm Springs, CA, Nov. 2013, pp. 207–215.

[18] T. K. Pong, P. Tseng, S. Ji, and J. Ye, "Trace norm regularization: Reformulations, algorithms, and multi-task learning," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 3465–3489, 2010.

[19] J. Yi, T. Yang, R. Jin, A. K. Jain, and M. Mahdavi, "Robust ensemble clustering by matrix completion," in *IEEE International Conference on Data Mining (ICDM)*, 2012.

[20] J. Yi, L. Zhang, R. Jin, Q. Qian, and A. Jain, "Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion," in *International Conference on Machine Learning (ICML)*, 2013, pp. 1400–1408.

[21] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for multi-label image classification." in *NIPS*, vol. 201, no. 1, 2011, p. 2.

[22] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu, "Transduction with matrix completion: Three birds with one stone," in *NIPS*, 2010, pp. 757–765.

[23] J. Yi, R. Jin, A. Jain, S. Jain, and T. Yang, "Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1781–1789.

[24] J. Yi, R. Jin, A. K. Jain, and S. Jain, "Crowdclustering with sparse pairwise labels: A matrix completion approach," in *AAAI workshop on Human Computation (HCOMP)*, 2012.

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

[26] I. Ronen, E. Shahar, S. Ur, E. Uziel, S. Yogev, N. Zwerdling, D. Carmel, I. Guy, N. Her'El, and S. Ofek-Koifman, "Social networks and discovery in the enterprise (SaND)," in *Proc. ACM SIGIR Int. Conf. Res. Dev. Inf. Retriev.*, Boston, MA, Jul. 2009, p. 836.

[27] 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.*, Barcelona, Spain, Jul. 2011, pp. 582–585.

[28] 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.*, Rio de Janiero, Brazil, May 2013, pp. 515–525.

[29] A. Yogev, I. Guy, I. Ronen, N. Zwerdling, and M. Barnea, "Social media-based expertise evidence," in *Proc. Eur. Conf. Comp. Support. Coop. Work*, Oslo, Norway, Sep. 2015, pp. 63–82.

[30] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[31] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[32] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, "SDPT3—a Matlab software package for semidefinite programming, version 1.3," *Optim. Meth. Softw.*, vol. 11, no. 1–4, pp. 545–581, 1999.

[33] A. Ganesh, Z. Lin, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast algorithms for recovering a corrupted low-rank matrix," in *IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process.*, Aruba, Dutch Antilles, Dec. 2009, pp. 213–216.

[34] J.-F. Cai and S. Osher, "Fast singular value thresholding without singular value decomposition," *Methods Appl. Anal.*, vol. 20, no. 4, pp. 335–352, 2013.

[35] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recogn. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[36] F. Nie, Y. Huang, and H. Huang, "Linear time solver for primal SVM," in *ICML*, 2014, pp. 505–513.

[37] S. J. Palmisano, "The globally integrated enterprise," *Foreign Aff.*, vol. 85, no. 3, pp. 127–136, May/Jun. 2006.

[38] A. Koohang, L. Riley, T. Smith, and K. Floyd, "Design of an information technology undergraduate program to produce IT versatilists," *J. Inf. Tech. Edu.*, vol. 9, no. 1, pp. 99–113, Jan. 2010.

[39] J. R. M. Hosking, "L-moments: Analysis and estimation of distributions using linear combinations of order statistics," *J. Royal Stat. Soc. Ser. B (Method.)*, vol. 52, no. 1, pp. 105–124, 1990.