Statistical methods for automated generation of service engagement staffing plans

J. Hu
B. K. Ray
M. Singh

In order to successfully deliver a labor-based professional service, the right people with the right skills must be available to deliver the service when it is needed. Meeting this objective requires a systematic, repeatable approach for determining the staffing requirements that enable informed staffing management decisions. We present a methodology developed for the Global Business Services (GBS) organization of IBM to enable automated generation of staffing plans involving specific job roles, skill sets, and employee experience levels. The staffing plan generation is based on key characteristics of the expected project as well as selection of a project type from a project taxonomy that maps to staffing requirements. The taxonomy is developed using statistical clustering techniques applied to labor records from a large number of historical GBS projects. We describe the steps necessary to process the labor records so that they are in a form suitable for analysis, as well as the clustering methods used for analysis, and the algorithm developed to dynamically generate a staffing plan based on a selected group. We also present results of applying the clustering and staffing plan generation methodologies to a variety of GBS projects.

1. Introduction

Service-related exchanges form a growing component of the global economy, often in the form of professional service engagements [1]. A service engagement is an activity in which one company provides a service—such as software implementation, business transformation consulting, or business process implementation—to another company during an agreed-upon amount of time in order to achieve a specified business objective at an agreed-upon price. The duration of a service engagement may vary from weeks to years. Each service engagement typically requires many different types of human resources, such as software architects, project managers, or SAP** specialists, and requirements may vary during the engagement. [The term *SAP*, used throughout this paper, refers to a business application and enterprise resource planning (ERP) solution software provider. SAP stands for systems, applications, and products in data processing.]

Once a service engagement contract is signed, the provider company must quickly assemble the resources required to carry out the project. A large company typically has multiple active service engagements at different stages of completion at any given time, with competing demands on its human resource supply. It is crucial to capture the expected staffing requirements for projects during the early stages of discussion with the client in order to enable informed staffing decisions, such as the need to hire new staff or retrain existing staff. However, generating detailed staffing plans for a project which is still in the early stages of client negotiation (and which may not even progress to contract signing) creates extra work for the sales people unless a project taxonomy exists that enables mapping of potential engagements to staffing needs. While this kind of mapping is relatively straightforward to do in the context of assembling off-the-shelf hardware to be used in an end-user system, a business consulting service engagement is typically considered a custom engagement, with little repeatability from project to project. This paper reports a procedure for identifying common structures within these service deployments in order to create a systematic method for

0018-8646/07/$5.00 © 2007 IBM

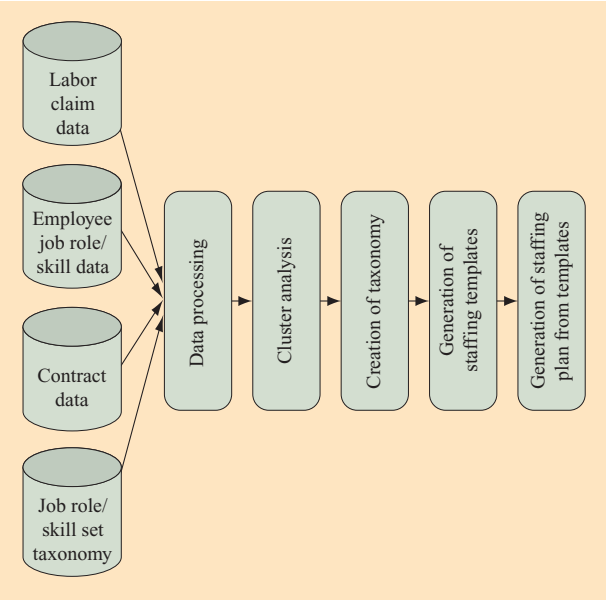

Figure 1

Flowchart illustrating the staffing plan creation process.

automatically estimating staffing requirements on the basis of key engagement characteristics. Such systematic and repeatable methods play a key role in driving the profitability of a service-oriented business [2, 3]. Although much research exists on resource estimation for IT projects, of which staffing is an integral component [4], these methods are typically quite complex and are used to create the detailed work tasks and associated staffing requirements needed for a statement-of-work document. To the best of our knowledge, the creation of reusable staffing templates for automated generation of simple staffing plans has not been well studied.

In this paper, we present a methodology developed for the Global Business Services (GBS) organization in order to identify commonality or structure within service deployments. We use that commonality to enable automated generation of *staffing plans*, i.e., the specification of staffing needs in terms of required hours of each skill each week for the planned project. The generation of the staffing plan is based on the choice of a project type from a taxonomy of project types, along with a few key additional characteristics of the potential engagement. Each project type has associated with it a staffing *template*, which is a summary of typical staffing needs specified in percentage terms. This paper focuses on the construction of the taxonomy and the associated templates and the use of the taxonomy to automatically generate staffing plans.

We describe a methodology based on statistical clustering techniques for generating groups of similarly staffed projects, using information on reported labor hours (labor-claim data) from a large number of historical GBS projects. Specifically, we detail the steps necessary to transform this labor-claim data into a form suitable for analysis. We also discuss the clustering methods developed in order to enable the grouping of projects based on staffing, along with the algorithms used to dynamically generate staffing plans based on the cluster results. The steps are illustrated using actual GBS project data. **Figure 1** is a high-level diagrammatic representation of the overall process of generating a staffing plan, each step of which is discussed in more detail throughout the remainder of the paper.

The remainder of the paper is organized as follows. Section 2 describes the data used to extract project groupings, as well as details on the data processing steps taken to manipulate the data into a format amenable to analysis. Section 3 describes the statistical clustering methodology used to create project groups having similar staffing needs, discusses the creation of an acceptable taxonomy for these groups, and presents the results for a subset of GBS projects. Cluster validation techniques are also briefly discussed. Section 4 describes the creation of staffing templates from the cluster results of Section 3 and provides an algorithm for generating a staffing plan from a staffing template. Section 5 discusses procedures for semiautomated taxonomy creation so that the methods described can be used by those with limited statistical skills, such as human resource managers. Section 6 summarizes and gives directions for further research.

2. Data description

Data sources

We obtained data from a labor-recording system that IBM uses to track information concerning the staffing of past projects. IBM consultants use this labor-claiming system each week to specify the number of hours worked that week on a particular project. The data for each consultant can then be linked to information concerning the consultant's professional development. This information includes the employee's primary and secondary job roles, primary and secondary skills, and experience level (or *band*) within the organization. A *work number*, identifying the particular project on which the consultant worked, is also associated with each record in the labor-claim system, allowing aggregation of individual labor-claim records at the project level of detail. The work number also serves to link staffing information to the ledger system in order to obtain actual revenue and cost information for the project. For our analysis, we extracted data on U.S. projects carried out by GBS

between January 1, 2004 and July 31, 2005. This resulted in approximately 2,270,000 records representing more than 9,000 projects.

IBM has developed a detailed taxonomy to describe its employees' job roles, skills, and skill levels, with each line of business maintaining a list of the current official job roles and skill sets. An example of a valid job-role/skill-set (JR/SS) combination for GBS is Application Architect/Enterprise Integration. The availability of such a taxonomy for labeling skills is crucial for application of the methodology described in this paper.

Each JR/SS combination is associated with only certain service areas of GBS, where *service area* is a term that denotes the type of business process best represented by the project. For instance, Application Architect/Enterprise Integration is a valid JR/SS for the Financial Management service area. Currently, there are more than 400 valid JR/SS combinations for GBS consultants, spanning seven service areas. Although the notion of service areas provides a high-level idea of the work involved in a project, it is not sufficiently specific to provide the detail needed for effective resource management.

At the time of our analysis, contractors were not required to record their skill information in the professional development database, so detailed information about JR/SS for contractor hours claimed against a project was unavailable. Additionally, consultants employed by global resource divisions, such as those in India or China, use a separate labor-claim system and professional development database. Current systems enable one to identify that a global resource was employed on a project, but do not provide the necessary information to link the global resource hours to the details of the employee professional development information, such as the primary JR/SS. These factors, along with other factors such as employee turnover prior to recording information in the professional development tool, resulted in some projects having a significant percentage of hours claimed for which no detailed JR/SS information could be determined. In our analysis, we omitted projects that had more than 10% of the total hours with missing JR/SS information. For the creation of staffing templates, specification of global resource skill percentages is not crucial, because the use of global resources is often dictated by client specifications and cannot be predicted from project staffing analyses. However, the omission of these projects does mean that projects that typically use global resources may not be adequately represented in the resulting taxonomy.

The current professional development database does not record the history of JR/SS changes, nor does it record the role an employee plays on a project. For the purposes of analysis, we assumed that the current JR/SS information was valid for the period spanned by the contracts under study and that a consultant's involvement in a project reflected his primary JR/SS capabilities.

Data processing

Given the discussion of the previous section, we reduced the set of projects obtained from our initial data extraction to a smaller set of projects specified by using three criteria. First, projects must have more than 40 hours claimed in total. Second, only short-term consulting projects are used, as opposed to projects concerning long-term business transformation outsourcing or application maintenance contracts. Third, projects must have less than 10% of their total hours associated with missing JR/SS information.

A duration of forty hours was used as a lower bound for projects of interest, because projects with durations below this threshold are typically staffed by a single person, for which creating a staffing plan is not an issue. Low-hour projects also often represent Project Change Requests, i.e., small additional work items tagged onto an existing project for which staffing is already in place.

We also restricted our analysis to consulting and systems integration projects of the type typically carried out by GBS. These projects are of an essentially different nature than long-term strategic or business-process outsourcing engagements, which typically involve transition of staff from client to provider, and business tasks that are repeated over the lifetime of the engagement. Projects having more than 10% missing observations were dropped because they do not provide useful information needed to generate a staffing template, as discussed above.

Various kinds of weekly variations over the project lifecycle appear to be driven, in part, by calendar effects, vacation schedules, and resource availability. For this reason, we chose to analyze projects at the aggregate level, i.e., by looking at similarities between skill distributions for the entire project and not considering variations over time. For input to the clustering algorithm described in Section 3, we characterized projects using an n-dimensional vector, where dimension j ($j = 1, \cdots, n$) denotes the percentage of total project hours claimed by consultants listed as having primary JR/SS j, and n is the number of JR/SS combinations.

As mentioned earlier, GBS has over 400 valid JR/SS categories, many of them seldom used. Additionally, many skill sets are directly linked to projects involving a particular software vendor, such as SAP, or a particular type of project. To reduce the high dimensionality of the clustering problem that results from so many rarely used JR/SS categories, we decided to group projects on the basis of similarity of job roles, not skill sets. Only those

Table 1 Percentage of total project hours within each Job Role category for a set of example SAP-related projects.

Project	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Application Architect	34	47	0	100	0	0	100	0
Application Developer	0	0	4	0	0	71	0	0
Architect–Other	0	0	0	0	0	0	0	0
Business Strategy Consultant	0	0	0	0	0	0	0	31
Business Transformation Consultant	0	0	0	0	0	0	0	0
Consultant–Other	0	1	0	0	0	0	0	0
Data Related	0	0	0	0	0	21	0	0
Knowledge and Learning	0	0	1	0	0	0	0	0
Missing	0	0	6	0	0	0	0	0
Other	1	42	4	0	0	1	0	0
Project Management	0	8	9	0	0	6	0	36
Package Solution Integration Consultant	65	3	77	0	100	0	0	33
Total (%)	100	100	100	100	100	100	100	100

job roles representing more than 95% of the total claimed hours for all projects under consideration were used for cluster analysis. Seldom-used job roles were placed in a single category labeled "Other." **Table 1** shows examples of several projects characterized in this way. Details of the clustering approach are given in the next section.

3. Cluster analysis

GBS partners suggested some initial groupings of projects by service area and inclusion or exclusion of software vendor products. For instance, partners suggested that projects involving the implementation or modification of a client's SAP system for enterprise resource planning may take one of only a few different forms, regardless of the particular service area represented by the SAP implementation (e.g., financial management or supply chain management). On the other hand, non-vendor-related projects may have staffing needs that are strongly correlated to the labeled service area of the engagement. (When we use the phrase *non-vendor-related projects*, we are referring to projects that do not involve software packages from non-IBM vendors such as SAP.) As a proof-of-concept, we first analyzed projects identified as related to a particular vendor for clients in a particular sector, regardless of a labeled service area, and then expanded our attention to projects for that vendor across all sectors. We then analyzed non-vendor-related projects, pre-grouped by service area. While grouping in this way provides some initial clusters, such clusters may exhibit large within-cluster variability because the labeled service area is not always accurate. Nevertheless, we applied formal statistical clustering methods to these pre-grouped sets of projects in order to identify more granular project groups showing the most similar JR distributions. The next section provides a brief description of the statistical clustering methods employed.

Statistical clustering methods

The goal of statistical clustering is to identify structure in a data set by organizing the set into homogeneous groups in which the within-group dissimilarity is minimized and the between-group dissimilarity is maximized. At the highest level, clustering algorithms can be divided into two categories: distance-based and model-based. Distance-based methods rely entirely on distance (or dissimilarity) measures between pairs of objects, whereas model-based methods assume a model for each of the clusters and attempt to infer the appropriate number of models and model parameters from the data. We use a distance-based approach for our application because it has the advantage that it is completely data-driven and does not require any prior assumptions regarding the structure or distribution of the data.

Two of the best-known methods for distance-based clustering are the k-means algorithm [5] and hierarchical clustering [6]. The k-means algorithm starts with an initial partition of data into k clusters and represents each cluster by the mean value of its objects. The membership of each object and the mean of each cluster are then adjusted iteratively in interlocking steps. A hierarchical clustering method works by organizing data objects into a tree. Two types of hierarchical clustering methods exist: agglomerative and divisive. Agglomerative methods start with each object being placed in its own cluster, and

the clusters are then merged repeatedly until certain termination conditions, such as a predetermined bound on a coherence measure, are reached. Divisive methods work in the opposite direction; that is, they start with all objects grouped into a single cluster, and the clusters are repeatedly divided until termination conditions are reached.

Although both of these clustering approaches have been widely used and proven to be highly effective in many applications, they have some drawbacks. The major drawbacks of the *k*-means algorithm are that 1) it assumes that the number of clusters is known beforehand, and 2) the final result is highly sensitive to the initial partition, since the method converges to a local optimal solution [7]. On the other hand, a purely hierarchical method suffers from the fact that it is a greedy approach. In other words, any merge or divide decision made early in the process cannot be adjusted later. Therefore, this method has no optimality guarantees. Numerous attempts have been made over the years to mitigate the deficiencies in both approaches. Methods for creating better initial partitions in *k*-means clustering include repeated sub-sampling [8], simulated annealing [9], and splitting based on the output of a Bayesian Information Criterion (BIC) evaluation [10]. Methods for improving the flexibility of the hierarchical approach include iterative relocation [11] and cluster refinement in relation to *k*-way graph partitioning [12].

We have adopted an approach similar to the hierarchical-*k*-means algorithm proposed by Chen et al. [13], which directly combines the two classic methods.

The approach starts with agglomerative clustering. First, each engagement is assigned to its own cluster, and the clusters are then merged iteratively. At each step, the two clusters with the smallest inter-cluster distance are merged into a new, larger cluster. In order to measure the distance between two clusters, we adopted the maximum-link formulation [6]. Using this formulation, the distance between two clusters A and B is defined as the maximum distance between any object in A and any object in B. Compared with alternative formulations, such as minimum link and average link, maximum link tends to achieve a higher degree of within-cluster homogeneity for a given number of clusters. This iterative process stops when the maximum within-cluster distance reaches a predetermined threshold. The resulting clusters are then used as the initial partition for a *k*-means procedure in which the cluster mean and membership are iteratively adjusted. At each step, a new mean is computed for each cluster, and each object is reassigned to the cluster with the closest mean. The procedure stops when there is no more improvement in average within-cluster distance.

For a distance-based approach, the choice of a between-object distance measure clearly has a significant impact on the clustering results. We are concerned with objects that are engagements represented by vectors with proportions assigned to different job role categories. We have investigated a set of standard distance measures that are applicable, including Euclidean (also known as "L_2 norm"), city block ("L_1 norm"), symmetric KL divergence (also known as Kullback–Leibler divergence), and cosine distance. The distance measures are defined as follows, given two proportion vectors $V_k = (p_{k1}, p_{k2}, \cdots, p_{kn})$ where $k = 1, 2$. (As a reminder, n refers to the clustering dimension, and p_{kj} refers to the proportion of total hours falling in category j for project k.)

Symmetric KL divergence:

$$D_{\mathrm{KL}}(V_1,\ V_2) = \sum_{i=1}^{n} p_{1i} \log \frac{p_{1i}}{p_{2i}} + \sum_{i=1}^{n} p_{2i} \log \frac{p_{2i}}{p_{1i}} .$$

Euclidean:

$$D_{\mathrm{e}}(V_1,\ V_2) = \sqrt{\sum_{i=1}^{n} (p_{1i} - p_{2i})^2} .$$

City block:

$$D_1(V_1,\ V_2) = \sum_{i=1}^{n} |p_{1i} - p_{2i}| .$$

Cosine distance:

$$D_{\mathrm{c}}(V_1,\ V_2) = 1 - \frac{\sum_{i=1}^{n} p_{1i}\, p_{2i}}{\sqrt{\sum_{i=1}^{n} p_{1i}^2} \sqrt{\sum_{i=1}^{n} p_{2i}^2}} .$$

The symmetric KL divergence approach has a problem dealing with 0 proportions. It is also difficult to interpret when mapped to the utilization of different job-role categories. Euclidean distance tends to place too great a penalty on large differences over a small number of categories. The city-block distance depends only on the sum of category-wise differences, with no regard to their distribution. Since in this particular application we are comparing proportions of hours distributed over different JR/SS categories, intuitively the appropriate distance should be one that measures the correlation between any two proportion vectors. We have thus adopted the use of the cosine distance because it is directly related to the Pearson correlation coefficient. The Pearson correlation coefficient represents the angular separation between two normalized data vectors measured from the mean, while the second term in the cosine distance measures the angular separation of two data vectors measured from zero. For future improvements, we plan to continue to investigate other possible distance measures, including correspondence measures such as the χ^2 distance [14].

Many different methods have been proposed in the literature concerning choice of the most appropriate number of clusters, including using some form of separation measure [15, 16], or BIC and related fitness measures [10, 17]. However, in practice the "right" number of clusters is often determined by the characteristics of the specific domain, and is typically somewhat subjective. The selection of an "optimal" number of clustering in general remains an unsolved problem. In our system the appropriate number of clusters to use depends not only on mathematical measurements, but also to a large degree on whether a business interpretation of the clusters can be found. We have thus taken the approach of generating a set of alternative clustering results using a range of different limits, then interacting with the business partners to determine an appropriate number in practice.

Creation of taxonomy

Once the statistical cluster analysis was complete, the next challenge was to create an appropriate project taxonomy from the results. This was accomplished using a two-step process. In Step 1, the distribution of values of project attributes for projects within the cluster were examined, and a name and description for each cluster were created. In Step 2, we validated cluster assignment and refined taxonomy labels and class descriptions through discussions with GBS subject matter experts.

The objective of the first step was to identify unique characteristics represented by each cluster. The predominant attribute values within a cluster served to provide alternative characterizations of projects, enabling the establishment of a relationship between project business attributes and project staffing requirements. Relevant attributes examined included the client's business sector (e.g., Industrial, Distribution), the service area most representative of the work (e.g., Customer Relationship Management, Supply Chain Management), the business unit of the project manager, the expected project revenue, and the way in which the project was priced (e.g., Time and Materials or Fixed Price). Information on the nature of the different projects within a cluster was also gleaned from the recorded project name and description. More formal methods could also be considered, such as fitting a multinomial generalized linear model using attribute values as predictor variables and cluster labels as dependent variables. The small size of some clusters, combined with the noise and unstructured information in the data, made this level of formalism impractical.

Step 1, involving the examination of project attributes and cluster naming, presented several challenges. First, the project descriptions were quite terse and noisy, and often did not contain much information about the project. Second, there was no data about the tools used or the platforms involved in a given project. Finally, no explicit information existed with respect to the deployment of software vendor products, although this information could sometimes be gleaned from the project description. We determined that a project was associated with a particular vendor by examining the predominant skill group used on the project.

In Step 2, domain experts were used to validate the various project types to ensure that each discovered project type was both meaningful, from a practitioner's viewpoint, and distinct, meaning that groups identified as *statistically* distinct in fact represented true variations in resource distributions due to differences in the types of projects implemented. We wanted to eliminate project clusters that were identified simply on the basis of small variations in the percentages of resources, such as may be caused by some unknown percentage of resources being supplied by the client or simply by noise in the data. We also wanted to identify in more concrete terms the type of work represented by a particular project cluster.

Cluster results for a set of 119 SAP projects from clients in the Industrial Sector are shown in **Table 2**. These five groups represent 105 of the 119 projects, or approximately 88% of the total. The remaining 14 projects were grouped into categories of size three or smaller, too small to consider as a repeated engagement type. For the five identified major project groups, no clear relationships between project attributes and project staffing requirements were readily apparent. Thus, a summary of the observed staffing requirements in each cluster was used in order to discuss the results with GBS domain experts, who were subsequently able to relate the observed staffing patterns to project type names typically used within GBS. For example, projects consisting primarily of project management and application architect resources represent the design phase of a project, commonly called a Phase 1/Blueprint project. Upon discussion with GBS partners, the three project types shown in italics in Table 2 were determined to be similar in terms of work content, differing only in the percentages of certain job roles supplied by the client. For purposes of the creation of a taxonomy, projects in these three groups were combined to form a single group, labeled "Package Configuration and Implementation."

Expanded analysis of SAP projects spanning all client sectors (**Table 3**) indicated that the main SAP project types identified for the Industrial Sector are also found in other sectors; i.e., SAP projects are invariant with respect to sector choice. However, certain types of SAP projects are more common in some sectors than others. Additional project types were identified for the expanded set, incorporating some of the projects initially labeled as outliers when the analysis was based on Industrial Sector projects only. Additionally, the grouping labeled

Table 2 Derived taxonomy for SAP-related projects for Industrial Sector clients. Projects shown in italics were combined to form a single group, as discussed in the text.

Group	*Commonalities*	*No. of projects*
Project Management Office	Dominated by Project Managers	11
Software Package Implementation	*Package Solution Integration Consultants*	*51*
Specialized Configuration	*Package Solution Integration Consultants, Project Managers, Application Architects, and Other*	*7*
Package Configuration and Implementation	*Package Solution Integration Consultants and Applications Developers, some Architects*	*26*
Transformation and Planning	Package Solution Integration Consultants, Business Transformation Consultants, Project Managers, and Other	10
Total across groups		105

Table 3 Derived taxonomy for SAP-related projects spanning all sectors. (PSIC: Package Solution Integration Consultant; PM: Project Manager.)

Group	*Commonalities*	*No. of projects*
Package Configuration and Implementation	Mix of Architect, Developer, PSIC, and PM hours, with some Specialized skills. Mix depends on percentage staffed by client	325
Planning/Phase 1/Blueprint	Predominantly staffed by PMs and PSICs	26
Business Case Development (Project Prep/Roadmap)	Packaged Solution Integration Consultants, Business Transformation Consultants, PMs, and Other	22
Migration/Architecture	Staffed predominantly by Application Architects	36
Pure Development	Staffed predominantly by Application Developers	23
Project Management Office	Most of the hours claimed by PMs	23
Total across groups		455

Transformation and Planning for the initial set was divided into two categories using the expanded set. The first category was labeled Business Case Development (Project Prep/Roadmap), incorporating Package Solution Integration Consultants (PSICs) with Business Transformation Consultants, Project Managers (PMs), and Other job roles; the second was labeled Phase 1/Blueprint, staffed primarily by project managers and package solution integration consultants.

Somewhat surprisingly, the same basic job-role-based staffing structures were apparent when all vendor-package-related projects, such as Siebel** or Oracle** projects, were analyzed; these projects varied primarily in the specific skill sets needed for project deployment. Since the GSB-defined skill sets are directly linked to a particular vendor (for instance, Package Solution Integration Consultant/PeopleSoft**), these results indicate that, given a selected job-role-based project type, appropriate skill sets can be generated simply by knowing the particular package and/or package module(s) to be involved in the engagement. However, these results were not evident for non-package-related projects.

In another example, analysis of non-vendor-package-related projects in the Human Capital Management (HCM) service area (**Table 4**) indicated some relationship between project groupings and a labeled offering name for a project. For example, On Demand Workplace* is a current IBM offering in its HCM service area, and these projects tend to involve more portal-related skills needed for development of web-based Workplace solutions. However, these relationships were not sufficiently consistent to create a taxonomy based solely on project attributes as currently captured in the opportunity and contract management systems of IBM.

Statistical validation

The quality of the project groupings uncovered by our methodology was validated in two ways. First, as described in Section 3, domain experts were asked to examine the various project clusters, along with the

Table 4 Derived taxonomy for non-software vendor-package-related Human Capital Management service area projects. (CRM: customer relationship management; PM: project manager, HR: human resources; HC: human capital. The service area is generally referred to as Human Capital Management, whereas the job role is called HR Process and Services Delivery.)

Group	*Commonalities*	*No. of projects*
Knowledge and Learning	Primarily learning consultants with application training skills	41
Business Case/Roadmap	Primarily Business Transformation Consultants (HR Process and Services Delivery/HC Strategy)	25
Project Management Office	Significant use of PMs (Custom Development skills)	11
Development/Support	Primarily application developers with portal skills	10
Portal Planning and Architecture Strategy	Business Transformation Consultants and Application Architects (Portal skills)	9
Training	Business Analysts and Support Personnel	9
Portal Design	Primarily Application Architects	9
Phase 1 Planning/Blueprint	Business Strategy Consultants (CRM Strategy/Analytical Methods), PMs	8
Knowledge and Learning Management	Business Transformation Consultants and PMs	7

associated job-role distributions, to determine whether they in fact represented distinct project types. Second, the quality of the resulting expert-refined taxonomy was validated objectively using both a metric computed from the original training data and a metric computed for a set of ten projects not operated on by the original clustering algorithm. We report here the results of this exercise applied to the 119 SAP Industrial-Sector projects whose cluster results are given in Table 2. For each project and associated cluster in the original training data, we computed the JR distribution obtained for the cluster when the observed JR distribution for that project was not included. The average job-role distribution computed from the entire data set (i.e., without any project clustering) was also computed for the project. We used the absolute deviation between estimated hours and observed hours, summed across all job roles and divided by twice the total project hours, to measure the quality, i.e., utility, of the project groupings discovered by our methodology. A second measure computed the reduction in error obtained by using the average JR distribution within an assigned group versus the overall average distribution to predict hours per job role.

In the case of the SAP Industrial Sector projects, a quality measure of 70–80% was achieved at the individual engagement level and a measure of 90–95% at an aggregate level for all projects in the cluster for major job roles using the raw clustering results (i.e., results before combination of certain statistically identified clusters). A value of 100% is perfect estimation. This represents a 36% relative reduction in error, on average, over the quality obtained using the non-clustered results. For the ten projects not included in the original analysis, job-role distributions were computed using each of the project groups, along with the average distribution across all projects in the original training set. The project group giving the smallest deviation from the actual distribution was used to assign a label to each project. The quality measure in this case was similar to that seen for the original training data, with some minor degradation for certain job roles.

For other classes of projects, we have validated the results with the domain experts but have not evaluated the quality statistically as described above. This is because domain experts expressed satisfaction regarding the project types discovered, and because the resultant templates are already being deployed in the field, which will provide more and better practical feedback from the actual users of these templates.

4. Generation of staffing plan

Using cluster results for creation of staffing "template"

As was done for the statistical validation exercise described above, a straightforward way to create a typical template from cluster results is to combine the hours in each resource category across all projects in the cluster and compute the category distribution on the basis of the total claimed hours for all projects in the cluster. While this provides a consistent distribution (percentages of total project hours across all categories sum to 100%), large projects, defined as projects with a large number of total hours, have an unduly large influence on the results. Additionally, certain categories may turn out to have very small, but nonzero, percentages because only a

few projects in the cluster have hours claimed in that category. In such cases, a more representative template should reflect the fact that the "typical" project of that type does not use resources from that category, yet also indicate that similar projects have occasionally used resources from that category in the past.

In order to create valid distributions with percentages that sum to 100% while still addressing the concerns mentioned above, we use the median proportion of hours for each category across all projects in a cluster to typify that job-role distribution. Because the medians across all categories do not necessarily sum to 1, we uniformly rescale each nonzero job-role category by the inverse of the sum to achieve this objective of medians summing to 100%. In other words, we divide each nonzero job-role percentage by the total percentage across all job roles to force the JR percentage to sum to 100%. We then convert the "Other" job-role category, which was created to reduce cluster dimensions but may represent a nontrivial percentage of hours for a particular cluster, to an official job-role category by determining the most common official job role in the Other category after scanning all projects in the cluster. To reduce the chance of not having a necessary resource when that resource has been used in a small percentage of projects, but whose median proportion is zero, we include the resource in the staffing template, but with zero hours initially allocated. In this way, the resource is suggested, but not required. We discuss this issue further in the following paragraphs.

Although the clustering was done at the job-role level, effective management of resources, and hence useful staffing plans, require resource predictions at the job-role, skill-set, and band levels. Given that in many cases a single skill set was matched with a particular job role within a cluster, specification of a template to the JR/SS level was achieved without additional formal clustering of project groupings by skill-set categories.

The template information is specified in a hierarchical fashion, i.e., first, the typical distribution of hours by job role is specified. Next, the typical distribution of skill sets within each job role is specified (usually only one or at most two skill sets). Finally, the typical distribution of bands within each skill set is specified. At each level of the hierarchy, the percentages of allocated hours sum to 100%. Band-level allocations for each JR/SS combination were determined by finding the typical distribution of bands for each skill set within each job role. Band levels typically depend on a particular job role, but not on a particular skill set. The resulting templates were each validated to ensure that the recommended JR/SS combinations represented valid JR/SS combinations, as discussed in Section 2.

To make the system inviting to practitioners, the templates also include job roles and/or skill sets that have occasionally been used in past projects, but with zero percentage of hours allocated. In this example, all SS/Band allocations for that job role are also zero, but each of these JR/Band/SS combinations results in a generated position in the staffing plan, each with zero hours per week. In this way, practitioners can quickly customize generated staffing plans to include nontypical staffing requirements specified using valid JR/SS combinations, without having to remember or search through the 400 potentially valid JR/SS combinations.

Creation of a staffing plan from a template

For the cluster results to be practically useful beyond providing insight into typical types of projects, a mechanism must exist for generating an actual staffing plan from the resulting templates. For new projects, an estimate of the total project hours is required. Although this can come directly from expert opinion, more typically new project opportunities have an expected revenue value associated with them. For time- and materials-based projects, revenue is equal to the project cost, which is entirely a labor cost for GBS projects, plus a specified markup to generate profit. This implies that revenue and hours are exactly correlated for projects having exactly the same distribution of resources down to the JR/SS/band/geographic-location level. We use this implied relationship to construct cluster-specific factors that can be used to estimate expected hours from an expected revenue, given that the clusters are constructed to have approximately the same distribution of resources at least at the job-role/skill-set level. Variation in bands, and the use of unknown global resource skills on projects falling in the same cluster, introduce some noise into this relationship.

We use a simple zero-intercept least-squares regression of actual claimed hours against actual achieved revenue to obtain the scaling factor for each project type. The zero-intercept model was chosen after examination of scatter plots of hours versus achieved revenue, which indicated no systematic fixed-cost component for time- and materials-based projects. Fixed-price contracts were not used to estimate the hours-per-revenue relationship because the historical examples showed no systematic patterns. On the other hand, the estimated time-and-materials relationship can be used as a rough check on the ultimate pricing of a fixed-price project.

Indicative of the different resource requirements and different cost rates across project types, the scaling factor for SAP Business-Case Development projects is 0.003743, while that for projects labeled as Pure Development is 0.007423. This difference reflects the fact that Business Case projects typically involve senior Business Transformation Consultants, whose cost per hour is at the upper end of the scale, while Pure Development

projects are staffed predominantly by software engineers, whose cost per hour is significantly less.

The staffing templates described in the previous subsection are used to generate a suggested distribution of work hours by JR/SS/Band for new projects, given an estimate of the total project hours for the new project (as discussed above) and assuming that the expected project duration (in weeks) is already specified. (The duration of a project is typically driven by client timelines.) The hours for each JR/SS/Band are then converted to FTE (full-time equivalents) on the basis of a factor specifying the typical number of hours worked per person per week. Additionally, for certain project types, we solicit input from the staffing plan creator as to whether certain of the resources are to be supplied by the client (see the discussion in the subsection on taxonomy creation in Section 3). This information is used to modify the selected template. Details are given in the algorithm outlined below:

1. Look up selected project type. Use scale factor associated with project type to estimate Total Hours from Expected Revenue.
2. Use information captured concerning percentage of client resources to be used for a particular JR to redistribute the JR percentages given in the lookup table. (In other words, we redistribute the hours associated with the JR percentages given in the lookup table in order to reflect the staffing to be supplied to GBS.) If relevant, use input concerning the modules of a software vendor's package that will be implemented to associate skill percentages with the job roles Package Solution Integration Consultant and Business Transformation Consultant. The skill sets are distributed uniformly among the selected modules.
3. Distribute Total Hours among JR/SS/Band combinations according to the percentages given in lookup table, modified according to Step 2. Note that Band and Skill Set allocations are assumed to be independent; thus, the number of Band/SS combinations for a Job Role is the number of Bands multiplied by the number of Skill Sets, with the Allocation equal to (Allocation for band) $\times$ (Allocation for skill set).
4. Multiply percentages in Step 3 by Total Hours (rounded to nearest integer) to obtain Total Hours per JR/SS/Band.
5. Compute Hours per week per position. The Total Hours are divided by project duration, in weeks, to convert to hours per week. Next, the hours per week are divided by U, a maximum hours per week threshold, and the result is rounded up to the next largest integer to obtain the number of positions needed for the JR/SS/Band.

The following example should elucidate the process mentioned above. For an SAP Package Configuration and Implementation project, whose initial (i.e., nonadjusted) job-role distribution is shown in **Table 5**, suppose the user indicates that 80% of Developers, 40% of Application Architects, and 100% of Learning Consultants will be supplied by the client. Then 11% is multiplied by (100 – 40)% to obtain 6.6% as the new value for Application Architects. In other words, 40% of Architects will be supplied by the client, and thus 60% will be supplied by GBS. The value 24% is multiplied by (100 – 80)% to obtain 4.8% as the new value for Application Architects; 0% is obtained for Learning Consultants. These new percentages are summed across all job roles, giving 71.4%. The percentage of hours to be redistributed is then (100 – 71.4)% = 28.6%. The remaining percentage of hours is reallocated across all job roles that did not have client-supplied resources, with the exception of the Project Manager job role. The project management need is assumed to remain unchanged no matter the staffing source, as recommended by GBS subject matter experts. The reallocation is done in proportion to the amount the remaining job roles contributed to the total before redistribution. In this example, all of the extra hours are allocated to the PSIC, increasing the PSIC percentage to 78.6%.

To compute staffing needs in terms of persons (FTEs), assume that application architect skills are needed for 400 hours over a four-week project, translating to 100 hours per week. Let $U = 45$. Then $(100/45) = 2.22$, implying that three positions are needed. We assume that the people in the first two positions work for 40 hours per week, which we use as the *typical* number of hours per week. The third position is assumed to have only ten hours per week to work, or $10 \times 4 = 40$ hours total. Instead of assigning a resource for ten hours per week over the life of the project, we assign a resource for 40 hours per week for the first week of the project only. The staffing plan creator can adjust the exact start and end dates for the resource as needed. If a position has between L and U hours per week, where L represents the minimum number of hours per week, the number of hours per week is not changed. We allowed staffing in increments of 0.5 FTEs only to reflect the way in which projects are staffed at IBM.

5. Automation

While much of the earlier work involving data processing, generation of taxonomy, and generation of templates from clusters to create a meaningful project taxonomy was done manually, we have since developed a

Table 5 Job Role distribution (%) for SAP Package Configuration and Implementation project. Five job roles are shown.

Application Architect	*Application Developer*	*Learning Consultant*	*Project Manager*	*Package Solution Integration Consultant*
11	24	5	10	50

semiautomated process that allows the analysis to be done much faster and with less subjectivity. Moreover, a semiautomated process allows us to parameterize the process, thereby making the analysis more flexible by allowing us to impose different constraints and try alternative approaches by simply changing the values of corresponding parameters. Finally, a semiautomated process allows the eventual transfer of the technology to GBS so that GBS professionals can use the same process to extend the analysis to other regions and service areas, or simply to redo the analysis with different data.

The automation has been done for two distinct phases of the analysis—data processing and template generation from clusters. The data processing phase (top table of **Figure 2**) allows claim data, with each record detailing the hours claimed by an individual employee on a particular project (thus, multiple records for each project), to be aggregated and transformed into data with one record per project along with the job-role distribution for that project (bottom table of Figure 2). Several different parameters are specified, including such selection criteria for the accounts as sector and/or service area. Another parameter describes constraints on project types—for example, that projects should require more than a certain number of hours. Filters exclude certain job roles or accounts based on different criteria, such as the percentage of missing values above a threshold, as discussed in Section 2. By changing the values of such parameters, data can be processed very rapidly in order to perform analysis for various kinds of projects or using different methodologies.

Similarly, the template-generation phase helps partially automate the process of generating staffing templates from the clusters (project categories) discovered via the clustering process. While different ways of assigning job roles, skill sets, and employee bands (and corresponding distributions) to each cluster were described in detail in Section 4, we note that the automation with accompanying parameterization helps speed the process while reducing the subjectivity of the manual process. As such, this phase involves the computation of various cluster statistics such as the median and means for the various job roles across all projects in a cluster, and use of the computed values to assign an appropriate job-role distribution to each cluster on the basis of various parameters. For example, one way of assigning job roles to a cluster may be to include job roles that have a median greater than zero, or that have a nonzero mean with sufficient coverage by the JR of the cluster constituents and/or suitable difference from the median. By setting parameters for such rules, the process can ensure consistency across all of the clusters and repeatability of the process, as well as allowing the transfer of the template-generation work to other, nontechnical users.

Additional automation could be undertaken to compute the distribution of various project attributes for each cluster. This would facilitate the identification of relationships between clusters and project characteristics. As with the automation described above, the choice of attributes as well as the choice of parameters for producing the distributions could be parameterized for rapid exploratory analysis.

6. Conclusions and future research

The results presented in this paper have been implemented in a pilot tool used by GBS partners to capture expected staffing requirements for projects in the early stages of client negotiation, before contract signing. When the attributes of a potential project meet a set of specified criteria, the project information is downloaded into the GBS tool, and a request to generate a staffing plan is sent to the partner responsible for the project. The partner then uses the tool to create a staffing plan, either manually or by selecting an appropriate project type from the project taxonomy and requesting automated staffing plan generation. The resulting staffing can be used as is or customized by the user to better reflect the unique characteristics of the service engagement opportunity. The tool is used to modify and track staffing for the project throughout the lifecycle of the project.

Initial feedback indicates that the practitioners find that the automated staffing plan creation is useful, and they have requested that the initial results be extended to cover all types of GBS projects. However, in order to provide a truly automated method for generating resource predictions, we need a way to automatically map potential engagements to staffing requirements. The results of the current analysis have provided suggested modifications to the engagement opportunity tracking system so that this objective can be achieved in the future.

We can envision other extensions to the current work in order to provide not only typical staffing plans for projects, but "optimal" staffing plans as well (for instance, we may wish to take into account staffing availability). Similarly, it may be feasible to limit projects used for staffing template construction to those showing satisfactory profit margins and high customer

Work number	…	Serial no.	Job role	Hours	…
wrkNum1	…	000111	Application Architect	1030.5	…
wrkNum1	…	000111	Application Architect	602.5	…
wrkNum1	…	000111	Application Architect	432	…
wrkNum1	…	000222	PSIC	126	…
wrkNum1	…	000222	PSIC	269	…
wrkNum1	…	000222	PSIC	3482.8	…
…	…	…	…	…	…
wrkNum2	…	…	…	…	…
wrkNum2	…	…	…	…	…

Work number	…	Total hours	Application Architect	Application Developer	…	Missing	Other	PSIC
wrkNum1	…	5985.8	2064.5	0	…	0	35.5	3877.8
wrkNum2	…	479	224	0	…	0	200	12
wrkNum3	…	25194.2	0	1016.5	…	1150	959	19464.7
…	…	…	…	…	…	…	…	…

Figure 2

Data transformation from hour-centric to job-role-centric. The work number identifies the particular project on which a consultant works and allows aggregation of individual labor claim records at the project level of detail. Dots in columns indicate that additional attributes exist, associated with a record, which we omit for clarity. Dots across rows indicate that additional records are not shown. (PSIC: Package Solution Integration Consultant; Serial no.: employee serial number.)

satisfaction. Exploration of the relationships between project staffing and project success is a subject for future research, as is a method for automated updating of staffing templates on a periodic basis. From a methodological perspective, we note that it is challenging to identify project clusters reflecting not only similarity of staffing at the aggregate level, but also over the course of the project. Work in this area is currently being explored and will be reported in a separate paper.

Acknowledgments

The authors wish to thank John M. Collins and the Enterprise On-Demand Transformation team at IBM for their support of these research efforts. We also thank the IBM GBS WW Demand Capture team for their help in validating the analysis results and their willingness to include the resulting staffing plan generation algorithms in their resource management tool. We thank Yang (Jessie) Liu for extensive data processing support. The reviewer comments, which contributed tremendously to improving the quality of the paper, are also gratefully acknowledged.

References

1. T. Deong and A. Nanda, *Professional Services Text and Cases*, McGraw-Hill, Boston, MA, 2003.
2. R. Melik, L. Melik, A. Bitton, G. Gerdebes, and A. Israilian, Professional Services Automation, *Optimizing Project and Service Oriented Organizations*, John Wiley and Sons, New York, 2002.
3. D. A. Aranda, "Service Operations Strategy, Flexibility and Performance in Engineering Consulting Firms," *Intl. J. Oper. & Production Manage*. **23**, No. 11, 1401–1421 (2003).

4. P. Coombs, *IT Project Estimation: A Practical Guide to the Costing of Software*, Cambridge University Press, New York, 2003.
5. J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 1967, pp. 281–297.
6. A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Upper Saddle River, NJ, 1988.
7. S. Z. Selim and M. A. Ismail, "*K*-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality," *IEEE Trans. Pattern Anal. & Machine Intell.* **6**, 81–87 (1984).
8. P. Bradley and U. Fayyad, "Refining Initial Points for *k*-Means Clustering," *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI, 1998, pp. 91–99.
9. D. E. Brown and C. L. Huntley, "A Practical Application of Simulated Annealing to Clustering," *Technical Report IPC-91-03*, University of Virginia, Charlottesville, 1990; see *http://portal.acm.org/citation.cfm?coll=GUIDE&dl=GUIDE&id=900571*.
10. D. Pelleg and A. Moore, "*X*-Means: Extending *K*-Means with Efficient Estimation of the Number of Clusters," *Proceedings of the 17th International Conference on Machine Learning*, San Francisco, CA, 2000, pp. 727–734.
11. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *Proceedings of the 1998 ACM–SIGMOD International Conference on Management of Data*, Seattle, WA, June 1998, pp. 73–84.
12. G. Karypis, E.-H. Han, and V. Kumar, "Multilevel Refinement for Hierarchical Clustering," *Technical Report 99-020*, Department of Computer Science, University of Minnesota, Minneapolis, 1999; see *http://glaros.dtc.umn.edu/gkhome/node/153*.
13. B. Chen, P. C. Tai, R. Harrison, and Y. Pan, "Novel Hybrid Hierarchical-*K*-Means Clustering Method (*H-K*-Means) for Microarray Analysis," *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops (BCSBW'05)*, Stanford, CA, 2005.
14. M. J. Greenacre, *Theory and Application of Correspondence Analysis*, Academic Press, London, 1984.
15. L. Kaufman and R. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, 1990.
16. R. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," *Proceedings of the 20th International Conference on VLDB*, Santiago, Chile, 1994, pp. 144–145.
17. C. Fraley and A. Raftery, "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," *The Computer J.* **14**, No. 8, 578–588 (1998).

Received September 14, 2006; accepted for publication October 15, 2006; Internet publication May 10, 2007

Jianying Hu *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (jyhu@us.ibm.com)*. Dr. Hu received her Ph.D. degree in computer science from the State University of New York at Stony Brook in 1993. From 1993 to 2000, she was a member of the technical staff at Bell Labs in Murray Hill, New Jersey. From 2001 to 2003, she was at Avaya Labs Research in Basking Ridge, New Jersey. In 2003 she joined the IBM Thomas J. Watson Research Center in Yorktown Heights, New York. Her main research interests include statistical pattern recognition and signal processing with applications to document image analysis, handwriting recognition, multimedia content analysis and retrieval, and business analytics. Dr. Hu is a Senior Member of the IEEE, chair of Technical Committee 11 (Reading Systems) of the International Association for Pattern Recognition, and a member of the IEEE Multimedia Signal Processing Technical Committee. She served as an associate editor of the *IEEE Transactions on Image Processing* from 2001 to 2005, was a guest editor for the *International Journal on Document Analysis and Recognition* in 2003, and is currently an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bonnie K. Ray *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (bonnier@us.ibm.com)*. Dr. Ray has been a Research Staff Member at IBM since 2001; she is currently Program Manager, Data and Analytics, at the IBM China Research Laboratory. She received a Ph.D. degree in statistics from Columbia University and a B.S. degree in mathematics from Baylor University. Her area of expertise is applied statistics, with a particular interest in the use of statistics for business analytics. Prior to joining IBM, Dr. Ray was a tenured faculty member in the Mathematics Department at the New Jersey Institute of Technology, and she has held visiting appointments at Stanford University, the University of Texas, and the Los Alamos National Laboratory. She is a Fellow of the American Statistical Association and currently serves as an associate editor for the *Journal of Computational and Graphical Statistics* and the *International Journal of Forecasting*.

Moninder Singh *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (moninderr@us.ibm.com)*. Dr. Singh joined IBM in 1998 after receiving his Ph.D. degree in computer and information science from the University of Pennsylvania. His research interests include the general areas of artificial intelligence, machine learning, and data mining.