

Sequence Mining for Business Analytics: Building Project Taxonomies for Resource Demand Forecasting

Ritendra Datta¹, Jianying Hu², and Bonnie Ray^{2,3}

¹ Department of Computer Science and Engineering
The Pennsylvania State University, University Park, USA
datta@cse.psu.edu

² Mathematical Sciences Department
IBM T.J. Watson Research Center, Yorktown Heights, USA
jyhu@us.ibm.com

³ Data and Analytics
IBM China Research Lab, Beijing, China PRC
bonnier@cn.ibm.com

Abstract. We develop techniques for mining labor records from a large number of historical IT consulting projects in order to discover clusters of projects exhibiting similar resource usage over the project life-cycle. The clustering results, together with domain expertise, are used to build a meaningful project taxonomy that can be linked to project resource requirements. Such a linkage is essential for project-based workforce demand forecasting, a key input for more advanced workforce management decision support. We formulate the problem as a sequence clustering problem where each sequence represents a project and each observation in the sequence represents the weekly distribution of project labor hours across job role categories. To solve the problem, we use a model-based clustering algorithm based on explicit state duration left-right hidden semi-Markov models (HsMM) capable of handling high-dimensional, sparse, and noisy Dirichlet-distributed observations and sequences of widely varying lengths. We then present an approach for using the underlying cluster models to estimate future staffing needs. The approach is applied to a set of 250 IT consulting projects and the results discussed.

1 Introduction

A good view into future resource needs is essential for driving profitability in a service-oriented businesses [5]. Large projects typically require multiple resources, each having different skills. The resource requirements are not static, instead varying over the life of a project as it enters different phases. Both lack of resources with the appropriate skills to carry out a project when needed as well as over-supply of resources who are under-utilized result in loss of profits to the business.

One approach to predicting future resource demands is to create a project categorization scheme that links a set of project attributes captured in the early

stages of negotiations with a client to typical resource requirements over the project life-cycle. In this paper, we present a model-based sequence clustering algorithm useful for finding groups of projects showing similar resource requirements over the project life cycle, and show how the resulting groups can be used to infer a project taxonomy. To automatically determine project groups, we use a hidden semi-Markov model (HsMM)-based clustering algorithm. Higher-level descriptions are associated with each cluster with the help of domain experts. The models describing each cluster are used to form templates representing typical staffing requirements for the different project types. These templates can then be used to generate forecasts of future staffing needs.

Three major contributions are given in this paper. First, we formalize the problem of project taxonomy building for workforce management as a sequence clustering problem. Second, we present a clustering algorithm that includes several new advances over past attempts at sequence modeling and clustering [8, 9, 1, 4, 12, 11, 7] to address challenges specific to the business problem. Third, we outline the use of the cluster model results for project-based resource demand forecasting. We demonstrate the effectiveness of the proposed approach on labor claim data from IBM Global Business Services (GBS) consulting engagements.

The remainder of the paper is structured as follows. In Section 2, we motivate formulation of the business problem as a sequence clustering problem and define some notation. In Section 3, we present the HsMM-based sequence clustering algorithm. In Section 4, we outline use of the cluster results for project-based demand forecasting. Section 5 presents results of application to real data from a set of IBM Global Business Services consulting projects. We conclude with a discussion, in Section 6.

2 Taxonomy Building as a Sequence Clustering Problem

Fig. 1 shows a high level workflow for using historical project labor data for resource demand forecasting. We concentrate here on the initial steps, i.e. processing and analyzing labor claim records to determine similarities between projects. We formalize the model-based clustering problem as follows. Suppose we have labor claim data for a set of m historical projects $\{P_1, P_2, \dots, P_m\}$ completed by an organization, where the resources completing the labor are each labeled according to an agreed upon job role categorization scheme. Let the total number of resource types available within the organization be D . A project P_i has a duration of T_i (in weeks), and for each week j (relative to the starting week), the *proportion vector* representing the distribution of resources across job role categories is specified by $O_{i,j} \in \mathbb{R}^D$. Since resource distributions are represented as proportions, for each week i we have $\sum_{k=1}^D O_{i,j}(k) = 1$, $O_{i,j}(k) \in \mathbb{R}$, where $O_{i,j}(k)$ (the k^{th} component of the vector) denotes the observed proportion of resource hours in job role category k in week j for project i . Thus the full specification of a project P_i , in terms of resource requirement distributions, is given by the sequence of proportion vectors, ordered by weeks: $P_i = (O_{i,1}, O_{i,2}, \dots, O_{i,T_i})$, $O_{i,j} \in \mathbb{R}^D$, $1 \leq j \leq T_i$. Since durations of projects

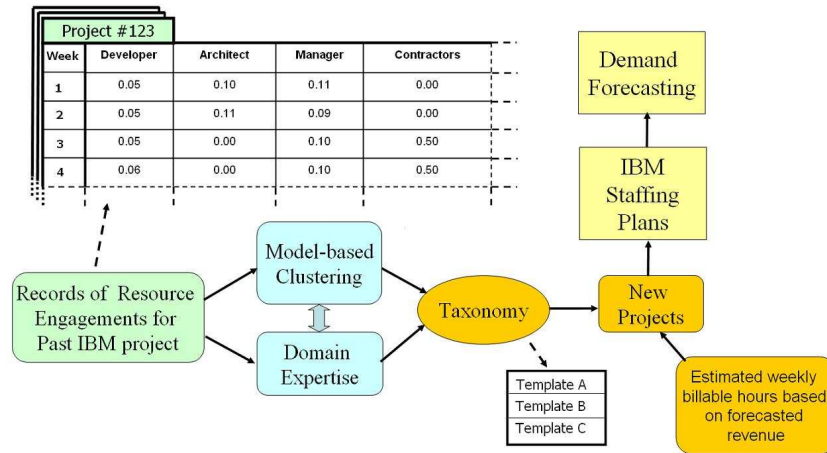


Fig. 1. A high-level view of how the sequence clustering fits into overall workflow for demand forecasting.

vary, T_i take different values for different projects. Therefore each P_i can be thought of as a variable-length multivariate sequence. Given that there are m such projects, we have a set of m multivariate sequences; we would like to find those sequences that exhibit similar behavior. More formally, the problem is to devise an algorithm for clustering variable-length multivariate sequences having some specific characteristics. In particular, we seek a *model-based* sequence clustering algorithm to capture the following characteristics of the business problem.

1. Projects are typically carried out in phases.
2. Temporal dependencies among project phases are common.
3. The observation vectors at each sequence time period represent proportions, so appropriate statistical distributions are necessary for their modeling.
4. The proportion vectors may be sparse, i.e., only a small set of the resources may be contributing to a project in a given week.
5. The number of clusters in the taxonomy is not known *a priori*.
6. There may be atypical weeks within projects, and atypical projects as a whole, both of which should be treated as outliers.

We are interested in inferring statistical models for each cluster, so that the resulting project taxonomy can be used to generate expected resource requirements given a selected project group. We call the model used to generate such a forecast a *staffing template*.

Note that alternative data mining algorithms, such as decision trees or item-set mining are not appropriate for this problem, as our objective is to group sets of projects that have no explicit labeling. The next section presents the HsMM-based sequence clustering algorithm.

3 HsMM-based Sequence Clustering

In this section, we provide motivation for the HsMM modeling choices used for analysis.

1. Proportion vectors are defined on a $(D - 1)$ -simplex, where D is the number of resources. Hence we use a Dirichlet distribution [6] for modeling the observation vectors, a standard distribution for describing multivariate proportion data. For x in the $(D - 1)$ -simplex, the probability density function (p.d.f.) for the D dimensional Dirichlet distribution is defined as

$$f_d(x|b) = \frac{1}{B(b)} \prod_{i=1}^D x_i^{b(i)-1}, \quad x \in \mathbb{R}^D, \quad (1)$$

where

$$B(b) = \frac{\prod_{i=1}^D \Gamma(b(i))}{\Gamma\left(\sum_{i=1}^D b(i)\right)} \quad (2)$$

and $\Gamma(\cdot)$ denotes the Gamma function.

2. Since phases within projects govern the resource requirements, we use a first-order hidden Markov model (HMM) [8] to characterize a project's transition through project phases over its life-cycle. An HMM is a finite-state probabilistic model governed by first-order state transitions.
3. Since project phases typically occur in a sequence, with the same phase not repeating itself, we restrict our model to (a) non-ergodic left-to-right HMMs, (b) having strict start and end states, with (c) state skipping being disallowed. These are essentially topological restrictions which yield a reasonable simplification to the model, without loss of information.
4. Because project phase durations may not necessarily follow a geometric distribution, which is implicit for a first-order HMM, we employ HMMs with state duration explicitly modeled by a Gamma distribution, allowing much more flexibility in the state durations. With explicit duration modeling, HMMs are no longer first-order models, and hence are referred to more accurately as hidden semi-Markov models (HsMM) [3]. The p.d.f. of the Gamma distribution is given by

$$f_g(x|\psi, \theta) = x^{\psi-1} \frac{e^{-x/\theta}}{\theta^\psi \Gamma(\psi)}, \quad x \in \mathbb{R}^+. \quad (3)$$

A schematic view of an N -state HsMM as described above is shown in Fig. 1. For clustering the project data under the assumption of an HsMM for each cluster, we use a modified version of the algorithm proposed in [1], designed to address challenges posed by the high dimensional, yet sparse, proportional nature of the observations, and the presence of noise both within a sequence and for a set of projects. The algorithm carries out **model construction** via robust estimation of HsMM parameters, with the number of HsMM states inferred iteratively

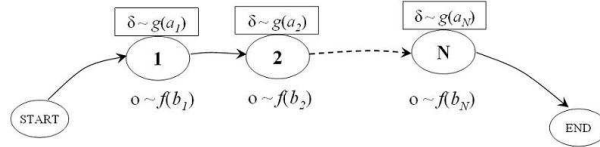


Fig. 2. The left-to-right HsMM topology used for sequence modeling. Symbol δ denotes state/phase duration, $f(\cdot)$ denotes the Dirichlet p.d.f., and $g(\cdot)$ denotes gamma p.d.f.

using the Bayesian Information Criteria (BIC). The number of clusters is also automatically inferred using a modified BIC, referred to as **partition_BIC**. The steps in the main sequence clustering algorithm are shown in Table 1.

The proposed model-based clustering algorithm extends the work of [1], primarily through additional focus on the modeling aspects of the work. The methods presented in [1] discretize the sequence data by performing an initial assignment of each observed time point to a state. An HMM model is then applied to cluster the sequences of discretized data, rather than fitting a distribution to directly characterize the observed vectors, as we have done here using a Dirichlet distribution. Here, the characterization of each state by a parametric model is useful for generation of staffing templates, as discussed in the next section. However, parametric modeling necessitates the development of a dimension reduction procedure for estimating the Dirichlet distributions, as the number of observation dimensions having non-zero proportions is typically small relative to the total number of dimensions. We leave further discussion of the dimension reduction technique to another paper. Lastly, we generalize the state duration distribution to better reflect the characteristics of project durations in practice.

4 Resource Demand Forecasting based on Cluster Results

In order to use the cluster results for resource demand forecasting, it is necessary to create a description of each cluster that reflects the typical resource requirements and project phase durations for the projects in that cluster. A straightforward way to obtain the typical resource distribution per phase for a cluster is to aggregate the hours in each resource category across all projects in the cluster for each identified project phase and compute the category distribution based on the total claimed hours for that phase. Note that this approach

Table 1. The proposed sequence clustering algorithm.

Assign all sequence to one cluster Apply model construction to the one cluster Compute partition_BIC While partition_BIC \geq old partition BIC Compute individual cluster_BIC values Split the weakest cluster in two using hierarchical clustering Apply model construction to the two new clusters While membership change continues Reassign each sequence to its most likely cluster Apply model construction to get a new set of clusters Re-compute partition_BIC Use Monte Carlo simulation to get cluster likelihood thresholds If a sequence has likelihood less than the threshold for its cluster Reject the sequence as noise/outlier Apply a final model construction to each cluster

results in certain categories having very small, but non-zero, percentages because only a few projects in the cluster have hours claimed in that category for a phase. Here, we take advantage of the models generated for each cluster by using the estimated, reduced-dimension Dirichlet distributions for each state to directly compute the mean resource distribution for each phase.

To obtain estimates of project phase duration, we distinguish between two different scenarios. 1) Individual phase durations and resultant overall project duration are estimated based on the mean calculated directly from the gamma distribution for each corresponding state. 2) In the case the project duration is specified *a priori*, for example, because of requirements from the customer, the duration of each phase can be estimated by scaling the mean phase durations generated in 1) by the fixed total project duration. Further discussion of template generation from cluster results is given in [2], where the focus is on project clustering at the aggregate level, i.e., the problem of sequence clustering is not addressed.

Once the statistical cluster analysis and template creation are complete, the next challenge is to create an appropriate project taxonomy from the results. In general, this can be accomplished using a two-step process. 1) For each cluster, examine a set of project attributes (other than resource requirements) whose distribution of values suggests a name and description for each cluster. 2) Validate and refine cluster taxonomy labels and class descriptions through discussions with subject matter experts.

The objective of the first step is to identify unique characteristics represented by each cluster. The predominant attribute values within a cluster serve to provide alternative characterizations of projects, enabling linkage of a project's business attributes and a project's staffing requirements. In Step 2, domain experts are used to validate the various project types to ensure that each discovered

project type is both meaningful, from a practitioner’s viewpoint, and distinct, meaning that groups identified as statistically distinct in fact represent true variations in resource distributions due to differences in the types of projects implemented.

The staffing templates can be used for generating project-based demand forecasts through application of the following steps.

1. Select a label for a project opportunity from the created project taxonomy.
2. Compute the duration of each project phase, either using a predetermined project duration or estimated project duration, as discussed above.
3. Estimate the number of project hours required for each project phase. This estimate can come directly from expert opinion, or be based on an established relationship between project revenue and project hours. For each project grouping, such a relationship can be established for each project phase through analysis of achieved revenue in each project phase relative to total project hours worked in that phase for the set of historical projects determined to group together. As for project durations, a typical percentage of revenue achieved for each phase in a project can be established. Then, regression analysis, for example, can be used to estimate a scaling factor to translate revenue to hours for each phase of the new project. Typically, new project opportunities have an expected revenue value associated with them early in discussions with the client.
4. Distribute the expected hours per phase across the project job roles according to the established Dirichlet distributions.

Estimated resource demands and their start/end dates are aggregated across all project opportunities at a weekly level to achieve a total view of expected resource requirements over a specific time interval.

5 Results on IBM Data

To test the efficacy of the proposed method, we initially conducted a number of experiments on synthetic sequences and found the clustering results satisfactory. Given our business goal, we then tested how well the algorithm was able to produce meaningful resource-based project taxonomies for real project staffing data. We applied the method to a set of 250 historical SAP-related IBM projects, each lasting at least 10 weeks. SAP is an Enterprise Resource Planning application; SAP-related engagements represent a large portion of IBM’s consulting projects and are thought to contain fairly predictable types of tasks, making their analysis easier. The resources on each project were labeled according to their primary job role using an IBM-defined taxonomy of job roles. Sixty-six different job roles were represented in these 250 projects, i.e., each observation vector was of dimension $n = 66$. The dimension reduction component of our algorithm selected seven out of these sixty-six job roles for explicit modeling, plus an “other” category representing the aggregation of all remaining job roles. Based on the iterative clustering algorithm given in Table 1, we obtained a set

of six project clusters. Fig. 3 shows the average percentage of resource hours in each of the eight job role categories relative to the total resource hours for the 250 projects. For each cluster, the graph shows estimates for each state in the constructed models, where the number of states are chosen automatically as part of modeling/clustering.

Roughly speaking, Cluster 1 can be said to represent projects consisting initially of Package Solution Integration Consultants (PSIC, Dimension 35) configuring and deploying particular SAP modules. In the second phase of the project, additional resources having different job roles, including a Project Manager (PM, Dimension 41) to coordinate the different activities and resources, are brought in and the role of the PSIC is reduced. These other roles include Application Developer (Dimension 4), Application Architecture (Dimension 2), Business Development Executive (Dimension 6), as well as Other. These types of projects are typical of many SAP engagements conducted by GBS. In contrast, Cluster 2 represents more of an initial roadmap/blueprint type of project, in which a project manager or partner is engaged as the primary resource in the initial stage of the project, while resources with additional skills are brought in only in the second phase, and the relative involvement of the PM is reduced substantially. Similar interpretations can be made for the other clusters. In discussion of these results with GBS domain experts, the identified clusters were found to be reasonable and representative of typical SAP engagements.

To create an appropriate project taxonomy from the results, we examined the distribution of additional project attributes (beyond resource distributions) for projects in each cluster, as discussed in Section 4. 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, the way the project was priced (e.g. Time and Materials, Fixed Price), etc. Information on the nature of the different projects within a cluster was also gleaned from the recorded project name and description. For the projects studied, 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 to discuss the results with GBS domain experts, who subsequently were able to relate the observed staffing patterns to project type names typically used within GBS. For example, Cluster 2 projects consisting primarily of project management and application architect were determined to represent the design phase of a project, commonly called a Phase 1/Blueprint project.

We do not give an example here to show the creation of staffing templates from these results and subsequent generation of staffing requirements for a newly identified project. See [2] for a detailed example of how this was done for projects clustered at an aggregate level.

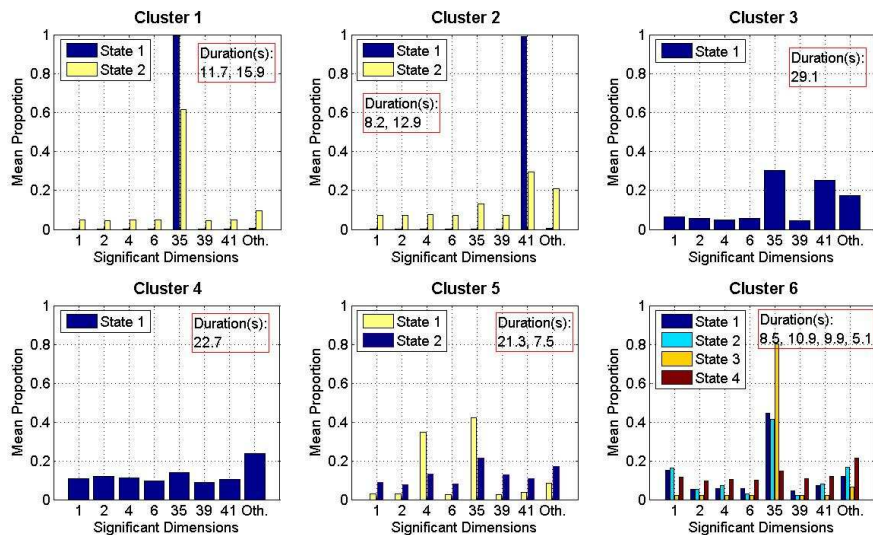


Fig. 3. The six clusters generated with the IBM GBS data. For each cluster, and for each state, the mean values of proportions (Y-axis) for the significant resources (X-axis) are shown, along with the mean state durations, calculated from parameter estimates.

6 Conclusions

A sequence clustering-based approach to building resource-based project taxonomies has been proposed, which handles noise, sparse high-dimensional observations, explicit state duration modeling, and lack of knowledge about the number of states and clusters. The clustering algorithm has been applied to IBM GBS project data, and the obtained clusters have been found to be reasonable representations of project types based on resource engagements, making the sequence clustering framework an attractive approach to taxonomy building.

The proposed sequence clustering approach is quite general, with potential applications in a wide variety of other problem domains, including video sequencing, gene expression clustering [10], etc. For example, video shots can be represented by allocations or proportions of a quantized color set. This way, video shots can be represented by sequences of proportional vector, making our Dirichlet-based sequence clustering algorithm appropriate. Moreover, data noise and sparsity typically occur in such sequences as well. We leave application of the technique to this and other applied problems for future work.

References

1. J. Hu, B. Ray, and L. Han, *An Interweaved HMM/DTW Approach to Robust Time Series Clustering*, Proc. Int. Conf. on Pattern Recognition, 145–148, 2006.

2. J. Hu, B. Ray, and M. Singh, *Statistical Methods for Automated Generation of Services Engagement Staffing Plans*, IBM Journal of Research and Development, 2007 (to appear).
3. S. E. Levinson, *Continuously Variable Duration Hidden Markov Models for Speech Analysis*, Proc. Int. Conf. Acoustics, Speech, Signal Processing, 1241–1244, 1986.
4. C. Li, G. Biswas, M. Dale, and P. Dale, *Matryoshka: A HMM based Temporal Data Clustering Methodology for Modeling System Dynamics*, Intelligent Data Analysis, 6(3):281–308, 2002.
5. R. Melik, L. Melik, A. Bitton, G. Gerdebes, and A. Israilian, *Professional Services Automation: Optimizing Project and Service Oriented Organizations*, J. Wiley and Sons, 2002.
6. T. Minka, *Estimating a Dirichlet Distribution*, Technical Report, M.I.T., 2000.
7. T. Oates, L. Firoiu, and P.R. Cohen, *Using Dynamic Time Warping to Bootstrap HMM-based Clustering of Time Series*, Sequence Learning, LNCS, 1828:35–52, 2000.
8. L. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, 77(2):257–285, 1989.
9. L. Rabiner and B.H. Juang, *Foundations of Speech Recognition*, Prentice Hall, 1993.
10. A. Schliep, A. Schonhuth, and C. Steinhoff, *Using Hidden Markov Models to Analyze Gene Expression Time Course Data*, Bioinformatics, 19:255–263, 2003.
11. P. Smyth, *Clustering Sequences with Hidden Markov Models*, Proc. Neural Information Processing Systems, 648–654, 1997.
12. S.Z. Yu and H. Kobayashi, *An Efficient Forward-Backward Algorithm for an Explicit-duration Hidden Markov Model*, IEEE Signal Processing Letters, 10(1):11–14, 2003.