Service analytics for IT service management

Y. Diao
E. Jan
Y. Li
D. Rosu
A. Sailer

Outsourcing enterprise IT service management is an increasingly challenging business. On one hand, service providers must deliver with respect to customer expectations of service quality and innovation. On the other hand, they must continuously seek competitive reductions in the costs of service delivery and management. These targets can be achieved with integration of innovative service management tools, automation, and advanced analytics. In this paper, we focus on service analytics, the subset of analytics problems and solutions concerning specific service delivery and management performance and cost optimization. The paper reviews various service analytics methods and technologies that have been developed and applied to enhance IT service management. We use our industrial experience to highlight the challenges faced in the development and adoption of service analytics, and we discuss open problems.

Introduction

In recent years, the information technology service management (ITSM) industry has faced continual competitive pressure to improve service quality, while simultaneously reducing service delivery and management costs. These objectives are particularly critical for IT service outsourcing, in which service providers manage the IT infrastructures, applications, and business processes on behalf of their customers. In IT service outsourcing, customers seek reduced costs, and accelerated time to market, along with external (i.e., outside of their company) expertise, assets, and/or intellectual property. Providers must efficiently address the scale and complexity of managed IT environment and the diversity of processes across all of their outsourcing customers. In current market conditions, the approach to achieving economy of scale through traditional workforce management and process standardization is less effective than it was in the past. Nowadays, IT service providers rely on innovative data analysis technologies to improve the efficiency of their service management processes. Valuable business insights, extracted from analysis of service management data, drive to higher IT service efficiencies and quality.

Service analytics identifies the subset of analytics problems and solutions concerning specific service delivery and management performance and cost optimization. Service analytics comprise a collection of data analysis, modeling, and optimization methods that aim to improve performance and competency of ITSM. **Figure 1** illustrates the high-level interactions between the service analytics architecture and the main ITSM processes (as per best industry practices Information Technology Infrastructure Library, or ITIL [1]), in the process of delivery to IT service customers. More specifically, ITSM comprises a set of management processes and technologies that enable service providers to manage the IT infrastructure and applications from the perspective of their customers and of their own business—following best practices, during the execution of ITSM processes. IT service providers track customer interactions and IT system performance, producing large volumes of data. This data is processed by service analytics to produce various types of insights for feeding back into the management of ITSM processes.

The objective for service analytics is to analyze these processes and related business data, and produce valuable insights related to service cost and quality. Analysis spans multiple dimensions of service delivery, such as service processes, workload types, customer domains, and delivery geographies. The analytic insights are consumed in many ways, including feedback for management of IT service processes, decision support for operating personnel, and real-time reports on service quality and

Digital Object Identifier: 10.1147/JRD.2016.2520620

0018-8646/16 © 2016 IBM

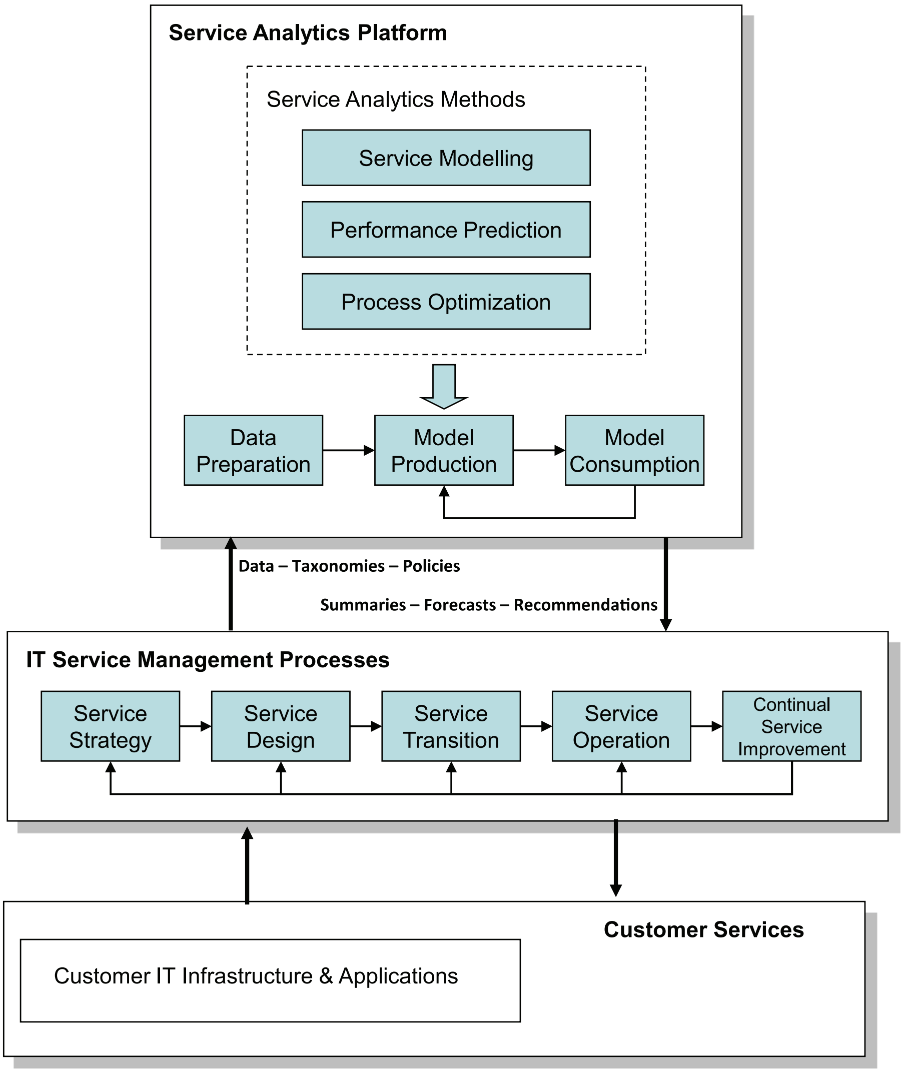
Service Analytics Platform
Service Analytics Methods
Service Modelling
Performance Prediction
Process Optimization
Data Preparation
Model Production
Model Consumption
Data – Taxonomies – Policies
Summaries – Forecasts – Recommendations
IT Service Management Processes
Service Strategy
Service Design
Service Transition
Service Operation
Continual Service Improvement
Customer Services
Customer IT Infrastructure & Applications

Figure 1

Service analytics for IT service management.

performance for IT service customers. Service analytics are materialized in a collection of tools, henceforth called the *service analytics platform*, which facilitates the extraction of analytic insights, from raw data collection and preparation through insight (i.e., model) production and consumption [2]. The service analytics platform is often designed as hybrid distributed systems with a cloud-based API (application programming interface) components platform [3–5]. The degree of integration across the analytics tools may vary widely. Similarly, the enterprise-level integration of content through taxonomies and representation standards varies widely as well. Automation and content integration are the key enablers of efficient real-time trustworthy business insight discovery.

A large set of service analytics methods have been studied and applied by both academic and industrial researchers to address specific problems encountered in ITSM. This paper is a review of service analytics research, taking a holistic perspective to classification of the contributions with respect to the target business problems and applied technical methods. Drawing on our industrial experience, we highlight domain-specific challenges that must be addressed by service analytics methods and platforms. We believe both of these contributions will help solidify the current service analytics research and set up a new ground to motivate future studies.

Previous overview studies in the area of ITSM are rather limited. Galup et al. [6, 7] focus on the architectural aspects, including ITSM process structure and standards. Galup et al. [6] also address the research methodology appropriate for services research. Analysis methodology, including the empirical variable approach of management information systems (MIS), and the process theory approach, is considered with respect to high-level business goals of quality management and process reengineering. In this work, we review analytic methods for ITSM processes with respect to specific low-level business problems of cost and efficiency. Ardagna et al. [8] review the analytics methods used for quality-of-service (QoS) management in cloud, classifying methods by the component that affects the QoS, i.e., workload and system modeling methods. While targeting a more limited scope of service management and different classification criteria, reference [8] complements our work, addressing specific cloud-management business problems.

Let us now discuss a *service analytics taxonomy*. The service analytics methods proposed in the literature can be categorized into three main categories:

1. *Service modeling*, used to understand features that characterize process execution as the basis for applying efficiency improvement actions.
2. *Performance prediction*, used to gain awareness on upcoming performance and workload events, and trigger proactive cost-efficient actions.
3. *Process optimization*, used to determine organization and execution models that maximize service management effectiveness.

These categories that span service analytics are applicable across all of the service management processes (see Figure 1). For instance, service modeling in *Service Operation* processes (see rectangle in Figure 1) relates to service requests, such as incidents and changes. Analytics help characterize operational aspects such as arrival patterns, handling procedures, service effort, human error, etc. These insights are used as input for process improvement actions such as personnel training [9, 10], service management tool tuning [11, 12], process automation [13], knowledge reuse [14–16], and many others. On the other hand, in *Service Design* (see rectangle in Figure 1),
service modeling relates to a high-level model of customer workload and infrastructure features that are further used to decide on the most efficient service-level agreements (SLAs) [17] and delivery organization models [18].

Performance prediction methods are often applied in service operations—for prevention of outages, SLA violations [19, 20] or for capacity planning. In *Service Transition*, prediction is used for decision support, guiding process-management decisions with a useful tradeoff between operational efficiency and performance risks [21].

Process optimization methods are often applied to processes that traditionally involve manual operations, including Service Operations processes, such as change request management [22] and workforce scheduling [23]. Typically, service analytics solutions integrate service modeling methods with prediction and optimization methods. For instance, the optimization of change scheduling employs information on the effort spent with the execution of various types of change operations, which is obtained with service modeling methods.

The remainder of this paper is organized as follows. First, we provide an overview of the main ITSM processes, which will introduce the reader to the processes themselves and to the opportunities within where analytics and optimization methods can be used. Next, we identify service analytics challenges that are specific to ITSM. Further, we discuss each of the categories of service analytics methods, providing examples of representative business problems and analytic tools. Also, we focus on the service analytics platform, discussing techniques and requirements specific to IT service provider domains. Finally, we conclude with a review of open problems.

Service management and analytics challenges

IT service management

ITSM, henceforth also called *IT services*, encompasses the processes involved in the management of IT for providing value to customers in the form of services. As we alluded to, the most adopted framework for ITSM is ITIL [1], which systematizes the planning, delivery and support of IT services around a five-stage service lifecycle. The lifecycle comprises Service Strategy, Service Design, Service Transition, Service Operation, and Continuous Service Improvement, as we have just discussed and as illustrated at the bottom layer in Figure 1. At each stage, the focus of service management changes, evolving from marketing and strategy, to delivery and operations, and closing the service lifecycle loop with continuous service improvement. Thus, the data associated with the service management at each stage changes as well, requiring specific analytics methods to handle stage specific data types and extract stage specific insights, as well as to integrate across stages, for global insights.

In *Service Strategy*, the focus is on transforming service management into a strategic asset by understanding which IT service offerings are of more interest to customers. This helps the service providers to have the most “fitted” service offerings to achieve customers’ required business outcomes. Business models and service value models are used in microeconomics to formalize and analyze the relationships between customers, providers, and services. The key Service Strategy processes relate to (i) *service portfolio management*, which ensures that the service provider has the right mix of services (i.e., service portfolio) to meet required business outcomes at an appropriate level of investment, (ii) *financial management*, which concerns the provider's own budget, accounting, and charging requirements, and (iii) *demand management* for understanding a customer's demands and provisioning to meet them. Service analytics for Service Strategy supports marketing decisions that have an impact on the design requirements and the go-to-market service catalog and prices [24]. Analytics for service modeling based on business and service request data provide the relevant features for the optimization of service portfolio and sales success [25]. Optimization techniques use strategy-specific cost/utility functions to represent the nonfunctional requirements and map the business requirements into IT requirements [14].

Service Design is focused on designing IT services to realize the strategy and ensure quality service delivery, customer satisfaction, and cost-effective service provision. Service Design processes include service catalogue management, service-level management, availability management, capacity management, IT service continuity management, information security management, and supplier management. Service modeling and process optimization service analytics methods are used in Service Design to develop more efficient organization models [17, 18] and related cost models; cost models must tie into the operational models in order for the designed service solutions to match in production the targeted service level agreements (SLAs) for service availability and support [26, 27].

The focus of *Service Transition* is to develop the capabilities to transit customer new and transformed services into operational, steady-state service delivery. Service Transition processes include change management, service asset and configuration management, knowledge management, transition planning and support, release and deployment management, service validation and testing, and evaluation. A wide range of service analytics methods are used in this stage in order to drive efficient execution and minimal risks. Sample problems addressed by analytics include optimization of project plan, risk prediction [21], workforce optimization, and optimal test coverage [28]. A newly emerged paradigm in the Service Transition area is the cloud transition modeling, which raises new types of problems for service analytics. For instance, performance prediction [29] and real-time anomaly detection [28] are paramount for minimization of go-to-market risks and service performance penalties.

Service Operation is focused on achieving effectiveness and efficiency in service delivery to ensure value for both customers and service providers. Key Service Operation processes include event management, incident management, problem management, request fulfillment, and access management. Specifically, the service requests (a.k.a., tickets) contain embedded knowledge about the service performance trends, workforce efficiency, capacity planning, and configuration compliance, to name a few. Service modeling methods are used to model request arrival [8], execution effort [9], errors users experience [30], process execution patterns [31, 32], and many other service operational features. These models are used to optimize performance in Service Operations and throughout the service lifecycle. Event and performance prediction is used to prevent outages [19, 33, 34], provision resources for workload peaks [35, 36], and dispatch incidents [37, 38]. Process optimization is used to schedule personnel based on workload fluctuations [23], bundle change execution [22], prioritize service request execution, allocate computational resources [36], and for many other purposes.

Finally, the focus of *Continuous Service Improvement* (see rectangle in Figure 1) is on creating and maintaining value for customers through better design and operation. Service analytics help us identify changes in service and request patterns—and tune service management tools and processes in order to attain the desired levels of productivity, and quality of service. Relevant service modeling methods include process behavior models for detection of deviations and defect similarity analysis for detection of repeated defects [33]. Prediction models are used to recommend infrastructure upgrades [19], automation solutions, knowledge transfers, or skill upgrades [10].

Although the above introduction to ITIL depicts the service management processes in a staged way, the implementation of these processes follows a networked model, where their roles intertwine and decisions made in earlier stages (e.g., design management and change management) have consequences in later stages (e.g., SLA management). Also, decision in one phase depends on service parameters that are managed and can only be collected or measurement in other phases. An integrated implementation of service analytics within the enterprise Service Analytics Platform is imperative for deployment of analytics solutions with high ROI (return on investment).

Challenges of service analytics

In addressing the diverse analytics opportunities arising in the context of ITIL processes, as described in the previous section, IT service provider organizations face challenges that stem from the specifics of large-scale IT service operations and have an impact on the ROI of service analytics technologies. At the core of these challenges is customer diversity, including service requirements, managed environments, geographic locations, and other service dimensions. This translates in diversity with respect to processes and tools. Another group of challenges derives from the SLA-driven business model, which drives to a fast-paced and risk-adverse work environment. The ROI of service analytic solutions is highly sensitive to managed environment—solutions developed for one customer or delivery environment might not work as well in other environments because of the differences in available content, analytic models, and model performance. In the following, we discuss the main challenges for development and adoption of service analytics solutions.

Diversity of processes

Most often, ITSM processes are implemented differently across service customers, geographies, and work cultures. While standardization is a well-known solution for effective IT services, it is often not an option or too expensive to implement as a large-scale transformation. Local or customer-specific implementation of IT management solutions lead to limitations for service analytics, such as lack of data from particular areas of the business, and differences in data semantics and data quality. This cascades into increased complexity of data preparation, limitations on choosing "universally" applied analytical methods, and lack of full consistency across the resulting models.

Diversity of tooling

Aside from process differences, service management tools used to perform a specific process component may be different with respect of the service record details that are available. For example, within the incident management process, multiple customer-specific tools might be used to capture incident records. Across tools, it is common to have different models to capture service details that are beyond what is required for basic incident tracking. Hence, details (such as the references to related servers or applications), or categorizations of incident root cause and resolution method, which can be highly relevant for service analytics, may be missing. Analytics that rely on these details would require more complex development effort to integrate across models. For instance, in order to extract the missing incident features, unstructured data analysis is used, which adds a high development overhead and quality risks as models must be tuned for each service domain.

Limited process integration

A typical approach across service lifecycle processes is to maintain per-stage ownership of service management tools. As a result, these tools are often developed and configured in a disconnected manner, i.e., without integration across processes. This makes it difficult to connect the insights about service data in related processes. Examples include the difficulty to correlate the performance in Service Operation processes and the performance in Transition and Transformation processes, or, even worse, the correlation of service requests and configuration management data within the Service Operation processes. Service analytics must include the development of dedicated tools for entity resolution in order to overcome this limitation. Another aspect of process integration relates to difference across geographical delivery center. Difference within same service lifecycle phase, like Service Operations, might occur because geography-level teams have autonomy in adoption and customization of service management tools and processes. Also, country-level privacy regulations limit the content availability for development of analytic solutions.

Limited expert availability

Typical IT service personnel operate in high volume and high risk situations, where risk draws from tightness of SLA constraints, high unpredictability of complex computing systems, and high costs of SLA failures [39, 40]. As a result, the service personnel subject matter experts (SMEs) have *limited time* to spend helping analysts with laborious labeling of observations for supervised learning methods [41]. Even more notable, the *quality of record details* acquired during the production processes is often low, as the personnel might rush to move to the next task. Service personnel might shorten the time spent navigating complex taxonomy hierarchies for picking the right category choice, or might choose avoiding provide details that are not mandatory, or include typos. For instance, a ticket's close time could be earlier than its open time, or a classification into service-specific taxonomies are missed or set to the default value of "other"—thus providing no insightful values for data analysis.

Risk-adverse attitude

The SLA-driven operation raises the propensity for risk-avoidance across service personnel and management staff. As a result, the adoption of service analytic solutions is highly dependent on their perceived impact on risk. The perceived impact depends on how analytics are used. The higher the risk, the higher the expectation of accuracy and the intuitive explanation for the analytics results. For instance, low risk is perceived for decision-support analytics in which service personnel are always available to override, or the cost of an error is small. Samples include recommendation for incident solution reuse [14] and request dispatching [38]. Other decision-support analytics may have higher costs of error, and thus higher risks, such as financial risk prediction [21] or server upgrade recommendations [19]. In risk-adverse processes, the analytic methods must provide high accuracy and provide an acceptable explanation, such as the identification of factors with largest impact on the analytic results [19]. Analytics can support autonomous execution of operations, such as image relocation [36].

Overall, these challenges are faced by most service analytic projects, and affect the ROI. Awareness of these challenges is critical for adopting the most appropriate analytics solutions and establishing an agile roadmap for enterprise-level integrated analytics platform.

Service analytics methods

Service modeling

The goal of service modeling is to model features that characterize customers and process behavior as the basis for understanding service requirements and delivery profiles. Service modeling methods are commonly applied in all phases of the ITIL service lifecycle and provide the enabling knowledge for efficiency improvement actions.

The most frequently occurring type of service modeling is related to workload, i.e., the request flow that a specific process or system component is observing. Workload characterization aims to model arrival patterns and service effort patterns that are related to various process features. In Service Operations, sample process feature include: (i) IT components such as servers, middleware, and business applications, (ii) service request types such as incident failure types, (iii) resolution methods such as the types of tools used, and (iv) process patterns, such as organization-level interactions. Workload characterization is fundamental to understanding how service-personnel effort is spent with respect to tasks and activities, and how customer experiences the service delivery. This enables an IT provider to solve various business problems contributing to cost savings, competitive analysis, and workforce and process optimization.

The following paragraphs provide a few examples of business problems in the scope of Incident Management processes that can be solved using service modeling methods. For example, one goal concerns the reduction of the occurrences of the most prevalent failure types. The approach to address this goal is for SMEs to identify the high-volume incident types, perform investigations, and eventually implement changes that will reduce the volume of those incident types. As incident classification into the enterprise-specific taxonomy is often not consistently capture in incident records, workload characterization methods help classify the requests by failure type. The classification model can be build using supervised data mining methods applied to structured and unstructured incident-record fields using sample content labeled by SMEs [38]. Alternative methods based on clustering can be applied when taxonomy is not known *a priori* [42, 43]. Identification of incident repeat occurrence and co-occurrence patterns [34] helps root cause analysis and more effective resolution.

A second goal is to reduce the risk of SLA violation. A sample approach to this problem is to ensure that each request is handled by the service personnel that are most efficient in solving the specific request type. To this end, workload characterization methods are used twofold. First, request types are classified into a specific taxonomy. Next, for each request types, efficiency profiles are created to model how efficient a service staff is at solving the request of the particular type. The request dispatching system will use the two pieces of information to map incoming requests to the service personnel likely to complete work the fastest [8]. Another instance of this problem is related to the cloud where image migration represents the main approach to controlling SLA violations in a likely overprovisioned infrastructure [36].

A third goal is to reduce the problem determination effort. An approach to this problem is to create a system for recommendation of solutions based on a characterization of request details. Feature profiling techniques and text analytics are used to characterize the historical content and create a solution catalog using request description and related system components [3, 30], [42]. A characterization of resolution type is used to trim the catalog by removing duplicate solutions. At request arrival time, the profiling technique is used to determine the type and retrieve relevant solution options.

A fourth goal is to improve the quality of change. An approach to this problem is to prevent change operations that are likely to lead to error. For instance, service personnel can analyze the error profile of specific change operations or installed software, and can avoid use alternate solutions if a specific operation has high likelihood of error. The error profile would allow answers to questions like “Which cloud management tools should I install such that we get the least amount of capacity events?” The error profiles can be created by integration across different types of service management content—such as change management, incident management, and configuration management. Entity resolution methods can be used to identify server of software/tools references in structured and unstructured change and incident text. Configuration records that related to servers can further provide features of interest such as installed middleware and applications as well as server purpose, manufacturer, and age [16]. These integrated records can be fed to search tools [44] or question answering [45] and accessed manually or automatically [20] upon change record creation.

Service modeling is performed by applying statistical and data mining methods to model the occurrence of target features within a selected scope (see [46, 47]). For a rich representation of the business scope, analysis is most often based on integration across several types of ITSM sources, such as customer management content, incident management, configuration management, etc.

The service analytics challenges discussed in the earlier section often result in missing or invalid references to service components, resources, or service features. As a result, extensive data cleaning [20, 34, 43, 48] and entity resolution [11, 19] are necessary to link across all data sources of interest. Such actions help us expand the set of records available for analysis, improve the accuracy of the service modeling, and provide knowledge beyond what is captured by the service management data models. For instance, in an sample scenario of operating system support services, extraction of server reference from the text of incident description and resolution increases the ability to match an incident with a server from ~65% to over 90% [19].

Another consequence of service analytics discussed above is dealing with missing information. A series of advanced data mining methods can be applied to create the missing information. One approach is to discover specific relationship across several structured attributes, and then use business rules to fix the incorrect or missing values [49]. Another approach is to identify the missing elements in text-based fields the record. Extraction can be done using dictionary-based identification, structural decomposition [16], semantic analysis [42], or topic discovery [43].

As illustrated by previous examples, text analysis is a highly valuable method, as service delivery best practices require that service records capture relevant details about the actions taken. A common approach to exploit the information captured by text fields is to expand the features set with elements extracted from or based on text fields [42], such as incident description and resolution. Natural language processing (NLP) tools are typically used. Challenges arise because typical text in IT service tickets is not as well-formed as newswire articles. The ticket-specific sub-language exhibits numerous ad-hoc abbreviations, limited details, grammar mistakes, and punctuation errors [50]. Text preprocessing and normalization are necessary for initial cleanup. In addition, compound word extraction and substitution, and alias identification and replacement, can give better accuracy, if SME input is available. There are multiple methods that can be used to extract text-based features. First, *n*-gram analysis can be extracted as collection of *n*-grams. Unigram decomposition is the most popular technique [51], but bigram and trigrams give best results for small and large corpora, respectively. Second, semantic token extraction can be used to extract keys and to link non-standard data templates based on domain specific rules [52]. Third, topic extraction can be used to extract new classification taxonomies. For simple domains, one can use rules and keywords provided by SMEs or extracted from the web. For more complex domains, machine learning methods—such as Latent Dirichlet Allocation (LDA) [53], Latent Semantic Analysis (LSA) [54], and Probabilistic Latent Semantic Analysis (PLSA) [55]—are fairly popular and can yield good results in handling the service data [56]. Alternatively, part-of-speech (POS) [57] analysis allows the identification of topics using specific POS patterns [42]. The challenge is that there are only limited SME resources for data annotation, so that methods must be specialized [16, 52]. Work-around approaches, although not desired, include the use of predefined rules for semi-automatic data labeling, followed by bootstrapping [58].

Once the set of features is cleansed and complete, descriptive data mining models can be created with clustering [10, 19, 30, 42, 43], summarization [31, 59], association discovery [60], classification [10, 12, 20, 38], and sequence discovery [34]. Most often, the goal of modeling is to label the records based on taxonomy of interest. When the taxonomy is unknown, clustering methods are used to discover the taxonomy and label the records accordingly. On the other hand, when taxonomy is known *a priori*, classification approaches are used for record labeling based on the learning from an input set of annotated records. When feature taxonomies are not of interest, yet records with a specific degree of similarity are sought, summarization, or text search methods can be used.

Summarization

Summarization methods are the most frequently used service analytics methods, First, summarization methods provide the data that service personnel use to monitor service performance and perform root-cause analysis [59], [61–64]. Specialized visualization, tailored to metrics of interest to service request management can further expand the tools that are available for process analysis. For instance, Cavalcante et al. [31] present a visualization method that illustrates the correlation between service request response times and SLA failures, which helps target process anomalies. Process Behavior Charts (PBC), which combine summarization with forecast, can be used for all of the processes and key performance indicators within the service delivery organization [33]. The PBC method has embedded rules for interpretation, based on the variation of running averages, providing consistent insights compared to leaving the user to perform the interpretation alone, as with other visualization tools. The method for organization of the data presented through visualization tools is highly relevant for service management, in general, and for cloud management, in particular, where the volume of data and number of metrics to be visualized is very high and where the user might not be always highly knowledgeable about the details of the managed system [65]. Moreover, summarization provides the information necessary to run process optimization methods [23], such as distributions of service request arrival times or resolution effort. Also, summarization is integrated in tools of automated service management [12], where configurable or adaptive monitoring profiles are used to trigger service management actions, such as file system clean-up or incident ticket generation.

Classification

Popular statistical classifiers such as Supporting Vector Machine (SVM) [66], Maximum Entropy Model (MaxEn) [67], Naive Bayesian with Maximum Likelihood, Discriminative Training (DT) [68], and *k*-Nearest Neighbor are commonly used for classification of ITSM records. Furthermore, feature selection methods are used to improve data classification accuracy, such as the Random Forest [69] (see [19, 32]), SVM (see [12]), or MaxEn (see [67]). Classification methods that integrate with on-line learning, such as the Perceptron learning algorithm used in [52], help overcome the specific IT service delivery challenges of limited SME availability and customer diversity. Similarly, benefits are brought by efficient tools for specification of classification rules by the service personnel, such as [41], or by methods based on weak supervision, such as proposed in [70].

Clustering

Some of the most widely used clustering methods are *k*-means, and hierarchical clustering (see [47] for details). The L2 Euclidean distance or Term Frequency Inverse Document Frequency (TF-IDF) weighted cosine distance (for text *n*-gram collections) [43] are the most common dissimilarity functions. Dissimilarity metrics specifically targeted to the service domain give better results. For instance, Li and Katircioglu [9] measure the dissimilarity of ticket clusters using a combination of the number of shared resources solving these tickets and the scale of sharing. An *n*-gram clustering technique is illustrated by Mani et al. [43], who apply hierarchical merging to the initial set of clusters defined by the principal components of the TF-IDF term matrix. Another techniques for text clustering is fuzzy match, using the edit-distance as a dissimilarity metric [i.e., how many changes must be performed to change one string (or set of tokens) into another], which allows the clustering of tickets despite typos.

Search

Information Retrieval (IR) methods can be used to determine the top-*K* most similar records given a target record. While TF-IDF [71] is the most prevalent similarity metrics, other metrics are used to better fit the domain. For instance, in [51] the search similarity metric is the correlation coefficient of topics structures—where the topics are extracted for each pair of documents using LDA or other equivalent method, and the correlation coefficient is computed considering an array with the common set of features.

Performance prediction

Performance prediction methods are aimed to help the service provider take early action towards prevention of failures or more efficient execution. Most frequent targets for prediction relate to process performance, service effort, and service workload. Such insights can be used to assist actions such as resource planning or staffing recommendation, process management, infrastructure configuration, and many others.

For example, in Application Management Service (AMS), it is estimated that organizations can spend as much as 80% of their application budgets on maintenance related activities [72]. AMS providers are interested in reducing the maintenance costs, and analysis of tickets trends helps forecast the type of skills that will be needed for the plan period as well as forecast the effectiveness of skill development actions and other process improvement actions [10].

Another application assessment is related to workload management in cloud environments. Workload forecasting [73] allows for automated migration of virtual machines, in order to minimize the resource utilization hot-spots [36, 73], or provisioning of additional virtual machines for services with increasing demand [35].

Other applications of prediction include prediction of delivery performance given customization of the SLA model [17], prediction of project performance based on a large set of qualitative factors collected at project start time [21], and prediction of ticket reduction upon system upgrades [19].

Overall, performance prediction is based on the process-specific records such as incident tickets, change request records, and monitoring events, which illustrate how service processes are managed as well as how organizations utilize their IT and human resources. In the following paragraphs, we discuss three targets for prediction in IT services.

Performance control

Measuring and monitoring process KPIs (key performance indicators) is the basis of any cost and quality management system. In IT services, each service management process has specific KPIs. For example, in Service Operation processes, KPIs typically quantify the volume, effort, and quality regarding incident resolution.

Other common KPIs include ticket volume per server, number of failed changes, mean time to repair, and backlog size. Statistical Process Control (SPC) is a widely adopted method for KPI tracking in services performance [74]. Its tools, such as cumulative sum charts (CUSUM) and c-chart, help tracking KPI evolution over time, detecting trends and anomalies, and enabling the business to take proactive actions and drive early root-cause analysis and remediation [75].

Effort estimation

Service effort estimation is the key element for assessing labor cost and driving effort-reduction programs. Per-activity effort indicates the actual amount of time that a system administrator spends on a service activity, such as incident ticket resolution, request auditing, and proactive maintenance. Nevertheless, the nature of IT processes makes it difficult to measure the service effort with high accuracy, because of the lack of standard effort-recording tools, the manual process needed for effort recording, and the common multi-tasking behavior due to the managed resource readiness or the required process transitions [39]. There indeed have been some efforts to address this issue such as the ticket priority-based effort estimation approach in [9] for modeling resources' multitasking behavior, and the time-volume capture tool developed in [76] along with time-motion studies.

Service forecasting

Service volume forecasting uses statistical regression models to predict the volume expected for certain types of service activities. Analytical methods combine volume forecasting with effort estimation to address business problems pertaining to staff planning, SLA feasibility assessment, return of investment (ROI) of automation projects, and contract negotiation. In addition, prediction of hardware maintenance costs or prediction of troubled customer engagements also offers significant value to IT service providers. The state-of-the-art methods in time series forecasting can usually be applied in the context of ITSM. Examples include autoregressive integrated moving average (ARIMA), neural nets, and autoregressive conditional heteroskedasticity (ARCH) models [77]. Other methods include stochastic processes with time-varying intensities such as non-homogeneous Poisson process and Poisson-Gamma processes for load forecasting of a telephone call center (see [78]) and for load simulation of an IT support organization (see [79]).

Process optimization

By analyzing various service data created and collected during service process execution, process optimization is aimed to improve IT management effectiveness through resource planning and staffing recommendation. Process optimization can be applied throughout the ITSM lifecycle. For example, in Service Strategy, optimizing the Demand Management process helps to meet the customer workload demands taking into consideration both SLA violation cost and delivery labor cost. Other process optimization scenarios include optimizing change request scheduling in the Change Management process to meet complex constraints and dispatching incident tickets in the Incident Management process to meet the desired SLA attainment level.

Compared to resource optimization problems commonly encountered in computing system performance management, lack of measurement data and limited data accuracy are two common challenges in service process optimization. Our experience shows that these limitations stem from the impact that the overhead of collecting actual work-times (albeit small, 1 min/record [76]) has on the productivity of service personnel; data collections are limited to 1 to 2 months per year, and errors often occur to mapping interval of times to the correct activity. To deal with the lack of data, one can use estimation from existing data using service modeling methods. When the amount of data necessary for modeling cannot be collected, process changes are necessary to acquire it. For example, even arduous, time-motion studies may need to be introduced and enforced for the system administrators to record where and how they are spending their time working on various service activities [76]. For data accuracy limitations, the two common solutions are: (i) to conduct rigorous data validation to identify problematic data and to remove statistical outliers, and (ii) to introduce feedback loops or feedback controllers to self-correct the modeling inaccuracy.

Generally, process optimization methods can be classified into two categories: workforce optimization and workload optimization, which are two of the most important factors in IT management processes.

With respect to workforce optimization, we note that managing service personnel labor cost is essential in ITSM. However, it is often difficult to decide the right staffing level due to the complex relationships among dynamic customer workload, strict service level constraints, and service personnel with diverse skill sets. For modeling simple service activities such as answering and making calls in call centers, analytical models or simple simulation models can work well [80, 81]. However, for more complex service operations, such as platform support and storage management, complicated discrete event simulation is typically required to model the service delivery environment with a relevant level of details [37, 79]. Once a simulation model is available, a simulation-optimization approach can be applied to minimize the total staffing related variable cost while considering the contractual service level constraints, the skills required to respond to different types of service requests, and the shift schedules that the service agents must follow [23, 82].

Another challenging workforce optimization problem relates to the best composition of service delivery teams. To accomplish this, a clustering-based approach can be used where service requests that require similar problem solving skills are first grouped into a single cluster using a statistical clustering technique [46]. Subsequently, associations are built between service request clusters and service agents with respective skills and confidence levels [10]. The insights can be extended to optimize skill management, encouraging service personnel to train in skills with limited coverage.

In the same scope area of skill management, Agarwal [83] presents a closed-formula solution for predicting the best workload assignment when service personnel can bid for it, based on their skills and preferences. These two approaches to skill development, one driven by the organization and the other by the individual wishing to gain more bids, complement each other towards the creation of highly effective resources.

With respect to workload optimization, we note that a large amount of service requests are being handled in service delivery centers on a daily basis, and service delivery teams must handle them in an order that optimizes across costs and SLAs. There are two common type of service workload: (i) requests that have dependencies among them (e.g., change requests) and (ii) requests that are independent but have a target finish time (e.g., incident tickets).

Workload optimization in the Change Management process relates to change request scheduling. All change requests need to be allocated to available change windows taking into account various constraints such as the temporal sequence of changes, change window times, and service level penalty for missing the change window or passing the deadline. Optimizing change request scheduling typically requires the use of an optimization model that takes into account both constraints and costs. While heuristics based optimization approaches are more commonly used (e.g., see [22]), the optimization model is formulated in a way that can be solved using standard mathematical programming techniques (i.e., mixed integer programming). This approach not only results in strictly optimal solutions, but provides a scalable tool for scheduling a large set of change requests with complex constraints and also performing sensitivity analysis.

In the Incident Management process, workload optimization is related to the SLAs on incident handling that are established at the time of contract negotiation as part of the service-level management process. Although many types of service level agreements exist, the most common type takes on the form of a tuple: (percentage attainment, scope, time frame, target time). For example, 95% (percentage attainment) of all severity-1 incident tickets (scope) that are opened over each one month period (time frame) must be resolved within 3 hours (target time). One will typically find a large number of service level agreements associated with each customer contract. The objective of optimizing incident management is to find the most cost-effective way to meet all relevant service level agreements while restoring normal service operations. In the case where data accuracy is a concern, a closed-loop performance management solution makes use of feedback controllers to dynamically adjust the priority of the incident tickets based on both statically defined contractual service attainment targets and dynamically measured (with measurement inaccuracy) service attainment levels [84].

Service analytics platform

Previous sections have identified multiple types of service analytics methods that can help IT service providers improve and maintain high levels of productivity and quality. In this section, we focus on how service analytics are implemented and delivered. We review a typical architecture and requirements that are drawn from the nature of IT services.

The service analytics platform comprises the collection of IT components—e.g., hardware, software, APIs, graphical user interface (GUI), and business artifacts (e.g., taxonomies, rules, processes—that are integrated for delivery of service analytics. The interactions of a service analytics platform with the IT delivery processes are illustrated in Figure 1 along with the main categories of components [2]:

- *Data preparation components*, which integrate with IT infrastructure and applications for collection, cleanse, and imputation of data.
- *Model production components*, which use service analytics methods to estimate and validate models for analysis and optimization the ITSM processes. Analytics methods typically draw from the categories analyzed in previous sections, including service modeling, performance prediction, and process optimization. These methods use the prepared data along with service specific artifacts, such as taxonomies and policies.
- *Model consumption components*, which use the analytics models in production environment in order to support service personnel or management to improve the quality of model production. Model consumption includes activities of model scoring for the generation of business recommendations, generation of business artifacts such as taxonomies or configuration policies, and generation of business metrics summaries and forecasts. A feedback loop, from model consumption to model production, is

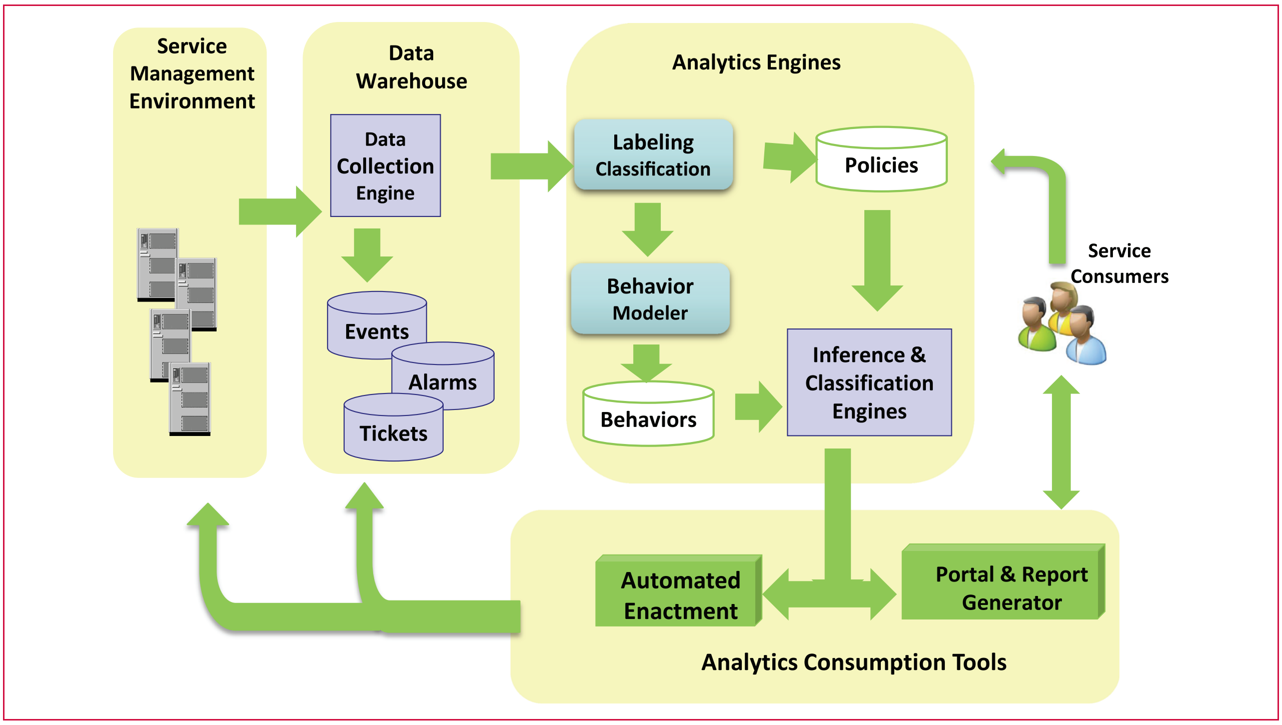

Figure 2

Illustrative service analytics platform architecture.

often necessary for model tuning and continual improvement [11].

Figure 2 illustrates the architecture of a typical service analytics platform (for other examples, refer to [11, 59]). The platform includes data preparation components for collection of service management environment data into databases or warehouses. The most common business artifacts considered for data collection include events, alarms, and tickets. Data collection may also happen through service management tools and infrastructures for running surveys and crowd-sourcing events [85]. The model production components are represented as the Analytics Engines (upper right in Figure 1), comprising a collections of analytics tools such as the behavior modeler to identify the service workload trend (e.g., using performance control and prediction methods) or to characterize the interactive relationship among multiple service components (e.g., process the results of process modeling). In addition, the labeling classification tool generates rules/policies to assist the service modeling methods, such as classification and clustering. The identified behavior models and policies feed into the inference and optimization engine, which produce recommendations, as model consumption artifacts. The model consumption artifacts are delivered to end-users through existing service management tools (e.g., incident management tool), dedicated portals [3, 59], or reporting engines [61, 62]. Also, artifacts can be delivered to automated enactment tools that adjust the service management environment (e.g., new service patches) or data collection engines (e.g., threshold for alert generation). An enterprise that exploits a wide range of service analytics typically comprises multiple instances of the pipeline illustrated in Figure 2, such as effort analysis [59], outage prediction [34], monitoring event optimization [11], recurring defect detection [59], and support for defect resolution [3].

Drawing from our experience with developing and deploying service analytics tools, we identify several requirements for the analytics-platform architecture including: data consistency, awareness of confidence levels, and efficient user interactions. All these contribute to increased usability, better adoption, and overall better ROI for service analytics projects. We discuss these requirements next.

With respect to data consistency, we note that across the multiple service analytics tools within the enterprise, and given the challenging diversity of customer environments and processes, data consistency is an imperative. Users expect that data that represents the same service features to be acquired through equivalent data collection and preparation procedures. Difference in processing may lead to different statistics and, possibly affect the business insights. Similar requirements apply to business metadata, such as taxonomies and business catalogs, which are used in processing the data. In order to address these requirements, model unification and standardization processing must be performed at the service-management data warehouse. Standardization relates to data cleaning and preparation, such as imputation of missing data, correction of input errors per specific taxonomies, entity resolution, and linking across content domains (e.g., tickets and servers) [86].

With respect to a requirement for awareness of confidence levels, note that previous research shows that service personnel performing server, application, and database administration have particularly high expectations for accuracy and trustworthiness, because of the high-risk nature of their work [39, 40]. This risk adversity is a challenge for service analytics that we identified in earlier sections. While a certain degree of error is intrinsic to data mining and process modeling, the end-users must be aware of the confidence level. Moreover, the implication of these errors must be carefully considered at time of model production and process design. For instance, if outage prediction has high likelihood of false positives, the daily workload will increase due to the added overhead of checking the system condition for false positive outage predictions. Similarly, confidence levels must be obvious or understandable for decision-support analytics, given that risk of the business decision compounds on the confidence of all related business insights. In other applications of service analytics, the impact of inaccuracy may be less critical. For instance, incident classification for workload dispatching may be erroneous at times without a critical impact unless the resolution deadline is very tight. As a result, throughout the analytics pipeline, from data preparation to model consumption, tools must provide visibility into model confidence levels in order to provide the users with relevant risk assessments. However, state of the art service analytics solutions are limited to providing accuracy estimations for the individual analytics results, such as classification and prediction, e.g., [21]. Methods and models for composition of accuracy estimations into overall risk assessments must be created, probably building on existing provenance and probabilistic database research.

With respect to a requirement for *efficient end-user interactions*, we note that efficient delivery of business insights is critical for service analytics adoption. A challenge discussed in previous sections, namely the SLA-driven service model, requires highly efficient tool interactions. Analytics tools for SLA-driven processes must add a minimal overhead to the existing work procedures in terms of wait times and tool interactions. However, for service analytics that are not constrained by SLA deadlines (such as analytics related to service design or change scheduling), given their response time is not immediate, the analytics consumption model must use asynchronous interactions in order to avoid slowing the end-user's overall activity.

With respect to agile and low cost development, service analytics solutions need to attain low costs for development and lifecycle management. The nature of IT service workloads, with high volumes and demanding SLA deadlines, drives the requirement for efficient human interactions, especially in the data collection and model consumption stages. For instance, analytical methods must be selected with the understanding that the experts (SMEs) are not readily available to annotate large volume of observations or revise extensive taxonomies. Interactions for collecting data input and insight delivery need to have low overhead and to possibly be asynchronous.

Agile and effective analytics platforms have been the focus of extensive research and development efforts in both industry and academia. In the recent past, platforms such as IBM Smarter Computing and Oracle Business Intelligence Foundations offered new hardware, software, and services technologies that facilitate the development and deployment of highly scalable service analytics platforms.

The emerging Cloud services platforms enable agile development of specific types of service analytics. For instance, the Google** Analytics API can be used for workload summarization analytics for management of web applications. The Google Cloud Platform Prediction API [5] allows the development of service analytics for performance prediction. Microsoft Azure** [3] provides built-in services for performance prediction analytics, but also a Cloud-integrated environment for development of custom big data analytics. IBM Bluemix* [4] provides similar services and, in addition, some machine translation services [45] that facilitate data preparation for service providers with a wide geographical footprint. Splunk** [76] is an integrated platform for operational analytics, covering most of the functions of a service analytics platform, including data collection, search and analysis, and exploration. Integration with Hadoop** provides for a wide range of analytics, from summarization to custom prediction methods. In addition, a wide variety of visualization tools, such as Google Charts, and Domo** [64] can be provided for agile customization of presentation tools.

A service provider can use cloud-based APIs when its service management tools are running in a private cloud or it has scalable and secure methods for data transfer into the cloud. Enterprise versions, such as offered by Splunk [76], allow for agile development of decision support tools. More extensive development efforts are required to integrate analytics into service management automation, or find optimal solutions using provider-specific simulation models. Hybrid solutions for the service analytics platform are required, integrating cloud APIs into in-house frameworks of ITSM process models and solution and content governance.

Flexible execution of analytics tools in dynamic computing environments is supported by standards and developed for specification of method components: features, models, and scores. For example, Predictive Model Markup Language (PMML) is addressing these elements with an XML (Extensible Markup Language)-based approach [87], and some of the cloud-based platforms, like Google Prediction, support it.

Overall, there are multiple alternative technologies and tools for development of a service analytics platform, ranging from the most agile (e.g., the cloud-based APIs), to the most comprehensive (e.g., enterprise-level data analytics frameworks). Several appropriate architecture choices must be made in order to satisfy a minimal set of usability requirements specific to service analytics.

Conclusion

The IT services industry faces continual pressure to improve the quality of its services while simultaneously reducing the service management cost. The scale, complexity, and diversity of the managed environment and processes make it difficult to effectively address these conflicting requirements. Service analytics represents an important research and application area, where a variety of data analysis, modeling, and optimization methods can improve ITSM performance and competency. This paper provided a review of service analytics research from the perspective of target business problems. The categories of service analytics include service modeling, performance prediction, and process optimization. In each of the categories, we reviewed relevant business questions and analytic methods that are used to answer them. Throughout the paper, we drew from our industrial experience, to highlight the challenges related to development and deployment of service analytics solutions. Indeed, some of the aforementioned service analytics research has resulted in significant value for the IT services business. For example, a modeling solution coming from the workforce optimization research has been deployed globally in IBM's service delivery centers covering 15,000 service personnel in more than 500 delivery teams.

While a substantial amount of service analytics research and applications exist today, there are several open problems that wait for even better solutions. One problem or challenge concerns the expansion of the service analytics coverage across the service lifecycle processes. The focus of service analytics is uneven across the service lifecycle. Service Operations and Service Design (Figure 1) (i.e., Service Steady State and Sales Engagement) have large coverage with analytics. We suggest that this is because these processes have potential for highly visible economic impact, and the related analytics benefit from the largest set of integrated service management tools and data sources, which provide a rich base for content for analytics. Other areas, such as Service Transition, are less represented, because many business problems require integrated data across the entire service lifecycle. For instance, to assess the risk associated with a specific service design solution requires integration across the sales (solution design), transition, and operations. This requires an integrated enterprise-level content model which is often missing in older service organizations and very expensive to deploy. The challenge is to *develop efficient and effective methods for content integration* across the various lifecycles of service management.

A second problem or challenge concerns the pervasive awareness of analytics confidence levels. As discussed previously, awareness of confidence levels of analytic insights is a requirement, intrinsic to the business model of IT service delivery. There is an open question on how to achieve this in a way that is flexible to match the variety of data sources, analytics methods, and service management processes that integrated them. Also, we must consider this agility and open standards, as Cloud-based analytics APIs are building momentum. Further relating to confidence awareness, the field evaluation of service analytic solutions must be integrated in the service analytics platform. Current approach for evaluation is for the data analyst to use the most appropriate data mining or statistical methods applied to a given training set. However, the performance of analytic solutions on actual content and operating conditions is not systematically collected and exploited. We suggest that in a large-scale, content-diverse ITSM domain, the service analytic solutions should be monitored for performance in a similar approach as we monitor the managed resources. The information will support realistic assessment of ROI, and provide valuable feedback for retraining the models or revising the methods.

A second challenge concerns automated identification of relevant analytic insights. The volume and diversity of data produced by service management infrastructures is continuously growing. Data analysts, either formally trained or ad hoc users, need to efficiently screen the large-volume and diverse content—and identify interesting cases for deeper investigation, rare cases, or new developments. Automated tools are necessary to guide their comprehensive investigation of the data. It is an open question on how to model a highly customizable, role-based insight profile, and how to scale the automated search for insights to the volume of operational data from large-scale IT service provider.

A fourth challenge involves scalable text analytics tools. The application of NLP to text descriptors in IT service records is difficult to scale, since each customer or operation domain uses its own abbreviations, notations, and protocols. This further complicates the development of a common cross customer/domain data set for performance evaluation. A set of industry specific tools and artifacts will enable efficient and repeatable extraction of entities and relationships—for instance, process-specific taxonomies and related pattern matching rules, dictionaries of terms and synonyms for software and hardware products, or process semantic representation. A highly relevant text analytics problem is anonymization for personal privacy reasons. As previously discussed, a challenge in ITSM organizations that span multiple countries, access to personal content is limited due to country-level regulations. Very often, this limits the access of analytics tools to service management records. Anonymization techniques are used to remove sensitive content, yet course-granularity patterns also discard elements that are relevant for analysis. Methods for efficient anonymization should be developed, customizable for maximum performance across a large variety of service management records.

We expect many of these open problems to be better resolved soon by development of novel artifacts for service process representation. Overall, the application and adaptation of NLP and machine learning methods to IT service analytics will continue bringing greater benefits to the management of IT services.

Acknowledgments

We would like to express our gratitude to Yu Deng, Vijay Naik, and Ronnie Sarkar, all employed by IBM, for many helpful and constructive discussions that helped us improve the quality of this paper.

*Trademark, service mark, or registered trademark of International Business Machines Corporation in the United States, other countries, or both.

**Trademark, service mark, or registered trademark of Google, Inc., Microsoft Corporation, Splunk Corporation, Apache Software Foundation, or Domo, Inc., in the United States, other countries, or both.

References

1. “IT infrastructure library. ITIL service support, version 3,” Office Government Commerce, London, U.K., 2007.
2. R. Grossman, “What is analytic infrastructure and why should you care?” *ACM SIGKDD Explorations*, vol. 11, pp. 5–9, 2009.
3. “Microsoft azure machine learning frequently asked questions,” Microsoft Corp., Redmond, WA, USA. [Online]. Available: http://azure.microsoft.com/en-us/documentation/articles/machine-learning-faq/
4. “Bluemix overview,” IBM Corp., New York, NY, USA. [Online]. Available: https://www.ng.bluemix.net/docs/#overview/overview.html#overview
5. “Google cloud platform. Prediction API,” Google, Mountain View, CA, USA. [Online]. Available: https://cloud.google.com/prediction/
6. S. Galup, J. J. Quan, R. Dattero, and S. Conger, “Information technology service management: An emerging area for academic research and pedagogical development,” in *Proc. ACM SIGMIS CPR*, 2007, pp. 46–52.
7. S. Galup, R. Dattero, J. J. Quan, and S. Conger, “An overview of IT service management,” *Commun. ACM*, vol. 52, no. 5, pp. 124–127, May 2009.
8. D. Ardagna, G. Casale, M. Ciavotta, J. F. Pérez, and W. Wang, “Quality-of-service in cloud computing: Modeling techniques and their applications,” *J. Internet Serv. Appl.*, vol. 5, no. 1, pp. 417–423, Jan. 2014.
9. Y. Li and K. Katircioglu, “Measuring and applying service request effort data in application management services,” in *Proc. IEEE Int. Conf. Serv. Comput.*, Jun. 2013, pp. 352–359.
10. Y. Li and K. Katircioglu, “Improving application management services through optimal clustering of service requests,” in *Proc. SRII Global Conf.*, 2012, pp. 885–894.
11. S. Kim, W. Cheng, S. Guo, L. Luan, D. Rosu, and A. Bose, “Polygraph: System for dynamic reduction of false alerts in large-scale IT service delivery environments,” in *Proc. Usenix*, 2011, pp. 24–28.
12. L. Tang, T. Li, L. Shwartz, F. Pinel, and G. Grabarnik, “An integrated framework for optimizing automatic monitoring system in large IT infrastructures,” in *Proc. KDD*, 2013, pp. 1249–1257.
13. R. Gupta, H. Karanam, L. Luan, D. Rosu, and C. Ward, “Multi-dimensional knowledge integration for efficient incident management in a services cloud,” in *Proc. IEEE SCC*, 2009, pp. 57–64.
14. R. Knapper, C. M. Flath, B. Blau, A. Sailer, and C. Weinhardt, “A multi-attribute service portfolio design problem,” in *Proc. Serv. Oriented Comput. Appl.*, 2011, pp. 1–7.
15. H. Li and M. Vukovic, “Intelligent solution discovery for self-service systems,” in *Proc. IEEE Int. Conf. Serv. Comput.*, 2009.
16. X. Wei, A. Sailer, and R. Mahindru, “Enhanced maintenance services with automatic structuring of it problem ticket data,” in *Proc. IEEE Int. Conf. Serv. Comput.*, 2008, pp. 621–624.
17. Y. Diao, L. Lam, L. Shwartz, and D. Northcutt, “Predicting service delivery cost for non-standard service level agreements,” in *Proc. IEEE/IFIP NOMS*, pp. 1–9.
18. A. Agarwal, R. Sindhgatta, and G. Dasgupta, “Does one-size-fit-all suffice for service delivery clients?” in *Proc. ICSOC*, 2013, pp. 177–191.
19. J. Bogojeska, D. Lanyi, I. Giurgiu, G. Stark, and D. Wiesmann, “Classifying server behavior and predicting impact of modernization actions,” in *Proc. CNSM*, 2013, pp. 59–66.
20. S. Güven, C. M. Barbu, D. Husemann, and D. Wiesmann, “Change risk expert: Leveraging advanced classification and risk management techniques for systematic change failure reduction,” in *Proc. IEEE/IFIP NOMS*, 2012, pp. 795–809.
21. S. Guven, M. Steiner, T. Ide, S. Makogon, and A. Venegas, “Mining for gold: How to predict service contract performance with optimal accuracy based on ordinal risk assessment data,” in *Proc. IEEE SCC*, 2014, pp. 315–322.
22. L. Zia, Y. Diao, D. Rosu, C. Ward, and K. Bhattacharya, “Optimizing change request scheduling in IT service management,” in *Proc. IEEE Int. Conf. Serv. Comput.*, 2008, pp. 41–48.
23. Y. Diao and A. Heching, “Staffing optimization in complex service delivery systems,” in *Proc. Int. Conf. Netw. Serv. Manage.*, 2011, pp. 1–9.
24. Y. Zhou, L. Liu, C. Perng, A. Sailer, I. Silva-Lepe, and Z. Su, “Ranking services by service network structure and service attributes,” in *Proc. IEEE 20th ICWS*, 2013, pp. 26–33.

25. R. Knapper, B. Blau, T. Conte, A. Sailer, A. Kochut, and A. Mohindra, “Efficient contracting in cloud service markets with asymmetric information—A screening approach,” in *Proc. IEEE 13th CEC*, 2011, pp. 236–243.
26. A. Sailer, M. R. Head, A. Kochut, and H. Shaikh, “Graph-based cloud service placement,” in *Proc. IEEE Int. Conf. SCC*, 2010, pp. 89–96.
27. A. Lall, A. Sailer, and M. Brodie, “A graphical approach to providing infrastructure recommendations for IT,” in *Proc. IEEE Int. Conf. SCC*, 2008, pp. 161–168.
28. M. R. Head, A. Sailer, H. Shaikh, and D. G. Shea, “Towards self-assisted troubleshooting for the deployment of private clouds,” in *Proc. IEEE CLOUD*, 2010, pp. 156–163.
29. B. Tak, C. Tang, H. Huang, and L. Wang, “PseudoApp: Performance prediction for application migration to cloud,” in *Proc. IFIP/IEEE Int. Symp. IM*, 2013, pp. 303–310.
30. Y. Song, A. Sailer, and H. Shaikh, “Hierarchical online problem classification for IT support services,” in *Proc. IEEE TSC*, 2011, pp. 345–357.
31. V. F. Cavalcante, C. S. Pinhanez, R. A. de Paula, C. S. Andrade, C. R. B. de Souza, and A. P. Appel, “Data-driven analytical tools for characterization of productivity and service quality issues in IT service factories,” *J. Serv. Res.*, vol. 16, pp. 295–310, 2013.
32. I. Giurgiu, J. Bogojeska, S. Nikolaiev, G. Stark, and D. Wiesmann, “Analysis of labor efforts and their impact factors to solve server incidents in datacenters,” in *Proc. Int. Symp. Cluster, Cloud Grid Comput.*, 2014, pp. 424–433.
33. G. Stark, “*Three Killer Graphics*,” Research Gate, 2014. [Online]. Available: http://www.researchgate.net/publication/259763980_Three_Killer_Graphics_v2
34. R. Liu and J. Lee, “IT incident management by analyzing incident relations,” in *Proc. Int. Conf. Serv. Oriented Comput.*, 2012, vol. 7636, pp. 631–638.
35. W. T. Tsai, P. Zhong, and J. Balasooriya, “Services utility prediction on a cloud,” in *Proc. IEEE Int. Conf. SOCA*, 2011, pp. 1–4.
36. S. Daniel and M. Kwon, “Prediction-based virtual instance migration for balanced workload in the cloud datacenters,” Dept. Comput. Sci. (GCCIS), Rochester Inst. Technol., Rochester, NY, USA, Tech. Rep., 2011.
37. H. S. Gupta and B. Sengupta, “Scheduling service tickets in shared delivery,” in *Proc. Int. Conf. Serv. Oriented Comput.*, 2012, pp. 79–95.
38. D. Loewenstern and Y. Diao, “A dynamic request dispatching system for IT service management,” in *Proc. IEEE CNSM*, 2012, pp. 271–275.
39. E. Haber and J. Bailey, “Design guidelines for system administration tools developed through ethnographic field studies,” in *Proc. ACM Symp. Comput. Hum. Interaction Manage. Inf. Technol.*, 2007.
40. N. F. Velasquez and A. Durcikova, “SysAdmins and the need for verification information,” in *Proc. ACM Symp. Comput. Hum. Interaction Manage. Inf. Technol.*, 2008, pp. 4:1–4:8.
41. Y. Diao, H. Jamjoom, and D. Loewenstern, “Rule-based problem classification in IT service management,” in *Proc. IEEE Int. Conf. Cloud Comput.*, 2009, pp. 221–228.
42. R. Potharaju, N. Jain, and C. Nita-Rotaru, “Juggling the Jigsaw: Towards automated problem inference from network trouble tickets,” in *Proc. NSDI*, 2013, pp. 127–141.
43. S. Mani, K. Sankaranarayanan, V. S. Sinha, and P. Devandu, “Panning requirement nuggets in stream of software maintenance tickets,” in *Proc. ACM SIGSOFT Int. Symp. FSE*, 2014, pp. 678–688.
44. J. Lenchner, D. Rosu, N. F. Velasquez, S. Guo, K. Christiance, D. DeFelice, P. M. Deshpande, K. Kummamuru, N. Kraus, L. Z. Luan, D. Majumdar, M. McLaughlin, S. Ofek-Koifman, D. P, C.-S. Perng, H. Roitman, C. Ward, and J. Young, “A service delivery platform for server management services,” *IBM J. Res. & Dev.*, vol. 53, no. 6, pp. 1–17, 2009.
45. “Watson services for bluemix,” IBM Corp., New York, NY, USA. [Online]. Available: https://console.ng.bluemix.net/&num;/solutions/solution=watson
46. C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2007.
47. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer-Verlag, 2011
48. E. Rahm and H. Do, “Data cleaning: Problems and current approaches,” *Bull. Techn. Committee Data Eng.*, vol. 23, no. 4, pp. 3–14, 2000.
49. C. Sapia, G. Höfling, M. Müller, C. Hausdorf, H. Stoyan, and U. Grimmer, “On supporting the data warehouse design by data mining techniques,” in *Proc. GI-Workshop Data Mining Data Warehousing*, 1999, pp. 1–8.
50. S. Symonenko, S. Rowe, and E. D. Liddy, “Illuminating trouble tickets with sublanguage theory,” in *Proc. Hum. Language Technol. Conf. NAACL*, 2006, pp. 169–172.
51. C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.
52. E.-E. Jan and B. Kingsbury, “Rapid and inexpensive development of speech action classifiers for natural language call routing system,” in *Proc. Spoken Language Technol.*, 2010, pp. 336–341.
53. D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
54. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
55. T. Hofmann, “Probabilistic latent semantic indexing,” in *Proc. SIGIR*, 1999, pp. 50–57.
56. G. Miao, Z. Guan, L. E. Moser, X. Yan, S. Tao, N. Anerousis, and J. Sun, “Latent association analysis of document pairs,” in *Proc. ACM SIGKDD Conf.*, 2012, pp. 1415–1423.
57. K. Toutanova, D. Klein, C. Manning, and Y. Singer. “Feature-rich part-of-speech tagging with a cyclic dependency network,” in *Proc. HLT-NAACL*, 2003, pp. 252–259.
58. E.-E. Jan, J. Ni, N. Ge, N. Ayachitula, and X. Zhang, “A statistical machine learning approach for ticket mining in IT service delivery,” in *Proc. Int. Symp. Integr. Netw. Manage.*, 2013, pp. 541–546.
59. Y. Li, T.-H. Li, R. Liu, J. Yang, and J. Lee, “Application management services analytics,” in *Proc. IEEE Int. Conf. Serv. Oper. Logistics, Informat.*, 2013, pp. 366–371.
60. R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” in *Proc. ACM SIGMOD*, 1993, pp. 207–216.
61. “Cognos software—Business intelligence and performance management,” IBM Corp., New York, NY, USA. [Online]. Available: http://www-01.ibm.com/software/analytics/cognos/
62. “Tableau server—Rapid-fire business intelligence, today,” Tableau Softw., Seattle, WA, USA. [Online]. Available: http://www.tableau.com/products/server
63. “What is Splunk?” Splunk, San Francisco, CA, USA. [Online]. Available: http://www.splunk.com
64. “What is domo?” Domo, American Fork, UT, USA. [Online]. Available: http://www.domo.com/
65. Y. Thoss, C. Pohl, M. Hoffmann, J. Spillner, and A. Schill, “User-friendly visualization of cloud quality.,” in *Proc. IEEE CLOUD*, 2014, pp. 890–897.
66. C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Knowl. Discovery Data Mining*, vol. 2, pp. 121–167, 1998.
67. A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, “A maximum entropy approach to natural language processing,” *Comput. Linguistics*, vol. 22, pp. 39–71, 1996.
68. H.-K. J. Kuo and C.-H. Lee, “Discriminative training of natural language call routers,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 1, pp. 24–35, Jan. 2003.
69. L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

70. C. Kadar, J. Iria, “Domain adaptation for text categorization by feature labeling,” in *Advances in Information Retrieval, Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2011, pp. 424–435.
71. H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, “Interpreting TF-IDF term weights as making relevance decisions,” *ACM Trans. Inf. Syst.*, vol. 26, no. 3, pp. 1–37, 2008.
72. C. Alexander and M. Marzullo, “Rethinking applications management: A best practice approach to managing your portfolio,” HP, Palo Alto, CA, USA, HP Viewpoint Paper, 2013. [Online]. Available: http://www8.hp.com/h20195/v2/GetPDF.aspx%2F4AA4-5494ENW.pdf
73. T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, “Black-box and gray-box strategies for virtual machine migration,” in *Proc. USENIX NSDI*, 2007, pp. 17–28.
74. D. Wheeler, *Understanding Statistical Process Control*, 3rd ed. Knoxville, TN, USA: Statist. Process Control Press, 2010.
75. J. Oakland, *Statistical Process Control*, 6th ed. Evanston, IL, USA: Routledge, 2007.
76. M. Buco, D. Rosu, D. Meliksetian, F. Wu, and N. Anerousis, “Effort instrumentation and management in service delivery environment,” in *Proc. Int. Conf. Netw. Serv. Manage.*, 2012, pp. 257–260.
77. J. De Gooijer and R. Hyndman, “25 years of time series forecasting,” *Int. J. Forecast.*, vol. 22, no. 3, pp. 443–473, 2006.
78. L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, “Statistical analysis of a telephone call center: A queueing science perspective,” *J. Amer. Statist. Assoc.*, vol. 100, no. 469, pp. 36–50, 2005.
79. C. Bartolini, C. Stefanelli, and M. Tortonesi, “SYMIAN: A simulation tool for the optimization of the IT incident management process,” in *Proc. IFIP/IEEE Int. Workshop Distrib. Syst., Oper. Manage.*, 2008, pp. 83–94.
80. Z. Aksin, M. Armony, and V. Mehrotra, “The modern call center: A multi-disciplinary perspective on operations management research,” *Prod. Oper. Manage.*, vol. 16, no. 6, pp. 665–688, Dec. 2007.
81. M. Cezik and P. LaEcuyer, “Staffing multi-skill call centers via linear programming and simulation,” *Manage. Sci.*, vol. 54, no. 2, pp. 310–323, Feb. 2008.
82. Z. Feldman and A. Mandelbaum, “Using simulation based stochastic approximation to optimize staffing of systems with skills based routing,” in *Proc. Winter Simul. Conf.*, 2010, pp. 3307–3317.
83. S. Agarwal, “Hybridsourcing: A novel work allocation mechanism to provide controlled autonomy to workers,” in *Proc. IEEE/IFIP NOMS*, 2014, pp. 1–8.
84. Y. Diao and A. Heching, “Closed loop performance management for service delivery systems,” in *Proc. IFIP/IEEE Netw. Oper. Manage. Symp.*, 2010, pp. 61–69.
85. M. Vukovic, J. Laredo, and S. Rajagopal, “Challenges and experiences in deploying enterprise crowdsourcing service,” in *Proc. Int. Conf. Web Eng.*, 2010, pp. 460–467.
86. D. Rosu, W. Cheng, E. Jan, and N. Ayachitula, “Connecting the dots in IT service delivery—From operations content to high-level business insights,” in *Proc. IEEE Int. Conf. Serv. Oper. Logistics, Informat.*, 2012, pp. 410–415.
87. *PMML 4.1—General Structure of a PMML Document*, D. M. Group. [Online]. Available: http://www.dmg.org/v4-1/GeneralStructure.html

Received March 3, 2015; accepted for publication April 3, 2015

Yixin Diao *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (diao@us.ibm.com).* Dr. Diao is a Research Staff Member at the IBM T. J. Watson Research Center. He received his Ph.D. degree in electrical engineering from Ohio State University in 2000. He has published more than 80 papers in systems and services management, and is coauthor of the book *Feedback Control of Computing Systems* (Wiley 2004). He received an IBM Outstanding Innovation Award in 2005, was named as IBM Master Inventor in 2007, and also received an IBM Outstanding Technical Achievement Award in 2013. He is the recipient of the following: the 2002 Best Paper Award at *IEEE/IFIP (Institute of Electrical and Electronics Engineers/International Federation for Information Processing) Network Operations and Management Symposium*, the 2002–2005 Theory Paper Prize from the International Federation of Automatic Control, the 2008 Best Paper Award at *IEEE International Conference on Services Computing*, and the Second Prize of the 2012 Innovation in Analytics Award from Institute for Operations Research and the Management Sciences. He served as Program Co-chair for the *6th International Conference on Network and Service Management* in 2010 and Program Co-chair for the 13th *IFIP/IEEE International Symposium on Integrated Network Management* in 2013. He is an Associate Editor of IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, *Journal of Network and Systems Management*, and *Computers & Electrical Engineering: An International Journal*.

Ea-Ee Jan *IBM Global Technology Services Division, Poughkeepsie, NY 12601 USA (ejan@us.ibm.com).* Dr. Jan is a Senior Data Scientist in the IBM Global Technology Services (GTS) Division. He is the technical lead for ATM predictive analytics services. Prior to joining GTS in 2015, he was a Research Staff Member at IBM Research. He has worked on machine learning and data mining in the areas of speech recognition, natural language processing, machine translation, and multi-lingual information retrieval for software research. He later worked on image analytics for cloud computing, and service analytics for IT delivery. Dr. Jan was a Research Assistant Professor at Rutgers University before joining IBM. He holds 13 U.S. patents and has published more than 50 technical papers. He is an IBM Master Inventor and was recently recognized by the IBM Corporate Data Science Board as one of the first data scientist practitioners in IBM.

Ying Li *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (yingliatus.ibm.com).* Dr. Li has been a Research Staff Member at the IBM T. J. Watson Research Center since 2003. She was with the Exploratory Computer Vision Group before she joined the Service Research area in 2011. Dr. Li's research interests include audiovisual content analysis and management, computer vision, business analytics, and service management. She has authored or coauthored 30 patents and approximately 50 peer-reviewed conference and journal papers, as well as five books and book chapters on various multimedia and computer vision related topics. Dr. Li received her M.S. and Ph.D. degrees from University of Southern California in 2001 and 2003, respectively. She is a senior member of the Electrical and Electronics Engineers (IEEE).

Daniela Rosu *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (drosu@us.ibm.com).* Dr. Rosu is a Research Staff Member in the Service Delivery Management and Analytics department of the IBM T. J. Watson Research Center. Her current research interests include Q&A systems for client-assist on IT trouble resolution. In the past, she worked in multiple areas including process modeling and optimization of IT Service Management processes, productivity tools for IT Service Operations, business goal-driven resource management in complex IT environments, operating systems support for high-performance web servers, distributed web caching infrastructures, and real-time operating systems. Dr. Rosu received a Ph.D. degree in computer science from the Georgia Institute of Technology in 1999, with a dissertation in the area of adaptation in complex real-time systems. She also holds an M.S. degree in computer science from Georgia Institute of Technology (1995) and an MS. degree in theoretical computer science from the Faculty of Mathematics, University of Bucharest, Romania (1987).

Anca Sailer *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (ancas@us.ibm.com)*. Dr. Sailer received her Ph.D. degree in computer science from UPMC Sorbonne Universités, France, in 2000. She was a Research Member at Bell Labs from 2001 to 2003, where she specialized in Internet services traffic engineering and monitoring. In 2003, Dr. Sailer joined the IBM T. J. Watson Research Center where she is currently a Research Staff Member and IBM Master Inventor in IBM Watson Health. She was the architect lead for the automation of the IBM SmartCloud* Enterprise Business Support Services, which earned her, in 2013, an IBM Outstanding Technical Achievement Award. Dr. Sailer holds more than two dozen patents, has coauthored numerous publications in IEEE and ACM-refereed journals and conferences, and co-edited three books on network and IT management topics. Her research interests include cloud computing business and operation management, self-healing technologies, and data mining. She is a senior member of the IEEE.