RC25275 (WAT1203-084) March 28, 2012
Computer Science

IBM Research Report

H-IQ: Human Intelligence for Quality of Service Delivery Data

Maja Vukovic, Jim Laredo, Valentina Salapura
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

h-IQ: Human Intelligence for Quality of Service Delivery Data

Maja Vukovic, Jim Laredo, Valentina Salapura
IBM T.J. Watson Research
19 Skyline Dr, Hawthorne NY 10532, USA
{maja, laredoj, salapura}@us.ibm.com

***Abstract*—Service delivery centers are extremely dynamic environments in which large numbers of globally distributed system administrators (SAs) manage a vast number of IT systems on behalf of customers. SAs are under significant time-pressure to efficiently resolve incoming customer requests, and may fall far short of accurately capturing the intricacies of technical problems, affecting the quality of ticket data. At the same time, various data stores and warehouses aggregating business insights about operations are only as reliable as their sources. Verifying such large data sets is a laborious and expensive task. In this paper we propose system h-IQ, which embeds a grading schema and an active learning mechanism, to identify most uncertain samples of data, and most suitable human expert(s) to validate them. Expert qualification is established based on server access logs and past tickets completed. We present the system and discuss the results of ticket assessment process.**

Keywords-component; service delivery, automation, automatic data quality evaluation, data quality, social networking

I. Introduction

Service delivery centers are large, complex and dynamic ecosystems, which engage 100000s of experts globally to manage 1000s of processes supporting 1000s of IT systems with 100s of configurations. While operations at service delivery centers are typically associated with back-end processes, its efficiency affects quality at front-end (e.g., client experience and satisfaction).

Multiple ticketing systems, data stores and warehouses trace the operations in service delivery centers. They capture practices of Subject Matter Experts (SMEs), who are typically System Administrators (SAs), and changes in the IT infrastructure (e.g. server decommissioning). These ticketing systems, and enterprise-level warehouses are only reliable as their sources, whether human-driven (tickets submitted by SAs) or system-driven (automated updates of server registries).

TABLE I. Categorization of 200,000 Unix Tickets

Unix tickets for 44 SO customers over 3 months time			
Maintenance	***Problem***	***Change***	***Empty***
2%	47%	7%	44%

Low quality of such data leads to inefficiencies in operations (e.g. incomplete tickets slow down the problem resolution process), or leads business analytics to reach wrong or suboptimal conclusions. Table 1 summarizes analysis of tickets created for 44 strategic outsourcing (SO) customer accounts over 3-month period, where 44% of data records are blank with insufficient data and as such unusable.

Accumulated problem resolution records contain tremendous source of information about the managed system, its efficiencies and weaknesses, and in addition to analytics, it is a valuable source for knowledge transfer and learning in attempt to train new administrators. The record data are also used for reporting and report generation in billing and service level agreement (SLA) measurements.

Moreover low quality of data affects the business decisions (e.g. leading to poor business insights when identifying opportunities for new service offerings, such as “show me the low utilization servers across the banking sector”). Business insights and problem resolution processes require careful quality assessment to build credibility with stakeholders and efficiently resolve problem tickets. Moreover in such volatile environments, quality of operations and business insights will vary depending on the corresponding data source.

Planning activities also depend on good quality data. Take for example server consolidation, where old servers or underutilized servers are migrated into virtual environments with newer hardware. Being able to understand the configuration information such as number of CPUs, speed, memory, operating system and software configured as well as resource information such as network bandwidth, disk and CPU utilization are all key to be able to prepare a plan that maps to proper sized servers. Bad quality data could easily derail a plan from improper source selection to bad target allocations.

Human computation is a rapidly growing area that combines the processing power of a multitude of humans (often SMEs) to solve different computational problems [1]. We are witnessing increasing number of application areas within and outside the enterprise where the intelligence of humans is collectively harvested [2,3].

In this paper we present a system h-IQ, which employs human intelligence to manage quality of data in services delivery. What differentiates our approach is the ability for a targeted group of experts (system administrators) to provide and validate knowledge about the tickets and infrastructure, based on their access rights to a given set of servers. We present the core elements of our system and discuss impact of this approach to the quality of data in services delivery.

The paper is structured as follows. In the next section we provide overview of data and processes in services delivery. Section 3 describes our system in terms of the components and their interfaces. Section 4 presents the ticket quality assessment process. Section 5 discusses the benefits of ticket analysis and challenges in engaging SMEs to validate the quality. Section 6 puts our work in the context of state of the art in human computation, ticket analysis and incentive mechanisms. Section 7 concludes and lays out future work.

II. BACKGROUND

Service systems can be conceptualized as a stage with front-end (client-facing) and back-end (operations) functions. As a result to meet the quality expectations and reduce operating costs the providers need to continuously improve services quality both at front-end (e.g., client experience and satisfaction), and back-end (e.g. production and delivery).

Service delivery centers tend to host large number of customers, each customer with its own policies, and regulations that they need to adhere to. Applications will vary, the software stacks in use will be different, and the processes to manage them will have their own adaptations. For the service provider that owns the delivery center there is the incentive to leverage economies of scale and seek to standardize wherever possible across as many accounts as possible. Such efforts may include common tools to manage environments, or consolidation into less platforms, or unification of processes to manage incidents, problems and changes, to name a few. Many of these rely on availability of quality data in order to make sensible decisions. One such data is configuration information. Discovery tooling if available and properly configured with credential information can scan the endpoints and report on configuration and dependencies.

However, as configuration changes, new scans are required, and if the credentials change or the agents fail it is possible to easily fall behind. Discovery information although potentially extensive is not able to capture the business purpose, or business requirements leaving it up to the application owner to provide it. As any user provided data, it is as transient as the configuration it describes. If any major change occurs it should be revised, including addition of new components or applications, repurposing the server, or decommissioning it altogether. Another source of data often used on consolidation decisions is resource management information. This information is captured by agents that regularly report on the utilization of resources such as memory, disk, and network, to name a few. If the agent ceases to execute, information will become stale very quickly. Moreover, unless there is good lifecycle management it may not be clear if when reporting stops is because of a failure or perhaps a customer decision or the server was just decommissioned. The mere fact that the agent needs to "steal" resources to execute is sometimes enough reason to shut it down in critical times or disallowing it for mission critical environments. Creating a situation, like all the ones just described above where the data quality about the environment starts to differ from server to server and account to account interferes with any effort to standardize the operations and cost effectiveness of the service delivery and associated services quality.

From the front-end perspective of operations, many companies have customer service department to provide customer service support. Every customer support request generates service logs and records. These records in the IT Service Support and Delivery organization are recorded by system administrators, and are referred to as ticket data, where every service request represent a ticket. A sample ticketing record is shown in Figure 1. Similarly, in customer support call centers; every interaction with customer and customer request is documented. The tickets contain information as reported by a customer describing experienced problem symptoms, or a new service needed, and represent a link between customers and the services infrastructure. Opened tickets are queued in the ticketing system, and dispatched to the appropriate system administrator, service center, or an agent for handling and resolution.

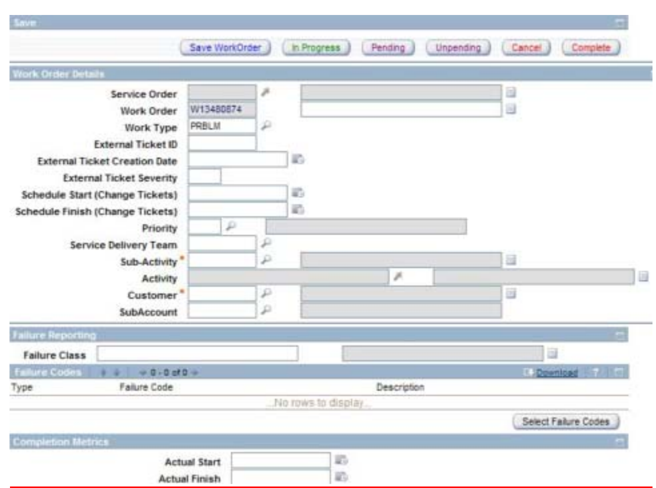

Figure 1. Ticketing tool

Service centers typically collect service data, which can then be used to assess and improve the performances of their representatives. In managing an IT data center, ticket data can be used to understand the distribution of underlying issues typically encountered in the operation of a data center for future planning. While underlying techniques and tools are applicable on different service centers, in this work we are focusing on managing an IT data centers. Ticket data spans from resolving issues to customer meetings and management cost around processes involved in operating a data center.

TABLE II. SAMPLE TICKET DATA

WONUM	***Work Order Description***
W9979846	Installation of Recover console on SV73412
Categorization	***Failure Class***
APPLICATION	OS / SYSTEM SOFTWARE
WORKTYPE	***Resolution***
SREQ	Recovery Console installed on the server

Typically, tickets contain both pre-defined, structured fields (e.g. problem type, support person/group handling the problem, ticket creation date, failure cause, failure class fields, etc.) as well as, unstructured fields (e.g. open ended text describing problems and solutions as entered by the support administrators), as shown in Figure 1. Table II illustrates a typical ticket data. It contains both pre-defined fields, as well as open ended fields. As can be seen, problem descriptions are about the descriptions of problems as supplied by customers, whereas the descriptions about problem solution were provided by a system administrator.

III. SYSTEM ARCHITECTURE

When a business analytics query is issued, or a ticket is submitted to the ticketing system it is important to provide a degree of confidence on the results provided or data captured. If the score provided is insufficient and creates too much risk on, the results should be discarded, unless there is a way to improve on their quality. The h-IQ system, shown in Figure 2., aims to evaluate the quality of a result and assist on improving it through the involvement of related parties that can attest for the data and may provide additional information to correct or improve existing data elements.

There are two entry points to the h-IQ system:

A. *Business Analyst submits a request for a certain business insight, based on existing data (Step 1a in Figure 2).*

B. *System Administrator submits a ticket to the system, which may trigger quality management process (Step 1b. in Figure 2.)*

At the core of h-IQ system there are two key components that allow for the assessment and continuous improvement of the data based on a given request, namely the quality assessor and the collective intelligence module.

Main components of our system are: analytics module, quality assessment module and human intelligence module. The system necessarily relies on existing data in ticketing systems, data stores and data warehouses.

Quality assessment module assigns a confidence score to tickets and other data (e.g. server utilization, server purpose, etc.). The overall confidence score also takes into the account reliability of the data source, whether human or system driven. E.g. confidence of the human input is derived from their expertise/familiarity with the given system and prior contributions. When confidence level is low, the system continuously seeks further input from other experts until satisfactory level is achieved. There may be cases when the desired quality level is not possible to be achieved, even with several expert validations, and such data elements are flagged accordingly. To bootstrap quality assessment, data elements are grouped by type, nature and their purpose to establish minimal expected confidence levels.

Collective intelligence module uses the data from access logs and tickets to identify the most suitable experts. It is applies a variation of multi-labeler active learning method [5], whose objective is to allow learning from multiple users, whose expertise across the data space may vary. We extend their model to include cost of expert engagement as another variable in optimization problem (i.e. selecting most uncertain sample and most suitable expert(s)). Cost is an important consideration in enterprise domain (e.g. you may not want to engage your top performers to evaluate data quality, when there is a Severity 1 problem for a customer).

When an analytics platform draws on a particular data set (Step 1c. in Figure 2), it is often able to assess which data points are better contributors than others to the overall model. Take a linear regression once the coefficients have been determined, it is possible to evaluate the test data and see the amount of error each data point is causing. These errors known as residuals when totaled are able to provide metrics to measure the quality of the model. Based on the performance of the individual data points it is possible to determine which data points to drop or better yet, need revision, and in which direction to improve the model. The Quality assessor will take such guidance and track the improvement, if any, of those data points as the collective Intelligence module takes over.

The collective intelligence module, upon request of the Quality assessor locates experts that may contribute to improve the quality of the data. We call an expert anyone that we can show is related to the data in question. We look at several IT artifacts to find these relationships. On server related inquiries we can look at access logs, who has accessed a given server most recently and for what purpose. On ticket related inquiries, depending on the nature of the data quality issue we have several options. If related to the resolution we could examine the ticket and find involved parties, if related to the type of ticket, searching for similar tickets would identify other experts that can assist, if related to some status, sever information leveraging the access logs is also a possibility. Pool memberships are also a source to mine experts, in part because they may have similar roles and access to the same infrastructure and may be familiar with the environment.

Once an expert or experts have been identified (Step 3 in Figure 2.), a task or set of tasks are presented to him with the specific inquiry. Expert is able to answer the task, defer the task, or invite others to participate, much how it is described in [4]. After reaching resolution the information is sent back to the Quality assessor module who re-tests for quality and decides whether or not to seek for additional assistance. If the data is deemed reliable, the quality assessor will ensure that the source system is updated, and the affected business inquiries are re-run.

h-IQ system is designed to automatically and systematically compute confidence level of service delivery data and engage the most qualified Subject Matter Expert (SME), typically System Administrator (SA), to validate it. Traditional organizational expertise repositories record only high-level job function, and possibly a set of skills associated with it (e.g. Network Administrator). In addition, information about who is managing servers for a given customer is often not exposed for compliance reasons, making locating the right employees almost intractable.

To identify the most qualified experts to validate the data, our system h-IQ relies on two sources: 1) ticketing data (e.g. reported problem, account, server affected, resolution applied, etc.) and 2) server access logs (e.g. who logged onto RHEL v5 instance and installed a security patch – who performed what, when and using which permission rights) to identify the most suitable SME.

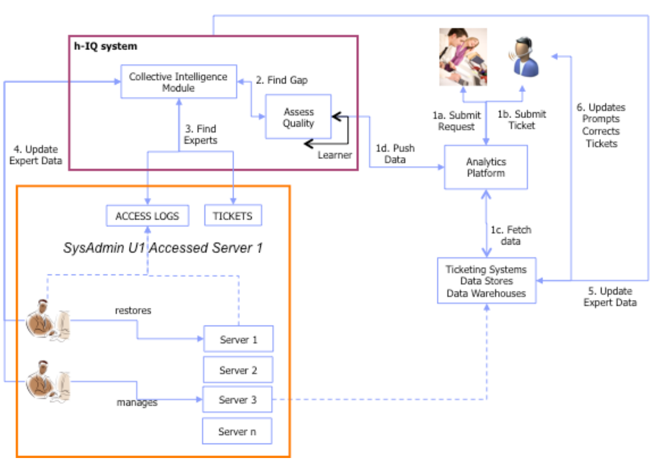

Figure 2. h-IQ system: overview of main operations.

IV. Ticket Quality Assessment

As a first step in building and evaluating our h-IQ system we focused on the quality of data in the problem and incident management domain. To develop our Quality Assessment component we have evaluated a set of 2595 tickets that were generated in the support center for a customer ACME over the course of a few months. The objective of Quality Assessment module is to separate data elements that require further validation by a human expert from usable data elements.

A. Ticket Grading

Our quality assessment module assigns a confidence score to incoming tickets giving each ticket a numeric score on a scale, and determining a threshold, which separates tickets into satisfactory and unsatisfactory confidence and quality level data, to assess the quality of incoming tickets, we took a two-stage approach. In the first stage, ticket quality was evaluated based on the completeness of following fields: failure class, failure symptom and worktype. These fields have pre-determined values which are selected from a drop down fields in our ticketing system. We decided to use these fields for our analysis because of the importance of information they carry. These fields specify if performed work was a part of service request, problem, or maintenance, and give information on the technical area the work was performed on – was it in the application domain, OS domain, hardware problem or a new server was built or migrated. Different studies might opt to use different fields.

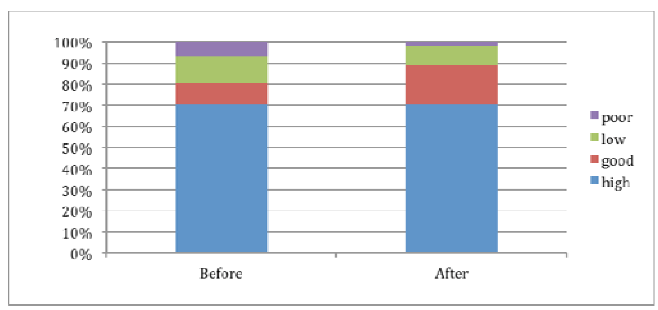

Figure 3. Ticket evaluation based on drop-down fields

A single point (grade) was assigned to a ticket for each one of the three desired fields. As shown in Figure 3, all tickets received a grade between 0 (none of the relevant fields are specified) to 3 (all relevant fields are specified). After re-categorizing the original data set, 88% of tickets were of sufficient value for SMEs to complete the requested work items.

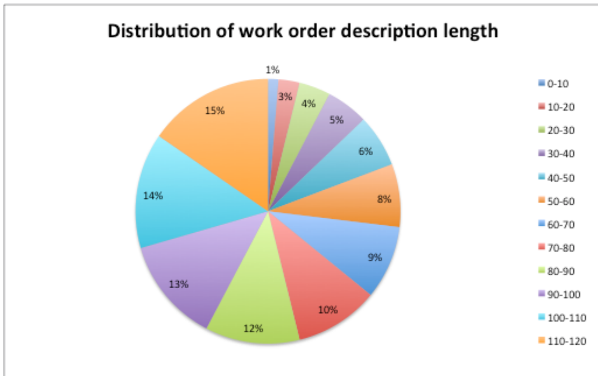

Figure 4. Length of text in work order description

We then incorporated the analysis of the work order description, which is an open-ended text field in the ticket. We have identified 29 keywords, such as meeting, network, build, etc., which describe different activity requests. First we observed the length of the work order description - shown in Figure 4- as an indicator of the ticket's potential quality. Majority of tickets has description of at least 20 characters, which is sufficient if the key activity is captured (e.g. build server).

Next, we evaluated the number of keywords that were found in each ticket, as shown in Figure 5. 57% of tickets had at least one keyword, and 85% has at least two matching keywords.

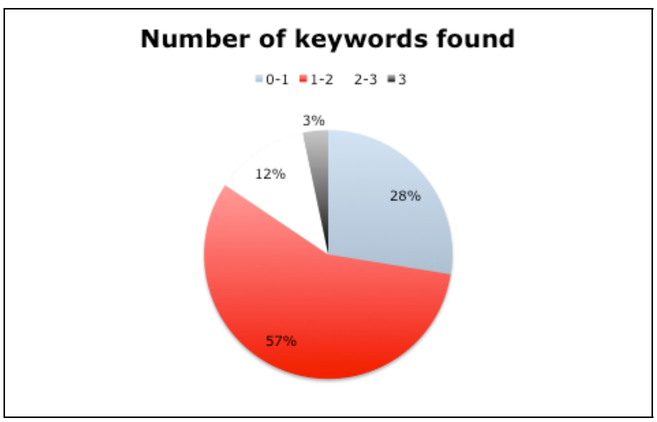

Figure 5. Number of keywords matched in the tickets

We have then extended the grading mechanism to include both drop-down values and the work order description. The final grade was assigned as follows. One grade would be assigned for availability of each value in the following fields: failure symptom, failure class and worktype.

The length of work-order type would be multiplied by a coefficient in order to normalize the string length against the ticket corpus. We selected the coefficient of 0.3 in order to provide a weight to the length of the work order description. This value was selected based on the maximum length of the work order description in the sample. Equation 1 shows that by adding a sum of grades in dropdown fields to the number of keywords occurring with the normalized string length results in the final grade.

Equation 1. Final grade computation

Final Grade = Sum(Drop-Down-Fields) +
C x Length(Work-Order-Description) +
Count(Keywords-Found)

Table 2 summarizes the results and distribution of grades before and after the inclusion of the text description field.

Table 2. Ticket grades and count

grade	0	1	2	3	4	5	6	7
before	179	320	270	1826	0	0	0	0
after	0	25	110	278	957	796	328	101

Figure 6 provides an overview of how we increased the grade of tickets once the work order description was processed. After we incorporated the description field usable / tickets of value increased to 95%.

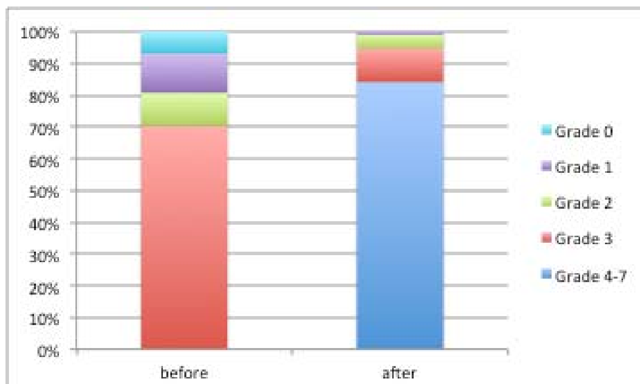

Figure 6. Ticket evaluation based on drop-down and description field

Figure 7 shows the basic building block of the quality assessor module, as we used it and the flow of operations. The tickets are inserted from the ticketing systems once system administrators have completed their work and closed the ticket. From all ticket data, only selected fields are parsed and sent to individual modules for grading. Finally, the overall score of the ticket is determined. To evaluate the performance a selected sample of tickets was later on manually graded, as discussed in Section 4B.

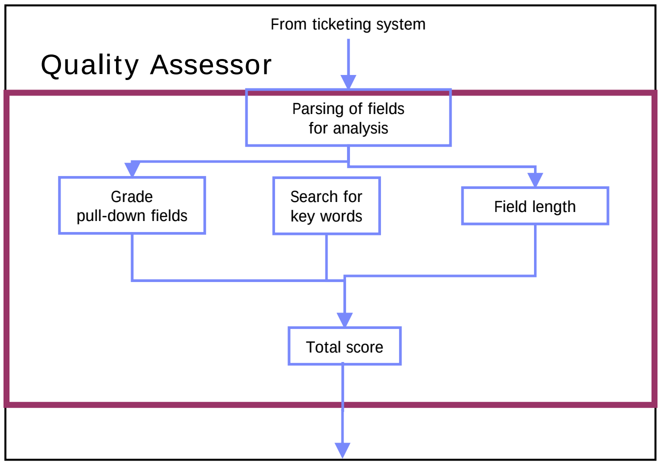

Figure 7. Quality assessor system: structure.

B. *Performance Evaluation*

To evaluate the effectiveness of our approach we have selected randomly a set of 92 tickets. We have then manually evaluated and rated using a 3-grade scale (1=poor, 2=acceptable, 3=good) in the context of:

- reporting value
- knowledge transfer value
- and prediction value (for future system design)

The quality value of ticket is threefold, as mentioned above. Firstly, tickets are used for reporting purposes and insights about the operations (e.g. peak hours in the support center). Secondly, tickets that have encoded a problem resolution approach are useful for knowledge transfer purposes. The solution to the ticket problem is often verified by knowledge administrator and eventually shaped into a documented best practice. In addition, such tickets can be used for training of new system administrators. Third benefit of high quality tickets is the ability to use them to provide predictive analytics on incoming tickets, to identify expected categorization, and/or suitable descriptions of the problems.

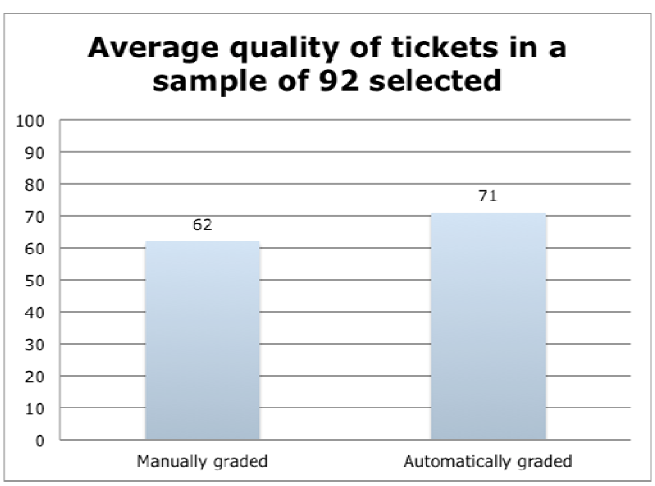

Figure 8. Performance of automated grading

Three subject matter experts evaluated a batch of 30-31 tickets each. We have then compared the manual and automated grading of selected 92 tickets to assess the performance of proposed method, as shown in Figure 8.

The average difference (in a sample of 92 tickets) is 9%; meaning that the performance of our claimed approach is accurate up to 91% (barring any perception differences arising from persons evaluating the tickets. Performance evaluation indicates that indeed tickets can be, at real time, analyzed for their quality.

V. Discussion

Once we have identified the low quality tickets the next step is to discover suitable SMEs to validate them in order to make them usable. Engaging experts based on their prior experience and access rights to the servers affected and reported on in the tickets is an effective approach to increase the data quality. For purposes of our evaluation, when working with data on a small account, SMEs may be easily and directly reachable due to the existing collaborations between the teams, and readily willing to help validate and classify a subset of tickets.

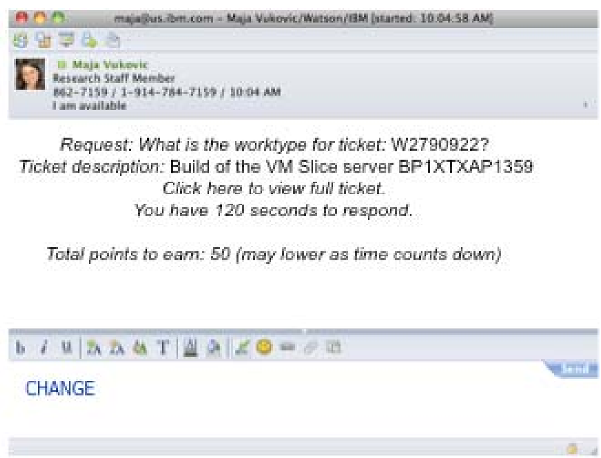

Figure 9. SME data validation

Figure 9. depicts how an SME can be engaged using instant messaging to validate the data in short amount of time.
Using the mechanism of sequential verification more SMEs can be engaged to cross-verify the data. For example, Figure 10. shows how another SME was verifying the data/answer provided by SME in the Figure 9.
SME in Figure 9 has assigned the worktype to be "CHANGE", and SME in Figure 1- was reviewing that.
Figures 9 and 10 show that SMEs were given a limited time to provide their answers, and were offered virtual points for their effort, which can be used to build leaderboard and represent a form of incentives.

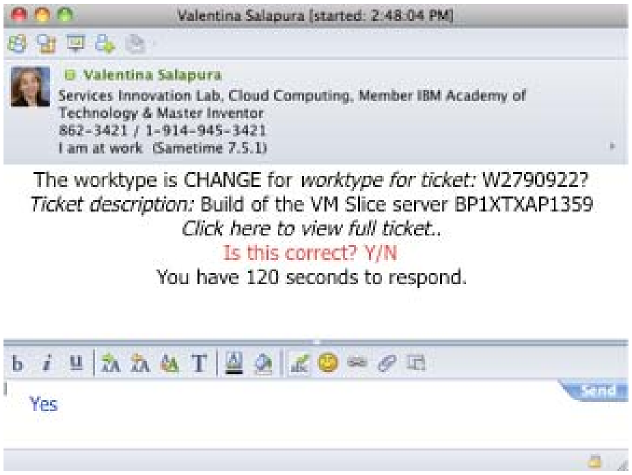

Figure 10. SME data validation – follow-up

SMEs that may need to be engaged are often over-utilized, having to deal with Severity 1 issues. As a result they often may not be available to answer the validation requests. In addition to an SME being unavailable they simply may not have sufficient interest or incentive to contribute to quality validations. This opens up a main question of how to build and maintain such a community over the time. In particular, in the beginning SMEs may be actively participating but over the time their engagement levels may decrease. Tangible and intangible incentives drive the success and failure in human computation systems.

In enterprise settings the design of incentives is challenged by additional considerations of human resource and business control guidelines. Services delivery experts may be employed at different conditions (e.g. permanent as opposed to a temporary contract) that may limit their eligibility for example for additional monetary incentives.
From our prior experience in engaging enterprise SMEs [6] we found that the traditional leaderboard and point mechanisms were not sufficient intangible incentives for participation. However, when contributors were offered access to a common knowledge base in return for their effort they found as an extra motivation to participate.

Stewart et al. [7] observe an inequality in enterprise participants in human computation systems and propose SCOUT model to drive the low contributors towards a sustainable crowd over the time.

In the context of our work we envision account or function-specific competitions enabling SMEs to build their reputation over the time in a) entering high-quality tickets and b) validating correctly incoming data.

VI. Related Work

With increase of scale and complexity of services delivery centers, IT Service Management gained momentum. In particular incident and problem management processes are design to detect, record, isolate and correct defects that occur in the services delivery environments. “Defect” can be defined as an instance or event that is not a satisfactory outcome, such as a malfunctioning server or a failed patch management process. Prior work in this area has focused on analysis of tickets that are raised at the time of defect, in order to automate resolution processes with adaptive dispatching (where tasks sent to experts based on domain and complexity of defect) [8]. Furthermore, as multiple tickets are often created both by humans and system to report on related or similar events a further analysis is critical in understanding the correlation of the events to streamline the resolution. Marcu et al. [9] correlate events by analyzing tickets along three dimensions: category classification, configuration setup and time.

Tickets drive the problem resolution and defect prevention processes, hence their quality is a key concern for standard operations in services delivery center, and for further optimizations. In order to automate the ticket classification and assessment of their value researchers often have to train large sets of data before running classification algorithms.
Large groups of (unknown) human, often non-experts, are increasingly being harvested to label the data sets and establish quality aiding automated recognition algorithms given the low-cost promise of collective intelligence [10,11]. Sheng et al. [10] demonstrated that repeated labeling can improve both the quality of the labeled data directly, and the quality of the models learned from the data. In contrast to the existing approaches, our system engages users who are known experts in their domain, and whose cost can be established according to the business goals, thereby driving the optimization formula for selection of a data element and a corresponding expert.

Enterprises are applying this human computation approach to gather, validate and improve data. Stewart et al. [12] use crowdsourcing to effectively tap into the collective intelligence of multilingual employees to translate sentences or correct machine translated sentences for improving translation accuracy and quality. Other approaches to managing data quality through human computation include majority voting and aggregation of contributions [13,14].

Recently human computation systems have started to embed elements of games, thereby creating “Games with a Purpose” (GWAP) [15] in order to harness useful work from humans for free in AI-hard problems. A key challenge in this approach is how to avoid unrelated contributions from users by allowing users to verify answers from others. Ho et al. [16] design a game for semantic annotation of images called the PhotoSlap game, demonstrating how they reach the subgame perfect Nash equilibrium with the target strategy when players are rational and do not engage in collusion.

The results show that the players can be kept u to date of a default strategy, the target strategy, in advance to satisfy subgame perfect equilibrium.

VII. Summary and Future Work

In this paper we proposed a system that automatically and systematically engages enterprise experts to validate quality of data in services delivery. It identifies the gaps needed to improve the data at real time, in order to leverage users that are closely related to assets to close this gap. Using multi-labeling probabilistic model system identifies low quality, incomplete, and inaccurate data elements and assigns a confidence level. System then assigns the data elements to corresponding experts for verification, based on expert's access history to relevant servers / IT assets.

As we evolve our system and evaluate its effectiveness, our next steps are to frame the data verification goal as a GWAP and explore incentive mechanisms.

Acknowledgment

We thank Alan Bivens for his input and feedback on ticket analysis and performance evaluation.

References

[1] J. Surowiecki, The Wisdom of Crowds. Anchor, 2005.

[2] D. Brabham, Crowdsourcing as a model for problem solving: An introduction and cases, "Convergence: The International Journal of Research into New Media Technologies", 14(1):75–90., 2008.

[3] M. Vukovic, V. Naik, "Managing IT Enterprise Systems using Online Communities", Proceedings of Service Computing Conference (SCC), 2011.

[4] J. Laredo, M. Vukovic, S. Rajagopa, "Service for Crowd-Driven Gathering of Non-Discoverable Knowledge". International Conference on Service Oriented Computing (2011).

[5] Y. Yan, R. Rosales, G. Fung, J. Dy, "Active Learning from Crowds," Proceedings of the International Conference on Machine Learning (ICML), 2011.

[6] Mariana Lopez, Maja Vukovic, Jim Laredo: PeopleCloud Service for Enterprise Crowdsourcing. IEEE SCC 2010: 538-545

[7] O. Stewart, D. Lubensky and J.M. Huerta J. M., "Crowdsourcing participation inequality: a SCOUT model for the enterprise domain." In Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10).

[8] L. Zia, Y. Diao, D. Rosu, C. Ward, K. Bhattacharya, "Optimizing Change Request Scheduling in IT Service Management", IEEE SCC, 2008.

[9] P. Marcu, G. Grabarnik, L. Z. Luan, D. Rosu, L. Shwartz, C. Ward, "Towards an optimized model of incident ticket correlation", Integrated Network Management, 2009.

[10] V. S. Sheng, F. Provost, P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers", In Proceedings of the International conference on Knowledge discovery and data mining (KDD), 2008.

[11] A. Bernstein, J. Li, "From active towards InterActive learning: using consideration information to improve labeling correctness", In Proceedings of the Workshop on Human Computation, 2009.

[12] O. Stewart, J. M. Huert, M. Sader, "Designing crowdsourcing community for the enterprise." In Proceedings of the ACM SIGKDD Workshop on Human Computation, 2009.

[13] A. Sorokin, D. Forsyth, "Utility data annotation with Amazon Mechanical Turk", In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society, Washington, WA, USA. 2008.

[14] R. Kern, H. Thies, C. Bauer, and G. Satzger, "Quality Assurance for Human-based Electronic Services: A Decision Matrix for Choosing the Right Approach". In Proceedings of First Enterprise Crowdsourcing Workshop in conjunction with ICWE 2010.

[15] L. von Ahn,"Games with a Purpose," *Computer* 39, 6, 92-94, 2006.

[16] C. Ho, T. Chang, and J. Hsu, "PhotoSlap: A Multiplayer Online Game for Semantic Annotation", Processings of the 22nd AAAI Conference/ 2007