

# Enabling Analysts in Managed Services for CRM Analytics

Indrajit Bhattacharya  
IBM Research  
indrajbh@in.ibm.com

Shantanu Godbole  
IBM Research  
shgodbol@in.ibm.com

Ajay Gupta  
IBM Research  
ajaygupta@in.ibm.com

Ashish Verma  
IBM Research  
vashish@in.ibm.com

Jeff Achtermann  
IBM MBPS  
jachterm@us.ibm.com

Kevin English  
IBM MBPS  
englishk@us.ibm.com

## ABSTRACT

Data analytics tools and frameworks abound, yet rapid deployment of analytics solutions that deliver actionable insights from business data remains a challenge. The primary reason is that on-field practitioners are required to be both technically proficient and knowledgeable about the business. The recent abundance of unstructured business data has thrown up new opportunities for analytics, but has also multiplied the deployment challenge, since interpretation of concepts derived from textual sources require a deep understanding of the business. In such a scenario, a managed service for analytics comes up as the best alternative. A managed analytics service is centered around a business analyst who acts as a liaison between the business and the technology. This calls for new tools that assist the analyst to be *efficient* in the tasks that she needs to execute. Also, the analytics needs to be *repeatable*, in that the delivered insights should not depend heavily on the expertise of specific analysts. These factors lead us to identify new areas that open up for KDD research in terms of ‘time-to-insight’ and repeatability for these analysts. We present our analytics framework in the form of a managed service offering for CRM analytics. We describe different analyst-centric tools using a case study from real-life engagements and demonstrate their effectiveness.

## Categories and Subject Descriptors

H.4.0 [Information Systems Applications]: General;  
I.7.0 [Document and Text Processing]: General

## General Terms

Design, Human Factors, Experimentation

## 1. INTRODUCTION

In recent years, organizations have been increasingly turning to data mining tools and techniques to gain insights into

their business and processes. In a crowded and competitive business landscape, data mining for insights has evolved from a matter of choice to a necessity. Generally speaking, delivering business insight for a company translates to figuring out using all available data, *what issues are hurting its products and services*, and *what action needs to be taken* to address these issues. In the customer relationship management (CRM) domain – our focus in this paper – a typical problem is understanding the key drivers of customer satisfaction (CSat) or dissatisfaction in a contact center, from customer survey comments, customer profiles, and various other data sources. Data mining tools can be used to observe patterns and correlations in the data with respect to target metrics. However preparing and transforming the data for analysis and translating the *findings* into *actionable insights* requires domain knowledge and human expertise. A typical finding in CSat analysis can be that low ratings are associated with long wait times before customers talk to agents and also with long call handling times. Clearly the recommendation here is not ramping up agent headcount, but better use of available agents. This human expertise is often the best asset in an analytics task, and this is extremely challenging, if not impossible, to automate. Accordingly, business analytics usually ends by presenting interesting patterns to the analyst in a manner that aids interpretation.

Deploying analytics tools in a business environment thus places a high requirement on the technical skill as well as the business expertise of users. As a result, rapid, successful deployments across domains has been a significant challenge; more so as traditional, retrospective, reactive, business intelligence[10] (1<sup>st</sup> generation BI) is being replaced by proactive, on-the-fly (2<sup>nd</sup> and 3<sup>rd</sup> generation[11]) BI.

A recent development that has further impacted the deployment issue is the explosion of unstructured and textual data available to businesses. Data in a CRM contact center scenario includes agent logs, billing notes, chat logs, emails, huge volumes of audio call data, customer feedback surveys, and also data from reviews, forums and blogs. Recent industry reports[12] suggest that volume of unstructured data can be as high of 80% of an organization’s data and this is rapidly growing. The value of unstructured data is often complimentary to that of structured business data – structured dimensions can indicate that a customer is likely to churn, while the customer communicates to the company, through surveys for example, the reason for her dissatisfaction. This opened up the market for analytics vendors (like SAS, SPSS and others) to venture into text analytics products. However, deployment of text mining solutions not only

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD’09, June 28–July 1, 2009, Paris, France.

Copyright 2009 ACM 978-1-60558-495-9/09/06 ...\$5.00.

needs knowledge of techniques like clustering, classification, annotation, but also needs in-depth business understanding. Designing class labels for classification, creating labeled training data[5], understanding clusters, or defining rule-sets for annotation are non-trivial at best for non-experts.

Since human expertise is unavoidable in delivering business insights, a company wishing to gain insights into its business may either develop a custom solution internally or invest in ‘setting up’ existing products. Both options require hiring professional analysts apart from investments in infrastructure. These analysts would be trained very specifically to understand the business of the organization and would evolve into the most appropriate domain experts. However such investments are not feasible for all organizations. The other alternative that many companies are turning to is the ‘analytics as a service’ model.

A managed analytics service<sup>1</sup> takes care of the end to end data ingestion, processing, and mining of client data in a secure environment. Such an offering has two main advantages against product based analytics deployments. First, it leverages consulting experience of business analysts who employ standard analytics methodologies (like CRISP-DM<sup>2</sup>) using proprietary and cutting-edge products. Secondly the client need only submit data, interact through a liaison, and receive actionable recommendations, without needing in-house data mining expertise or investments in tools and other infrastructure. The central entity in a managed analytics service offering is the analyst, who liaises between technical aspects of the analytics and the client’s business.

One apparent disadvantage of such an offering compared to a home-grown analytics solution, is that this analyst may not have the level of expertise or knowledge about the business of the specific company that comes from training or experience within that company. This can adversely affect her *time-to-insight*. The second challenge is ensuring *repeatability of service* across engagements. Though the role of the business analyst is central to the services model, the delivered insights should not depend heavily on the expertise and domain knowledge of specific analysts in the team. The research agenda in a managed analytics service scenario is thus different from that of the ‘install and use’ environment common for products. The evolved goal of research is to design the right set of tools that: 1) reduce the ‘time to insight’ for an analyst, and 2) enable ‘repeatability across analysts’ by bringing the most important insights to light.

In this paper we describe our experience in designing and developing the IBM Voice Of Customer Analytics *hosted, asset-based, managed service offering for CRM analytics* (referred to as *IVOCA* in this paper). *IVOCA* is an evolving ongoing service offering. We investigate the evolved research landscape to identify novel challenges for the KDD community. We first identify the different tasks that the business analyst needs to perform in a managed service for analytics, and list out new research challenges that come out around ‘time to insight’ and repeatability(Section 2). We highlight the role of research in this new setting through various *IVOCA* tools and components (Section 3). We present results(Section 4), experiences, and lessons learned (Section 5) from a real-world CRM analytics service engagement with a large automobile company.

## 2. RESEARCH CHALLENGES

As researchers working in the design of a managed service offering, our *IVOCA* experience has provided us with first hand knowledge about the challenges and demands of a managed analytics service. Instead of bypassing human intervention through the design of automatic solutions, we would like to “enable analysts”, by (a) reducing their ‘time to insight’ and (b) aiding ‘repeatability of service’ across analysts. The key to reducing ‘time-to-insight’ is engaging the analyst in the process minimally. Research has looked at some aspects of ‘optimal engagement’ or putting the human in the loop such as active learning[3]. The repeatability aspect, in contrast, has not received much attention in the academic literature, though it has been stressed by many practitioners [11]. Also, the analyst in a managed service does a lot more than labeling data instances as in active learning. In the rest of this section we identify the other critical tasks that the analyst performs, where similar ideas of *optimal engagement* and *repeatability* need to be defined and addressed. We have addressed some of these issues in the current *IVOCA* solution, but we believe these are larger challenges which open up new research opportunities.

### *Analyzing Noisy Unstructured Text.*

In the CRM setting, large volumes of customer communication coming from emails, chat logs, transcripts of phone conversations, and surveys, are available as unstructured voice-of-customer (VoC) data, and are critical for understanding what the customers are dissatisfied (or satisfied) about. However it is almost impossible to get actionable insights from VOC data without the analyst’s involvement beyond finding simple patterns. This data is extremely noisy; feature noise due to misspellings, abbreviations, slang, shorthand, and other errors, blow up the dimensionality of the data. As a result, completely unsupervised analysis or clustering does not go very far. Also, the supervised alternative is not straight forward to implement, since neither crisply defined label sets nor consistently labeled training instances are easily obtainable [5]. To reduce time-to-insight, an analyst typically designs label-sets using a top down approach, where she comes up with a list of concepts or issues that are likely to be mentioned in the documents using her domain knowledge. She can also provide a set of keywords to define each of these concepts. However, the keyword set is often analyst-specific and affects repeatability across analysts. Also, she is typically unsure whether all of them are relevant for a particular dataset or whether her list of concepts is exhaustive.

Providing a labeled training set of any significant size for each of these concepts is again a significantly time consuming task for an analyst. Outsourcing the labeling task to non-experts does not work as an easy fix and tends to hurt repeatability - experts and non-experts often do not agree with each other (and sometimes with themselves) on label assignments to documents [5, 13]. Given this background, a rule-based classifier/annotator is the most applicable technique to use. But the accuracy of such approaches, both in terms of precision and recall, is typically not very high, as we will see in our experiments (§4). This is because that the keywords that the analyst provides for each concept are not exhaustive (affecting recall) and are more descriptive than discriminative for the concepts involved (affecting precision). Finally, the unstructured text data has to be used in con-

<sup>1</sup>[http://en.wikipedia.org/wiki/Managed\\_services](http://en.wikipedia.org/wiki/Managed_services)

<sup>2</sup><http://www.crisp-dm.org>

junction with all the other data that is available to detect meaningful patterns. We address some of these challenges using concept-discovery and building a Synonym Finder tool in Section 3.3.

### *Assisting Insight Generation.*

An analyst's search for insights begins by identifying patterns or correlations in the data between business dimensions like CSat and other groups of structured and unstructured attributes. In the CRM setting, "*Customers have complained in surveys about agents ABC, DEF, and GHI using unacceptably rude language during calls; all of these agents report to team leader JKL who has the lowest average CSat score amongst team leaders*", is an insightful pattern. However, the dimensionality of the space that needs to be searched by the analyst for patterns is still unmanageably high – scores of dimensions from unstructured data add to the many structured dimensions.

Our experience is that this search is one of the most time-consuming aspects of a managed analytics service and some degree of automation would hugely benefit the analyst and the efficiency of the overall process. Clearly, it would be an extremely hard problem to come up with a completely automated solution for this problem that has both high precision and recall. Therefore it is crucial to leverage the domain expertise of the analyst. The problem can be broadly categorized into two groups - 'pattern identification' and 'pattern discovery'. The first is a relatively easier problem, where the analyst knows what she is searching for. For example, 'lack of knowledge' for an agent might be an expected issue affecting CSat, but the extent of its impact on CSat may be unknown. This problem is easier to automate, but it would still be cumbersome if the analyst has to list down all the expected patterns, each possibly involving multiple dimensions. In the second problem, the analyst is looking for interesting but unexpected patterns. For example, the analyst is typically not expecting the issue of 'language barrier' to come up when all agents are native speakers of the language as against off-shore agents.

The search for unexpected patterns is typically driven by the intuition, instinct and experience of individual analysts, and leads to repeatability concerns. This aspect of the problem would benefit the most from automation that throws out all meaningful patterns at the analyst, but is also significantly harder given the high dimensionality of the search space. We present Correlation Finder in Section 3.5 that assists the analyst's search for insightful patterns. We found it to work satisfactorily for reasonably large data sets. However, we are still far from addressing the longer term challenge of scalable insight search to reduce 'time to insight'. Learning from past analyst behavior in similar engagements is one possible avenue to explore.

### *Creation of Derived attributes.*

A task that is very related to assisted insight generation is creation of derived attributes or dimensions from the original data. When classes are not linearly separable, for example, transformations such as kernel methods[7] are popular in the machine learning literature. But finding the right transformation automatically is a severe practical challenge. Also, not all data dimensions need to be transformed similarly. This is another area where the domain expertise of the analyst is invaluable. It is easy for the analyst to know,

for example, that the dissatisfaction of the customer would depend on the time taken to resolve her issue rather than the actual dates of reporting and resolution of the problem. This becomes a non-trivial task and a repeatability challenge in the presence of multiple data sources as in *IVOCA*. To help the analyst in this role, a framework for learning and suggesting derived attributes would be of immense value. We are yet to address the bigger problem but have created a Derived Attributes Framework for *IVOCA* presented in Section 3.4, which analysts found to be very helpful.

### *Reporting.*

The managed service approach is unique in its report delivery model through the liaison between technology and business. The analyst needs to provide actionable insights and supporting material in concise visually appealing form, and typically spends a significant amount of time in preparing reports (graphs). The skill and expertise of individual analysts are often critical for this stage and ensuring repeatability across analysts is a significant challenge. OLAP cubes (typically a handful of dimensions) have found favour with BI practitioners because they capture the idea of patterns in multi-dimensional data, but are unwieldy and unmanageable when the number of interesting dimensions becomes large. Learning *what reports need to be generated for what data and what is the visual representation of an actionable insight*, is an open research problem. We address some aspect of injecting reporting automation into data flows, in Section 3.6, thereby significantly reducing 'time to insight' and aiding repeatability.

### *Assetization and Reuse.*

In analytics service offerings, no two engagements are identical. Data sources, data forms, domain knowledge, use cases, delivery models, business metrics, all can vary. This is the primary reason why off-the-shelf tools and frameworks do not work as-is in these scenarios and customized solutions need to be designed. This is also the reason why service engagements can take weeks to deliver results. While it is true that two engagements are almost never the same, when we consider the analytics work-flow, many components are similar across work-flows, with varying degrees of similarity. For components that involve the analyst, identifying and leveraging the similarities across work-flows, has the potential of drastically cutting down on the time and effort that the analyst has to spend in any particular engagement. Therefore, assetization and reuse of developed assets forms a crucial task in the analytics services business. Also, very significantly, the same analyst may not be involved in all engagements. Assets are critical for transferring analyst expertise across engagements for ensuring repeatability.

As an example, a significant time is spent by the analyst in identifying, defining and refining the domain concepts in any engagement, which are implemented as annotators. As can be imagined, most of the concepts and annotators developed for an automotive client, would be very different from those required for a tele-communications client; they will have to be created from scratch. However, what is more noteworthy, is that *many* customer-related issues and concepts would be similar, such as dissatisfaction with service, warranty issues etc., and when customer-care is involved, most agent and contact center related issues. The similarities would be more pronounced when moving from one automotive client to an-

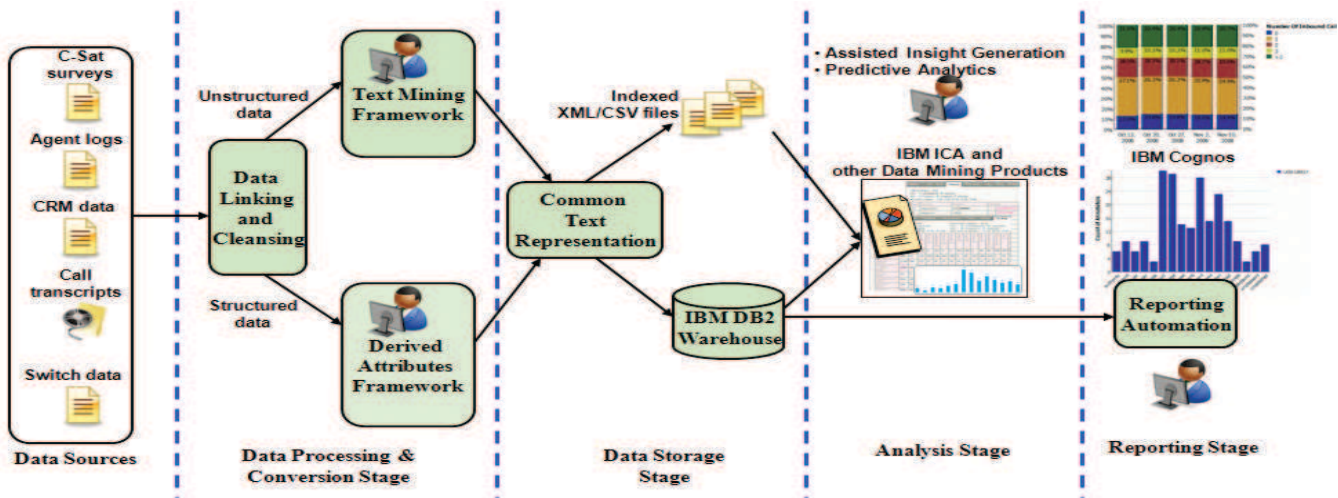


Figure 1: Data flows in *IVOCA*

other, and most annotators that were developed for the first one are expected to be applicable again. It therefore makes sense to create industry assets out of ‘domain-independent’ and ‘domain-specific’ annotators which can then be reused with minimal refinement when necessary. Other aspects of the process, apart from creating annotators, that would benefit from assetization and reuse are data flows and predictive models. This opens up new challenges for the areas of transfer learning and domain adaptation. We revisit assetization in Section 5.

We have attempted to address the above challenges with reasonable levels of success in our *IVOCA* framework. The next section describes the different tools we have built.

### 3. IVOCA COMPONENTS

In this section we describe the architectural flows and components of *IVOCA*. We emphasize on the tools developed with the aim of enhancing the productivity of the analyst in a managed service engagement for CRM analytics. We describe next a typical analytics process centered around the analyst, correlating it with the data flows shown in Figure 1. *IVOCA* is designed to ingest data (structured as well as unstructured documents) from various *data sources* (Section 3.1). These are first cleansed and integrated to create a single record for each business entity (customers, transactions, problem tickets) in the *Data Linking and Cleansing step* (Section 3.2). The unified record provides a single view of the entity across the different disparate data sources. Next is the *Data Processing and Conversion stage* where the structure and unstructured data are processed differently with the analyst involved as shown at various points in Figure 1. The Derived Attributes Framework described in Section 3.4 is used to create new attributes from existing structured attributes. The unstructured data passes through the Text Mining Framework where various tools described in Section 3.3 assist the analyst create meaningful structured summaries from the unstructured data. The enriched records containing several new derived attributes, enter the *Data Storage stage* and are stored in a IBM DB2 database as well as indexed semi-structured XML or comma separated CSV files for upstream applications. We do not

describe the data storage stage further since it does not involve the analyst (who is our focus).

The *Analysis stage* consists of 1) IBM Content Analyzer (ICA) built on UIMA<sup>3</sup> as the text processing and mining engine, 2) Cognos as the BI and reporting engine, and 3) industry standard data mining products (such as SPSS Clementine or SAS Enterprise Miner) for data mining and predictive modeling. The analyst uses all these products and other tools described in Section 3.5 to derive business insights by analyzing the data. Cognos also doubles up as our reporting platform in the Reporting stage; we describe some automation efforts to aid the analyst in Section 3.6. All IBM products mentioned are from the Information Management software family<sup>4</sup>. The architecture is not wedded to any of these tools and is intended to adapt to any other combination of products as the situation demands. Gluing different products to work together seamlessly is non-trivial in itself, requiring design of data interchange formats and tooling, but is again not analyst centric. The data and process flows around these products are enabled by custom developed Java utilities and data interchange takes place using XML/CSV files.

#### 3.1 Data Sources

In *IVOCA* we dealt with a variety of data sources common to CRM analytics. We list some of the contact center data sources and their characteristics from our automotive client shown in Figure 1:

- Case management data from a CRM system (such as Siebel) about customer-agent interactions. This contains structured demographic and interaction data.
- Detailed agent logs in the form of activity records. This contains structured as well as unstructured text data; structured attributes are activity types, people handling activities, date and time stamps; the text logs are extremely detailed notes about each agent activity, closely recreating the exact turn by turn interaction.
- Transcripts of recorded phone conversations suitably converted by automatic speech recognition[14] (ASR).

<sup>3</sup><http://www.research.ibm.com/uima>

<sup>4</sup><http://www.ibm.com/software/data>

This conversion, as is well known, involves a word error rate (WER) of 25% or more [16].

- CSat survey records of a small sample of customers. Third party vendors usually perform surveys and record structured scores and free-text comments of customers.
- Telephone switch data that contains metadata about calls, hold times, and transfers between agents.

All these data sources are exported and available as CSV files, a standard export option, for subsequent *IVOCA* components. For the automotive contact center, a month's data was to the tune of 50,000 new cases and their many associated activities, calls, and surveys.

## 3.2 Data Linking & Cleansing

Integration of data from various sources is a challenge in *IVOCA*, as in many data mining tasks. However in most of our engagements, creating a unified record centered around entities was easy as almost all data sources could be linked using unique identifiers (case ids) maintained across data sources. Call transcripts, however, is one data-source where linking records based on names and numbers is not possible due to high WER. For such cases, we had to resort to more sophisticated linking technology based on inexact matches [16]. Telephone switch data was another hard to link data source; network level data about hardware (like phones, switches, hold times, durations) did not contain application level data like case ids or agent ids. Approximate linking based on start and end times of calls was done by linking with the times recorded in the CRM system. Some cleaning was also required across file types including cleaning special characters, handling delimiters and newlines. We do not go into further details of this step since it is not focused on the analyst.

## 3.3 Text Mining Framework

Mining information from different unstructured or textual data sources forms one of the most critical components of *IVOCA*. We use two types of concept extractors for *IVOCA*. The first is *concept annotation*, where *segments* of text such as words and phrases are tagged. The second is at the document level, where an *entire document* is tagged as belonging to one of multiple categories (using text classification), which we call *category annotation*. The categories are usually poorly defined and the training data is noisy since it is hard for an analyst working under time constraints to prepare a training set for supervised learning of a set of concepts. Supervised classification fits in our text mining framework as the situation sometimes demands, and it has been reported earlier in detail[5]; we do not discuss it here further. Rule-based concept annotators are easily applicable and we focus next on this task, from concept discovery to improving rule-based annotators.

### *Concept Discovery.*

Typically, coming up with an exhaustive list of relevant concepts for any domain is challenging even for an experienced analyst without inspecting all of the documents. A concept can be defined as a group or cluster of semantically related words and as such may be discovered automatically by clustering together words that occur in similar contexts or in the same documents. There are myriad commercial (SAS, SPSS) and open-source (Weka, CLUTO) products

and workbenches for exploration and clustering over text corpora which include techniques like k-means and EM.

We use an internal text mining workbench for clustering[15] in *IVOCA*. This workbench provides for user-guided feature engineering, k-means clustering, and visualization that shows distribution of indicative cluster words contrasted against their background distributions. While the clustering process can jump-start the concept definition process, the domain knowledge of the analyst is critical in 1) assigning semantic labels to word clusters, 2) discarding trivial, irrelevant or un-interpretable clusters, and 3) pruning irrelevant or incorrect words from individual clusters. The discovered clusters are presented to the analyst and she can perform each of the above operations with ease. Once the clusters have been refined to the analyst's satisfaction, they can be used to define concepts for annotating the textual documents.

### *Improving Rule-based Annotators.*

We use dictionary and rule-based annotators for tagging documents with each concept provided by the analyst in the form of a small dictionary. The initial dictionary is created from domain knowledge and using the most descriptive words in clusters from the concept discovery phase above. In our example of finding reasons for low CSat of customers of the automobile company, let us imagine that the analyst finds a concept around *discounts* i.e. customers complaining of not receiving, or not receiving enough, discounts in the form of rebates, coupons, certificates etc. She creates a dictionary for this concept containing the words *discount*, *loyalty*, *certificate*, *coupon*. The problem with this annotator is that it will have high precision, but very low recall, stemming from the inability of the analyst to create an exhaustive dictionary for this concept. The low recall is made worse by very high levels of textual noise arising from the use of abbreviations, slang, shorthand, and misspellings.

We created a *Synonym Finder* tool for the analyst that starts from a small list of seed words in an initial dictionary and automatically expands it by finding semantically similar words from the document collection. This is done by finding words that are contextually similar to the seed dictionary words. We first build a context vector for each word by finding words that occur within a window of length  $k$  around the word, create TF-IDF weights for the context vector, and use cosine similarity to find words that have similarity above a certain threshold to the seed dictionary words. This is similar in spirit to earlier literature [8], and Google Sets<sup>5</sup>. The tool supports unigrams, bigrams and trigrams. Since the analyst uses the tool in an on-line manner, the tool has to be efficient in coming up with the expanded dictionary. This is achieved using an inverted index on the context vector. The recall of the concept annotators can be greatly increased by the use of this tool as shown in Section 4. Of course, the tool cannot be completely accurate in coming up with synonyms. This is where the domain knowledge of the analyst again comes in handy. The tool returns a ranked list of synonyms to the analyst, who discards the incorrect synonyms. Note that discarding from a ranked list is a significantly easier task for the analyst than creating an exhaustive dictionary from scratch in the presence of noise. For example, some of the unigrams in the final 'discounts' dictionary were *loyalty*, *certificate*, *coupon*, *discount*, *rebate*,

---

<sup>5</sup><http://labs.google.com/sets>

*discont, reb, cpn, money, dollar, reimb, credit, notice, cash;* note the noisy variants found.

Starting from clustering, to reviewing, refining, enhancing the set of concepts from domain knowledge, and using Synonym Finder to expand the list of n-grams, the analyst is now able to construct powerful concept annotators. Annotators are deployed as *patterns* in ICA, though any rule-based tagger using dictionaries could be used (ICA additionally has indexing and a visualization UI). Hence annotations is one of the primary ways the *IVOCA* analyst derives meaning and structure out of unstructured text. Annotation hits or counts at a document level are then used as attributes derived from text.

### 3.4 Derived Attributes Framework

*IVOCA* provides a configurable framework for defining derived attributes on structured data as shown in Figure 1. As an example, derived attributes were crucial for our activity logs data source described in Section 3.1. The raw data contains many activity records for one particular case describing the activities an agent needs to perform to resolve the case in question. The structured attributes need to be transformed to be helpful (refer Section 2) and three kinds of transformations are typically needed: 1) groupings of activities by type and their summary, 2) time elapsed (average, minimum, maximum) between two specific (groups of) activities, and 3) count, average, and other aggregates of specific (groups of) activities. The framework is an efficient configurable implementation that scans the input file linearly, reads all activities of a particular case, and implements the derived attributes based on configurable inputs. For example, some complex derived attributes from the activities structured data were a) time in hours elapsed between an inbound customer call and the next outbound agent call to that customer, b) number of times ownership of a case is changed from agent to agent. Some of these attributes can be computed from structured data in a database using SQL self-joins, but some require more complex aggregation functions than those typical available in SQL. The effectiveness of our framework in creating greatly beneficial derived attributes for predictive modeling is discussed in Section 4.

### 3.5 Analysis Stage

Next we describe assisted insight generation and predictive modeling in the Analysis stage of Figure 1.

#### *Assisting Insight Generation.*

The role of tooling in insight generation is to present all interesting correlations among data dimensions to the analyst for interpretation. The goal is to increase recall and aid repeatability by guaranteeing that important obvious insights are not overlooked by any analyst.

To this end we built a web-based *Correlation Finder* utility for the *IVOCA* analyst to point out all interesting correlations spread across data sources. Lots of original and derived attributes are available from structured as well as unstructured data as described previously. The analyst now points out target attributes of interest such as CSat or satisfaction with agent to *Correlation Finder*. Visually checking all possible slices of data in a cubing sense is infeasible, but a program can easily find striking patterns such as unusually high correlation in the cubes. *Correlation Finder* takes the target attribute of interest and computes correlation (or

other measures) with all other structured/unstructured derived attributes, logically ‘AND’ing up to 2 or 3 attributes (more complex combinations may not be *interpretable*). The web-based output is a ranked list of correlations or patterns that exist in the huge volume of data. The analyst can sift through these visually, inspect data as evidence, and decide whether they are good candidate insights to be developed further. The analyst can sort the output by correlation values, types of structured attributes, annotator types, and frequencies. Such a tool is intended to assist *recall* of candidate insights in exploring a dataset by pointing out correlations an analyst is likely to miss in manual exploration. Thus this tool aids repeatability across analysts, as will become clear with the example presented in Section 4.

#### *Predictive Modeling.*

Predictive modeling using an industry standard data mining product is one of the major tools of the *IVOCA* analyst. The goal typically is to detect undesirable events early based on historical data, so that appropriate action may be taken to prevent them from happening. An example use-case of predictive modeling in our example of CSat analysis is first *predicting CSat for open cases*, so that cases where the customer is likely to give a poor rating can be pro-actively salvaged. Historic training data is available in the case management system. Along with a rich set of derived attributes from structured and unstructured data described in previous sections, we can learn to predict whether open cases will give rise to satisfied/unsatisfied customers (results in Section 4.1).

The role of the analyst in interpreting results of data mining tools and bridging the technology vs. business gap becomes important at this point in the managed service analytics process. The analyst almost always chose predictive models such as decision trees over complex models because of interpretability of the ‘if-then-else’ rules output. The other significant role of the analyst in predictive modeling was choosing parameters such as mis-classification costs in skewed data distributions, sampling strategy for time varying data, and feature engineering.

### 3.6 Reporting stage

A typical *IVOCA* engagement cycle ends with the analyst presenting a set of reports and actionable recommendations to the client. A lot of time is usually spent in creating an effective set of reports (graphs) that convey the impact and validity of the insights. One of the components of *IVOCA* is a reporting automation and streamlining function that is simple yet practical, bringing down the analyst’s time spent on report preparation from days to hours. The reporting component of *IVOCA* uses a set of Cognos templates that are standardized by industry type and customized once for every engagement. The templates recognize that most reports plot a data series (some combination of dimensions) against a measure (such as mean CSat score). *Correlation Finder* also populates template tables in the database and reports are automatically generated using the Cognos SDK. The analyst visually inspects the many automatically generated graphs and starts the observation → insight derivation activity from there. Predictive analytics flows are also streamlined to work seamlessly with Cognos; scores output from tasks such as CSat prediction are output to database template tables and reports are automatically generated.

Sat	C	U	C+U	A	A+C	A+U	A+C+U
62.4	73.5	71.3	75.6	80.5	81.6	80.3	81.8

**Table 1: CSat prediction accuracy using various structured and unstructured attributes**

## 4. EXPERIMENTAL RESULTS

In this section we investigate the effectiveness of the tools developed in *IVOCA* in addressing the problem that we had set up right at the beginning - namely, reducing the business analyst’s time to insight. Measuring repeatability is harder; we consider it indirectly through the usefulness of the assets that we developed as part of the solution. We report results from an *IVOCA* engagement with an automotive client. While time-to-insight is reduced, correctness of the delivered insights cannot be compromised. So we first measure the accuracy of our CSat predictive modeling task using the various data sources as input. The value of unstructured data sources for analytics tasks is often debated in the community. We do a careful study of the various attributes derived from structured and unstructured sources in terms of their usefulness in our prediction task. Correctness of overall insights is again hard to measure. Instead, we provide examples of interesting hidden insights that were found by our Correlation Finder tool and were deemed to be extremely useful by our analysts. We measure accuracy at the level of the concept annotators for which ground-truth is relatively easier to prepare. These experiments bring out the usefulness of our annotators in greatly reducing time-to-insight without significant deterioration of accuracy.

### 4.1 Predicting CSat

First, we consider the CSat prediction task and the usefulness of the different attributes in the data. We set up a binary CSat prediction task (satisfied and dissatisfied customers) on one week of data from the automotive contact center. The goal is to learn to output a list of customers who are most likely to be dissatisfied, so that follow-up action may then be taken to address their grievances. There are more satisfied cases in this dataset than dissatisfied cases (skewed class distribution), mirrored in a dummy predictor (denoted *Sat*) always predicting satisfied. We used J48 decision trees, naive Bayes, logistic regression, and SVM classifiers in Weka<sup>6</sup> with default parameters to predict CSat. Weka is not a part of *IVOCA* which uses other commercial data mining products, but we report with Weka for accessibility. Interpretability of results being critical for the analysts, they preferred decision trees; results were not very different for other classifiers. 10-fold cross-validation accuracies over 8200 labeled examples with decision trees are reported in Table 1.

The table groups the input attributes under three categories to predict CSat. The first category (denoted *C*) was derived from the structured CRM data, and included attributes such as customer/agent/contact center information, attributes about the customer’s vehicle, and reason codes assigned by the agent to the call. The second category (denoted *U*) was derived from the unstructured data sources such as textual agent logs in activity data using various concept annotators and is represented as annotator counts

<sup>6</sup><http://www.cs.waikato.ac.nz/ml/weka/>

described in Section 3.3. The third category (denoted *A*) includes complex derived attributes from structured activity data as described with examples in Section 3.4. Combinations of the three types are labeled appropriately. Comparison of the individual data sources shows that *C* is only slightly more useful than *U*, while *A* is the most accurate predictor individually. This is readily explained by the richness and level of detail in the agent activity logs and the complex attributes that could be derived from it. The results show that for the specific task of CSat prediction in this engagement, there is no significant additional value in the unstructured sources beyond what is already available from *A*. However, such rich detailed structured records are not always available, and in such scenarios, we need to fall back on the textual and other data sources which do quite well beyond the baseline predictor. Moreover, while unstructured data may not be the best predictor for CSat, insights however are very impactful when derived from both structured and unstructured data sources. More importantly, the value of the unstructured data is clearly brought out in the Insight Generation task, which we address next.

### 4.2 Insight Generation

The second task of an analyst is to identify interesting patterns in the data and determine what actionable insights can be derived to correct problems such as low CSat or customer churn. In this regard, the Correlation Finder tool was able to make significant contributions beyond ICA and the data mining tool used for predictive modeling. Correlation Finder is able to discover patterns across data sources and presents them to the analyst in a ranked order, and, as such, begins to address the hard challenge of insight discovery mentioned in Section 2. In our engagement with the automobile company, Correlation Finder was able to find several unexpected insights which the analysts acknowledge would have been hard to identify otherwise. We list some sample patterns below:

- There was high correlation between low CSat score and complaints about dealerships in the CSat surveys
- There was high occurrence of mentions of dealers and dealerships in customer agent call transcripts
- The activity logs of the agents recorded detailed customer complaints about dealers and dealerships

All these three observations are meaningful candidate insights by themselves. What is striking is that the same dealership annotator finds these high correlations with low CSat ratings. Taken together, these lead to a much stronger impact insight, raising the priority of investigating or recommending corrective actions against problematic dealerships (identifiable from structured data). The Correlation Finder output, sorted by annotator type and correlation values (Section 3.5), immediately throws up this striking trio of observations. This insight is now likely to be repeatably discovered by any analyst; without this tool these observations would have been lost among many other observations.

### 4.3 Precision and recall of annotators

Finally, we address the twin issues of accuracy and time to insight. Overall, the analysts in our team are of the opinion that the insight discovery process has been reduced to hours from days using the tools currently provided in *IVOCA*. Accuracy is hard to quantify at the insight level. Instead, as

Concept	Annotator	Dictionary	Rules	Precision	Recall
Timeliness		hours, minutes	number followed by dictionary mention	60.4%	7.2%
		add 15 synonyms	longer phrases	55.4%	9.4%
		all synonyms	combine dictionaries, out-of-order mentions	48.3%	24%
		Manual Ref	as above	45.9%	58.7%
Parts		brake,tire,engine>window,motor,light	dictionary mention	88.2%	24.3%
		add 15 synonyms	longer phrases	74.7%	37.1%
		all synonyms	combine dictionaries, out-of-order mentions	77.5%	49.1%
		Manual Ref	as above	80.5%	63.2%
Agent		helpful,polite,nice,friendly,pleasant	dictionary mention	94.7%	29.8%
		add 15 synonyms	longer phrases	93.2%	48%
		all synonyms	combine dictionaries, out-of-order mentions	82.7%	57.5%
		Manual Ref	as above	75.4%	64%

**Table 2: Annotator precision and recall**

a substitute, we take a closer look at the accuracy of the various concept annotators that were created as part of the engagement and the reduction in effort that came from using tools like Synonym Finder.

We picked three annotators for this experiment as a representative sample: 1) timeliness issues, 2) mentions of positive attributes of agents by the customer, and 3) problems with car parts. Car parts is a very domain specific concept and also relatively easy to define. Agent appreciation is generally decoupled from the domain and is also not too hard to understand as a concept. The timeliness concept is also domain independent, but significantly harder to define objectively; humans disagreed about whether mere mentions of any length of time in the comments should be tagged as a timeliness issue. A set of 630 customer survey comments were manually tagged (labeled) with these annotators.

We measure precision and recall at various stages of the annotator development process. In the first step, the analyst inspects clusters and picks the top 5 words from the cluster that best describe the concept and includes them in the annotator dictionary (only hours and minutes are used to seed timeliness). In the second step, the dictionaries are expanded using the top 15 synonyms from Synonym Finder (examples in Section 3.3) and by adding rules to handle longer n-gram synonyms. In the third step, all synonyms of the seed words returned by the Synonym Finder that have similarity above a cosine threshold of 0.2 are added to the dictionaries after analyst review, and rules are added to include out-of-order mentions for phrases, and combining mentions from multiple dictionaries. For example, one rule for identifying a positive agent comment is: a word/phrase from the ‘agent’ dictionary, one wildcard, followed by a word from the positive attributes dictionary. This rule identifies the phrases ‘guy was polite’ and ‘agent is nice’.

For reference, we also include a *manual reference* step, where the dictionaries and annotators are built by going over the manually tagged annotations. Clearly this can pick up rare synonyms and misspellings that Synonym Finder can miss. To avoid overfitting, only words/phrases occurring in manual annotations more than twice were included in the dictionaries (thus moderate recall).

First, we look at the accuracies achieved by these different steps in Table 2. We can see that, as expected, recall is increased at the cost of precision, over steps 1 to 3. Notably, the accuracies at step 3 are quite close to that achieved using manual review of the labeled data. The precision levels are quite similar, and with the exception of the ‘timeliness’ an-

notator, recall levels are only marginally higher after manual refinement. The reason for the difference is that ‘timeliness’ is an inherently difficult concept to define using simple rules and due to sense ambiguity of words. For such cases, where achieving good precision and recall simultaneously is hard, the analyst has the freedom to choose the operation point.

Coming to ‘time-to-insight’, the manual tagging and manual reference annotator construction took over 15 hours. In contrast, the first three steps involving tool execution times, analyst review of output, writing rule patterns, and running ICA patterns took a little under an hour. This shows that our annotation construction process that engages the analyst optimally and provides him with tools to work efficiently is able to create annotations that are quite close to that obtainable from an analyst-intensive approach in terms of accuracy, and the savings in time can be quite substantial.

In summary, the results show that we have been reasonably successful in *enabling IVOCA* analysts in providing actionable business insights quickly, accurately, and repeatably over various data sources.

## 5. EXPERIENCES

Finally, we review where *IVOCA* stands as a managed service offering for CRM analytics, by describing the nature of ongoing and new client engagements, and contrasting with earlier efforts in CRM settings. We also recount our experiences and lessons learned from the design and development of *IVOCA*. Lastly we outline future research directions.

After successful pilots, *IVOCA* is now available as a managed service. Business analysts have been able to efficiently use the various tools we developed and found that they reduce the ‘time to insight’ from *days to hours*. The analysts have been effectively trained to inter-operate the various products glued together by tools to derive actionable insights around areas of CSat drivers, issue resolution hurdles for agents, and customer retention efforts from the data sources in client contact centers. The actionable insights have led to recommendations such as process changes in the contact centers, agent lists for targeted training, and customer lists for prioritized issue resolution among others. A good amount of repeatability has also emerged in the analytics due to the tools we developed for the analysts. These factors show services like *IVOCA* as the best alternative to organizations investing internally for CRM analytics.

### *Related efforts.*

CRM analytics has been an important application area



in the rapidly growing services industry. A wide range of business problems such as CSat prediction, churn management, customer retention, cross-sell & up-sell opportunities have attracted attention from customer facing arms of organizations. Solutions have been centered around analytical entities such as products, marketing, sales, and interactions[1], or have applied data mining to various aspects of the customer[6]. The data spectrum has seen analytics efforts depending on whether the data is structured data residing in databases, unstructured text data, or large volumes of conversational recorded speech data. Traditional BI solutions concentrate on one or two kinds of data. They either provide reporting and dashboarding solutions on CRM data warehouses[1], or they provide text analytics solutions on mails and surveys[4, 5, 2], or there are speech analytics solutions such as sentiment extraction or issue identification in customer calls[9, 16].

These efforts have been undertaken by organizations and researchers mostly as proof of concept pilots or by applying commercial data mining products to solve business problems. Though such efforts succeed with time and effort, *IVOCA* presents a wholly new setting of a completely managed service approach. We believe a lot of high impact insights are only evident when gleaned from many data sources (Section 4). Our focus is not on particular CRM analytics tasks but enabling a services organization to make viable analytics offerings, and help business analysts achieve this by engaging them in optimal feedback conversations.

### Lessons learned.

As we worked on building *IVOCA*, it was evident that the success of a managed service approach to analytics depends on how asset-based, repeatable, and adaptable the *IVOCA* platform is to different needs in the CRM analytics domain. One of our main learnings was the stress imparted on assets by our service provider arm offering *IVOCA*. One of the biggest challenges lies in being to able to seamlessly utilize various data forms and sources. From the analyst's point of view the ability to transcend the structured-unstructured data chasm with ease was the greatest improvement. The analyst greatly appreciated our tools that helped her use text mining and derive meaningful concepts and structure out of myriad unstructured sources.

Repeatable was another desirable aspect in the solution design – an analyst should not only be able to repeat similar analytics tasks, say CSat prediction or issue identification, in different engagements using the tools at hand, but different analysts should also be driven toward a similar set of insights on a particular dataset. The right set of tools should bring up the significant observations in the data to all analysts, and thus assist them equally in reaching the same insights and recommendations. It was revealing to see how some annotators (contact center and agent specific ones) from the automotive domain were directly applicable with little adaption to a consumer electronics client's data. With minimum modification, the CSat predictive modeling flows are also usable for other clients.

A major component of *IVOCA* was enabling the services arm to carry out complex analytics engagements with little intervention from researchers. This was a maturing of sorts from research prototype projects, where deployments and successes were not guaranteed for lack of business understanding[5, 16]. However as the *IVOCA* offering devel-

oped, the segregation of the roles of research, development, and service offering became clear. The researchers focused mainly on the tasks highlighted in the paper. The development team excelled and software development, testing, and maintenance – tasks they do best. The service offering led by the business analyst liaised between the technical (research and development) teams and the client to meet expectations and drive actionable insight generation.

### Future work.

*IVOCA* is work in progress and, as stressed, there is ample room for improvement. Our experience has brought out two important directions for further research to ensure analytics deployments become more widespread. First, research and tooling to help reduce an analyst's 'time to insight' will be useful. We believe the analyst should concentrate on insight postulation and development rather than spending time only exploring data. Second, an new research opportunity is presented by the assetization angle to aid 'repeatability' across analysts and engagements.

## 6. REFERENCES

- [1] SAP customer relationship management. [www.sap.com/solutions/business-suite/crm/index.epx](http://www.sap.com/solutions/business-suite/crm/index.epx).
- [2] M. Bhide, A. Gupta, R. Gupta, P. Roy, M. Mohanaia, and Z. Ichhaporia. Liptus: Associating structured and unstructured information in a banking environment. In *Proc. of SIGMOD*, 2007.
- [3] D. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. In *Proc. of NIPS*, 1995.
- [4] D. Garg, N. Kambhatla, and M. Vukovic. Mining top issues from contact center logs for self help portals. In *Proc. of SCC*, 2008.
- [5] S. Godbole and S. Roy. Text classification business intelligence and interactivity: Automating CSat analysis for services industry. In *Proc. of SIGKDD*, 2008.
- [6] G. Hershel. Magic quadrant for customer data mining applications, 2008.
- [7] T. Joachims. *Advances in Kernel Methods: Support Vector Learning*, Editors B.Scholkopf, C.J.C. Burgess, and A.J. Smola, chapter Making large-scale SVM learning practical. MIT Press, Cambridge, MA, USA, 1999.
- [8] R. Jones, A. McCallum, K. Nigam, and E. Riloff. Bootstrapping for text learning tasks. In *IJCAI Workshop on Text Mining*, 1999.
- [9] C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005.
- [10] H. P. Luhn. A business intelligence system. In *IBM Journal*, 1958.
- [11] C. Monash. *Analytic Business Processes: The Third Generation*. In *White paper available at <http://www.monash.com/whitepapers.html>*, 2004.
- [12] S. O'Dowd. Unstructured data and text analytics in capital markets. In *Financial Insights Report - IDC report*, 2008.
- [13] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proc. of KDD*, 2008.
- [14] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig. The IBM 2004 conversational telephony system for rich transcription. In *Proc. of ICASSP*, 2005.
- [15] S. Spangler and J. Kreulen. Interactive methods for taxonomy editing and validation. In *Proc. of CIKM*, 2002.
- [16] L. V. Subramaniam, T. Faruque, S. Iqbal, S. Godbole, and M. Mohanaia. Business intelligence from voice of customer. In *Proc. of ICDE*, 2009.