

LIPTUS: Associating Structured and Unstructured Information in a Banking Environment

M. Bhide¹ A. Gupta¹ R. Gupta¹ P. Roy¹ M. Mohania¹ Z. Ichhaporia²

¹IBM India Research Lab, New Delhi, India

²HDFC Bank Ltd., Mumbai, India

{abmanish, ajaygupta, rahulgupta, prasanr, mk mukesh}@in.ibm.com
zenita.ichhaporia@hdfcbank.com

ABSTRACT

Growing competition has made today's banks understand the value of knowing their customers better. In this paper, we describe a tool, LIPTUS, that associates the customer interactions (emails and transcribed phone calls) with customer and account profiles stored in an existing data warehouse. The associations discovered by LIPTUS enable analytics spanning the customer and account profiles on one hand and the meta-data associated or derived from the interaction (using text mining techniques) on the other. We illustrate the value derived from this consolidated analysis through specific customer intelligence applications. LIPTUS is today being extensively used in a large bank in India. A highlight of this paper is a discussion of the technical challenges encountered while building LIPTUS and deploying it on real-life customer data.

Categories and Subject Descriptors: H.2 [Database Management]: Systems - *Textual Databases*

General Terms: Algorithms, Design, Experimentation

Keywords: Customer Intelligence, Customer Support, Information Integration

1. INTRODUCTION

Growing competition has made the today's banks understand the value of knowing their customers. They are eager to understand the customers' concerns so that they can serve them better. If a customer leaves, they want to know what the complaint was, so that they can prevent any further attrition the best they can. They want to understand the changing needs of the customers in a timely manner, and use it to introduce new products and services, as well as to improve and personalize the existing ones. A bank typically has a "customer intelligence" setup that tries to mine such information from the available structured data such as the customer's account balance, transaction frequency, product holdings, demographics, etc. While such data is helpful, it is

essentially indirect in nature and therefore unable to provide a complete picture.

Customers, on the other hand, regularly interact with the bank by sending emails, calling up or walking in a bank branch and meeting a banker. Most banks have a "customer support" setup that takes care of such interactions, which could for instance involve complaints about the service, or inquiry about a new product being introduced. For their own records, the banks typically consolidate and archive these interactions; once archived, however, these interactions are put to little use.

Ideally, the customer intelligence analytics should be able to exploit the valuable customer interactions available with customer support. The reason why this does not happen is purely technical. First, since the customer interactions are not tagged with customer or account ids, there is no direct way to "join" an interaction with a customer or an account. Second, the customer intelligence analytics works on clean, structured information, while the customer interactions that are available with the customer support are essentially free-flow unstructured text.

In this paper, we describe a tool, LIPTUS,¹ that addresses these issues. LIPTUS automatically associates the customer interactions (emails and transcribed phone calls) with customer and account profiles stored in an existing database. The associations discovered by LIPTUS enable analytics spanning the customer and account profiles on one hand and the meta-data associated or derived from the interaction (using text mining techniques) on the other. We illustrate the value derived from this consolidated analysis through specific customer intelligence applications. LIPTUS is today being extensively used in a large bank in India. A highlight of this paper is a discussion of the technical challenges encountered while building LIPTUS and deploying it on real-life customer data.

Overview. The various components of LIPTUS and the corresponding process flow is shown in Figure 1. LIPTUS takes as input the customer interactions, available as text files stored in a content management system, and heuristically extracts the customer and account identifiers mentioned in the text. These extracted identifiers are then matched with the identifiers present in the customer and account profiles (such as customer ids, credit card or bank account numbers) and the best matching profile is then linked with the interaction. This linkage consolidates the infor-

¹Linking and Processing Tool for Unstructured and Structured information

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.

SIGMOD'07, June 12–14, 2007, Beijing, China.

Copyright 2007 ACM 978-1-59593-686-8/07/0006 ...\$5.00.

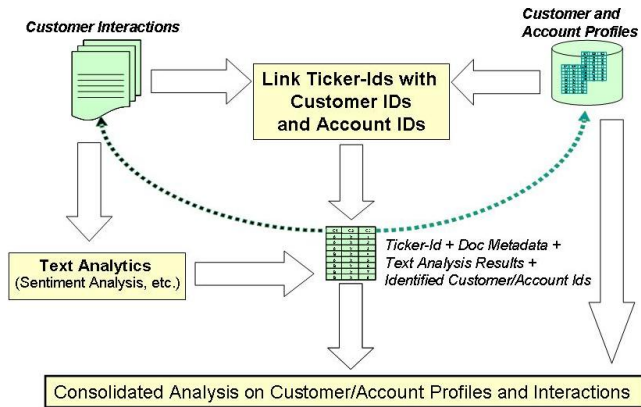


Figure 1: LIPTUS Overview

mation available in the customer profile (customer product holding, profitability, etc.) and account profile (account type, usage, loyalty, age, etc.) with the information available with the interaction (date, purpose, etc.). In addition, as shown in the figure, LIPTUS also applies text classification and information extraction techniques (such as sentiment analysis, keyword extraction) to mine additional information from the interaction text. This combined information can be used by a variety of applications, including standard OLAP applications, to perform customer intelligence analysis not possible earlier.

Organization. The remainder of the paper is organized as follows. In Section 2, we provide details about the structured data (customer and account profiles) and unstructured data (customer interactions) available to the system. This is followed by a description of how LIPTUS finds the links between the customer interactions and related customer and account profiles in Section 3. Next, in Section 4 we describe the text analysis LIPTUS performs on the customer interactions, resulting in interesting characteristics of the interaction (e.g. satisfaction level) and, by association, of the customer that are not available otherwise. In Section 5, we discuss a few real-world use cases showing how LIPTUS is being deployed in a real-life environment. The prior work related to LIPTUS, both in research and industry, is discussed in Section 6. Finally, in Section 7, we present the conclusions.

2. INFORMATION INFRASTRUCTURE

In this section, we describe the structured and unstructured data sources that contain customer information in the banking environment we were engaged with.

Customer Profiles (Structured Data)

A customer may have multiple accounts with the bank, these accounts could either be in the same product line or across different product lines (current and savings bank accounts, credit cards, housing loans, mortgages, automobile loans, personal loans, mutual fund accounts, trading accounts, etc.). The customer information for each product line is stored in a different system.

Our environment had an elaborate setup that incrementally extracted the customer and account information from each of these underlying data sources and consolidated in a

“master” data warehouse. This resulting master customer profile not only included attributes such as customer name, address, contact, profession, geography, number of dependents, marital status etc., but also aggregates such as the set of accounts held by each customer, and the customer’s overall profitability across all these accounts. For each account, similarly, the account profile included detailed information about the account. For a savings bank account, for instance, the profile included the date of opening, average quarterly balance, date of last activity, fees charged till date, interest paid till date, etc. This consolidated information is updated once a month, and is regularly used to generate a variety of business intelligence reports for the marketing team as well as for other decision makers within the organization, and also for ad-hoc OLAP analytics.

The customer information present in the underlying data sources was provided by the customer at the time of opening the respective accounts, and part of this information could become stale over time. As the information across the different sources is aggregated, inconsistencies abound, compromising the quality of the aggregated information. Some of these inconsistencies are resolved by assuming that the most recently provided information is correct (the data available for the most recently opened account supersedes any conflicting data). For the remaining, ad-hoc heuristics are deployed, or all versions are maintained. Furthermore, some attributes in the customer profile have no data at all; these are the optional attributes in the applications that are rarely filled up by the customer. These issues of inconsistent and missing data make LIPTUS’s task of matching interactions with the customer profiles challenging; however, as we discuss later, the linking strategy in LIPTUS is designed to be robust despite such issues.

Customer Interactions (Unstructured Data)

Customer interactions, stored as text documents, form the unstructured data of interest. These could either be in terms of emails received directly from the customer, transcribed phone calls, or notes written by bankers on behalf of the customer (in case of customer walking in the bank, or sending a handwritten letter or fax).

Each interaction is identified by a “ticker-id”. A unique ticker-id is generated for the email or phone-call that initiates the interaction, and all related subsequent exchanges between the bank and the customer are threaded together using this ticker-id. In addition, as a part of the process, each ticker-id is manually classified into one or more of predefined categories; the categories assigned to an interaction identify its purpose (such as “credit card inquiry”, “cheque status inquiry”, “charge dispute”, “change of address request”, etc.).

The customer interactions are essentially free-flow text, and meant for human consumption. They can include significant amount of text that has no bearing to the discussion at hand. For instance, a mail sent to the customer may include an advertisement for a product recently launched by the bank; similarly, a mail sent by the customer through a free email service may include an advertisement as well. Equally useless are the standard “polite” phrases included in the bank’s responses to every mail it receives from the customer. Moreover, as the emails are exchanged, the history text is seldom deleted and therefore each email from either side has the text of the prior emails. All this redundant con-

tent tends to overwhelm the interaction content, and identifying the informative content in an interaction consisting of multiple mails is a nontrivial challenge.

The issues mentioned above are relevant to the bona-fide customer interactions. The customer support email address, being publicly known, gets messages from non-customers as well. Some of these non-customer messages could be potential sales leads, and can not be ignored. LIPTUS, as a side-effect of its linking process, is able to separate out the customer messages from the non-customer messages to a reasonable extent. The customer support receives junk mails (including job requests and resumes) as well; thankfully, these mails are eliminated from consideration as they are processed by customer support, and LIPTUS does not need to handle such mails.

3. LINKING CUSTOMER PROFILES WITH INTERACTIONS

In this section, we describe how LIPTUS associates the customer and account profiles (identified by the customer and account ids) with customer interactions (identified by the ticker-ids). To make the linking procedure more effective, however, LIPTUS first needs to “clean” the interaction text. The details of this cleaning step are described in Section 3.1. LIPTUS then matches the customer and account profiles with the cleaned interactions, linking each interaction with the right match; this step is described in Section 3.2.

3.1 Cleaning the Customer Interaction Text

The customer interactions contain a significant amount of irrelevant and redundant text (including irrelevant advertisements, disclaimers, canned greetings, text of earlier messages repeated as history, etc.). This useless additional text makes analysis of the interaction content not only slower, but also less effective since it tends to obscure the actual information contained in the interaction. In this section, we describe the cleaning steps that try to identify and remove the irrelevant and redundant text present in the transactions.

Given the absence of structure in the interaction text, it is hard to devise a perfect procedure for the cleaning task. Aiming for a best-effort efficient solution, LIPTUS deploys a handful of simple-minded heuristics that try to exploit the hints present in the text to identify the text to remove. Some of these heuristics are listed below. These heuristics worked very well on the interactions we analyzed, but we emphasize that these heuristics are fine-tuned for email interactions,² and might need to be modified for other type of interactions.

- *Remove the stock replies:* When the customer sends a message to the bank, the customer support immediately responds acknowledging the receipt of the mail, and ensuring prompt response. Such stock replies do not contain any useful information and can be safely removed from the interaction. Given their standard content, such messages are very easy to identify.
- *Remove the history text:* The customer often includes the history of conversation as she replies to the emails

²Other interactions, such as phone-call transcriptions, tend to be succinct enough.

sent by the bank as a part of the interaction. This history serves as the context for a particular email message, but is redundant when the entire interaction thread is available already. This history text is identified by looking for characters such as “>” at the beginning of the lines in the text, or identifying standard phrases such as “On <date>, <name> wrote:” (or its variations).

- *Remove the advertisements and disclaimers:* The email messages often have irrelevant text such as advertisements and disclaimers attached to them. In the emails sent by the bank, identifying such text is relatively easy – it is the same across multiple interactions and consists of standard phrases that can be compiled beforehand by manually analyzing a small sample of the emails (this set of phrases can change over time, though, and needs to be updated regularly). In the emails sent by customers, no such commonality exists, making the task much harder – at the moment, the advertisements and disclaimers in such emails are not removed.

3.2 The Linking Procedure

We now describe the procedure used in LIPTUS to link the (cleaned) customer interactions with the best matching customer and account profiles. The procedure consists of two steps. In the first step, the customer and account ids mentioned in the customer interactions are extracted. In the second step, these ids are used to identify and link with the relevant customer and account profiles in the database. We describe these steps in turn below.

Extracting Customer and Account Ids. This step takes as input the cleaned interactions, and extracts the customer and account ids present therein. Note that the interactions that are generated by the bank staff (transcribed phone calls or emails sent by a personal banker on behalf of the customer) are relatively structured – they usually have the customer and account ids already present as a meta-data; the information needed being already there, such interactions can bypass the extraction step described in this section. In contrast, in an email sent directly by the customer, these ids are mentioned in free-flow, unstructured manner, and are hard to trace automatically. The techniques mentioned in this section, therefore, are specifically geared towards emails messages.

This task is far more difficult than merely looking for numeric sequences in the text and then disambiguating these sequences based on the number of digits, prefix sequences and other patterns. This is because of a variety of reasons, some of which are listed below.

- The customer and account ids are formatted in a variety of ways in the email texts. For instance, the bank account and credit card numbers are often stated with hyphens or spaces in between. Hyphens and whitespace may also appear in case the id is split across two lines in the text.
- We know that the customer ids have six digits, bank account ids have nine digits, credit card numbers have sixteen digits, and so on. However, sometimes the customer chooses to omit the leading zeroes of her account number (the bank account id 000321675 appears as

321675); this means that the length of the numeric sequence is not a reasonable hint and it is hard to tell a bank account number from a customer id or even a currency value just by looking at the numeric sequence itself.

- The first few digits of a numeric sequence can be used as a hint for identifying the type of the number. The first four digits of a credit card number, for instance, are usually unique for a bank and the card type (Visa or Mastercard). The first three digits of a customer identify the branch where the customer first opened an account, and so on. However, these can lead to false positives – the system still cannot distinguish between customer id 110022 from the postal code 110022.

LIPTUS uses annotators based on the Unstructured Information Management Architecture (UIMA) [6] to identify the customer and account ids. At its simplest, an annotator tokenizes the text and applies pattern-based rules on the token sequence obtained to identify the interesting tokens (customer and account ids in our case). These rules combine the hints mentioned above (size of the numeric sequence, identifying prefixes) and take the presence of hyphens and whitespaces into account as well. Moreover, they also take hints from the surrounding text to identify the type of the id identified (for instance, a credit card number could be surrounded by the words such as “visa”, “mastercard”, and “expiry”). The annotator also takes hints from the category the interaction is associated with (“credit card inquiry”, “cheque status inquiry”, “premium payment”) to identify a small set of alternatives; a cheque status inquiry, for instance, can only relate to a savings or current account.

We again emphasize that this extraction process is essentially a best-effort solution, and there is a possibility of an incorrect sequence being extracted as a customer or account id, as well as of a valid customer or account id not being extracted. On the interactions we considered, however, we found that these simple heuristics performed well enough.

Joining Customer Interactions with Customer and Account Profiles. The extraction step outlined above identifies the set of customer ids and account ids (along with the corresponding account types) mentioned in each interaction. Further, LIPTUS validates each customer and account id identified in an interaction by checking whether or not it corresponds to a customer or account (of the given type) in the database; if a customer or account id is not found valid, it is discarded.

If only one customer id (and no account id) remains for the interaction after the pre-processing, then we do not have a choice and this customer id is considered the most relevant. Similarly, if only one account id (and no customer id) remains for the interaction, then this account id is considered the most relevant. The interesting case occurs when multiple customer and account ids remain.

A naive procedure would link the interaction with all the multiple customer and account ids present. But this would not be correct if, for instance, the customer interaction mentions money transfer (or cheque payment) from the her account to another customer’s account – we would not like this interaction to be linked to the latter customer’s profile. LIPTUS’s solution is to gather support for each customer or account id mentioned from the remaining information present in the interaction (customer name and other customer and

account ids mentioned) and eliminating the customer or account ids that do not have any support; the details follow.

LIPTUS first builds up the context of the given interaction as the set of valid customer and account ids identified as above, along with the name of the customer obtained from the email header (or the appropriate metadata in case the interaction is not an email). It also builds up the context of each customer id by querying the database and extracting the name of the customer and the ids of each account held by the customer. Similarly, it builds up the context of each account id by querying the database and extracting the customer ids and names of the account holders.

The support of a customer or account id in the interaction is computed as the size of intersection of the id’s context with the context of the given interaction. Clearly, the greater the support of an id, the more relevant it can be assumed to be to the given interaction. LIPTUS eliminates the customer and account ids with zero support and, among the remaining, identifies those ids with the greatest support as the most relevant to the given interaction.

The discovered links between the interactions (identified by their ticker-ids) and the customer and account ids are populated in a table within the database. This enables consolidated analysis on both the customer profiles and interactions, which can be exploited in a variety of ways as discussed in Section 5.

Performance Results

LIPTUS was run on 1.3 million customer interactions (1.2 million customer emails and 100,000 transcribed phone-calls). LIPTUS was able to link around 80% of the customer emails with the customer profiles. A careful analysis of the 20% of the data which LIPTUS was not able to link, revealed that they were junk emails that had escaped the spam filter. Out of the valid set of customer emails, LIPTUS was able to link more than 98% of the emails correctly. The accuracy of the transcribed phone-calls was also similar, with LIPTUS being able to link more than 95% of the customer complaints. Moreover, the total time taken across all the 1.3 million interactions was only about a couple of hours, which is very reasonable.

4. LEARNING MORE FROM THE TEXT

The linking of customer profiles with customer interactions brings together the factual information about the customer (such as the customer’s demographics, profitability, product holdings) with the factual information about the interaction (purpose of the interaction, the product or service it concerned, etc.). However, useful additional information can be gained by analyzing the content of the interaction. In this section, we describe the text analysis LIPTUS performs on the customer interactions. This analysis pulls out a variety of interesting characteristics of the interaction and, by association, of the customer that are not available otherwise.

For instance, information such as events of interest (travel outside the country), relationship with a competitor, etc. can be useful for targeted marketing (cross-sell and up-sell) based on the needs of the customer, identifying new product and service markets, identifying the market trends, behavioral analysis, etc. As we shall see, the customer interactions can be effectively mined to infer the customer’s satisfaction level for the services and products she avails and things she feels bitter about – getting such feedback without the need

of extensive customer surveys is indeed of significant value to the organization.

4.1 Extracting Events

Customer interactions often convey, either directly or indirectly, events happening with the customer. Such events can often be of significant use since they present immediate business opportunities with the customer.

In our case, we found several cases wherein the customer requests online banking password resets while on foreign travel. The marketing teams are very interested about such information since it opens up avenues for targeted marketing (the customers on foreign travel could be a target for foreign exchange products, offers from partner hotel chains, airlines, etc.). However, the metadata for such interactions does not capture this interesting fact about the customer being on a foreign travel, since this is of little consequence to the customer support.

LIPTUS uses a classifier [14] that identifies the customer interactions based on the presence of suggestive keywords such as “abroad”, “outside <country name>”, “currently in <country name>”, etc. in the interaction body. These keywords are identified by manually going through a small sample of relevant interactions. While more sophisticated solutions are possible, we decided to use this simple classifier because of (a) its simplicity and ease of implementation, and also (b) the unavailability of enough training data that a more sophisticated classifier would have required. Moreover, the rule-based classifier provided very reasonable results on our sample datasets.

4.2 Extracting Competitor Product Holdings

Knowledge of the competitor products held by a customer can be invaluable for an organization – it clearly conveys their products’ standing in the market against the competing products. Moreover, it gives the current snapshot of the needs of the customer and her preferences.

Let us first consider the kind of interactions that tend to contain such information. Customers send in emails for a variety of reasons which could include problems in cheque processing, credit card charges, complaint about services etc. In many cases the customers refer to the service or products of other banks in such emails. For instance, a customer could mention that due to delay in processing of a cheque, the customer was unable to pay an installment towards repayment of a loan she has from some other bank. Customers also often complain about a service saying that they have had better experiences with the competition.

This information can be used to understand the what products the customer holds, beyond the relationship the customer has with the bank. This tells the bank what they are up against – that is, the alternatives for the customer they are competing with. A proactive marketing strategy team might want to incorporate such data in their competitive analysis and to design their marketing campaigns.

LIPTUS uses a UIMA annotator [6] to identify the competing products mentioned in the mail. The annotator takes as input a dictionary of the competing product names, and identifies these names in the interaction text. The annotator uses standard dictionary-based named-entity recognition techniques to perform the task [15]. This simplistic solution could be misleading at times, however. For instance, the customer may just mention “cheque drawn on XYZ Bank”

– this does not mean that the customer has an account in XYZ bank. To eliminate such false positives, the annotator would have to apply natural language understanding techniques [11]; this is a part of our future work.

4.3 Extracting Customer Signature

Customer emails sent using the customer’s work address often include the customer’s signature. LIPTUS identifies and analyzes such signatures, extracting useful information that can be used to update and improve the customer profile.

We first discuss the issue of identifying the location of the signature in the email text. While sophisticated alternatives exist [4], LIPTUS uses a very simple heuristic that seems to work well – the idea is to first extract the customer name (either from the “From” field of the email header, or from the linked customer profile) and then search for it towards the end in the body of the email.

Once the position of the signature is identified, LIPTUS tries to parse this signature and extract information of interest from the same. The signature may include a variety of information, including the customer’s name, contact number, designation, employer’s name, contact number, postal address, etc. LIPTUS currently extracts only the contact number and employer’s name as these were considered more important by the customer intelligence teams. We consider these in turn below.

LIPTUS finds the location of the employer’s name in the signature by looking for keywords such as “Corporation”, “Ltd.” and “Inc.” If this fails, LIPTUS tries the slower option of matching the terms in the signature with a dictionary of company names; this dictionary of company names is constructed apriori by collecting the unique company names present across the customer profiles in the database. In our interaction sample, we found that most of the company names started on a new line and the name of the company is generally present in the first word on the line; we utilize this observation to avoid matching each term in the signature with the dictionary, making the overall procedure efficient.

To identify the customer contact number in the signature, the primary challenge is to identify the phone number from other numbers present in the signature, such as the postal code, street or house number. We use rules that use a number of simple patterns such as the presence of leading “+” signs (the standard international format for specifying phone numbers), leading zeroes (long distance calls in India need to be dialed beginning with a zero followed by the area code), the presence of phrases such as “Phone”, “Contact Number”, etc. Such simplistic ideas worked reasonably well on the datasets we had.

4.4 Estimating Customer Satisfaction Levels

Companies spend significant time and effort gauging how satisfied their customers are with the services and products they avail. In this section, we describe techniques used in LIPTUS for estimating customer satisfaction levels from the customer interactions [17]. These estimates, coming from direct customer interactions, are likely to be more accurate and timely than, for instance, the more traditional customer surveys companies routinely spend significant time and effort on. Moreover, LIPTUS is able to get the satisfaction levels for each individual customer and even at the level of each individual account held by the customer – a granularity that the traditional customer survey techniques can proba-

bly never reach. These estimates can be used, for instance, to evaluate the efficacy of the customer support by comparing the satisfaction of the customer in the first and last email sent by the customer in an interaction. Individual customer satisfaction levels can also form an important input towards predicting the set of customers who are likely to defect in the near future.

LIPTUS considers customer satisfaction at only two levels – either the customer is satisfied, or dissatisfied. This reduces the problem to binary classification with the two labels “satisfied” and “dissatisfied”. LIPTUS uses a naive-Bayes classifier because its training time is linear in the corpus size and also because more sophisticated classifiers were found out to be only marginally better on the given dataset. In the discussion below, we present the issues involved in performing this classification task on the customer interactions available, and also present the approaches used by LIPTUS to tackle those issues.

Insufficient training data. Unsupervised classifiers need to be trained using statistically significant amounts of training data (also called labeled data), to achieve high classification accuracy. A major challenge we faced while building the classifier was the lack of any training data.

LIPTUS addresses this issue using bootstrapping techniques. The idea is to manually build an initial sample, and then have the classifier “bootstrap” on this sample [7]. We took a sample of 1000 customer interactions, and manually tagged each interaction with the appropriate label, based on whether the customer was satisfied or dissatisfied. LIPTUS learns a classifier using this initial training set and applies it to the entire collection of customer interactions – this results in a classification of additional documents. The interactions that get classified with high confidence are added to the training set. This increases the size of the labeled dataset, but possibly makes it dirty. LIPTUS continues this process for more iterations and assigns progressively decreasing weights to interactions added in later iterations. The process ends when no further interactions are classified with high confidence.

Skew in the training data. Most classification algorithms assume that the training data has Classification performs best when all the classes are represented by an equal proportion of high quality training examples. In the training sample we had (ref. the discussion above), 68% of the interactions were labeled “satisfied” and the remaining 32% were labeled “dissatisfied.”

LIPTUS addresses this issue by giving high weights to features (discriminating words in the text) that are more likely to appear in the “dissatisfied” interactions than in the “satisfied” interactions. We are currently exploring more sophisticated ways of handling this problem [3, 12].

Ungrammatical text. A customer service executive speedily transcribing a phone-call while on call with a customer has grammar, spelling and punctuation as the least of her concerns. A variety of abbreviations occur, and often we found that entire messages are written in a single case. Similar issues exist in the messages sent by bank staff (in case of customer walk in). Customer emails are relatively cleaner, but not always so. These issues make the task of identifying interesting keywords in the text extremely difficult. For instance, since the case information is not reliable, it is difficult to differentiate misspelled words from proper nouns.

LIPTUS addresses this problem by focusing on words that occur statistically significant number of times, and which are discriminative of the class of the document [16]. This helps LIPTUS to eliminate a large number of misspelled words and infrequent proper nouns.

Complex phrases. Traditional text classification techniques [14] model documents as a bag of “ n -grams” (n -word sequences appearing in the document). Typically, unigrams (1-grams) or bigrams (2-grams) are considered appropriate. However, consider the bigram “close account”. This bigram rarely appears in an interaction, but its variants like “close the account” and “closed my bank account” are frequent. In general, we found that restricting to unigrams and bigrams does not lead to good features. Using trigrams (3-grams) fared better, but then since the number of possible trigrams in the document is large, it is hard to reliably estimate their frequency based on the limited training set available. This can lead to missed features – an informative trigram can be pruned out during feature selection just because it was not frequent enough in the given training corpus.

LIPTUS effectively avoids such issues by using *long-range features* [10] instead of n -grams. Long-range features consist of at most w words that occur in a window of size l in the text (w and l are parameters). In LIPTUS, we fix $w = 2$ and $l = 10$. Note that, unlike n -grams, the constituent words of a long-range feature need not occur consecutively. In the example above, the long-range feature “close..account” works better than choosing the bigram “close account”. LIPTUS uses efficient algorithms to compute these long-range features efficiently [10].

Performance Results

We considered two versions of the classifier – one that used trigrams, and another that used long-range features instead. Each classifier was run on ten independent random splits of the corpus, where each split consisted of 90% of the corpus as training data and the remaining 10% as validation data. We found that, on average, the version based on trigrams could find 73% of the “dissatisfied” interactions while the version based on long-range features could find 80% of such interactions, a significant improvement.

5. APPLICATIONS

In this section, we describe example applications where the linking of customer interactions with customer and account ids, enabled by LIPTUS, proves useful. Some of these applications have already appeared as motivations for the material in earlier sections.

We first present examples showing how the linking can help provide better understanding of the customers’ overall concerns and help identify trends in their behavior and preferences. Next, we show how to use the available information to gather additional insights about each individual customer; these insights are of immense value in predictive analytics (such as customer attrition analysis), generating personalized marketing campaigns, etc. Finally, we present applications directed towards improving the quality of the data constituting the customer profiles.

5.1 Aggregate Customer Analytics

The customer interactions and associated metadata (including derived features such as the satisfaction level) are

now available in the data warehouse alongside, and linked to, the customer and account profiles. This enables interesting analytical queries that involve predicates and groupings based on both “kinds” of attributes, and their combinations; for instance:

- *What are the ten categories that over the past month have received the greatest upsurge in “dissatisfied” complaints from the most profitable (top band) customers?* This analysis gives the bank insights about the customers’ concerns in a timely manner. Filtering on the satisfaction level allows the bank to identify and focus on the more important issues – the bank might be receiving a number of minor complaints about online banking, but the more serious complaints could be about delays in processing cheques.
- *Which product category has been receiving most inquiries from salaried customers between 25 and 35 years of age?* This information is useful, for instance, in creating campaigns directed to the specific segment (salaried people between 25 and 35 years).
- *What are the most common phrases appearing in the interactions in each category?* Monitoring the most common phrases used in the customer complaints (more importantly, the dissatisfied ones) is likely to help identify problems that are more specific than the available set of category labels. For instance, complaints on the internet website being excessively slow would be classified under a “technical problems” or “miscellaneous” category, which is not very informative. Recall that these common phrases are computed by the classifier as a set of features (ref. long-range features discussed in Section 4.4).

The examples mentioned above are only a sampler; in general, it is clear that the linking enabled by LIPTUS results in insights that are crucial for almost all the customer-facing aspects of the business. Interestingly, LIPTUS allows the bank to tap such insights from the information it already has (the customer interaction) without having to spend time, effort and money in gathering the data through explicit customer surveys.

5.2 Individual Customer Analytics

In this section, we show how the linking information can be used to gather insights about an individual customer and her relationship with the bank. These insights can be effectively used not only for designing campaigns, but also for identifying and optimizing the set of target customers for the campaigns. Such insights can also be helpful to customer service executive when she is on call with the customer; these insights increase the executive’s perspective about the customer and help her attune the interaction to the customer needs as much as possible.

In Section 4, we discussed how the individual interactions can be analyzed to identify interesting opportunities for marketing products to the associated customer. For instance, if the interaction contains hints about the customer being on a foreign travel, then the bank can offer the customer foreign exchange, money transfer, and online bill payment services. Similarly, if the interaction contains clues about the customer holding competing products from another bank, then the customer can be targeted for a per-

sonalized campaign that highlights features of the bank’s products as compared to such competing products. Further, the category assign to an interaction identifies the concerns of the customer, which the bank can exploit for cross-selling other products. For instance, a customer complaining about the charges penalizing the low balance in her account can be offered a waiver if she invests a certain sum as fixed deposit with the bank.

Even deeper insights about a customer can be obtained by analyzing the entire history of the interactions on record for the customer – this history can be reconstructed by consolidating all available interactions linked to a given customer or her accounts. A consolidated analysis of the interactions in this history allows us to derive interesting insights for each customer and her relationship with the bank; for instance:

- *Has the customer been upset in the (recent) past?* The customer might not have been upset in the last few interactions; worse, she could have been sarcastic (“My cheque delayed again—what an excellent service!”), a fact which is very hard to detect. Looking at the history would show that the customer has been very upset in the past and suggest that all may not be well.
- *What is the frequency of the customer’s interaction with the bank? Are they inquiries or complaints?* It helps to identify a customer who is *not* indifferent towards the bank. A customer who complains excessively needs special attention to prevent him from leaving the bank; this could be important if she is a highly profitable customer. On the other hand, an existing customer who keeps on inquiring about additional services related to the accounts she holds, or additional products is obviously a dream target for the marketing and presents an opportunity that cannot be missed.
- *On the average, what is the duration of the interaction with a given customer? How many messages on an average are exchanged per interaction?* This information could be used to evaluate the efficiency of customer support, and in case of a problem, help identify the cause.
- *Are the interactions (especially the “dissatisfied” ones) focused on a single category?* If the customer has been interacting over a particular topic again and again, either the problem is chronic, or it is not being solved properly – in either case, this should be a serious cause of concern to the bank.
- *If the customer holds multiple products, what is the spread of her interactions across these products?* If the customer holds five different products, but complains only about one of them, then she is satisfied with the bank in general, but not with the product. Such a customer could be a good source of constructive feedback.

So far we have only considered the interactions with the customers. As apart of the linking procedure (Section 3.2), LIPTUS separates out the interactions that could not be linked to a customer or account profile; these interactions include inquiries from non-customers. Event extraction (Section 4.1) and competing product identification (Section 4.2) can be applied to such interactions, as earlier, and the derived information can be used to identify promising marketing leads among the senders of such interactions as well.

5.3 Updating Stale Customer Profiles

The customer profiles on record with the bank may become stale with time, and need to be updated pro-actively by the customer when she changes address, changes the employer, etc. LIPTUS can help figure when the customer profile becomes stale – the customer can then be contacted and asked to update the profile. For instance, if a customer uses a mail id different from what is available in her profile, or if the customer’s employer on record is different from the one found from the signature mentioned in the latest interaction (Section 4.3), then there is a possibility that the current customer profile is stale, and needs to be updated.

In the given dataset, it was found that 23% of customers interacting were flagged non-contactable through any means (stale or no email id, stale postal address and invalid contact number). Even for the contactable customers, the analysis of the emails showed that around 17% of the customers who had sent emails did not have any email id in the data warehouse. Further around 21% of the customers used an email id which was different than that given in the data warehouse. Linking the interactions to the customer profiles allowed the bank to note the email addresses of such customers as their alternate contacts, used to send an request asking them to update their contact information.

6. RELATED WORK

Linking of unstructured and structured information has been explored in our prior work, SCORE [13] and EROCS [2]. SCORE enhances structured data retrieval by associating additional documents relevant to the user context with the query result. EROCS is closer to the problem addressed in LIPTUS. However, EROCS is designed to be a generic solution, and is an overkill for the data targeted by LIPTUS. Specifically, EROCS views the database as a set of entities, and identifies the entities that best match a given document – it performs the matching even if the identifier of the entity does not appear in the document text, and allows different segments in the document to match different entities. The customer interactions LIPTUS is designed to work with are much simpler; a typical interaction has the customer or account id (or both) explicitly mentioned in the text, and relates to a single customer or account.

LIPTUS also performs text analysis over the customer interactions, such as analyzing customer satisfaction levels, extracting competitor product holdings etc. The task of extracting satisfaction levels from documents (sentiment mining) has received attention in the past [8, 17]. Bootstrapping techniques to cope with small training data size while constructing classifiers has also been studied earlier [7]. Identifying company names and other useful information in text falls under the category of Named Entity Recognition (NER) [1, 15]. Cohen and Sarawagi [5] propose techniques of improving NER techniques by using an external dictionary; this is similar to the problem addressed in Section 4.3.

Overall, even though significantly more sophisticated solutions are possible for almost all problems addressed by LIPTUS [14, 15], we used the simplest solutions that worked on the datasets we had. This was necessary since the requirement was to keep the complexity of the solution as low as possible, while achieving scalability to work on tens of thousands of interactions per day.

7. CONCLUSION

In this paper we have presented LIPTUS, a tool to link unstructured customer interactions with structured customer and account profiles. Unstructured information, such as these customer interactions, exist as silos with limited use in marketing, business intelligence etc. which are based on structured information. LIPTUS bridges this gap, enabling consolidated analysis of both the structured and unstructured data. A major challenge faced by LIPTUS was to work effectively in presence of the extensive amount of repeated, irrelevant text, disclaimers, advertisements, etc. present in the customer interactions, and the incomplete and inconsistent information present in the customer and account profiles. LIPTUS exploits a mix of principled ideas and ad-hoc hacks to counter these challenges. As mentioned earlier, LIPTUS has been deployed in a real banking customer intelligence setup, where it is gradually finding good use [9].

In summary, we think that LIPTUS is a first of its kind tool, that tries to solve an interesting but hard problem in as effective a way as possible given the constraints on complexity and scalability of the solution. Even though LIPTUS was developed for a specific domain, we hope that the overall utility of such a tool would appeal to practitioners in other domains as well.

Acknowledgments

We would like to thank Neisha Sen, Swarup Chaudhary, Raghuram Krishnapuram, Ponani Gopalakrishnan, Daniel Dias, Nelson Mattos, Laura Haas and the FOAK Board of IBM for their help and encouragement. We are also grateful to C. N. Ram, T. R. Deepak, Harish Shetty, Ajay Kelkar, Lata Murjwani, Suryakant Shelar and Gopal Vasudevan from HDFC Bank for their support.

8. REFERENCES

- [1] BORTHWICK, A., STERLING, J., AGICHTEIN, E., AND GRISHMAN, R. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Workshop on Very Large Corpora* (1998).
- [2] CHAKARAVARTHY, V., GUPTA, H., ROY, P., AND MOHANIA, M. Efficiently linking text documents with relevant structured information. In *VLDB* (2006).
- [3] CHAWLA, N., JAPKOWICZ, N., AND KOTCZ, A. Editorial: Special issue on learning from imbalanced data sets. In *SIGKDD Explorations* (2004).
- [4] CHEN, H., HU, J., AND SPROAT, R. W. Integrating geometrical and linguistic analysis for email signature block parsing. *ACM Trans. Inf. Syst.* 17, 4 (1999).
- [5] COHEN, W., AND SARAWAGI, S. Exploiting dictionaries in named entity extraction: Combining semi-markov extraction process and data integration methods. In *SIGKDD* (2004).
- [6] GOTZ, T., AND SUHRE, O. Design and implementation of the UIMA common analysis system. *IBM Systems Journal* 43, 3 (2004).
- [7] HAMAMOTO, Y., UCHIMURA, S., AND TOMITA, S. A bootstrap technique for nearest neighbor classifier design. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (1997), 73–79.
- [8] HU, M., AND LIU, B. Mining and summarizing customer reviews. In *SIGKDD* (2004).

- [9] IBM. Made in IBM Labs: IBM Helps HDFC Bank Extract Information Insight to Enhance Customer Care. <http://www.ibm.com/press/us/en/pressrelease/20729.wss>.
- [10] JOSHI, S., RAMAKRISHNAN, G., BALAKRISHNAN, S., AND SRINIVASAN, A. Aggregating contextual patterns for information extraction. In *IJCAI 2007 Workshop on Text Mining and Link Analysis* (2007).
- [11] MANNING, C., AND SCHUTZE, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
- [12] MLADENIC, D., AND GROBELNIK, M. Feature selection for unbalanced class distribution and naive bayes. In *ICML* (1999).
- [13] ROY, P., MOHANIA, M., BAMBA, B. AND RAMAN, S. Associating relevant unstructured content with structured database query results. In *ACM CIKM* (2005).
- [14] SEBASTIANI, F. Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1 (2002).
- [15] TURMO, J., AGENO, A., AND CATAL, N. Adaptive information extraction. *ACM Computing Surveys* 38, 2 (2006).
- [16] YANG, Y., AND PEDERSEN, J. A comparative study on feature selection in text categorization. In *ICML* (1997).
- [17] YI, J., AND NIBLACK, W. Sentiment mining in web-fountain. In *ICDE* (2005).