Semantic Mining on Customer Survey

Yang Yu*
The MathWorks
1 Apple Hill Dr, Natick, MA 01760
yang.yu@mathworks.com

Jiangbo Dang
Siemens Corporate Research
755 College Road East, Princeton, NJ 08540
jiangbo.dang@siemens.com

ABSTRACT

Business intelligence aims to support better business decision-making. Customer survey is priceless asset for intelligent business decision-making. However, business analysts usually have to read hundreds of textual comments and tabular data in survey to manually dig out the necessary information to feed business intelligence models and tools. This paper introduces a business intelligence system to solve this problem by extensively utilizing Semantic Web technologies. Ontology based knowledge extraction is the key to extract interesting terms and understand the logic concept of them. All knowledge extracted forms a semantic knowledge base. Flexible user queries and intelligent analysis can be easily issued to the system over the semantic data store through standard protocol. Besides resolving problems in theory, we designed a flexible, intuitive user interaction interface to explain and present the analysis result for business analysts. Through the real usage of this system, it is validated that our system gives good solution for semantic mining on customer survey for business intelligence.

Categories and Subject Descriptors

D.2.8 [**Computing Methodologies**]: ARTIFICIAL INTELLIGENCE—*Knowledge acquisition*

Keywords

Ontology based Knowledge Extraction, Ontology based Text Mining, Sentiment Analysis, User Interaction, Visualization

1. INTRODUCTION

Business intelligence (BI) can be defined as the process of finding, gathering, aggregating, and analysing information for decision making. Customer survey is a priceless asset for any company/organization conducting intelligent business analysis. The primary goal of conducting customer survey is to increase the quality of the products and services of the companies they serve. To know customers' current level of satisfaction and to realize if changes in their opinion do occur, we need a measure that accurately assesses customer attitudes. When developing questionnaires, it is important to ensure that the data obtained from them reflect reliable and valid information. An user friendly survey questionnaire is an necessary condition to achieve accurate customer attitudes. While structured feedback in a survey questionnaire can be more easily digested by machines, the degree of expressiveness allowed by structured customer forms appears largely limited. Therefore it implies that a good questionnaire usually is hard for machine to automatically process and understand. Thus there is a gap between acquiring true information from customers and retrieving / understanding knowledge in survey feedbacks automatically. The knowledge in question include products, services, customer's opinions, etc. This problem is related to the typical conventional bottleneck of knowledge acquisition and is the first main research problem the system focuses on. Another challenge this work tries to overcome is how to support flexible usage / presentation of the acquired knowledge, e.g. what we can summarize / analyze from the knowledge in survey and how to demonstrate them.

Knowing of these research difficulties, the research goals (also the contributions) of this work are as follows;

1. extract information related to products and services in survey;
2. extract customer sentiment on related products and services;
3. design a platform supporting flexible intelligent business analysis over extracted knowledge;
4. intuitively present the mining and analysis results.

The keys to attain these goals are to extract useful knowledge, understand its semantics and intelligently answering queries using it. Although knowledge extraction (KE) is methodically similar to information extraction (IE), the main criteria is that the extraction result goes beyond the creation of structured information or the transformation into a relational schema. It requires either the reuse of existing formal knowledge (reusing identifiers or ontologies) or the generation of a schema based on the source data. According to the definition, one of the key function of our system is KE. Without a KE component, business analysts carrying out BI activities would have to read hundreds of textual comments and tabular data in survey to manually dig out the necessary information to feed BI models and tools. To represent the structured information and for reusing in-

*This work was done when the first author was on an internship at Siemens Corporate Research.

formation, Semantic Web technologies are perfect choices, because the Semantic Web is designed for reuse, integrate heterogeneous information and support semantics specified entailment. When solving knowledge management problem by using Semantic Web technologies, ontologies can be used in order to provide formal grounding for representing the semantics of knowledge elements; they can guide creation of semantic annotations constituting a set of all meta-level characterizations easing knowledge source description, evaluation, and access. Therefore we propose a system based on semantic technology to accomplish these goals.

2. RELATED WORK

Exploiting the help of Semantic Web technologies is the trend of information extraction research. More importantly, Semantic Web technologies can truly transform "information" to "knowledge", because they aim to attach data structure, typed links, and axiomatically represented implicit facts to other forms of data. Among many existing works on ontology-based information extraction (OBIE), we can categorize them by the following dimensions.

The first dimension is the method of extraction. There are four kinds of methods. 1) Linguistic rules. Embley [5] and Maedche et al. [8] both consider the linguistic rules used for information extractions a part of "extraction ontologies". This means that a person or a group of persons have to read all the documents of the corpus in concern and figure out suitable extraction rules. It can be seen that this a tedious and time consuming exercise which does not scale well. Besides systems that use manually identified linguistic rules, some systems have aimed to automatically mine extraction rules from text. Vargas-Vera et al. [11] have designed and implemented an OBIE system that operates on these principles. However, even if two documents have similar content, they might use different expressions or linguistic structures. Therefore linguistic rules are usually hard to cover all types of documents. 2) Gazetteer lists. This technique relies on finite-state automate just like linguistic rules but recognizes individual words or phrases instead of patterns. These systems often use gazetteer lists containing all the instances of some classes of the ontology. They have been used in the SOBA system [3] to get details about soccer games and in the implementation by Saggion et al. [10] to get details about countries and regions. 3) Classification techniques. It is also a common practice to convert an information extraction task into a set of binary classification tasks. For example, Li et al. [7] developed a system using uneven margins SVM and perceptron techniques, uses one binary classifier to decide whether a word token is a start of an entity and uses another to detect the end token. 4) Web-based search. Cimiano et al. [4] have implemented an OBIE system named PANKOW that semantically annotates a given web page using web-based searches only. It conducts web searches for every combination of identified proper nouns in the document with all the concepts of the ontology for a set of linguistic patterns.

The second dimension is ontology construction and update. OBIE systems can be classified based on the manner in which they acquire the ontology to be used for information extraction. One approach is to consider the ontology as an input to the system. Under this approach, the ontology can be constructed manually. Such systems include SOBA [3], the implementation by Li and Bontcheva [7], the implementation by Saggion et al. [10] and PANKOW [4]. The other approach is to construct an ontology as a part of the OBIE process. Kylin (through Kylin Ontology Generator [12]) and the implementation by Maedche et al. [8] construct an ontology as a part of the process, although their main aim is to identify new instances for the concepts of the ontology.

3. SYSTEM OVERVIEW

Before jumping into each component of the system, we give Figure 1 which shows the overall architecture of our system. The input is customer survey raw data files which may be in text or spreadsheet form (we use spreadsheet in this work). Taking this input, the system first parse each field into corresponding field in a data structure representing a survey questionnaire. There are several APIs for parsing a spreadsheet file and we use SAX [2]. The next stage is knowledge extraction which extracts all related knowledge that business analysts may be interested in, such as products (including name, series, etc.), services (including human service, parts supplying, etc.) and the sentiment related expressions. This component also uses other tools and input, such as Named Entity Recognizer, Natural Language Processing parser and the ontology. The details of knowledge extraction will be introduced in Section 4. After the knowledge is extracted, an immediate question is how to represent and store it in order to support flexible query and intelligent business analysis. We exploit the strong capability provided by semantic technology. The whole knowledge base in stored in a semantic data store with fully flexible query interface. More importantly, the unique power of logic reasoning provided by Semantic Web technologies can greatly improve the intelligence of business analysis. Based upon this knowledge base, the system provides various analysis functioning similar to traditional text mining tools, such as classification, clustering and feature selection. The details of semantic mining on the extracted knowledge is given in section 5. Finally, since it is not straightforward to let business analysts easily understand the systematic analysis results, we visualize most of the semantic mining results in a animated, interactive and graphical user interface. Section 6 is mainly about this visualization design.

4. KNOWLEDGE EXTRACTION

4.1 Ontology

As we discussed in related work section, there are a number of methods to build a ontology for helping knowledge extraction. The most popular one is based on prior knowledge which means that an domain expert, e.g. the survey designer or business analyst in this problem domain, defines all possible related concepts/vocabularies in the ontology. The advantage of this approach is that the concept and relationship between them can be defined precisely. But the deficiency is that the vocabulary in the ontology may be limited compared to large amount of terms possibly appeared in various questionnaire answers. The second one is based upon learning process which is to use a portion of survey data to learn the concepts that will be useful in a ontology. This approach has the advantage and deficiency contrary to the previous approach. Thus we choose to combine the two approaches by defining some higher level concepts and some

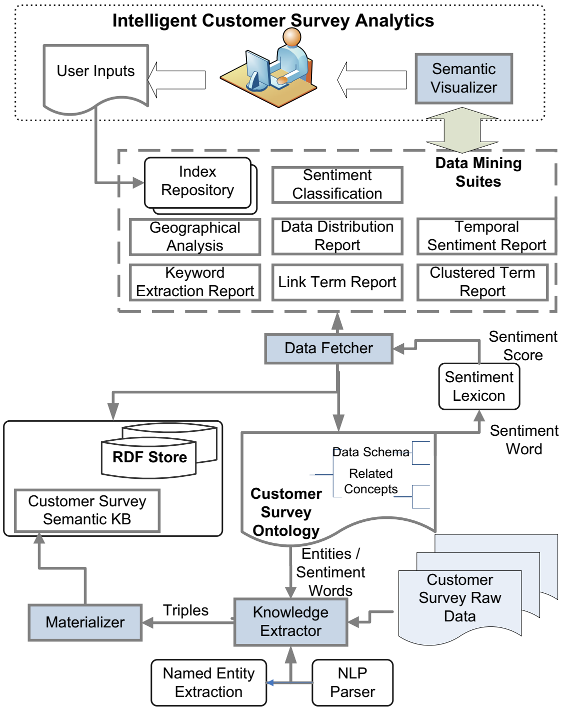


Figure 1: System overview.

of their descendant concepts as base in advance and then using Natural Language Processing (NLP) tools to learn and populate the ontology with more terms appeared in actual survey answers.

There are four categories of high level concepts/terms related to surveys: *product*, *service*, *sentiment* and *schema*. For schema related concept, the terms are designed to represent the relationship between survey questions and between the knowledge and the survey record where the knowledge is extracted. Using this, any form of survey can be easily incorporated into system given the mapping between fields in raw survey data and the conceptualized concepts in ontology. The class *product* is designed to represent all instances of certain type of products. The class *service* is designed to represent all instances of certain service that a company provides. The class *sentiment* includes various terms or phrases that could show user's sentiment.

4.2 Entity Recognition

Although we have decided to use ontology as a guide on extraction, it is impractical to find exact match of words/phrases against the ontology, because there are countless specific words/phrases. This section discusses how to find some interesting words/phrases as the seed for helping run time matching against the ontology. To this end, the system needs to automatically recognize entities (or generally interesting sequence of words) mentioned in survey. Named Entity Recognition (NER) has been researched intensively in NLP domain and a number of such tools show decent performance, such as Standford NER, OpenNLP name finder, LingPipe NER. We compared these popular tools on some sample text from our problem domain (shown in Table 1). The words/phrases in bolt in sample texts are the terms need to be extracted (i.e. ground truth). We noticed that no one can perform as what we expected in terms of both precision and recall. In addition to named entities, we also need to extract sentiment related terms which usually are adjectives and are not entities. So we chose a more general method that use intermediate parsing results of a NLP parser and then manually judge some phrases reported by this tool. Although this process turns into semi-automatically, it greatly improves the performance and practical for real problem in terms of efficiency. Because this learning process is expected to be conducted once and only on a small portion of sample data drawn from the whole data set, this semi-automatic process is limited.

We use Stanford Parser to return the intermediate parsing results of text. Because we notice that the quality of text would affect the parsing results much, we tried to remedy text quality in the following three steps before send it into the parser. First, if a piece of text does not have any period punctuation, we add a period at the end of text. Second, if the text are all capitalized, we convert it into all lower case. Third, if the first letter of the first word is not capitalized, we make it upper case.

Now compared with using those NER tools only, we can see that our method performs much better both in recall and precision. This automatic extraction gives candidate terms in our predefined four categories of higher level concept. Then domain expert evaluate these terms and put them as appropriate subclasses in our survey ontology.

4.3 Semantic Knowledge Extraction

The previous subsection discussed how we build the ontology and input initial portion of the ontology by learning from small portion of survey data. For larger amount of survey data, the system needs to automatically extract semantic knowledge using the survey ontology built by above process. Essentially, the differences between this automatic extraction process and the above semi-automatic process are as follows. First, we need an extra automatic solution to determine the appropriate type of certain concept appeared in survey. Second, we need to construct the extracted knowledge into semantic web form, i.e. OWL/RDF format. So we mainly describe these two aspects in this subsection.

The phrases determined by NLP parser are mapped to the concepts in the survey ontology to determine their corresponding ontology concepts. As described in Algorithm 1, If there is an exact match between the phrases and an ontology instance, the instance is returned as the result. Otherwise if the noun phrase is composed of multiple tokens of words, the system generates two subphrases by removing the first and last word from the phrase. We replace the original phrase with these two noun phrases and repeat the phrase mapping process until there is a match or there is no more token left. Our algorithm favors the longer and the right side phrase in the case there is a tie.

4.4 Materialization

After all knowledge is extracted, the system represent these knowledge into the Semantic Web form and store it in a semantic knowledge base. For each survey record, we create objects of a customer, survey and location. For fields

Table 1: Comparison on entity recognition tools.

Example Text	Stanford NER	OpenNLP	LingPipe NER	Stanford Parser
MARK IS ALWAYS HERE, HE IS **GREAT**. MY PROBLEM IS WHEN I CALL FOR **TECH** HELP AT SUPPORT SERVICE I FEEL THAT THEY ARE NOT QUITE ON THE BALL. SINCE **SIEMENS** HAS TAKEN OVER IT SEEMS THAT I AM ON HOLD FOREVER. IT WAS NEVER AS LONG AS IT USED TO BE, I HAVE NOTICED	TECH, SIEMENS			MARK, GREAT, SIEMENS
More timely delivery on ordered **MIC plates** ... I order them, but then they don't arrive until the following.	MIC	MIC	MIC, I	More timely, MIC, plates
it's about the **Advia 2120** being able to adjust the **CBC Differentials**	Advia, CBC Differentials	CBC Differentials		Advia 2120, CBC Differentials

Algorithm 1 *extractEntities*(*phrase*), *phrase* is a sequence of words separated by space.

```
if $queryOntology(phrase) > 0$ then
    return $phrase$
else
    if $length(phrase) = 1$ then
        return NULL
    else
        $P_r \leftarrow extractEntities(removeFirstToken(phrase))$
        $P_l \leftarrow extractEntities(removeLastToken(phrase))$
        if $length(P_r) \geq length(P_l)$ then
            return $P_r$
        else
            return $P_l$
```

related to customer profile, the system maps them to the properties in the survey ontology whose domains are class customer. For fields related to geographical locations, the system maps them to the properties in the survey ontology whose domains are class *location*. The class *location* and some properties about it are imported from the popular used geographical ontology, GeoNames. For all other fields that are about survey itself are mapped to the properties whose domains are class survey. Finally, the system connects these three objects: customer, survey and location by appropriate properties. In addition to the original fields in a survey, all other knowledge, i.e. concepts and sentiment related terms extracted, is connected to the survey object by using property *hasKeyword*. Thus all knowledge in a survey is fully and semantically represented in the Semantic Web form.

There are a number of choices to store these extracted Semantic Web data. Compared to traditional databases or other solutions, RDF store has the biggest advantage that it naturally support heterogeneous survey data. Because when information resources commit to the same ontology then the same meaning is anticipated for any term from that ontology. Even data that commit to different ontologies can be integrated in a information repository, as long as the ontologies have certain relationships, e.g. their concepts are defined in terms of a common ontology or an alignment is provided. Another obvious advantage of a RDF store is that it can give more flexible query support through SPARQL query interface. SPARQL is standardized by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium, and is considered a key semantic web technology.

5. SEMANTIC MINING

5.1 Classification

The first question that needs data mining techniques to answer is if there are patterns in survey for customer's satisfaction. From the classification perspective, given the categories of satisfaction level in survey, can we use answers in survey to predict them? In other words, the class type to be classified is the satisfaction level of customer shown in survey and the features used to classify are answers of all fields in survey. If we can find such pattern, we can know which kind of answers are more important, as they affect the overall customer satisfaction more. Then next time, if we see such answers, we can predict the customer satisfaction to some extent.

Given such traditional classification problem, we test several algorithms, such as decision tree, Naive Bayes, etc. Here Table 2 shows the result of decision tree classification. Through the table, we see the accuracy is about 76%.

5.2 Reducing Question Space

Few customers like to answer too many questions in a survey. It is a research problem that how to reduce questions that are required to answer in a survey while keeping the same amount of information that a company wants to acquire. This problem can be mapped to the feature selection problem in machine learning theory. In machine learning and statistics, feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique of selecting a subset of relevant features for building robust learning models. In other words, the problem is: given all the features (fields) in a survey, which features are key features that can determine the class type that we want to predict. Corresponding to different purposes of a survey, we test on using several different questions as the classes to be predicted, such as overall satisfaction, satisfaction with product, satisfaction with service, repurchase product, repurchase service, etc. Some of example results are shown in Table 3. The algorithm starts with empty set of features and greedily choose every single attribute in forward direction.

5.3 Sentiment Analysis

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. That is also the capability that we want the system to have, i.e. given comments or any piece of free text input in a survey, the capability to determine a rough overall sentiment of it.

Table 2: The classification result of C4.5 decision tree on survey data.

	eight	nine	ten	five	six	seven	one	four	two	three	Sum
eight	442	0	0	0	0	30	0	0	0	0	472
nine	6	80	522	0	0	0	0	0	0	0	608
ten	19	84	2119	0	0	0	0	0	0	0	2222
five	0	0	0	88	0	0	0	0	0	0	88
six	0	0	0	60	0	0	0	0	0	0	60
seven	132	0	0	0	0	16	0	0	0	0	148
one	0	0	0	17	0	0	0	0	0	0	17
four	0	0	0	16	0	0	0	0	0	0	16
two	0	0	0	9	0	0	0	0	0	0	9
three	0	0	0	9	0	0	0	0	0	0	9
Sum	599	164	2641	199	0	46	0	0	0	0	3649

Table 3: The feature selection result.

Class to be predicted	Selected features
OVERALL SATISFACTION	SEGMENT, SATISFACTION WITH PRODUCT, RECOMMEND SIEMENS PRODUCTS, REPURCHASE SIEMENS PRODUCTS, REPURCHASE COMPETITOR PRODUCTS, SERVICE AND SUPPORT, RECOMMEND SIEMENS SERVICE, SIEMENS TREATS ME WITH RESPECT, SIEMENS VALUES ME AS A CUSTOMER, SIEMENS ENSURES A TRUSTWORTHY RELATIONSHIP, COMPETITOR TREATS ME WITH RESPECT, AVOID DOWNTIME THROUGH PROACTIVE SERVICE, TIME TO CALL BACK
SATISFACTION WITH PRODUCT	OVERALL SATISFACTION, RECOMMEND SIEMENS PRODUCTS, REPURCHASE SIEMENS PRODUCTS
REPURCHASE SIEMENS PRODUCTS	RECOMMEND SIEMENS PRODUCTS
REPURCHASE SIEMENS SERVICE	RECOMMEND SIEMENS PRODUCTS, REPURCHASE SIEMENS PRODUCTS, SERVICE AND SUPPORT, RECOMMEND SIEMENS SERVICE, SIEMENS ENSURES A TRUSTWORTHY RELATIONSHIP

It is traditional in information retrieval to represent a piece of text as a feature vector wherein the entries correspond to individual terms. One influential finding in the sentiment-analysis area is as follows. Term frequencies have traditionally been important in standard IR, as the popularity of tf-idf weighting shows; but in contrast, Pang et al. [9] obtained better performance using presence rather than frequency. That is, binary-valued feature vectors in which the entries merely indicate whether a term occurs (value 1) or not (value 0) form a more effective basis for review polarity classification than does real-valued feature vectors in which entry values increase with the occurrence frequency of the corresponding term. This finding may be indicative of an interesting difference between typical topic-based text categorization and polarity classification: while a topic is more likely to be emphasized by frequent occurrences of certain keywords, overall sentiment may not usually be highlighted through repeated use of the same terms.

Based on analysis in related work introduced above, we utilized a third party resource, SentiWordNet [6], to determine which term/word of appearance is indicating certain sentiment. SentiWordNet is a publicly available lexical resource for opinion mining which provides the values PosScore and NegScore for each term in WordNet. The two values are the positivity and negativity score assigned by SentiWordNet to the synset. The objectivity score can be calculated as: ObjScore = 1 - (PosScore + NegScore). Referring to this resource, each keyword that is an instance of certain sentiment class defined in survey ontology can be assigned the overall sentiment score.

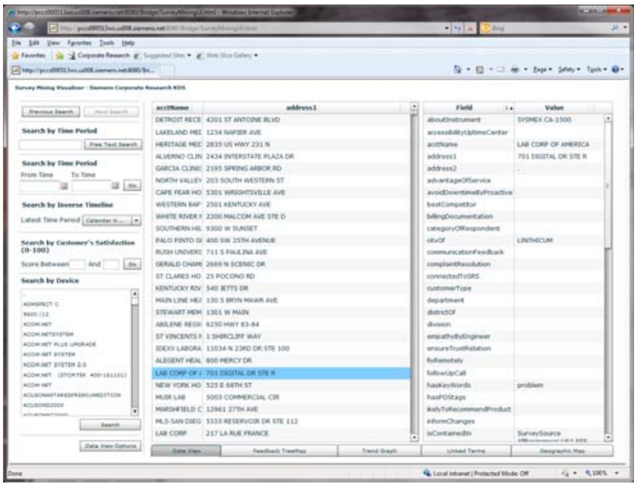
6. SEARCH AND EXPLORATION

Without a good user friendly interface, most of the system functions designed for business analysis would not work ideally. As shown in Fig. 1, the system has a data mining suites which analyzes business intelligence queries. Due to space limit, in this section, we only introduce the visualization for some of them in the system.

6.1 Faceted Search

The first and basic function of the system interface should be exploring the survey raw data and extracted knowledge. Given a large amount of data, it is a research problem that how to ease users to look into the data from a ideal perspective or point of view while not overwhelming users. One of such the state-of-the-art browsing techniques is faceted search, also called faceted navigation or faceted browsing. Faceted search is a technique for accessing a collection of information represented using a faceted classification, allowing users to explore by filtering available information. A faceted classification system allows the assignment of multiple classifications to an object, enabling the classifications to be ordered in multiple ways, rather than in a single, predetermined, taxonomic order. Each facet typically corresponds to the possible values of a property common to a set of objects. Facets are often derived by analysis of the text of an item using entity extraction techniques or from pre-existing fields in the database such as author, descriptor, language, and format. This approach permits existing web-pages, product descriptions or articles to have this extra metadata extracted and presented as a navigation facet.

In our system, through discussion with professional business analysts, we observed such primary facets as: time, keyword,



acctName	address1
DETROIT RECE	4201 ST ANTOINE BLVD
LAKELAND MEI	1234 NAPIER AVE
HERITAGE MEC	2833 US HWY 231 N
ALVERNO CLIN	2434 INTERSTATE PLAZA DR
GARCIA CLINIC	2195 SPRING ARBOR RD
NORTH VALLEY	203 SOUTH WESTERN ST
CAPE FEAR HO	5301 WRIGHTSVILLE AVE
WESTERN BAP	2501 KENTUCKY AVE
WHITE RIVER F	2200 MALCOM AVE STE D
SOUTHERN HIL	9300 W SUNSET
PALO PINTO GI	400 SW 25TH AVENUE
RUSH UNIVERS	711 S PAULINA AVE
GERALD CHAM	2669 N SCENIC DR
ST CLARES HO	25 POCONO RD
KENTUCKY RIV	540 JETTS DR
MAIN LINE HEA	130 S BRYN MAWR AVE
STEWART MEM	1301 W MAIN
ABILENE REGI	6250 HWY 83-84
ST VINCENTS I	1 SHIRCLIFF WAY
IDEXX LABORA	11034 N 23RD DR STE 100
ALEGENT HEAL	800 MERCY DR
LAB CORP OF /	701 DIGITAL DR STE R
NEW YORK HO	525 E 68TH ST
MUIR LAB	5003 COMMERCIAL CIR
MARSHFIELD C	12961 27TH AVE
MLS-SAN DIEG	5335 RESERVOIR DR STE 112
LAB CORP	217 LA RUE FRANCE

Field	Value
aboutInstrument	SYSMEX CA-1500
accessibilityUptimeCenter	
acctName	LAB CORP OF AMERICA
address1	701 DIGITAL DR STE R
address2	-
advantageOfService	
avoidDowntimeByProactive	
bestCompetitor	
billingDocumentation	
categoryOfRespondent	
cityOf	LINTHICUM
communicationFeedback	
complaintResolution	
connectedToSRS	
customerType	
department	
districtOf	
division	
empathyByEngineer	
ensureTrustRelation	
fixRemotely	
followUpCall	
hasKeyWords	problem
hasPOStage	
likelyToRecommendProduct	
informChanges	
isContainedIn	SurveySource

Figure 2: The main interface of the system.

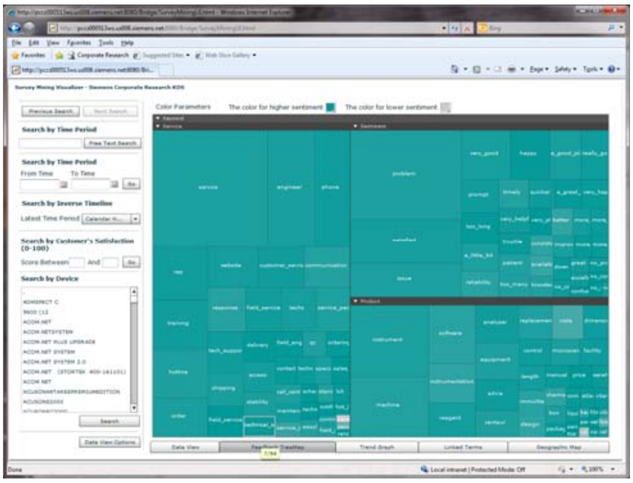


Figure 3: The treemap view of the system.

product, customer satisfaction. Figure 2 shows the main screen of the system. The interface has two main parts. The left part is navigation area where the facets to navigate exploration are listed and user can input. The right part is mainly the presenting panel where different views of data or analysis are shown.

In the left navigation panel, the keyword search box is on the top. This keyword search will match any fields in a survey against the input words. If the input contains multiple words, the field to be matched in a survey should contain all of them.

The second group of facets on the left of screen is about time. The group for time facet includes a period of latest week/month/quarter/half year/year and a period between any specified time.

The third facet is about customer satisfaction. Using this facet, business analyst can directly retrieve all surveys where customer's satisfaction is within a specified range.

The last facet is about products. The purpose of this facet is to enable analysis on survey focusing on certain product.

Having these facets and a user's input, the survey data can be retrieved and guided explored. However since it may still be a large amount of records, the data view we designed on the right of interface has two levels. One is a record list view and the other is a detailed view of certain selected record. In the record list table, the system provides customizable ability to let users specify which columns to show. These columns selected serve as summaries/preview of record details. This function facilitates users to easily group or compare records on certain columns. In addition, the record list can be ordered on any column.

6.2 Hierarchical Sentiment Exploration

The data view introduced above simply serves as a direct exploring of survey data that matches certain conditions, i.e. there is not much analytic or aggregated knowledge in it. Since this subsection, we will discuss how to give aggregated analysis over the faceted search results. The first analytic view we will introduce is treemap view. Treemaps are often used for space-constrained visualization of hierarchies and it also has been utilized on exploring customer feedback data before [13]. The survey ontology we build contains a natural hierarchy of products, services, sentiment concepts and their relationships. Thus using treemap can give a direct sense from the perspective of the elements in each hierarchy tree. Shown in Fig. 3, there are three big blocks for each category of concepts. Taking the sentiment as example, the box representing sentiment would contain a number of boxes representing various instances of sentiment concepts appeared in the survey data that is selected through faceted search. The size of each box indicates the frequency of this concept in selected survey data and the color of each box indicates the average satisfaction of selected survey records that contains this term. Therefore, through this graphic view, users can easily found which concept is mentioned more often (by viewing its size) and which concept may often related to (un)satisfied records (by viewing its color). By relating the term with these graphic features, users of the system can intuitively interpret the story of the survey records. In addition, the treemap view can also be further explored. If the user wants to take a closer look at certain part, a simple click on that part will navigate the treemap to only show that part and may bring more details of that part onto surface. Besides the graphic representation, to let the user know the precise value for each concept in a box, the system gives a hint popup window showing the number of frequency and the average satisfaction.

6.3 Product Trend Exploration

Data that is arranged in columns or rows on a table can be plotted in an area chart. Area charts emphasize the magnitude of change over time and can be used to draw attention to the total value across a trend. A stacked area chart also shows the relationship of parts to a whole. In the system, we developed stacked area chart to show the relationship of each product to all products. Each area chart for a product shows the customer sentiment trend about this product. The x axis of the chart is the time and the y axis is the average satisfaction score for products. Initially, this stacked area

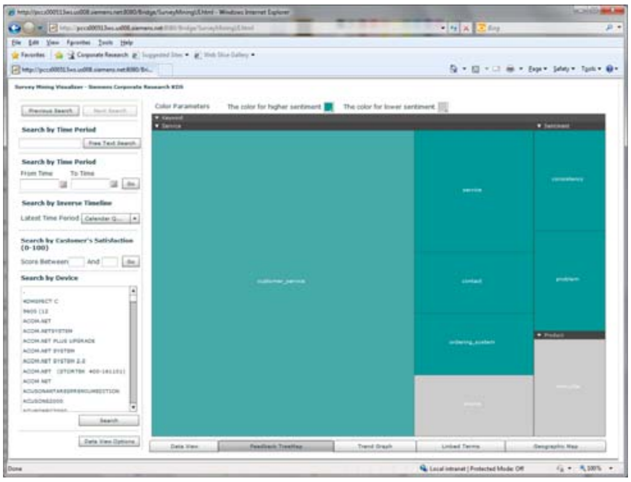


Figure 4: The treemap view of the system after certain item is clicked.

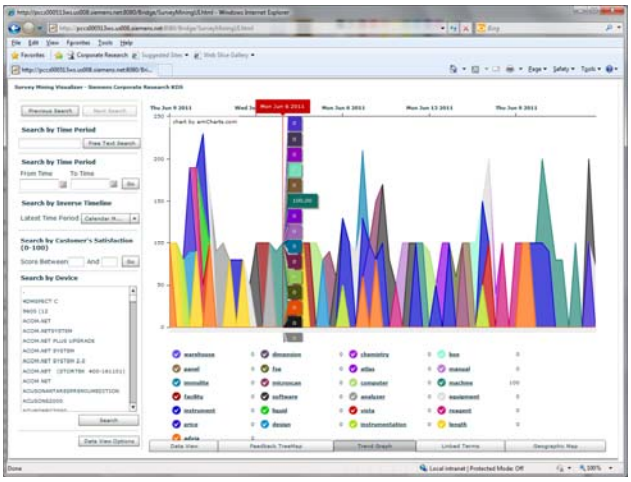

Survey Mining Visualizer - Siemens Corporate Research KDS
Search by Time Period
Search by Inverse Timeline
Search by Customer's Satisfaction (0-100)
Search by Device
chart by amCharts.com
Data View
Feedback TreeMap
Trend Graph
Linked Terms
Geographic Map

Figure 5: The trend graph view of the system.

chart is an analysis over the selected survey data filtered by facets. Users can further explore some portion of this chart by narrowing the time range or just clicking certain product line. A popup window shows the detailed numbers in a vertical line for each product when the user moves mouse on certain time point on x axis.

6.4 Hidden Keyword Relatedness

In a nutshell, terms, stems, and concepts that co-occur more frequently tend to be related. For instance, when we hear the term "aloha" we immediately think of Hawaii, not of Montana or Indiana. This is a semantic association between these two terms. According to Fuzzy Set Theory [1], the degree of term co-ocurrence in a database is a measure of semantic connectivity and can be used to build thesaurus for the database. Why should we care about co-occurrence? Some compelling reasons are: Co-citation of products and services, etc. To make business analysts easily understood, we call it Linked Term Network. The networks can depict different information as there are multiple types of the network actor, e.g. locations, organizations, people, products, services and sentiment, to reflect knowledge structures in micro-, meso-, and macro-levels, respectively. To avoid overwhelming users at the initial screen and easy navigation, we group products, services and sentiment terms in three categories as recognized entities. Through this linked term network, business analysts can clearly know the general relationship between three groups and specific relationship between any specific two terms as well. When any term is clicked, all links and other terms linked to by this term can be highlighted for focused review. When mouse is over any link, there is a tip window showing the two terms linked and the number of co-occurrence of them.

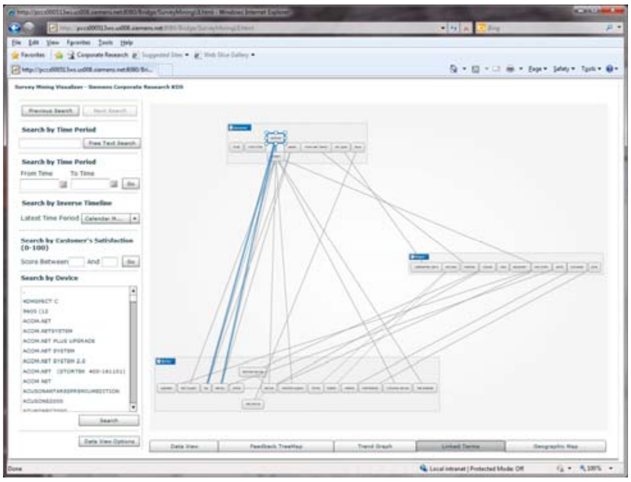

Survey Mining Visualizer - Siemens Corporate Research KDS
Search by Time Period
Search by Inverse Timeline
Search by Customer's Satisfaction (0-100)
Search by Device
Data View
Feedback TreeMap
Trend Graph
Linked Terms
Geographic Map

Figure 6: The interface showing linked term network.

7. EXPERIMENTS

Besides some experiments for keyword extraction and classification (I introduced in section 5), we also tested the whole system. In the experiments, we collected about 18k survey records from Siemens customers during the time between 2005 and 2011. Using our ontologies, we transformed these survey into about 200k RDF triples. Based on these triples, our experiments tested the system performance on answering various business intelligence related queries and compare them using different techniques.

First, we compared two choices of RDF store. Among others, AllegroGraph[1] and OWLIM[2] are two most popular RDF stores. AllegroGraph provides intuitive system management interface and so make it easy maintainable by non-expert of the Semantic Web. However it relatively requires higher performance machine to support efficient querying. Comparatively, OWLIM is more suitable for technical professionals and it is more efficient on query answering, since it is more memory based. Specifically, comparing on typical business intelligence queries required for our system's graphical exploration (discussed in Section 6), AllegroGraph usually needs 5 times time than OWLIM on a Duo core personal computer. Thus we choose OWLIM for all the following experiments.

[1] http://www.franz.com/agraph/allegrograph/

[2] http://www.ontotext.com/owlim

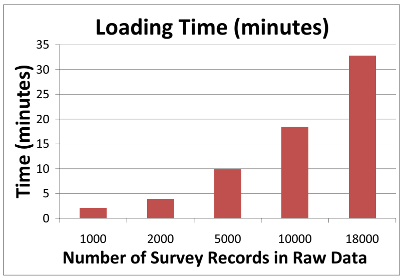


Figure 7: Loading time of the system.

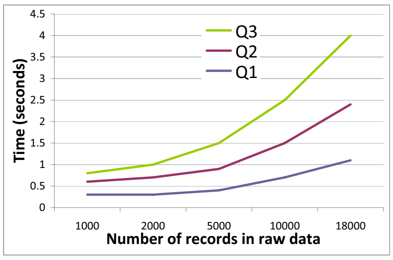


Figure 8: Query time of the system.

In the first experiment, we investigated the time complexity of "loading" original raw data into our knowledge base. Because it is not only simply "loading" the data, but also, more importantly, it retrieves the knowledge and makes them integrated through Semantic Web technologies (discussed in Section 4), this step could affect system user's satisfiability. Fig. 7 shows the performance trend when the size of original data varies. it can be seen that the time cost is linear with the size of raw data.

In the second experiment, we investigated the performance of query answering. Since there are numerous queries that can be issued to the system, we tested several typical business intelligence queries as follows.

1. Q1: The satisfaction of customers from a given state on a given product and its related services within a given half year.
2. Q2: The trend of keywords and their frequencies in comments given by customers with certain level of satisfaction in latest half year.
3. Q3: The common association between products and a given service class mentioned by customers and their potential to continue using the products.

All these queries are not only based on structural information given in raw data but also require semantic knowledge extracted from raw data. Shown as Fig. 8, all these queries can be retrieved efficiently in terms of running time.

8. CONCLUSION

To increase the quality of the products and services of the companies they serve, we propose a system semantically mining customer surveys. This system tries to solve the bottle-neck of retrieving the information in surveys input by customers in various forms. The four main research problems solved by this system are: 1) extract information on products or services in survey; 2) extract customer sentiment on related products or services; 3) design a platform supporting flexible intelligent business analysis using extracted knowledge; 4) intuitively present the mining and analysis results. The whole system is not just designed purely from a research perspective, because it operates easily given any type of input raw survey data, makes users understood the output. Not only the system but also the methodology have been adopted into real world production in world wide divisions of Siemens Corp.

9. REFERENCES

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[2] D. Brownell. *SAX2*. O'Reilly Series. O'Reilly, 2002.

[3] P. Buitelaar, P. Cimiano, S. Racioppa, and M. Siegel. *Ontology-based Information Extraction with SOBA*, pages 2321–2324. Citeseer, 2006.

[4] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. WWW '04, pages 462–471, New York, NY, USA, 2004. ACM.

[5] D. W. Embley. Toward semantic understanding: an approach based on information extraction ontologies. In *Proceedings of the 15th Australasian database conference*, pages 3–12, Darlinghurst, Australia, 2004.

[6] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422, 2006.

[7] Y. Li and K. Bontcheva. Hierarchical, perceptron-like learning for ontology-based information extraction. WWW '07, pages 777–786, New York, USA, 2007.

[8] A. Maedche, G. Neumann, and S. Staab. Intelligent exploration of the web. pages 345–359. 2003.

[9] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Stroudsburg, PA, USA, 2002.

[10] H. Saggion, A. Funk, D. Maynard, and K. Bontcheva. Ontology-based information extraction for business intelligence. In *Proceedings of the 6th international The semantic web*, ISWC'07, pages 843–856, Berlin, Heidelberg, 2007. Springer-Verlag.

[11] M. Vargas-Vera, E. Motta, J. Domingue, S. Buckingham Shum, and M. Lanzoni. Knowledge extraction by using an ontology-based annotation tool. *Knowledge Creation Diffusion Utilization*, pages 5–12, 2001.

[12] F. Wu and D. S. Weld. Automatically refining the wikipedia infobox ontology. WWW '08, pages 635–644, New York, NY, USA, 2008. ACM.

[13] C.-N. Ziegler, M. Skubacz, and M. Viermetz. Mining and exploring unstructured customer feedback data using language models and treemap visualizations. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '08, pages 932–937, Washington, DC, USA, 2008.