Application of CRISP-DM to Voice of Customer and BI integration

Lucie Sperkova
Department of Information Technologies, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic, lucie.sperkova@vse.cz

George Feuerlicht
Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia, Unicorn College, Prague, Czech Republic, george.feuerlicht@uts.edu.au

Abstract

Consumers are increasingly using social networks to share their opinions about products and services generating large amounts of unstructured Voice of Customer (VoC) data that can be utilized to significantly enhance the knowledge of likely future customer behavior and to analyze their buying intensions. In this paper we propose a methodology that helps organizations to systematically incorporate VoC data into their existing data warehouse. We have chosen the CRISP-DM data mining methodology as the basis of our approach for VoC integration. We focus on VoC data gained mostly from Web 2.0 sources, as this is still an open research problem. The paper proposes a reference model for the analyses of VoC content using a BI (Business Intelligence) process.

Keywords: *CRISP-DM, unstructured data, Voice of Customer, Business Intelligence, customer analytics, opinion mining*

1. Introduction

Current marketing focuses on customer interactions on the Web attempting to derive useful information about consumer behavior by storing this data in databases and by using Web analytics (e.g. Google Analytics) to analyze clickstreams. However, consumers are increasingly using social networks to share their opinions about products and services generating large amounts of unstructured *Voice of Customer* (VoC) data that can be utilized to significantly enhance the knowledge of likely future customer behavior and to analyze their buying intensions. The integration of this unstructured data with Web analytics data and data stored in enterprise applications, e.g. CRM (Customer Relationship Management) systems remains a challenge. Data generated by social networks and similar platforms present an enormous potential for knowledge acquisition via customer analytics and promise a more effective approach to marketing. Content analysis of VoC data can help to prioritize customer interactions, identify customer needs, and detect problems relating to products and services. Analysis of VoC data provides insights into customer sentiment and can assist in streamlining communication with customers, and lead to improved quality of products and services. This creates an opportunity to fine-tune marketing practices to establish a consistent approach to customers across all sales channels and lead to increase in the ROI (Return on Investment).

Manual analysis of unstructured data (e.g. customer reviews) is slow and error-prone and needs to be replaced by automated data-driven approaches. In this paper we propose a methodology that helps organizations to systematically incorporate VoC data into their existing data warehouse avoiding inconsistencies. We have chosen the CRISP-DM [1] data mining methodology as the basis of our approach for VoC integration with the Customer Analytics model. The methodology as an artefact of Business Intelligence processes for VoC data is applicable to various business domains. CRISP-DM is currently used by organizations across different industries to define their customer-focused business models [2]. Data mining is a widely accepted approach for the optimization of organizational business models, and as the CRISP-DM reference model is sufficiently generic and transferable we use this approach for VoC data integration.

doi: 10.4156/pica.vol5.issue1.2

In section II we discuss related work that deals with integration of unstructured data with existing data warehouses, in section III we describe the reference model for integration of VoC and in section IV we propose a methodology based on CRISP-DM. Section V illustrates usage of VoC analyses combined with e-commerce structured data.

2. Related Work

The emergence of social networks has enabled the collective intelligence of online users [3], which impacts on individual decisions and affects the operation of organizations. Social network VoC has been gaining increasing attention from managers as a new type of Business Intelligence [4]. Industry practitioners and researchers have started to address the problem of integration of unstructured data into existing BI (Business Intelligence) solutions. Kantardzic [5] argues that the analysis of structured and unstructured data is essential to provide valid insights into current business developments. In our previous research [6], [7] we have identified a gap between the business perspective on VoC and that of data of computer science. The marketing approach is to attempt to automate VoC content analysis, but business researches in general lack the required computing expertize, while computer scientists do not have sufficient marketing knowledge to systematically incorpotate VoC content analysis into marketing. To bridge this gap we propose the integration of VoC analysis with BI processes. Inmon et al. have proposed a methodology for integration of unstructured data and textual analytics into BI [8], but, they do not include VoC as a potential source of data. Sukumaran and Sureka [9] used text tagging and annotation techniques to integrate structured and unstructured data. Yaakub et al. [10] use a similar approach based on factual data and descriptors to integrate unstructured and structured data into a Customer Analytics model. The authors propose integration of an ontology model for opinions about products based on customer reviews into a multidimensional model.

BI integrated framework that includes unstructured data was constructed by Baars and Kemper [11], but the authors focus only on traditional enterprise data and data derived from a CRM application. Similar to [8] they do not include VoC as a potential source of data for the BI process. The authors use three approaches to integrating structured and unstructured data embedded in the three-layer BI framework. The first approach uses a portal add-on to build a bridge between the logic layer and the access layer that links search function and data navigation. The second approach uses tools to integrate components from the data layer to analyze content and extract metadata. The third approach involves a middleware hub at the logic layer for the exchange of the results and templates between the analytical systems and distributed knowledge components.

Chau and Xu [12] designed a generic framework for content analysis of blogs for BI purposes. A hybrid framework that integrates domain knowledge with a data-driven approach to analyze semi-structured customer data was proposed by Peng et al. [13]. Varadarajan and Soundarapandian describe a methodology for the refinement of text mining and opinion mining tool for domain knowledge gaining accurate sentiment from data through different Web 2.0 tools [14].

3. Reference Model For VoC Integration

For purposes of analysis of the VoC content using BI, information obtained from VoC has to be combined with information from existing structured data stored in a data warehouse. Figure 1 shows a simple reference model for VoC analysis that transfers comments from customer text files and loads them into a multidimensional data warehouse, structured as fact and dimension tables. Existing analytical methods and techniques are used for content and sentiment analyses of customer posts stored in the data warehouse combined with structured data derived from enterprise applications, e.g. CRM.

4. CRISP-DM for VoC Integration

VoC integration can be used to enhance data from a CRM system using a five-step procedure: Retrieve, Clean, Store, Update and Distribute [15]. At the end of the integration process unstructured and structured data are stored in different dimensions of the data warehouse. Inmon [8] has proposed basic steps for setting up and running an iterative process of visualization of unstructured data that includes the following steps: initial setup, source setup, integration, verification of results, loading of data into a relational database and visualization. Fig. 2 illustrates how tasks from [8] and [15] are combined with the CRISP-DM method.

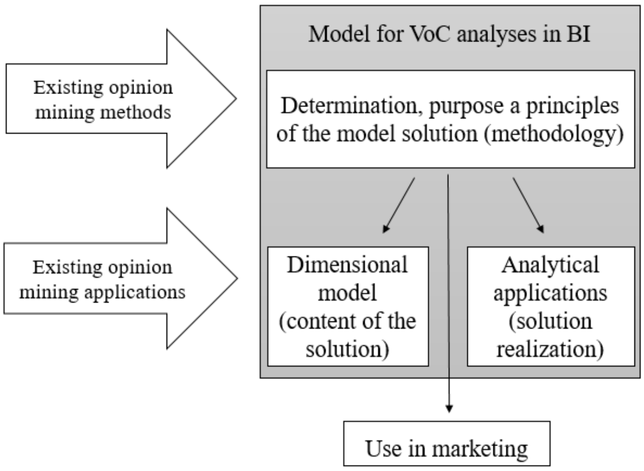
Existing opinion mining methods
Existing opinion mining applications
Model for VoC analyses in BI
Determination, purpose a principles of the model solution (methodology)
Dimensional model (content of the solution)
Analytical applications (solution realization)
Use in marketing

Figure 1. Reference model for VoC analysis using BI processes.

The goal of the new methodology based on CRISP-DM is to integrate existing knowledge into a comprehensive VoC analytics model, mapping the VoC entities to the corresponding concepts in the CRM customer database, storing the data in a fact table showing the associations between VoC and other data dimensions.

The methodology can be divided into three parts. Technical data processing is basically textual ETL that includes Data Understanding and Data Preparation phases of the CRISP-DM methodology. This task includes data transfer, formal check and preliminary survey. According to Rainardi [15] the following steps are required: Retrieve, Clean and Store from CRM, data integration, and source setup task derived from [8]. The analytical part uses opinion mining algorithms performed on the extracted data from textual ETL processes and their evaluation for accuracy. Business processing part appears at the beginning and at the end of the whole cycle to meet the desired business objectives. The final distribution and deployment tasks involves running visualizations, reporting and creating data cubes [10]. The entire data mining cycle is iterative and can return to the previous stages to improve the accuracy of results. The arrows indicate the most significant and frequent dependencies between phases.

4.1 Business Understanding

Business understanding is an initial phase of the methodology. Data analyst must understand the objectives from a business perspective and know what kind of results marketing needs in a particular domain. The cooperation of domain experts is required to determine and to define dimensions that are to be evaluated and to determine the subjects and notions (words, phrases) related to the problem domain.

4.2 Data Understanding

The main goal of this phase is to integrate data for the analytics of unstructured data. It includes *VoC collection*: collection, indexation in the repository and filtering VoC-related content from the web.

4.2.1 Collection of initial data

The first step in VoC integration defines the sources of data, determining if the unstructured documents have relevance to the business and estimating the cost of acquiring data. It is important to select a reasonably homogenous data set across a specific set of documents. Data transfer is typically implemented using a standard format, e.g. CSV and should meet the security standards of the organization. All data records should have unique identifiers and a timestamp and the text fields should be correctly coded. The raw text data is entered into the integration software where several kinds of editing are performed using data extractors over pre-defined fields using pre-defined method of processing, combining this data with customer data extracted from a CRM system.

4.2.2 Data description

Pre-reading of the data and *domain definition*. The domain definition involves consultations with the domain experts and it includes the development of dictionaries needed for the analytics. More specifically, it includes he the following tasks:

- Synonym determination and development of synonym vocabulary: this task creates a list of synonyms linking slang words with formal language expressions and abbreviations with their full meaning
- Stop-words list development.
- Homographic determination and development of homographic vocabulary
- The definition of external categories and grouping of words according to the categories

4.2.3 Data exploration

This phase involves explorative analysis of the data. Visualization tools can be used for clustering and histograms.

Word/phrase histograms: Set of unique words across a collection of documents is ranked by their relative frequencies. A word histogram does not check if each word is valid in the particular language, all strings and characters like misspellings, personal names and other words not found in the dictionary are included, helping to identify common typographical errors.

Clustering analysis: Data clustering is used to identify the most discussed entities. The clustering analysis involves the determination of the words or phrases and their frequency and relationships.

4.2.4 Verification of Data Quality

Assessment of the quality of collected data.

4.3 Data Preparation

Data preparation involves transformation and loading of the textual ETL. It involves the transformation of the data to the format that will be loaded into the data warehouse. This phase involves *parsing the data and the data type definition* for the individual fields.

4.3.1 Data selection

Determining if customer comments have any relevance to the business is another important step. If the comment is not relevant to the business problem, the comment is not included in the data warehouse for textual analytics. The word histogram, clustering analysis and domain definitions are used as the input to this subtask.

4.3.2 Data cleaning

Information extraction and Natural language processing using the text fields (textual ETL) has to be performed first. The general group of activities include the setting of the word parameters:

- word stemming,
- negativity exclusion,
- removal of stop-words,
- removal of often repeating words,
- synonym editing where synonyms can be either replaced or concatenated,
- homographic editing – multiple meaning of the word is broken into different possibilities,
- grouping of words according to external categories,
- part of speech tagging - to recognize appraisal words in a sentence,
- automatic indexing enables search over a data corpus

4.3.3 Data construction

Subject collection (feature extraction): determines if the comments relate a business issue - search for the occurrence of specific terms (e.g. name of the brand, service components, or phrases entered by users). Additional descriptive attributes for each word or phrase including length, document frequency (i.e. number of documents it appeared one or more times in), global frequency (number of times it appeared in the entire collection), and part of speech.

Subject filtering (feature selection): filtering the terms (i.e. words, phrases) from the content. Terms are compared with the database of existing categories.

Synonyms/homonym assignment to filtered subjects

Categorization: of subjects to predefined dimensions and to new dimensions recognized by clustering.

4.3.4 Data integration

This task involves the integration of the text data and the creation of the data model for the data warehouse. Based on filtered subjects and features with the combination of domain knowledge an ontology is developed. The sentiment words (appraisal words) are identified. All opinion sentences are loaded into the fact tables using their interaction identifier. This fact table is associated with other dimensions in the data warehouse and with the customer database. All other information about selected subjects, features, lemmas stems, etc. are also stored in tables and related to the table with opinion sentences with the interaction identifier.

4.3.5 Data formatting

After all relevant data is loaded into the data warehouse, the format can be modified without changing the meaning of the data, so that the opinion mining algorithms can be run over the VoC data.

4.4 Modelling

Modelling is the analytical part of the methodology. It involves *deep* analysis of unstructured data and opinion mining.

4.4.1 Selection of modelling technique

In this task, the appropriate opinion mining method is chosen. It includes the setting of the appraisal attributes analysis rules (appraisal word is a word which is the subject of the estimate).

4.4.2 Generate Test Design

This involve the separation of the dataset into the training and testing data. The model is built using the training data set and its correctness is verified using the test data set.

4.4.3 Build Model Parameter Settings

Appraisal word monitoring - calculation of appraisal words weights.

Subject and feature matching: matching the sentences according to the collected subjects and features.

- *Appraisal sentences evaluation*
    1. *appraisal sentence marking*: sentences that include appraisal words are identified out and marked as appraisal sentences
    2. *appraisal sentence analysis*: adjectives and adverbs are then examined to see whether they are positive or negative, and whether they have been negated (i.e. occurrence of words like “not”)
    3. *determination of sentiment* of the claims about the service components/attributes
- *Descriptive frequency statistics analysis*: to normalize the results for synthesis with the results gained by other components metrics - measurement of the subjects in claims, assigning weights to individual subjects and dimensions and their sentiment

4.4.4 Evaluate the Model

Evaluation of the model quality using the test data set measures the accuracy of the results from the opinion mining based on statistical tests.

4.5 Evaluation

Evaluation of model by domain experts and conclusion; the domain experts check if the results are sensible and identify the shortcomings that need to be improved in next iteration.

4.6 Deployment

The deployment phase includes reporting and data visualization. The combination of customer entity and opinion extraction is prepared for the design of metrics for customer analysis. Different metrics are evaluated across different dimensions. The findings are then assigned to particular marketing activities, which are proposed in next section.

The proposed methodology is an iterative process with business roles assigned responsibility for individual phases, and should be used in the long term continually improving over the business cycle. The methodology is transferable within the same business domain.

5. Design of VoC Usage in Customer Analytics

When integrating VoC into the BI process, it is necessary to define the purpose for which the data will be collected and analyzed. This proposal focuses on customer analytics:

- Detection of products that customer speak positively about and subsequent targeting of advertising. It may be possible to use cross-sell and up-sell activities if the identity of purchasing customer is known.
- Prediction of whether a customer stays or churns.
- Customer segmentation for targeted advertising.
- Identification of the sentiment of direct communication for handling requests by priority customers.
- Targeting customer's friends who speak positively about the product through social networks.
- Correlation of the development of new customers with the overall attitude of VoC.
- Correlation of the development of sales with the overall attitude of VoC.
- Forecast which product can be more profitable than other products.
- Analysis of the influence calculated over Customer Lifetime Value according to his VoC.
- Redistribution of the campaigns budget according to the attribution of natural *Word of Mouth*
- Detection of opinion makers (i.e. people that influence opinions of others) and the subsequent targeting of advertising.

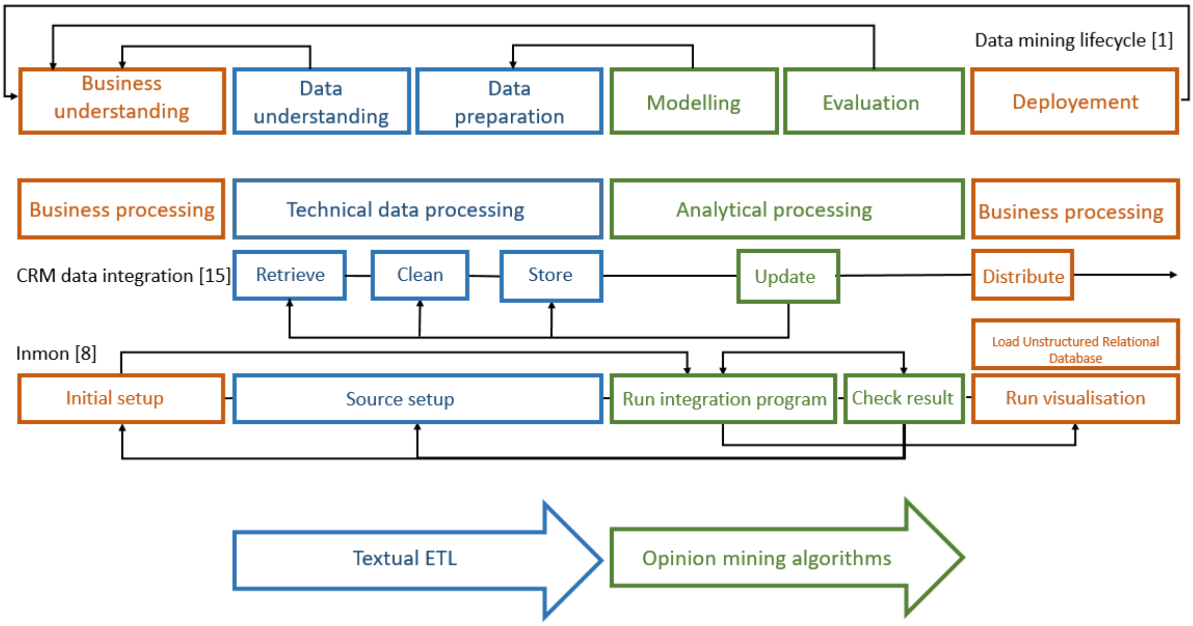

Figure 2. CRISP-DM for integration of VoC to the Customer Analytics.

6. Conclusion

In order to assist marketing professionals who typically lack the knowledge of data-driven analyses of VoC, we propose the integration of VoC to the BI processes using a methodology based on CRISP-DM. This involves the definition of the key indicators and dimensions that are used for analytics and opinion mining. The proposed methodology is generic and can be utilized in different business domains to assist in marketing products and services. A reference model was described that can be used as a guideline for the implementation of a specific solution. Our future research efforts will focus on providing more detailed description of each phase of the methodology.

Acknowledgment

This paper was prepared thanks to the IGA grant VSE IGS IG406025 and with the contribution of long-term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

References

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, *CRISP-DM 1.0 Step-by-step data mining guide*, 2000.

[2] K. K. Tsiptsis and A. Chorianopoulos, *Data mining techniques in CRM: inside customer segmentation*, John Wiley & Sons, 2011.

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

[4] H. Chen, "Business and market intelligence 2.0, Part 2", *Intelligent Systems, IEEE*, vol. 25 no. 2, 2010, pp. 74-82.

[5] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*, Wiley-Interscience, 2003.

[6] F. Vencovský, L. Šperková, "IT Service Quality Model: Evaluation of Quality In Use", *Proc. 16th European Conference on Knowledge Management*, 03.09.2015 – 04.09.2015, Udine, Academic Conferences and Publishing International Limited Reading, 2015, pp. 821–827.

[7] L. Šperková, P. Škola, Petr, T. Bruckner, "Evaluation of e-Word-of-Mouth through Business Intelligence *processes in banking domain", Journal* of Intelligence Studies in Business, 2015, vol. 5, no. 2, pp. 36–47.

[8] W. H. Inmon and A. Nesavich, "Tapping into unstructured data: integrating unstructured data and textual analytics into business intelligence", Pearson Education, 2007.

[9] S. Sukumaran and A. Sureka, "Integrating structured and unstructured data using text tagging and annotation", *Business Intelligence Journal*, vol. 11 no. 2, 2006, pp. 8-17.

[10] M.R. Yaakub, Y. Li, A. Algarni and B. Peng, 2012, December). "Integration of opinion into customer analysis model", *Proc. IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 03, IEEE Computer Society, Dec. 2012, pp. 164-168.

[11] H. Baars and H. G. Kemper, "Management support with structured and unstructured data—an integrated business intelligence framework", *Information Systems Management*, 2008, vol. 25 no. 2, pp. 132-148.

[12] M. Chau and J. Xu, Jennifer, "Business intelligence in blogs: Understanding consumer interactions and communities", *MIS quarterly*, 2012, vol. 36 no. 4, pp. 1189-1216.

[13] W. Peng, T. Sun, S. Revankar and T. Li, "Mining the "voice of the customer" for business prioritization", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3 no. 2, 2012, pp. 38.

[14] S. Varadarajan and Soundarapandian, "Maximizing Insight from Unstructured Data", *Business Intelligence Journal*, vol. 18 no.3, 2013, pp. 17–26.

[15] V. Rainardi. Building a Data Warehouse with Examples in SQL Server. Apress, 2008.

Lucie Sperkova is a PhD candidate at the Department of Information Technologies, Faculty of Informatics and Statistics, University of Economics in Prague where she also gained her master degree. Her main field of research is unstructured data use in Business Intelligence and Customer Analysis. Ing. Sperkova worked as a Business Intelligence analyst in banking, e-commerce and digital marketing in Czech Republic. She presented multi-channel attribution modelling at the largest international marketing conference in Czech Republic, Marketing Festival 2016. Currently is a visiting research student at the University of Technology, Sydney.

George Feuerlicht is a Research Associate at the School of Software, Faculty of Engineering and Information Technology, UTS. George is the author of over 70 publications across a range of topics in information systems and computer science, including recent publications on enterprise architectures, SOA, and Cloud Computing models. George is a member of ACM and a number of conference organizing and program committees. He holds a PhD from the Imperial College, London University, U.K.