TOWARDS INTEGRATING ONLINE SOCIAL NETWORKS AND BUSINESS INTELLIGENCE

Paulo R. S. Costa
Universidade Federal de Pernambuco
prsc@cin.ufpe.br

Fernando F. Souza
Universidade Federal de Pernambuco
fdfd@cin.ufpe.br

Valéria C. Times
Universidade Federal de Pernambuco
vct@cin.ufpe.br

Fabrício Benevenuto
Universidade Federal de Ouro Preto
fabricio@iceb.ufop.br

ABSTRACT

The growing use of Online Social Networks (OSN) in recent years is attracting the interest of the corporate world, where departments interested in analyzing their contents have been dealing with such technology. There are a number of proposals in the literature illustrating algorithms for social network analysis and sentiment analysis to discover, respectively, patterns of relationships between individuals and qualitative aspects in recorded statements. Despite the importance of integrating social network and sentiment analysis with user's decision making processes, there is a lack of research aimed at achieving such integration as far as OSNs are concerned. This paper proposes a Business Intelligence Architecture, called OSNBIA, to achieve such integration. A case study has been developed to illustrate the proposed architecture. It extracts data from Twitter and applies social network and sentiment analysis to generate a data warehouse, enabling thus new possibilities of OSN's data manipulation through Business Intelligence technology.

KEYWORDS

Online Social Networks, Business Intelligence, Social Network Analysis, Sentiment Analysis.

1. INTRODUCTION

Online social networks have mobilized an increasing integration of individuals around the planet. Structured on Twitter (twitter.com), Facebook (www.facebook.com), Myspace (www.myspace.com), Bebo (www.bebo.com) and LinkedIn (www.linkedin.com), among many others, comprise millions of dispersed and interconnected users attracted by some kind of affinity (e.g. political, commercial, religious, recreational, educational, emotional or professional).

The 22% of the time in which a user is online is dedicated to the interaction with social networks, corresponding to 110 billion minutes (Nielsen, 2010). Aspects inherent to human nature, as the natural interest to relate with their peers, sharing ideas and reviews (Curran, et al., 2010;D'Andrea, et al., 2010), coupled with major technological advances in the direction of Web2.0, through which the user will now have much more control on creating and sharing contents (Oreilly, 2007), contribute towards online social networks achieving new levels of popularity, thitherto unimagined.

The corporate world has followed the popularity and growth of online social networks, identifying a fertile ground for the dissemination of products and services, to strengthen their brands, to monitor concurrence, as well as prospecting new customers (Dawson, 2003;Weber, 2009;Mackelworth, 2007).

There is also a huge interest in the content inherent to social networks, inducing corporations to seek a better understanding of what is registered in them. The Gartner Group (2010) reinforces the importance of this initiative, from the moment that among the ten technologies considered strategic for corporations for the next three years, it cites the importance of online social networks and analysis of its contents. In this context, questions like: What are you talking about my brand, my products or services? What about the incidence of positive and negative statements? Which aspects are most relevant in relation to the statements found? Who (profile) are making these statements? Where are they? Are there influential users in this net of relationships? Among many others, these are topics of interest to the corporate world in order to obtain intelligence from online social networks, contributing to the development of marketing strategies.

The analysis of information from online social networks highlights two areas of research: Social Network Analysis (link mining) and Sentiment Analysis (opinion mining). The first is the result of a set of research in Social Networks, Link Analysis, Hypertext and Web Mining, with the intention of analyzing patterns of descriptive or predictive relationships between the elements of social networks (Curran, et al., 2010; Han & Kamber, 2006).

Sentiment Analysis refers to the computational treatment of a text in order to identify whether it represents a positive, negative or neutral statement about a given topic, including areas such as Information Retrieval, Natural Language Processing and Text Mining (Pang & Lee, 2008; Liu, 2008).

Based on the analytical needs of the business world, as previously mentioned, one can deduce that both link mining and opinion mining could contribute to more comprehensive analysis of online social networks. In this sense, a computing environment that integrates such technologies, allowing both structural analysis of relationships and qualitative analysis of recorded testimonies, should become valuable for corporations. There has not been observed in the literature, studies dealing with such integration. This article proposes a decision support environment integration, through Business Intelligence (BI) technology, data from online social networks (over which link mining and opinion mining algorithms are applied) to corporate relational structured data. This article is organized as follows: in Section 2, the main concepts related to this work are discussed; Section 3 presents and analyzes related works to the theme of this article; the proposed architecture, called Online Social Networks Business Intelligence Architecture (OSNBIA) is detailed in Section 4; in Section 5, a case study is presented, illustrating the proposed architecture. It is based on the extraction of data from Twitter, incorporating link and opinion mining to tweets, and also integrating the social data warehouse to a corporate warehouse. The resulting data warehouse is then integrated into a BI tool. In the same section, we present some answers to questions that can be asked by business users about the data warehouse implemented; conclusions and suggestions for future works are presented in Section 6.

2. GENERAL CONCEPTS

This paper proposes an architecture that integrates technologies, involving Online Social Networks, Link Mining, Opinion Mining and Business Intelligence. In this sense, it is necessary to contextualize such technologies, providing the basic foundations for understanding this proposal.

2.1 Online Social Networks

Social Network theory elements are found since the ancient Greeks. It´s credited to John A. Barnes, James C. Mitchell and Elizabeth B. Spillius the first fieldworks related to social network analysis. Barnes (1954) investigated social groups at Bremner (Norway), reporting that the connections between individuals were motivated by common affinities (not only by kinship or friendship), forming cohesive groups, and that such connections could transcend the limits of the village, having a direct impact on decision making and individual motivation. Mitchell (1969) defines social networks as an interconnected set of individuals whose behavior could be understood through the characteristics of their links as a whole.

A social network, from the perspective of data mining area, can be defined as a set of heterogeneous data related to each other represented by a graph (Han & Kamber, 2006). The structure of a graph consists of nodes and links, the latter showing the relationships between nodes, either uni or bidirectional. In this sense, social network does not mean that are necessarily composed by individuals. In the real world, there are numerous examples of social networks linked to other areas (e.g. biology, economics and technology). One can cite examples such as graphs representing telephone calls, signaling the spread of disease, diagramming the flow of e-mails exchanged between users, among others. This paper will analyze social networks of individuals structured by Social Network Sites (SNS), called Online Social Networks (OSN).

According to Boyd & Ellison (2007), Social Network Sites are Web-based services that allow individuals to characterize their profiles and articulate relationships with other users in order to share information, allowing them to view and traverse their direct and indirect relationships. Despite the subtle conceptual difference between SNS and OSN, the term adopted throughout this article will be OSN, because it is widely used in related works (Benevenuto, 2010; Mislove, et al., 2007; Cachia, et al., 2007).

Online social networks have common features, especially, according to Benevenuto (2010): user profiles - aspects related to user characteristics such as demographics (location, age, gender, education) and issues of interest (religion, sports, politics , music, literature). Such a profile can act as an integration element with other individuals, due to a strong relationship between the real profile and the one registered in social networks (Boyd, 2008); updates - in order to motivate the use, new content placed on social networks are updated in real time, being visible to all users that are part of the direct or indirect relationship of an individual; comments - content entered by a user can be commented by other members of the social network; evaluations - a user can classify the content posted by others (e.g. "like" on Facebook, or "like this" on Youtube); favorite lists - allow the user to better organize his/her topics of interest and may serve as recommendations for others ; top lists - evidences hot topics being mentioned in a given period, which may serve as an instrument for the dissemination of knowledge; metadata - possibility to create references to user content (e.g. title, description, category and keyword found in Youtube, # hashtags in the case of Twitter).

2.2 Social Network Analysis (Link Mining)

Social Network Analysis identifies patterns of relationships between individuals in social networks, assuming that these patterns represent important aspects of their lives. It is believed that the way in which an individual lives depends to a great extent on how he/she presents him/herself connected to a wider social network. From the 70's, with the evolution of Graph Theory and the emergence of computers widely available for research in this area, social network analysis emerged as an interdisciplinary knowledge area. In this sense, its application has been used in organizational behavior, relations between organizations, analysis of the spread of contagious diseases, among many other areas (Freeman, 2004).

From the perspective of data mining area, social network analysis is known as link mining, including a convergence of research in social networks, link analysis, hypertext and web mining, graph mining, relational learning and inductive logic programming, providing both descriptive and predictive analysis scenarios. Some are related to link mining algorithms: link-based object classification identifies the category of a node in the network not only by its attributes, but also by its relationships (links) and the attributes of the related nodes; object type prediction is similar to the above, but referring to the type of the node; link type prediction~~:~~ identifies the type or purpose of a link, based on the properties of the nodes involved; predicting link existence assesses whether two nodes have some kind of connection; link cardinality estimation provides the number of links of a node or the number of intermediate nodes between two others; object reconciliation assesses if two nodes are identical, according to their attributes and relationships; group detection identifies the existence of groups (cluster) of nodes with common structural characteristics; and sub-graph detection finds sub-graphs in existing networks (Han & Kamber, 2006; Gettor & Diehl, 2005; Getoor, 2003).

Some measures can be obtained from the application of link mining algorithms: betweenness - degree of connectivity from one node to their neighbors, possessing greater importance nodes interconnecting clusters. A node with high betweenness has great influence over what flows in the network; degree - number of direct connections a node has. Individuals with high degree are called hubs or connectors; closeness - degree of direct or indirect proximity of one node to others in the other network. Individuals with a good degree of closeness are close to any network node, having a clear view of what flows in it; centralization - centralized networks are characterized by a dependence on one or a few central nodes. A centralized network around a hub node is susceptible to failure from the moment the respective node is removed; reach - the degree to which a network node can reach other members of the network; density - a high density suggests that the number of links between the nodes is close to its maximum; clustering coefficient - measures the tendency of a graph to form clusters (Mislove, et al., 2007; Müller-Prothmann, 2008).

2.3 Sentiment Analysis (Opinion Mining)

Sentiment Analysis (opinion mining) evaluates, computationally, opinions, emotions and sentiments expressed in a text. It tries to automate the retrieval process from relevant sources of information, extracting relevant sentences, interpreting its contents and summarizing/presenting the results in a friendly way. There is an increasing growth of studies in recent years that has its origins in the late '70s and early '80s.

According to Liu (2008), despite the great importance of opinion mining, there were a small number of researches in this area before the advent of the World Wide Web. This fact refers to restrictions imposed to the collection of opinions in the past. The explosion of the generation of opinions on the web, through product review sites, web feeds, blogs, forums, discussion groups and social networks, as well as advances in machine learning methods applied to natural language processing and information retrieval, is driving researches in opinion mining and motivating the interest of the corporate world on types of information that can be obtained from these media.

Liu (2010) presents the steps that comprise the sentiment analysis process: opinion identification - retrieval of relevant opinions; feature extraction - identifying objects and features over the opinions to which they refer; classification of feelings - determining the opinion's polarity; and visualization - presentation of results in a friendly way to the decision maker.

2.4 Business Intelligence

The term Business Intelligence System is credited to Luhn (1958), who defines Business as a set of activities for any purposes (e.g. industry, commerce and government), being provided by an Intelligent System able to assimilate interrelationships between these facts, in order to guide actions to achieve a desired goal.

Corporations need an increasing intelligence capability to be competitive, as they need to anticipate and react to changes that occur in the context in which they operate. Business Intelligence fills the need that many companies have today: finding the right information; understanding what it means for business; and putting it in the hands of the right people, in order that decisions can be made at higher condition of certainty and minimum risk (Gilad, 1988).

According to Sallam et al (2011), the market for BI solutions continues with one of the highest growth rates in the software market (7% per year until 2014). Some aspects will be critical to expanding the use of BI solutions on the market: more intuitive and simple interfaces; support to mobility; good performance when dealing with the expansion of the data volume; ability to handle unstructured data; incorporation of features capable of handling data from social networks; greater integration to business processes; features for simulation and predictive analysis; support to collaborative decision making processes; and easier ways to integrate departmental silos of information to the corporate context.

Business Intelligence basically comprises: (a) the extraction, transformation and loading of data (ETL) from structured sources (e.g. ERP, CRM, SCM and Legacy Systems) and/or unstructured data (e.g. Online Social Networks, Blogs, Videos, E-mails, Text Documents, Chat, among many others), resulting a data warehouse. This includes a corporate data repository that is topic-oriented, integrated, time-variant and nonvolatile (Inmon, 1996); (b) the use of analytical tools, integrated to the data warehouse, for the analysis and dissemination of knowledge (On-line Analytical Processing, Ad hoc Querying, Reporting, Data Mining, Dashboarding and Alerts).

The integrating nature of a data warehouse, the fact that its modeling is directed to the decision making process, the possibility of integration with user friendly analytical tools, the business need of a better understanding of unstructured data that are found in online social networks and the existence of mechanisms for its analysis provide the basis for proposing an architecture involving all these technologies. The goal to be achieved is a decision making environment able to deal with unstructured (e.g. OSNs) and structured data (e.g. corporate data warehouse) in a flexible, user friendly and dynamic way (BI technology), enriched by qualitative and quantitative perceptions of unstructured data (Opinion and Link Mining technology). In this way, as it can be seen in the next sections, there are opportunities for investigations focused in this approach.

3. RELATED WORKS

The importance of unstructured data for decision making process has been evidenced in numerous works (Bhide, et al., 2008; Perez, et al., 2007; Park & Song, 2011; Moya, et al., 2011). According to them, a small percentage of corporate data is structured and stored in relational databases; while the vast majority is unstructured, registered in e-mails, memos, call centers notes, online social networks, web forums and chat rooms.

The incorporation of unstructured data to a decision making environment represents a business challenge, because current techniques and technologies of Business Intelligence are not adequate to deal with it.

3.1 EROCS

Bhide et al. (2008) propose the integration of text documents to relational databases through a system called EROCS (Entity Recognition in the Context of Structured Data).

Information sources for EROCS include a set of emails containing customer complaints about various issues and a data warehouse covering corporate information about the business. Each e-mail is submitted to an UIMA Annotator (uima.apache.org), responsible for identifying relevant entities contained therein and based on the entities that comprise the data warehouse dimensional model (e.g. Clients, Shops, Products and Suppliers) produces, as a result, a Link Table associating some text elements to entities of the dimensional model.

The architecture also provides the possibility to incorporate, into the Link Table, opinions resulting from the application of opinion mining algorithms over the e-mails sent by customers.

As a result, OLAP cubes are built from the Link Table, allowing the implementation of MDX queries to answer questions like "How many complaints about product X, grouped by store do we have?", as well as providing for the end-user features typical to OLAP technology (slice, dice and drill down/up).

3.2 Contextualized Warehouse

Perez et al. (2007) present data integration architecture (contextualized warehouse), whose main components are the corporate data warehouse, the XML document warehouse and the fact extractor module.

Initially, the end user has a business context to be analyzed (set of keywords) with multidimensional expressions showing the dimensions and measures of interest.

Using techniques of Information Retrieval plus their relevance, documents are retrieved from the XML document warehouse. The Fact Extractor Module performs the parsing of the obtained documents, returning the set of facts described therein, added with their frequency. The multidimensional expression is submitted to the corporate warehouse and the resulting facts are associated with the documents retrieved in the previous step.

The results are embodied in a R-cube (Relevance Cube), where OLAP operations are available, involving dimensions and measures within it.

The study case of the application of this architecture took into consideration a data warehouse consisting of the historical evolution of some indicators of global market stocks, as well as a set of business news rescued from international newspapers. From the combination of these knowledge bases, facts that may have influenced the growth or decline of market stocks in a given region could be assessed.

3.3 Total Business Intelligence Platform

Park and Song (2011) provide the integration of structured and unstructured data, enabling a Total Business Intelligence Platform, using technologies like Information Retrieval (retrieval of documents based on keywords provided by a user query); Text Mining (extraction of the main keywords, summarization, classification and clustering of documents); and Information Extraction (extraction of structured information based on a schema provided by the user) and OLAP.

Text OLAP (multidimensional analysis of textual documents) integrated to Relational OLAP foster the creation of a Consolidation OLAP, able to handle both structured and unstructured data. The integration between the OLAP and Relational OLAP Text is done by means of shared dimensions.

The analysis can be initiated either by the Relational OLAP or from Text OLAP. From Relational OLAP, aspects like where, when, how and who performed what can be rescued. If there is a need for an analysis of the background facts involving the rescued facts, document reviews are performed by using Text OLAP in search of reasons for the occurrence of such events. In the opposite direction, documents can be redeemed through Text OLAP, and in a supplementary form, business facts that occurred in the same period can be found.

3.4 Web Feeds and Corporate Warehouse

Moya et al. (2011) have proposed an integration of feelings, expressed through web feeds, to a corporate data warehouse, enabling OLAP analysis to be made.

Taking as starting point a comment made by a user about a product through a web forum, opinion mining is applied in two levels of granularity: the feed as a whole and to the aspects retrieved from the feed.

The integration among the opinions (Sentiment Model) and the corporate data warehouse (Corporate Model) was achieved by shared dimensions like Product, Time and Location.

In this approach, we highlight the possibility of sentiment analysis at the aspects level of granularity. In this sense, it could be considered the impact of higher sales of a particular product, taking into account the incidence of negative opinions, positive and neutral, and also relevant aspects of the product evaluated.

3.5 Concluding Remarks

It can be seen, through the works analyzed, the importance and corporate interest in the integration of structured information to unstructured ones.

However, the information scope was based on data derived only from unstructured documents (e.g. news, e-mails and memos), except with Moya et al (2011). Table 1 summarizes some aspects of the previous architectures, based on the technologies proposed in this article:

Proposed Architecture	**Information Scope**	**Link Mining**	**Opinion Mining**	**Business Intelligence**
EROCS	Structured and Unstructured	Not Supported	Supported	Supported
Contextualized Warehouse	Structured and Unstructured	Not Supported	Not Supported	Supported
Total BI Platform	Structured and Unstructured	Not Supported	Not Supported	Supported
Web Feeds & Corporate Warehouse	Structured and Unstructured	Not Supported	Supported	Supported

Table 1 – Comparison of Architectures

At this time, we identify an opportunity of research involving the integration of technologies like Online Social Networks, Opinion Mining, Link Mining and Business Intelligence, exploring them broadly.

4. PROPOSED ARCHITECTURE

The software architecture proposed in this paper, named Online Social Networks Business Intelligence Architecture (OSNBIA) is focused on the feasibility of a Business Intelligence environment capable to support organizational departments of Marketing and Social Media for a better interpretation of topics of interest recorded in online social networks. Such environment can allow such events to be related to corporate data (structured data).

This proposal fills up a gap in the literature regarding the analysis of data from online social networks, and the integration of Online Social Networks, Opinion Mining, Link Mining and Business Intelligence.

Figure 1 show the architectural components, which are described in subsequent sections.

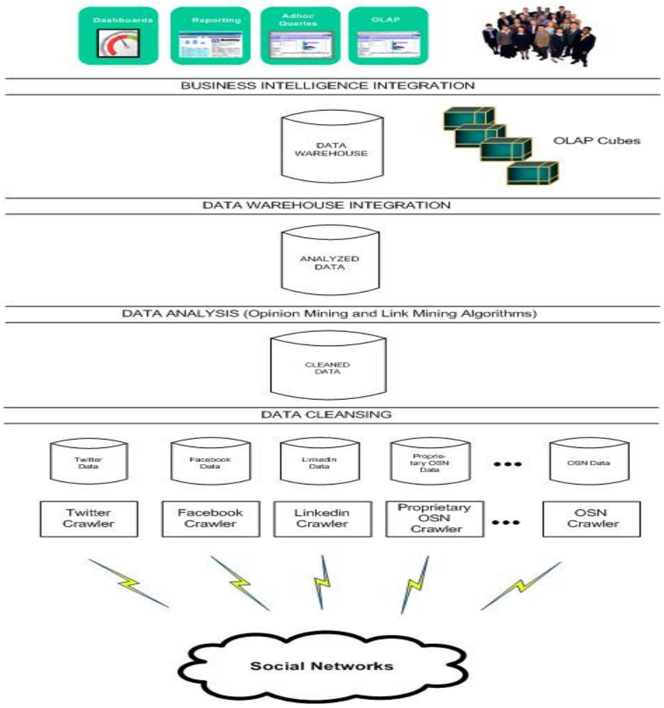


Figure 1 - Online Social Networks Business Intelligence Architecture

4.1 Social Networks Crawling

Using Application Programming Interfaces (API) provided by Social Network Sites, crawlers can be implemented through various possibilities of programming languages (e.g. PHP, Python, Java, among others), with support to various result types of data formats (e.g. XML and JSON).

The architecture also provides the possibility of extracting data from online social networks developed by the corporation itself (e.g. British Social Telephone Network (Sass,2010)) as well as online Decentralized Social Networks based on peer-to-peer architecture (Berners-lee et al, 2009).

The extraction of this data should follow a standardized and comprehensive template~~,~~ to meet all social networks covered by the architecture, originating flat files to be used by the Data Cleansing stage.

The frequency of data extraction will depend on the limitations imposed by the online social network, the availability of hardware to process large volumes of data and management's interest for more frequent data.

4.2 Data Cleansing

Once extracted, crawled data are submitted to quality operations. Because of the restrictions imposed by the online social networks (e.g. connection time to carry out transactions), it´s suggested a separate module to handle such issues. The main goal is to correct inconsistencies of data before transferring them to the next phase, enabling the generation of a Cleaned Data Repository.

Aspects such as completeness, consistency, validity, conformity, accuracy and integrity are treated in this phase (Singh & Singh, 2010). In the case of missing data, for example, arising from the impossibility of extracting some attributes (due to the lack of the same data in a social network) or for the lack of content (for non-registration), the NOT AVAILABLE constant should be used to fill these attributes, avoiding the existence of void content.

Spam/Spammers detection may be addressed in this phase (Benevenuto et al., 2010), as well as location normalization (e.g. in case of Twitter, the location attribute is filled in free form or contains the latitude and longitude of the posted tweet).

4.3 Data Analysis

It is the application of appropriate link mining and opinion mining algorithms (over the Cleaned Data), based on the investigative needs and peculiarities of the source data.

New attributes are added to the tables to be inserted into the data warehouse (e.g. polarity of the sentence and degree of influence /popularity of a user) enabling the data repository called Analyzed Data.

4.4 Data Warehouse Integration

It is the incorporation of Analyzed Data into the data warehouse, taking into account: the generation of surrogate keys, treatment of Slowly Changing Dimensions, Late Arriving Facts and updating/generating aggregate tables if needed. The changing character of online social networks imposes the correct contextualization of events over time, thereby avoiding analytical distortions.

Besides the generation of the data warehouse, OLAP cubes may be processed, dashboard´s key performance indicators may be calculated and business rules associated to alert tools may be processed, notifying end users proactively. This set of operations depends on the portfolio of BI technologies available at the corporation.

In addition to data from online social networks, data from transactional systems are integrated into the enterprise data warehouse (e.g. through a time dimension), allowing impressions obtained from social networks to be compared with events recorded in corporate databases (e.g. sales to customers).

4.5 Business Intelligence Integration

It is the integration and availability of the data warehouse to the presentation layer tools (OLAP, SOLAP, Ad hoc Querying, Reporting, Data Mining, Dashboards and Alerts).

5. CASE STUDY

The proposed architecture was submitted to a case study involving data from Twitter, over which were applied opinion mining and link mining. Twitter is an online social network focused on the sharing of short text messages (up to 140characters), used by approximately 175 million users spread around the globe. Access to its facilities can be made by computers, cellular phones or tablets.

The social networks that are structured on Twitter have asymmetrical characteristics, and are directed with a high degree of dissemination of information (Haewoon et al., 2010), which makes Twitter a social network important for scientific research.

Users have followers and followees, without the requirement of reciprocity. Tweets can be sent or forwarded (retweeted) to all his/her followers; be directed towards specific users; mention users in its content; and may contain hashtags (initiated by the "#" character) in order to categorize its content. Each user has a profile, comprising a basic set of information.

5.1 Twitter Crawling

The Twitter Crawler of this case study used a Twitter database previously extracted and stored at the Max Planck Institute for Software Systems (www.mpi-sws.org).

The choice for this database is due to the following aspects: access of one of the researchers involved in this work to the MPI-SWS; the fact that the database included data since the beginning of Twitter (July/ 2006 until July/2009), totaling 17 billion tweets from 54 million users; the interest of having a dataset over 18 months in order to do comparative analysis and evaluate historical trends; the impossibility to reproduce historical data for 18 months in a short period of time.

The period chosen for data collection in our research was January/2008 to July/2009, period considered very active in terms of use by Twitter´s community. Since the research interest was to perform the analysis of data from online social networks over a brand, product or service, the tweets collected during this period were based on the text "lenovo thinkpad". Based on the ranking of the top 100 technological products of the year 2008 by PCWorld (Sullivan, 2008) and by comparative analysis of the frequency of Google searches for each of the five top ranked products (via Google Trends), we concluded that "lenovo thinkpad" would be a good choice. We started from the assumption that if it was very searched, there should be much talked about it on social networks.

Twitter Crawler, applied to the dataset of the MPI-SWS obtained 77,429 tweets from 32,924 users, who have registered some comments about "lenovo thinkpad".

5.2 Data Cleansing

Using programs developed in Python, we treated some situations: missing attributes were identified and treated appropriately (e.g. location unknown), avoiding the appearance of null identifiers in the data warehouse; featuring spam tweets were identified and ignored in the extraction process (tweets from the same user with more than one occurrence, which did not represent retweets); and standardization and hierarchization of user´s location on three levels (Great Region, Region and Sub-Region), through reverse geocoding (using Google Geocoding API).

After the application of data cleansing processes, mass data came to 58,906 tweets associated with 26,122 different users.

5.3 Data Analysis

The 58,906 tweets were submitted to an opinion mining algorithm suggested by Go et al (2009). The strategy adopted in this implementation was suitable to Twitter by the time the sentiment classifier was trained with tweets that had emoticons. Moreover, because no human intervention was required to label the tweets (through the use of distant supervising learning), the number of records to train the classifier were significantly higher.

Still, according to Go et al (2009), this method of implementation ensures the classifier accuracy above 80%. These authors also made available an API (TwitterSentiment at http://tinyurl.com/3qxevxg) to be used by researchers who want to develop applications referencing the classifier. In this case study, we developed a Python program, to access this API, in order to identify the feelings of 58,906 tweets.

Regarding the application of link mining in this case study, the degree metric was taken directly from the Twitter user profile (Twitter Crawler). Having the amount of followers and followees, we included the indegree and outdegree measures respectively, the first one indicating the popularity of a user (Meeyoung et al., 2010). In addition to these indicators, we used the Klout Score (del Campo-Ávila et al., 2011; Vega et al., 2010) that measures how successful a user is to engage his/her audience and the degree of impact that their messages have on other users. This indicator is calculated by the combination of three measures: true reach - it takes into account how active is the network of followers of the user; amplification ability - the probability of the user message to generate retweets or to start a conversation; and network influence - how influential are the users that retweet, mention and follow the user. The algorithm returns values between

0 - 100 (for each Twitter user) so that the higher this value, the more influential the user is. Obtaining Klout score for each user was done through the implementation of a Python program accessing the Klout API (developer.klout.com/api_gallery).

5.4 Data Warehouse and Business Intelligence Integration

The tweets properly addressed by the previous components of the architecture, but still available in the form of "flat files", were inserted into an Oracle (www.oracle.com) relational database data warehouse through IBM Cognos Data Manager (tinyurl.com/8a9xyak).

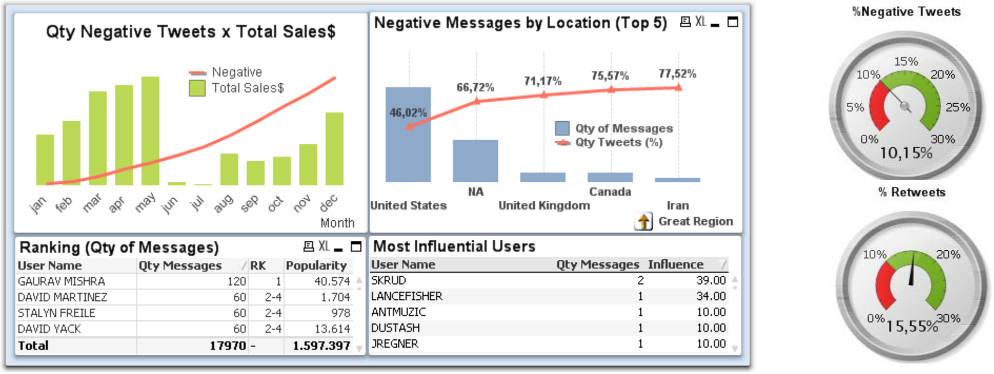
Data Manager is an ETL tool, capable of dealing with various aspects related to the settlement of a data warehouse (e.g. generation of surrogate keys, treatment of slowly changing dimensions and late arriving facts) was also used.

In addition to the data obtained from Twitter, an Oracle data mart was implemented, in order to represent Lenovo Thinkpad sales during 2008 and 2009. This initiative allowed, for example, the analysis of sentiments expressed in tweets compared with sales performance of Lenovo Thinkpad.

The main goal to be achieved by incorporating the data mart sales in the OSNBI was to prove that unstructured data (coming from online social networks) could be compared to structured data (coming from corporate management systems). This linkage was made through a shared dimension (Time).

Once generated the data warehouse, we used the QlikView (www.qlikview.com) for the development of a Business Intelligence analytical application, providing greater flexibility for scenario analysis. With an OLAP tool, operations such as slice, dice, pivot, drill down, drill up and rollup could be applied to the data context of online social networks (e.g. feelings, tweets, geographic location, popular/influence users) and enterprise data sales (e.g. time, customers); Dashboards could be implemented (e.g. percentage of negative opinions against the total number of tweets, retweets ratio), and Reports were developed showing the impact of negative tweets over the sales.

Figures 2 presents a screenshot of a business scenario developed in this application.



Ranking (Qty of Messages)

User Name	Qty Messages	RK	Popularity
GAURAV MISHRA	120	1	40.574
DAVID MARTINEZ	60	2-4	1.704
STALYN FREILE	60	2-4	978
DAVID YACK	60	2-4	13.614
Total	**17970**	-	**1.597.397**

Most Influential Users

User Name	Qty Messages	Influence
SKRUD	2	39.00
LANCEFISHER	1	34.00
ANTMUZIC	1	10.00
DUSTASH	1	10.00
JREGNER	1	10.00

Figure 2 – Analysis of negative tweets

Based on figure 2, we compared monthly sales over the year 2008, with the evolution of negative tweets posted at the same period of time. Also, it was noticed that the incidence of negative tweets was most concentrated in the USA country (46%). Users who posted more negative tweets in year 2008 were also shown, as well as their popularity (indegree), which would allow marketing efforts directed to them, in order to identify the reasons of the negative testimonies (trying to reverse it). Similar strategies could be adopted for the most influential users (Klout score) who posted negative statements. A dashboard was implemented, with some key performance indicators such as the incidence of negative tweets and retweets.

OSNBIA architecture represents a breakthrough in the aspect of integration of unstructured data from social networks and structured data from enterprise systems. We, in this sense, highlight also the following aspects: (a) viability, in the same environment, of mechanisms to analyze both structural patterns of social networks (link mining) and sentiments expressed in the statements (opinion mining); (b) possibility for a gradual and modular expansion of the social networks that could integrate the decision support environment; (c) possibility of adjustment/replacement of architectural components (crawling, data cleansing, opinion mining and link mining), as needs are identified (e.g. better performance and resolvability); and independence of BI presentation layer, since the intelligence resides in the data warehouse.

6. CONCLUSIONS AND FUTURE WORKS

The mix of technologies such as Online Social Networks, Opinion Mining, Link Mining and Business Intelligence at the same environment provides a different perspective for the analysis of unstructured and structured combined data.

Supported on the results of the case study, the proposed architecture allows: dynamism for the activities of data manipulation and scenario creation, as being supported by business intelligence technologies; marketing and social media departments can have better instruments to support planning, monitoring, analysis and execution of actions over online social networks; the expansion of the scope of online social networks being treated, through the implementation of new crawlers, specific link mining, opinion mining and data cleansing algorithms, aligned to the peculiarities of the social network.

In the future, we intend to expand the range of social networks analyzed; to adapt algorithms for data cleansing, opinion mining and link mining; to improve the data warehouse model in order to accommodate the differences between the OSNs; to consider ways to integrate a spatial data warehouse, allowing the use of SOLAP specific operators; and to instantiate the developed prototype in other business segments (e.g. political and health), thus evaluating its portability.

REFERENCES

Barnes, J.A., 1954. Class and Committees in a Norwegian Island Parish. *Human Relation*, 7, pp.39-58.

Barnes, A.J., 1954. Class and Committees in a Norwegian Island Parish. *Human Relation*, pp.39-58.

Benevenuto, F., 2010. Redes Sociais Online: Técnicas de Coleta, Abordagens de Medição e Desafios Futuros. In *Simpósios SBSC, Webmedia, IHC e SBBD*. Belo Horizonte, 2010.

Benevenuto, F., Magno, G., Rodrigues, T. & Almeida, V., 2010. Detecting Spammers on Twitter. In *Seventh Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS 2010)*. Redmond, 2010.

Berners-lee, T. et al., 2009. *Decentralization: The Future of Online Social Networking*. [Online] Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.190.1487 [Accessed 7 Dec 2011].

Bhide, M. et al., 2008. Enhanced Business Intelligence using EROCS. In *IEEE 24th International Conference on Data Engineering (ICDE '08)*. Cancun, 2008. IEEE Press.

Boyd, D., 2008. Why Youth Social Network Sites: The Role of Networked Publics in Teenage Social Life. *Youth, Identity, and Digital Media*, pp.119-42.

Boyd, D.M. & Ellison, N.B., 2007. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13.

Cachia, R., Compañó, R. & Da Costa, O., 2007. Grasping the potential of online social networks for foresight. *Technological Forecasting and Social Change*, 74, pp.1179-203.

Curran, K., O'Kane, P., McGinley, R. & Kelly, O., 2010. Social Networking. In S. Dasgupta, ed. *Social Computing: Concepts, Methodologies, Tools, and Applications*. Hershey: IGI Global. pp.156-68.

D'Andrea, A., Ferri, F. & Grifoni, P., 2010. An Overview of Methods for Virtual Social Networks Analysis. In A. Abraham, E.A. Hassanien & V. Snášel, eds. *Computational Social Network Analysis*. London: Springer. pp.3-25.

Dawson, R., 2003. *Living Networks: Leading Your Company, Customers, and Partners in the Hyper-connected Economy*. New Jersey: Prentice Hall.

del Campo-Ávila, J., Moreno-Vergara, N. & Trella-López, M., 2011. Analizying Factors to Increase the Influence of a Twitter User. In J. Pérez et al., eds. *Highlights in Practical Applications of Agents and Multiagent Systems*. Heidelberg: Springer. pp.69-76.

Freeman, L.C., 2004. *The development of social network analysis*. Vancouver: Empirical Press.

Gartner, 2010. *Gartner Identifies the Top 10 Strategic Technologies for 2011*. [Online] Available at: http://www.gartner.com/it/page.jsp?id=1454221 [Accessed 15 Jan 2011].

Getoor, L., 2003. Link mining: a new data mining challenge. *SIGKDD Explor. Newsl.*, pp.84-89.

Gettor, L. & Diehl, C., 2005. Link Mining: A Survey. *SigKDD Explorations Special Issue on Link Mining*, 7, pp.3-12.

Gilad, B., 1988. *The Business Intelligence System : A new tool for competitive advantage*. New York: AMACOM.

Go, A., Bhayani, R. & Huang, L., 2009. *Twitter sentiment classifcation using distant supervision*. Technical report, Stanford Digital Library Technologies Project. Stanford: Stanford University.

Haewoon, K., Changhyun, L., Hosung, P. & Moon, S., 2010. What is Twitter, a social network or a news media? In *19th international conference on World Wide Web*. Raleigh, 2010. ACM Press.

Han, J. & Kamber, M., 2006. *Data Mining - Concepts and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann.

Inmon, W.H., 1996. *Building the data warehouse*. 2nd ed. Indianapolis: Wiley Publishing.

Liu, B., 2008. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Berlin: Springer-Verlag.

Liu, B., 2010. *Sentiment Analysis and Subjectivity*. [Online] Available at: http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf [Accessed 7 Nov 2011].

Luhn, H.P., 1958. A Business Intelligence System. *IBM Journal of Research and Development*, 2, pp.314-19.

Mackelworth, T., 2007. *Social Networks: Evolution of the Marketing Paradigm*. [Online] Available at: http://www.amacltd.com/pdf/SocialNetworksWhitePaper.pdf [Accessed 12 Sep 2011].

Meeyoung, C., Haddad, H., Benevenuto, F. & Gummadi, K.P., 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media*. Washington, 2010. AAAI Press.

Mislove, A. et al., 2007. Measurement and Analysis of Online Social Networks. In *7th ACM SIGCOMM conference on Internet measurement (IMC '07)*. San Diego, 2007. ACM Press.

Mitchell, C.J., 1969. The Concept and Use of Social Networks. In J.C. Mitchell, ed. *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns*. Manchester: The University Press. pp.1-50.

Moya, L.G., Kudama, S., Cabo, A.M.J. & Llavori, R.B., 2011. Integrating Web Feed Opinions into a Corporate Data Warehouse. In *2nd International Workshop on Business intelligence and the WEB (BEWEB '11)*. Uppsala, 2011. ACM Press.

Müller-Prothmann, T., 2008. Use and Methods of Social Network Analysis in Knowledge Management. In E.M. Jennex, ed. *Knowledge Management: Concepts, Methodologies, Tools, and Applications*. Hershey: IGI Global. pp.1096-106.

Nielsen, 2010. *Social Networks/Blogs Now Account for One in Every Four and Half Minutes Online*. [Online] Available at: http://blog.nielsen.com/nielsenwire/global/social-media-accounts-for-22-percent-of-time-online/ [Accessed 15 Jan 2011].

Oreilly, T., 2007. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, 1, p.17.

Pang, B. & Lee, L., 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2, pp.1-135.

Park, B.-K. & Song, I.-Y., 2011. Toward Total Business Intelligence Incorporating Structured and Unstructured Data. In *2nd International Workshop on Business intelligence and the WEB (BEWEB*. Uppsala, 2011. ACM Press.

Perez, J.M., Berlanga, R., Aramburu, M.J. & Pedersen, T.B., 2007. R-Cubes: OLAP Cubes Contextualized with Documents. In *IEEE 23rd International Conference on Data Engineering (ICDE '07)*. Istanbul, 2007. IEEE Press.

Sallam, R.L., Richardson, J., Hagerty, J. & Hostmann, B., 2011. *Magic Quadrant for Business Intelligence Platforms*. [Online] Gartner Available at: http://www.board.com/download/press/EN/Gartner_BI_MagicQuadrant_2011.pdf [Accessed 12 Nov 2011].

Sass, E., 2010. *Businesses Create Proprietary Social Networks*. [Online] Available at: http://www.mediapost.com/publications/article/129479/ [Accessed 14 Jan 2011].

Singh, R. & Singh, K., 2010. A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing. *International Journal of Computer Science Issues*, 7, pp.41-50.

Sullivan, M., 2008. *The 100 Best Products of 2008*. [Online] PCWorld Available at: http://www.pcworld.com/article/146161-12/the_100_best_products_of_2008.html [Accessed 2011 Nov 8].

Vega, E., Parthasarath, R. & Torres, J., 2010. *Where are my tweeps?: Twitter Usage at Conferences*. [Online] Available at: http://www.socialcouch.com/demos/final_paper_twitter.pdf [Accessed 3 Dec 2011].

Weber, L., 2009. *Marketing to the Social Web*. Hoboken: John Wiley & Sons.