Scientific Bulletin of the „Petru Maior" University of Tîrgu Mureş
Vol. 9 (XXVI) no. 2, 2012, ISSN 2285 – 438X (Online), ISSN–L 1841 – 9267

UNSTRUCTURED SOCIAL NETWORKS DATA FOR BUSINESS CONTEXT ANALYSIS

[1]Walter MAIER, [2]Dumitru RĂDOIU

Department of Informatics, Faculty of Sciences and Letters
„Petru Maior" University of Tîrgu Mureş
Nicolae Iorga Street, no. 1, 540088, Mureş County, Romania
[1]walter.maier@upm.ro
[2]dumitru. radoiu@upm.ro

ABSTRACT

Communications technology has enabled new approaches to business context understanding. The paper proposes and explores a new mechanism through which unstructured social networks data about companies is gathered, aggregated and presented. The authors assume that collected data, interpreted through an adequate metrics, may be used as a tool for better understanding the business model of a company, its health and/or its sustainability. The paper does not address the issue of adequacy of the tool to the problem but the technical details to collect, aggregate and present unstructured social networking data for business context analysis. The proposed solution is a preliminary work
.

1. Introduction

Social networking sites have been growing in popularity over the past five years and lately, business people expect to see clear ROI (return on investment) for every other channel of online marketing starting with email, search, and display advertising and finishing with Youtube, Facebook and Twitter. Companies like ValueOfALike.com [6] quantify even the business value of a "Like". With decent analytic software (like Google Analytics and HobSpot's Marketing Analytics) marketers can track traffic from social networks and assign lead or customer acquisition values [7],[8].

Professionals use social networks to discuss industry related issues, companies have active presence on the web, and employees have a virtual life on social networks so there is a wealth of information which can be gathered to understand a business in a context. Thus, the data, the conversations, the tweets and all kind of measurements of web activity can be used for a more accurate image of the context in which a company is doing business.

Companies (and its employee) use the Web and social networking to achieve different goals, like marketing through discussions on social networks (e.g. Twitter), websites as channels towards different market segments (company websites), human resource recruiting through postings or discussions (e.g. LinkedIn), or broadcasting financial data (e.g. national/international legal entities). Unstructured data from companies are continuously published and in most of the time, hard to collect, aggregate and used to make decisions or to understand the context in which a company is running its business.

The goal of the paper is to present a (short) list of the unstructured data which can be found on the internet, and to present an architecture which allows collection of such data, aggregation and presentation.

Even if the authors acknowledge that the models and metrics used to produce decisions based on these data are flawed in many ways, we believe that this data is valuable by providing the context of a business and such, giving a new dimension to any business related analysis.

2. The problem

In order to understand the context, we need to analyze a massive amount of data from different places:

— Public accounting data (financial status of a company usually updated once a year, so data can be relevant, but outdated)

— Company website (updated data regarding location, contact info, customers and its present activity)

— Twitter accounts (for Company or employees generated data: employees opinion about the company or products of the company; information about social impact of company or moods/sentiments expressed by employees or customers in tweets)

— Facebook accounts (for Company data or employee's generated data). E.g. Image can be extracted from popularity on this social media platform and we can use as popularity metrics of a Facebook page through number of likes or number of talking about.

— LinkedIn accounts (job postings, employee profiles, company profile, connections, skills, expertise, references)

Here is a very short list of data which may interest business people next to financial "hard" figures from the company "data room':

Table 1 Short list of business context variables for an IT company

Company_Name	Buying_Guide__Product_Related	Other_Social_Networking_S
Industry	Ability_To_View_Price_Information	ites_Links
Postal_Address	Ability_To_Download_The_Product	Forum
Phone_Number_On_Homepage	Ability_To_View_Recommendation	Rss
Call_Back_Function	Ability_To_Write_Product_Rev	Chat
Job_Vacancies	Blogs	Product_Faq
Website_URL	Customer_Reviews	Cost_Calculator
Browser_Satisfaction	E_Mail_A_Friend	Site_Personalization
Monthly_Visits	Mobile_Commerce	Social_Networking
Monthly_Unique_Visitors	Product_Comparisons	Videocasts
Date_Of_Registration_Website	Product_Customization	What_S_New
Number_Of_Month_Online	Product_Ratings	Live_Chat_E_Mail
Ability_To_View_Company_News	Product_Recommendations	Rich_Media
General_E_Mail_Address_Contact_Form	Product_Wikis	Security_Certifications
Subscription_To_An_E_Mail_Newsletter	Registry	Site_Search
Ability_To_Report_ProductFailures_On_	Facebook_Friends	Web_Analytics
Line	Twitter_Followers	Google_Maps
Ability_To_Follow_The_Status_Of_Rep	Flicker_Fans	Youtube_Subscribers
ort		Myspace_Friends
Expert_Advice__Use_Related		

Table 1 has been build for an IT company whose value proposition is both products and services. Because of the specificity of each industry and company value chain creation, value proposition, market segment, channels and relations with the users and customers, the above list (Table 1) must be reviewed.

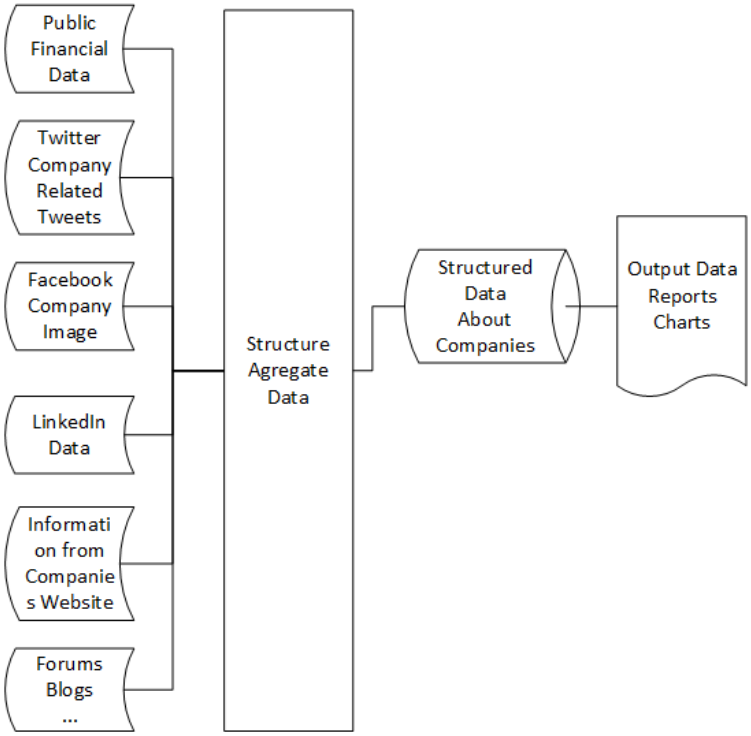

Fig. 1 High level flow of unstructured data

Figure 1 represents a high level view of the flow of unstructured data through an application from different sources to the end user.

The problem is how can we collect, filter, organize, and aggregate data from so many different sources, so that the context in which the company is doing business is revealed and how to present so much information to the end users, in a visual manner, so that the relations among different groups of data is grasped.

3. The Proposed Solution

We begin by observing that some data must be collected manually from the company site (or company related sites, like blogs of the professionals in the industry), some data could be extracted automatically from the company sites (e.g. traffic data, unique visitors) and some will be gathered via crawlers from social media sites via APIs. For each source, a data model must be developed and a full mapping between the list of sought for variables and their respective source must be completed.

Table 2 Mapping the variables to their sources

Variable	**Source**	**Extraction**	**Observation**
Employee Twitter account	Twitter	Twitter API	Automated

E.g. Figure 2 represents the data model for the account information we could extract from Twitter, Facebook and Youtube.

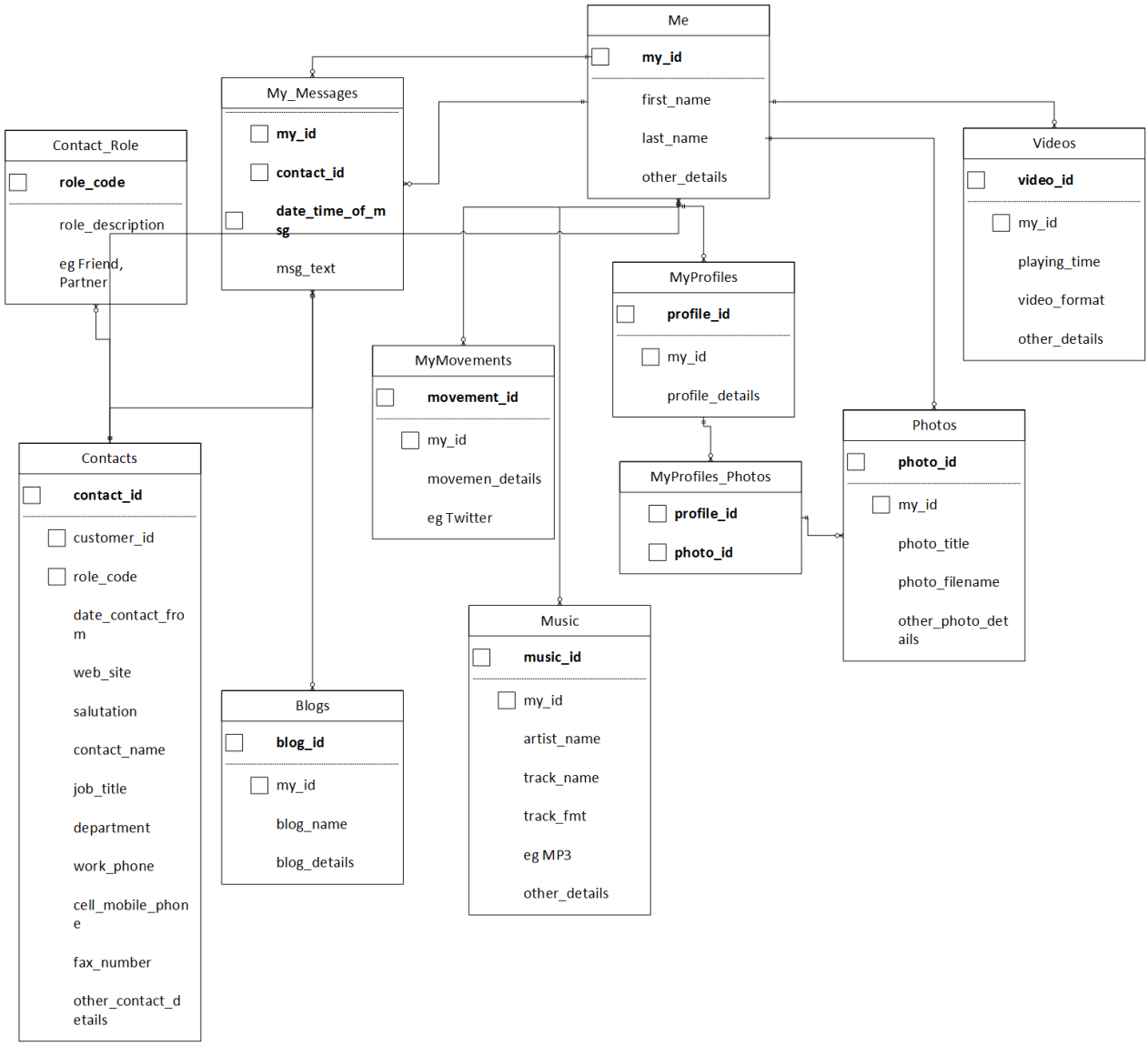

Fig. 2 Data Model for social networks [5]

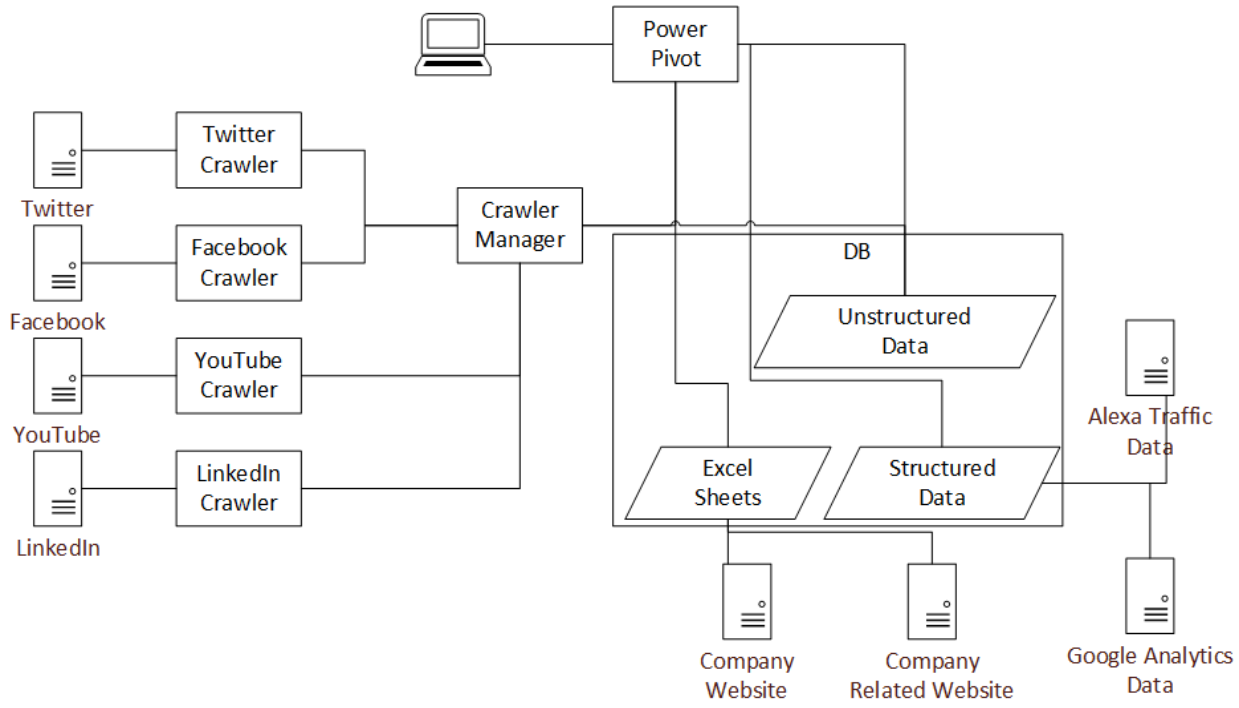

Fig. 3 High level architecture of the proposed solution

"When the gap between unstructured data and structured data is bridged, an entirely new world of possibilities and opportunity for information systems opens up." (Inmon, 2007). We call aggregation the process of organizing unstructured data and structured data from multiple sources and bring it to a form that can be analyzed.

A minimal platform to satisfy the goals (automate extraction, aggregation and presentation of unstructured data) should contain:

— Modules which collect data via API from Twitter, Facebook, Youtube and LinkedIn

— Crawler Manager for handling social media data; Crawler Manager will keep track of extractions that are made from different platforms and will take care of the limitations, so that the extraction process will not be interrupted by different events (e.g. bans for exceeded number of queries) and will optimize the process of collecting and storing data.

— Modules which collect data about the company website, from Alexa or Google Analytics

— Module used to facilitate extraction of data manually using company website content or company related websites (e.g. brand website, blogs)

— Data base which will keep three types of data

— A module for integrating and presenting and sharing the data. E.g off the shelf PowerPivot [4], an advanced analytic tool. PowerPivot allows manual creation of relationships between entities from different sources of data (e.g. One can connect data from different sources like SharePoint, SQL, or Access even if there are hundreds of millions of rows of data. Data from different sources can be related manually, visualized in PowerPivot, and finally published to SharePoint.

Part of this architecture has been discussed by [1] and [2]. The possibility to use MongoDB as a data repository is dependent on the flexibility of PowerPivot working with the NoSQL database.

4. Further work

Because we want to integrate in this solutions data from social media, using their API, we have a few limitations, for instance Twitter limits the queries to 350/hour, the same thing is available for Facebook and LinkedIn.

However finding Facebook employees and extracting their data is a bit more difficult than in case of Twitter or LinkedIn, because in order to have access to this type of personal data, you need permissions from the user.

Also a challenge could be crawling the website of a company, as you may know this is completely unstructured and it is harder to analyze than a tweet or data retrieved through API.

5. References

[1] Beres Adela – Oana, Sentiment Analysis Framework based on Twitter Corpus Data, *Scientific Bulletin of the „Petru Maior" University of Tîrgu Mureş*, Vol. 9 (XXVI) no. 1, 2012, ISSN 2285 – 438X (Online), ISSN–L 1841 – 9267 (2012)

[2] Goga Claudia, Johann Stan, A framework for Aspect-Based Opinion Mining, *Scientific Bulletin of the „Petru Maior" University of Tîrgu Mureş*, Vol. 9 (XXVI) no. 1, 2012, ISSN 2285 – 438X (Online), ISSN–L 1841 – 9267

[3] William H. Inmon and Anthony Nesavich, *Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence*, Prentice Hall PTR Upper Saddle River, NJ, USA ©2007 ISBN:0132360292

[4] http://office.microsoft.com/en-us/excel-help/whats-new-in-powerpivot-in-excel-2013-HA102893837.aspx, retrieved 03.12.2012

[5] http://www.databaseanswers.org/data_models/social_networking/index.htm, retrieved 03.12.2012

[6] http://valueofalike.com/ retrieved 03.12.2012

[7] http://blogs.hbr.org/cs/2012/11/how_to_calculate_the_value_of.html?cm_mmc=email-_-newsletter-_-weekly_hotlist-_-hotlist120312&referral=00202&utm_source=newsletter_weekly_hotlist&utm_medium=email&utm_campaign=hotlist120312, retrieved 03.12.2012

[8] http://hbr.org/special-collections/insight/putting-social-media-to-work, retrieved 03.12.2012