University of Economics in Prague

Faculty of Informatics and Statistics

Department of Information Technologies

Study programme: Applied Informatics

Field: Information Systems and Technologies

Using social networks for Competitive Intelligence

Author: **Bc. Tomáš Feige**

Thesis Supervisor: **prof. Ing. Zdeněk Molnár, CSc.**

Opponent of the thesis: **Ing. Martin Švík, Ph.D.**

Year 2012

Statement

I hereby declare that I worked on this thesis alone and that I honestly quoted all the sources and literature, which I used.

In Prague, May 1st, 2012

……..

Signed

Acknowledgment

I would hereby like to thank the supervisor of my thesis prof. Zdeněk Molnár, who provided me with valuable guidance and experience throughout the whole time I was working on the thesis.

Big thanks also go to all the great people at IBM, who were excellent to work with and were always there and willing to help whenever I needed them.

Contents

Abstrakt

Práce se zaměřuje na oblast Competitive Intelligence (konkurenční zpravodajství) s důrazem na nové možnosti ve vztahu k moderním sociálním sítím. Nejprve vyhodnocuje obecnou analýzu aktuálního stavu trhu a následně se podrobně věnuje jednotlivým předním hráčům a jejich klíčovým produktům, čímž poskytuje detailní pohled na celou oblast Competitive Intelligence. V závěru teoretické části se poté soustředí na možnosti využití moderních sociálních sítí a dalších *sociálních* a *soft* zdrojů právě pro Competitive Intelligence.

Praktická část práce je věnována reálnému projektu, který se uskutečnil ve spolupráci s IBM zkraje roku 2012. Na tomto projektu jsou demonstrovány teoretické poznatky sepsané v úvodní části práce. Vzhledem k absenci obdobných zdrojů může být celá kapitola použita jako referenční model pro budoucí projekty shodného zaměření.

Klíčová slova

Competitive Intelligence, konkurenční zpravodajství, sociální sítě, Social Content Management, řízení znalostí, nestrukturovaná data, Facebook analýza, IBM.

Executive Summary

This thesis focuses on the area of competitive intelligence with the emphasis on new possibilities and opportunities in relation to modern social networks. First it gives general analysis of the current state of competitive intelligence market as a whole and then deals with individual major leaders and their products, thus providing detailed overview of this business segment. It also discusses the possibilities of using social networks and other *social* or *soft* sources for competitive intelligence.

Practical part of the thesis then demonstrates the theoretical knowledge on a real life CI project, which took place in early 2012 in cooperation with experts from IBM, including some interesting results and findings in appendix. The whole chapter can be used as a reference model for future projects with similar goals.

Keywords

Competitive Intelligence, Social networks, Social Content Management, Knowledge Management, unstructured data, Facebook analysis, IBM.

Chapter 1 – Theoretical part

1 Prologue

Since the boom of various social networks headed by Facebook around five years ago, internet underwent a change in paradigm of how it is used by general population. Nowadays everything revolves around sharing, linking, socializing and communicating.

With numbers of active users growing every day[1], each *social* site is becoming a treasure box of knowledge. In line with current trends and customs users are willing to give up and share an incredible amount of information – often personal and delicate. No wonder that today, many subjects are trying hard to get to this information through various – even illegal[2] – ways and use it either for scientific purposes or to start a business and generate profit.

At Social Media Conference 2012, which took place at Charles University in Prague, was said that *modern social networks have a potential to become an exact scientific platform liberal arts and social science have been waiting for*. With no doubt, social networks have become a new phenomenon of current days.

In relation to academic sphere, the whole area inspires individuals to write their thesis and papers. E.g. (Juchelka, 2012) from Masaryk University examined in detail various possibilities of data extraction from two largest global social networks – Facebook and Twitter. Others such as (Berger, 2011) devoted whole books to various media analysis techniques and current trends in area of social networks and web communication.

And not just individuals, recently the whole teams of academics and experts joined together on projects aimed at extracting added value from modern social networks. One example is the whole Social Media Research Foundation and their open *NodeXL* project capable of importing and visualization of social data from various networks into MS Excel. (Social Media Research Foundation, 2012) Another project worth mentioning is *Sentiment140* from Stanford University capable of doing sentiment analysis on Twitter social network in English and Spanish. (Go, et al., 2009) Or on a related note Twitter Earth side project from the same authors aiming at exploiting Google Earth API to visualize Twitter activity on a world scale. (Go, 2009)

[1] See <http://en.wikipedia.org/wiki/List_of_social_networking_websites> for comprehensive list of social networks updated on regular basis.

[2] In late 2011 there was a cause in UK with *Ramnit trojan horse/virus* stealing login information of over 45,000 UK Facebook users. (Daily Mail, 2012)

With all big companies turning their heads towards social networks, the whole area becomes also very interesting from the point of view of competitive intelligence. Whoever will be the first one to master mining, visualization and analysis of these new *fuzzy* data sources will gain significant competitive advantage. And why limit oneself only to external third-party social networks? Why not to apply the same principles inside one's company too and benefit from advantages of the new *social era* also internally?

However there is still a long way to this *grail*. What is the current state in this area in general and where the *new* competitive intelligence is heading is one of the questions this thesis is trying to answer.

2 Area of interest

As technologies, social customs and paradigms evolve, new situations bring new ways of gathering, analyzing and exploiting information, which can be used for building corporate knowledge base and help companies keep their positions among competitors or discover new opportunities and gain some advantage.

The main goal of this thesis is to describe and discuss new trends and possibilities Competitive Intelligence has in relation to area of modern social networks and social content management.

First part of the thesis provides reader with general knowledge of current state of competitive intelligence market and its leaders and their products. After *preparing the ground* it discusses contemporary trends and possibilities in relation to emerging new business areas such as Social Content Management.

Second – practical – part of the thesis then focuses on detailed description of recent real life project done in cooperation with IBM. This project shows some real life implications of theoretical assumptions and expectations and addresses some issues and problems when working with fuzzy *social* data in *unsupported* environment. Outputs of this step-by-step project report may be further used as *best practice* guidelines.

2.1 Competitive Intelligence

Competitive Intelligence (CI) is rather young member of *Intelligence* family[3]. But even with its quite short existence beginning somewhere around 70s and 80s of previous century with publishing of Porter's *Competitive-Strategy: Techniques for Analyzing Industries and Competitors*, it is built on solid foundations – be it foundation of both American *Society of Competitive Intelligence Professionals (SCIP)*[4] in 1986 and specialized *CERAM Business School* in France with master programme in Economic Intelligence and Knowledge Management in 1995; or foundation of *Institute for Competitive Intelligence* in 2004 or *Centre for Global Intelligence and Influence* in 2011.

But as the whole area underwent a rather rapid development to keep up with the rest of the world and its changing principles and paradigms, same has done the definition of Competitive Intelligence itself. This problem of ever-changing and evolving meaning of CI

[3] Where belong intelligent services, business intelligence or even counterintelligence.

[4] Renamed to *Strategic and Competitive Intelligence Professionals* in 2010. See <http://www.scip.org/files/SCIPNameChangeRelease.pdf?navItemNumber=12756>

and what CI actually is and what is its scope, was discussed e.g. by Roberta Brody in *Journal of Competitive Intelligence and Management* a few years ago. (Brody, 2008)

To be able to talk further about CI and its connections to other areas and topics, four definitions of CI from various sources follow:

First one is a short general definition by SCIP: *"CI is a necessary, ethical business discipline for decision making based on understanding the competitive environment."* (SCIP, 2007)

John E. Prescott, professor of business administration at the University of Pittsburg defines CI as a *"formalized, yet continuously evolving process by which the management team assesses the evolution of its industry and the capabilities and behavior of its current and potential competitors to assist in maintaining or developing a competitive advantage."* (Prescott, et al., 1993)

Canadian professor Craig Fleischer offers following definition of CI: *"Competitive Intelligence is a process, by which organizations gather actionable information about competitors and the competitive environment and, ideally, apply it to their planning processes and decision-making in order to improve their enterprise's performance"* (Fleisher, et al., 2010)

Lastly Journal of Competitive Intelligence and Management defines CI as: *"The systematic and ethical process for gathering, analyzing, and managing information that can impact an organization's operations and plans. Competitive intelligence is a necessary, ethical business discipline for decision-making based on understanding the competitive environment."* (Competitive Intelligence Foundation, 2008)

This thesis will be working with the last mentioned definition found in Journal of Competitive Intelligence and Management mainly due to its emphasis on "*managing information that can impact an organization's operations and plans*" and "*understanding the competitive environment*," which both go the most in line with areas such as Social Content Management and presented findings of recent real life analytical project, that are all discussed further in the thesis.

2.2 Social Content Management

Social Content Management (SCM) is a new approach to content management with origins as a part of the new generation Enterprise Content Management (ECM) model introduced by IBM a few years ago (see Image 1). But same as it is with CI and its origins as a part of Business Intelligence and *Intelligence* family in general, it can be viewed alone as a whole new standalone area.

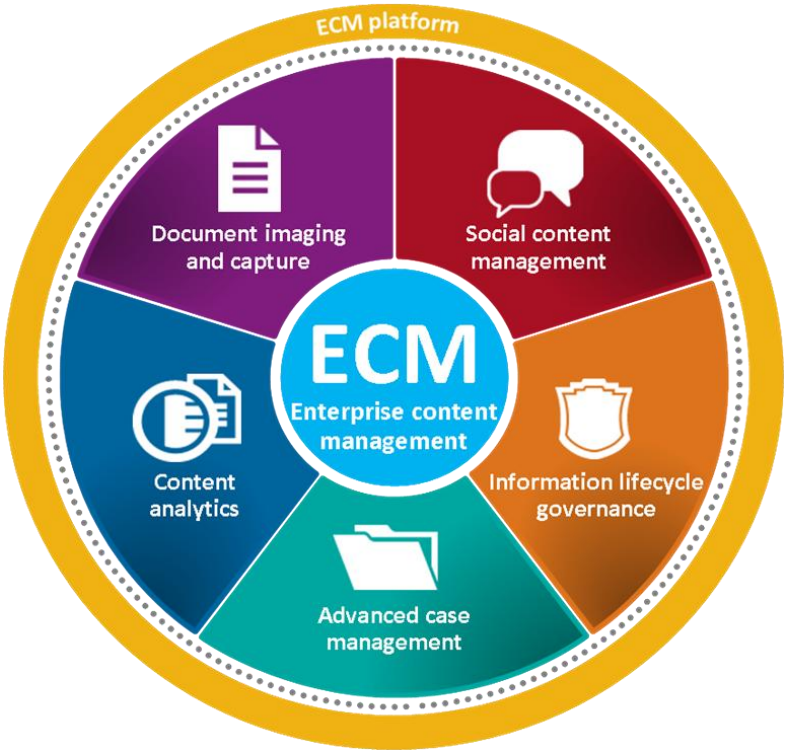


Image 1 - Enterprise Content Management platform model (source: (IBM, 2012))

Although there is no official definition, the aim of the SCM can be summarized as an effort to create an environment that encourages people in generating *social*[5] content through interaction and collaboration while providing platform owners (companies) with means of managing this content to be effectively used later.

In relation to Competitive Intelligence and Knowledge Management this means to focus on probably the most problematic part of every analysis – understanding the unstructured data. While normally when doing an unstructured data analysis one has to first gather the data and then try to parse it, process it and understand it; concept of SCM focuses on understanding the data and its meaning at the time it is created and accompanies it with enough describing metadata so it is not lost once saved and stored *for later*. This goes along

[5] Social content being results of human interaction, communication and collaboration.

with principles described by Zdeněk Molnár, professor at the University of Economics in Prague:

"Main goal of disciplines around CI is to build efficient bridge between those who know (Knowledge Owners) and those who decide (Decision Makers)." (Molnár, 2009)

Various aspects of Social Content Management and its connection to Competitive Intelligence are discussed in more detail in Chapter 5.

3 State of Competitive Intelligence market

To be able to talk about individual major business leaders or deal with various possibilities and new trends in competitive intelligence area, it is necessary to first give a general overview of the environment. Possibly the best way to provide reader with basic understanding of a current state of competitive intelligence market is to do a general analysis using a well known Porter's model of *Five competitive forces*. This model was formulated by Harvard professor Michael E. Porter in 1979 and provides basic framework both for industry analysis and business strategy development. (How Competitive Forces Shape Strategy, 1979)

Image 2 – Porter model of five forces (Wikipedia Contributors, 2006)

Although not without its critics[6], for purposes of this work it is more than sufficient. As illustrated on Image 2, model is based on 5 different forces which together depict the current state of the market. These are following:

- Threat of new competition,
- Threat of substitute products or services,
- Bargaining power of customers,
- Bargaining power of suppliers and
- Intensity of competitive rivalry.

Following chapters shortly characterize each individual force so it is possible to form a solid conclusion about the current state of CI market and its main players.

[6] For instance Kevin P. Coyne and Somu Subramaniam have stated that the model is based on dubious assumptions. (Coyne, et al., 1996)

3.1 Threat of new competition

Intensity of this force is based above all on overall attractiveness of the field, number of incumbents, height of entrance costs and demands on know-how. While there are several companies providing basic consulting services, there is only a handful of subjects currently having the necessary technology and know-how to create, sell and maintain complex BI and CI analytical tools. Therefore both demands on know-how and rather high entrance costs make the threat on new competition rather low despite the current attractiveness of the whole business analytical sector.

3.2 Threat of substitute products or services

This force can almost be omitted completely in our case, because there are hardly any substitutes for competitive intelligence field – at least from the point of view of this thesis. Possibly some complex robust BI solution might be able to provide company with some form of external data analysis, but that is just one dimension of multidimensional competitive intelligence area of focus.

3.3 Bargaining power of customers

Although BI and CI tools can prove useful even to small companies, they show their full potential only when used by large corporations. Because of that the total number of potential customers is not so big and one must be very careful when negotiating with them. Although there are not many possibilities for customers to switch providers, still their negotiating power is not to be underestimated, especially since *back-office* functionality of individual products doesn't differ much[7] and therefore switching costs can be evaluated as rather low. And after all – every customer counts.

3.4 Bargaining power of suppliers

The intensity of this force and its impact on the overall market competitiveness force varies a lot with each individual company. If CI vendor focuses only on providing end-customer solution and is therefore dependant on data suppliers providing large structured databases, then this force is not to be taken lightly. Although some data can be obtained for free and on global scale there are several such free databases (such as Freebase), when it comes to our local Czech market and *professional* data suppliers, there are only a few local

[7] Tools usually store gathered data in some structured form either as XML, JSON or CSV file format or in some kind of *data warehouse* database. Such structured data can easily be transferred between different products.

company data providers (like e.g. Čekia or CreditCzechBurreau) and so their bargaining power is rather high.

If on the other hand one possesses his own data channels and is able to create complex solution capable of both data mining and later analyzing and forecasting, his dependency on such providers is nonexistent or very low.

3.5 Intensity of competitive rivalry

As was already mentioned in the customers' section there is only a few potential customers among large corporations and since basically every leading global IT company nowadays has its portfolio of CI products, the competitive rivalry in the field is rather intense. Not that there were some aggressive marketing campaigns, but all firms feel that this field is something new with big business potential and so they proactively try to address each individual potential customer with their offers and as stated above – every customer counts.

3.6 Analysis conclusion

After the general analysis each competitive force was appropriately evaluated. Individual values and resulting total market force value can be seen in Table 1. While threat of new competition can be marked as low and not dangerous, and threat of possible substitutes even as next-to-nonexistent, both bargaining power of customers and intensity of current competition are rather high. Actual strength of bargaining power of suppliers is then dependent on chosen approach – whether company covers the whole analytical process including primary data mining, or if it uses external suppliers.

Table 1 - Porter analysis of Competitive Intelligence market (source: Author)

Market force	Intensity				
Threat of new competition	1	**2**	3	4	5
Threat of substitute products or services	**1**	2	3	4	5
Bargaining power of customers	1	2	3	**4**	5
Bargaining power of suppliers	1	2	**3**	4	5
Intensity of competitive rivalry	1	2	3	4	**5**
Total market force	1	2	**3**	4	5

As a conclusion of this analysis, competitive intelligence market can be described as a dynamic and rather hostile environment with intense competition – especially in relation to new potentially lucrative areas such as modern social networks.

4 Overview of main CI players

When trying to describe the current state of competitive intelligence market in connection to social networks, one must not omit main players. Following pages focus on giving a brief description of the most dominant companies in business, that were identified in cooperation with IBM experts as a main competitors and current CI market leaders. Specifically these are: IBM, SAP, Hewlett Packard, Google[8] and – when local market conditions were taken into account – also Tovek.

This is by no means a comprehensive list of all subjects. One should not forget about Microsoft and its *Fast* search module included as a part of MS SharePoint platform or already mentioned *NodeXL* project as a plugin for MS Excel; and there are of course other big companies such as SAS or Adobe, which too have their own products that could be also included into CI area. But the five companies listed above form a current leading group – at least in European conditions – and therefore it is them and their products, which will be given special interest on the following pages.

4.1 IBM Content Analytics

This giant international company with more than a hundred year tradition in IT business is well known to everybody and its position on the ICT market amongst top leaders is also unquestionable. No wonder then, that its area of interest covers almost every IT branches and fields including business intelligence, competitive intelligence and social networking and social content management.

IBM Content Analytics (ICA) is a so called company flagship in the area of business (and competitive) intelligence. This tool is capable of performing complex analysis on large amount of structured, semi-structured and unstructured data and was built around several open source standards.

ICA stands on two main pillar applications – Text Miner and Administration Console, and its *core* can be divided into several specialized components – Crawlers, Document Processors, Indexers, Text Analytic Collections and Search Runtimes.

While Text Miner serves as an end-user client analytical tool for the actual analyzing and *knowledge working*; Administration Console is a pure administering platform used for

[8] In case of Google the term *market leader* is not correct, because Google doesn't really have any professional CI product like others. However its free data refinery tool Google Refine deserves special interest and was therefore included into the thesis.

configuration and monitoring of the system. The whole architecture and functionality of each component is described in detail in Chapter 6.1.

4.1.1 ICA Extensions

ICA currently supports 11 main world languages and a variety of custom extensions. It is also delivered with some of its own such as IBM LanguageWare or Watson core components – mainly to further smoothen its natural language processing capabilities.

- **LanguageWare**

  LanguageWare is a natural language processing (NLP) technology developed for and included exclusively into IBM Content Analytics. It is an Unstructured Information Management Architecture (UIMA) compliant developer tool allowing companies to create and implement into ICA their own custom annotators and facets to meet their industry-specific needs. (IBM, 2011)

- **Watson core components**

  As a part of IBM Content and Predictive Analytics for Healthcare (ICPAH) project engineers from IBM integrated core components from well-known supercomputer Watson[9] to be able to perform more thorough natural language processing and predictive root cause analysis. See Image 3 for an overview of ICPAH architecture.

[9] This natural language processing supercomputer even won a famous TV show *Jeopardy!* against the two best players of all time. See <http://www.thedailybeast.com/videos/2011/02/16/watson-wins-on-jeopardy.html>

Image 3 - IBM Content and Predictive Analytics for Healthcare architecture (source: (IBM, 2012))

4.2 SAP Text Analysis

In the field of Business Intelligence SAP possesses a large family of products and various tools called SAP Business Objects. Business Objects was originally a French BI company that was acquired by SAP for $6.8B in 2008 to *"accelerate its growth in the Business User segment, while complementing the company's successful organic growth strategy"*. (Business Week, 2007)

Part of this standalone BI division of SAP is also a powerful text analytical tool SAP Text Analysis. This product capable of thorough text analysis over unstructured text was originally developed by Californian company Inxight Software which was then acquired by Business Objects in 2007. (KM World, 2007)

While Business Objects belonged to the world leaders in the area of structured data analysis, Inxight Software on the other hand focused on analyzing and understanding of unstructured text both from internal sources and *open* web. Merging these two companies led Business Objects to be *"[…] the first to provide customers with a BI platform that can streamline all of their internal and external information assets—both structured and unstructured data."* (KM World, 2007)

Following acquisition by SAP in 2008 led to integration of this powerful tool into broad SAP product platform, as can be seen in Figure 1.

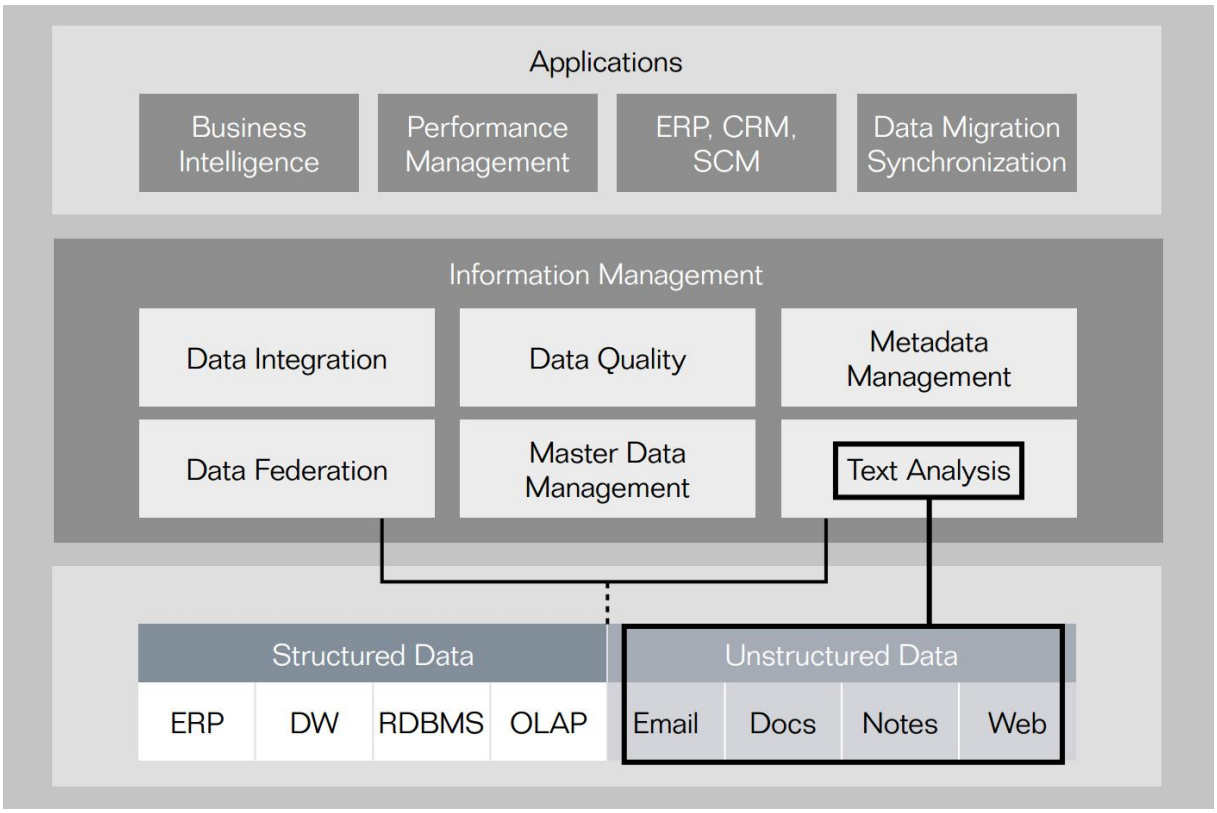


Figure 1 - Integration of SAP Text Analysis into SAP Framework (source: (SAP, 2009))

As was said, key function SAP Text Analysis delivers to Business Objects family is the ability to analyze unstructured text. This means for computer to understand the text and properly extract information from it, not just create a list of keywords. In contrast to standard search engines, which simply crawl through the given text and count occurrences of each *important*[10] word, Text Analysis uses natural language processing features to actually understand the meaning of each sentence. Two key features are language dictionaries and taxonomic trees.

- **Language dictionaries**

  Provide the knowledge base for the parsing machine giving it the ability to correctly identify language of the text and its lexical class together with its possible forms and variations.

- **Taxonomic trees**

  All *whats*, *wheres* and *whos* extracted from the text are identified according to the context and sorted based on their meaning to corresponding place in a preset taxonomic tree.

[10] It's common practice to exclude from indexing prepositions, conjunctions or articles because they don't alone carry any significant information.

How the whole natural language processing system actually works is illustrated in Image 4.

The proposed merger between Mega, Inc. and CNA Systems, Incorporated, has been postponed, Mega CEO Joe Smith said in an analyst call. "CNA's 1st quarter revenue dropped by 32%, and they lost 23 million dollars," Smith explained. CNA Systems sources blame weak sales in China. CNA shares (CNAI) fell 47 percent to $9.84 on May 12, the first trading day after the announcement.

Company	Mega, Inc., CNA Systems, Incorporated
Date	May 12
Person	Joe Smith
Person Position	Mega CEO
Currency	23 million dollars, $9.84
Measurement	32%, 47 percent
Country	China
Concept	proposed merger, analyst call, 1st quarter revenue weak sales, first trading day
Event: M&A	The proposed merger between Mega, Inc. and CNA Systems, Inc. has been postponed

Image 4 - Natural Language Processing via SAP Text Analysis (source: (SAP, 2011))

SAP Text Analysis currently supports over thirty world languages including not only the big ones like English, Chinese, Spanish or Arabic but also the small ones such as Czech or even Slovak. It also has more than 35 out-of-the-box predefined entities, relations and events to work with and categorize the extracted information into. Both these features can be further extended using custom or third-party packages and databases to satisfy the needs of industry-specific demands and related projects.

4.2.1 SAP Social Media Analytics

While already strongly positioned in the area of business intelligence and natural language processing, in December 2011 SAP proved its interest also in Social Competitive Intelligence and Social Content Management by joining forces with NetBase to provide a new service. By merging NetBase's natural language processing (NLP) engine with SAP's ConsumerBase a new product came to be: A cloud-based SAP Social Media Analytics to *"[…] conduct on-the-fly analysis on any subject of interest […]"*. (SAP, 2011)

"Before social media, businesses were in control of their brand messaging. Now, the market is saying what brands are really about. To keep ahead of the curve, next-generation companies need to adopt a new customer-to-business (C2B) operating model that runs at the speed of social media and can deliver what the market wants." (Peter Caswell, CEO at NetBase)

Because the event is relatively fresh in the time of writing this thesis, little to nothing is known about the actual capabilities of the product. But judging from demonstrational video available at the brand's official YouTube profile[11], SAP Social Media Analytics try to become a tool directly aimed at social networks with built-in support and connectors and thus be a complementary product to standard analytical tools in SAP Business Objects family.

4.3 HP Autonomy

Another big leader in area of content analysis and data understanding is Autonomy Corporation. This Cambridge spin-off company founded in 1996 focuses on something called Meaning Based Computing (MBC). Meaning of this rather ambiguous term hides different approach to dealing with data and searching for information. Instead of *traditional* keyword searching Autonomy platform focuses on understanding the meaning of the text be it in whatever form – structured, semi-structured or unstructured. This helps to dramatically reduce the number of false positive search results and return more accurate results including synonyms and documents with similar meaning rather than content.

This concept helps to automate processes that until now required human action. Autonomy doesn't aim to completely replace human work in area of sorting, approving and manipulating with documents, but highly reduces demands on labor. And although not without flaws[12], as Mike Lynch said in the interview on computing.co.uk (Computing, 2006):

"[Both MBC] technology and people are going to make mistakes, but the question is who will make the most mistakes. A human being using a keyword technology will often miss a lot more than MBC systems." (Mike Lynch, CEO at Autonomy)

Meaning Based Computing systems are not limited only to written text. Autonomy focuses also on the area of audio and video analysis and understanding, and delivers high end solution (for evolution overview see Image 5). The core of the whole platform, IDOL[13] server, is capable of performing conceptual analysis over audiovisual data, deep video indexing, trajectory analysis or optical character recognition. (Autonomy, 2012) This together with the ability to automatically cross-reference extracted information with other documents creates a very powerful analytical platform.

[11] See <http://www.youtube.com/watch?v=jMfYX3LgPJ0>

[12] Meaning recognition software such as Autonomy tend to have problems understanding sarcasm, irony, humor, etc. which produces negative or false positive results.

[13] Intelligent Data Operating Layer

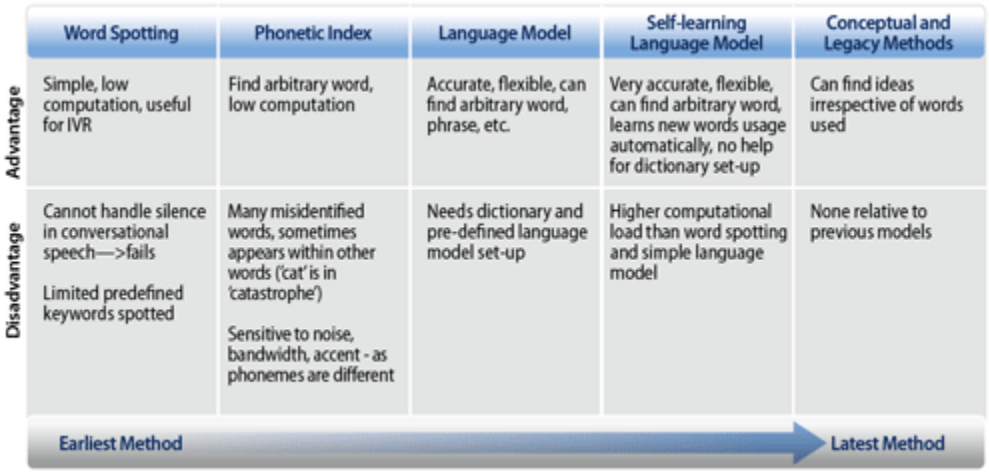


Image 5 - Evolution of speech recognition technology (source: Autonomy.com)

4.3.1 Acquisition

In October 2011 HP announced acquisition of over 200 million (87 %) of Autonomy Corporation shares at a price of $42 per share. (Autonomy, 2011) Full implications of this $8.5B investment are hard to predict and yet to be seen. However it is clear that HP, which currently undergoes some sort of business transformation and reorganization, too feels the importance and lucrativeness of this market area and wishes to stay amongst the leaders in this field.

On ZDNet Dennis Howlett offered the following commentary and explanation of events of the last year (Howlett, 2011):

"The fact Autonomy is transitioning to a SaaS model will provide HP's leadership with opportunities to learn from the experience of a company that so far has been successful in making the change." (Dennis Howlett, advisor to Constellation Research)

4.4 Google Refine

Last but not least of the important players in the area of business intelligence and data analysis is a well known Silicon Valley company Google.

Although known for internally dealing with large data for a long time (Feige, 2011 p. 8), this IT giant entered the public part of lucrative field of competitive intelligence and large data manipulation only recently with the latest improvement of its legacy software Google Refine – *"a power tool for working with messy data, cleaning it up, transforming it from one*

format into another, extending it with web services, and linking it to databases." (Google, 2011)

When Google acquired Metaweb in 2010 with a goal to *"improve search and make the web richer and more meaningful for everyone,"* (Google, 2010) together with full access to Metaweb's open internet database Freebase containing more than 12 million entries on various things such as books, movies, celebrities, companies, etc., it also gained Metaweb's open source tool for cleaning and enhancing large data sets – Freebase Gridworks. It wasn't long until the tool was renamed to Google Refine and moved under Google Project Hosting. (Google, 2010)

Unlike its competitors, Google Refine is open source, community-based and free of charge. And similarly to other products in Google family it is controlled fully via standard internet browser.[14]

Functionally it stands on three key features – cleaning messy data, transforming it to different formats and augmenting it, enriching it and extending it with the help of online web services. Thorough detailed description is out of the scope of this thesis, following paragraphs offer at least a brief overview based on official documentation and own personal experience. (Google, 2011)

4.4.1 Data cleaning

Main function of Google Refine tool is providing means of dealing with *messy* data, e.g. structured datasets consisting of various duplicates and inconsistencies. The simplest way to do that is by using facets, which group data based on the same value in the selected column, and group text operations like trimming whitespaces.

More advanced (and more powerful) feature is clustering, which allows forming data groups based not only on exact but also on loose similarities. This helps dealing with synonyms, abbreviations and other inconsistencies in naming. Google Refine also provides built-in mechanisms which form the possible clusters themselves based on heuristics mechanisms and offer them to user. That makes basic data refining fast, simple and quite intuitive.

[14] Although controlled via browser, Google Refine is a standalone product, not a web service. It must be downloaded, installed locally and run – either online or offline. For user convenience it only uses as a client GUI standard browser.

4.4.2 Data transformation

When it comes to dealing with data, cleaning and grouping is not the only function Google Refine offers. It also possesses variety of tools to help user transform dataset from one format into another. Program has built-in ability to recognize various standard data formats such as XML, CSV, RSS, JSON and other with possibility for user to specify and create a new structure template.

With this set of tools it is not difficult to work with differently structured, semi-structured or even unstructured data and transform them and merge them into large sets in the exact form one currently needs.

Also quite handy is the complex Undo/Redo function which allows for fast and easy traversing through all transformation steps back and forth without the fear of making some damaging changes to important data. Through this tool it is also possible to export all changes as a template and store them for future use on similar datasets.

4.4.3 Data augmentation

The most advanced features of Google Refine represent ability to exploit online web services and to use built-in extensions for data reconciliation. Both require advanced knowledge of both the tool and the way web services work (i.e. Knowledge of REST and JSON is almost mandatory) and will be therefore hidden to most users and used only through pre-configured templates. However when used correctly it can turn Google Refine from rather simple data refinery to powerful competitive intelligence tool.

Google Refine consists of built-in Freebase reconciliation plug-in and supports installation of other 3rd party extensions. With the help of its own heuristic algorithms it tries to *guess* the meaning of the data and link it to online database such as Freebase. With a bit of help from the user side one can easily transform for instance simple names' list of business partners into a complex dataset where each company or partner is linked to an online source with detailed and always up-to-date information.

With no doubt the most powerful feature of Google Refine is its ability to access web services and exchange data with various online sources. Every web service provides its users with an URL to its API through which they can access it and exploit it and these URLs can be filled into Google Refine and used in an automated data transformation and augmentation process.

While this might not seem that powerful at first glance, one must realize that there are hundreds and thousands of various web services accessible freely or for small payment today and more emerge every day. And while individually each web service offers only a limited functionality, when combined together they offer almost limitless possibilities. For instance when analyzing company Twitter account with aim to identifying key customers amongst *followers*[15] one might simply combine official Twitter API providing stream of latest *tweets* and statistics of *retweets* on specific account (e.g. company Twitter account) with another web service such as Topsy.com[16] which evaluates influence of individual users on Twitter network. Merging the results from these web services in Google Refine, company is able to get quick overview of social status and influence of its followers and aim its communication and advertising accordingly.

And this is just the beginning. One might not want to limit oneself only to one social network but may join the gathered results with other sources such as LinkedIn or Facebook, both of which also provide externally accessible APIs and gain complex and detailed database of his or hers customers, business partners and other notable subjects. Full strength of this feature is yet to be seen when more and more user-created templates are shared via community sites.

4.4.4 Summary

While simple functions like grouping data into facets create outcomes no better than *classic* pivot tables as seen e.g. in Microsoft Excel, the true power of the tool becomes visible when it comes to clustering data based on heuristics rules and augmenting it with the help of web services many of which (such as Freebase Web-API) are directly implemented into program's interface. As a downside of this otherwise nice and handy tool one might point out its complexity and demand on programming skills[17] which makes it more suitable for data mining experts and IT professionals than for managers and business-personal in need of help in their day-to-day business decision routine.

[15] On Twitter social network registered subscribers to specific Twitter account are called followers.

[16] Topsy is a search engine focused on social networks. All results can be accessed both via standard web interface and through RESTful "otter API" hosted at Google Project Hosting. (Topsy, 2011)

[17] GUI supports only basic operations. Advanced actions require use of built-in terminal and construction of set of logical functions and arguments.

Although not as polished as its commercial competitors and with somewhat limited capability of dealing with very large data sets[18] and its visualization, thanks to its zero cost and active community, Google Refine stands as a powerful tool and, if not a full replacement, then at least a solid complement to other structured data-focused analytical products currently on the market.

4.5 Tovek Tools

As was said earlier, although this work focuses primarily on global market and its current conditions, it has a small emphasis on local (Czech) market and its locally dominant players. And when one speaks about Czech CI market leaders, one must certainly not forget to mention Tovek – a small Czech company with almost twenty years of experience in the field of business intelligence and CI.

Since the division of Czechoslovakia in 1993 Tovek has been focusing on delivering business analytic solution for local companies both as a technology partner and distributor of Autonomy platform[19], and as a developer and vendor of its own custom solution – Tovek Tools.

Tovek Tools serve as a standalone analytical tool comparable to other major market players and their products. It is built around Verity core search engine[20] and offers a variety of data processing functions and features such as automated language detection, tokenization and stemming in five major European languages[21] with support for custom lexicons and dictionaries. (Tovek, 2011)

Tovek platform consists of two main parts – Tovek Server and Tovek Tools. While Server serves as a brain and central point of the system where the data are gathered, Tools represent a thick-client comprehensive analytical tool able to perform deep analysis over both structured and unstructured text. Amongst the main functions of Tovek Tools belong:

- **Topics categorization**

  When trying to extract information from given set of data, similar to other natural language processors each important term is properly evaluated, named, addressed to a given preset topic and categorized. This procedure turns

[18] When it comes to refining large sets with more than 100,000 rows and 10 columns, tool starts to be somewhat *clumsy* and overall slow.

[19] For introductory description see chapter 4.3.

[20] Verity was in late 2005 acquired by Autonomy. (Autonomy, 2005)

[21] Tovek Tools currently supports English, French, German, Russian and Spanish; and of course Czech.

formerly unstructured text into a set of structured meta-data and data, thus extracting from the text its actual value. This set of topics and key words can be then further worked with or even visualized.

- **Signal monitoring**

  Under the term of *signal monitoring* is hidden capability to periodically monitor given data sources for any unexpected changes in frequency of known topics. Such change in frequency can be a *signal* for some bigger event worth special interest and deeper analysis. Signal monitoring should help to reveal such important events right at the time they occur.

- **Entity extraction**

  Important *knowledge mining* tool is so called entity extraction feature, which allows for searching given datasets for any entity currently of interest. E.g. contact information on people connected with competitor.

- **Sentiment analysis**

  One of the *buzz words* of today is certainly a sentiment analysis. Trying to properly evaluate the social climate and sentiment of given web page, news article or community forum is something every major player on CI market is currently doing. And Tovek is no exception. By searching the text for *detractors* and *promoters*[22] their system tries to estimate the overall mood of the source.

As was already mentioned, core of Tovek Tools is based on Verity search engine. As such allows for complex searching over stored data using the modified in-house version of official Verity Query Language. Compared to e.g. Autonomy platform, Tovek Tools is more IT expert-oriented. While Autonomy primarily serves as a black box end-user analytical tool, Tovek Tools offer more open and deeper ways of configuration and customization, but for a price of higher demands on technical knowledge of its users.

Also in contrast to other products mentioned above, Tovek Tools is not primarily a data searching tool. While it can to some extent perform external data searching such as web sites crawling, it is still primarily an internal enterprise analytical tool.

[22] Detractors for negative sentiment and Promoters for positive sentiment.

To compensate for this, Tovek Server allows for connection of variety of different 3rd party databases and other data sources, and Tovek itself provides services for its customers in connecting their BI platforms to all main supported data suppliers such as:

- Systém sledování vazeb[23] (Creditinfo Solutions, s.r.o.) – Database of local economic subjects and their detailed information including interesting connections and links.
- Magnus (Czech capital information agency, a.s.) – Database of local economic subjects, stocks and important economic events.
- Anopress (Anopress IT, a.s.) – Daily refreshed media database of news, articles and broadcast transcripts from more than 1.500 sources.
- Economia (Economia, a.s.) – Database of economic magazines and articles.
- HeadlineReader (Anneca, s.r.o.) – Hourly refreshed database system monitoring more than a thousand different web sources, portals, news servers and press notes.

Although not as big as other major players, Tovek still plays an important role in the area of Competitive Intelligence – especially on central European scale. With almost twenty years of experience it rightfully belongs amongst the market leaders.

4.6 Conclusion

This chapter offered a list of dominant players on competitive intelligence market and description of their main products and features. Although not every company involved in competitive intelligence business was included and even those enlisted were given different level of interest with not all aspects being described in deep detail, this part of the thesis still fully serves as a complement to performed Porter competitive analysis from chapter 3.

As a final summary, following three facts are worth noting:

- SAP with its new cloud-based service Social Media Analytics appears to be the only company from mentioned *big five* to be directly interested and involved in social networks. While other companies are currently trying to extend their products for social networks support, by acquiring NetBase SAP took more direct and faster approach than any of its competitors.

[23] Loosely translated as System for Monitoring Connections

- Although Google doesn't have its own BI *flagship*, Google product family[24] offers several smaller tools which combined together create functioning platform capable of quite complex data analysis and visualization. Thanks to its open source approach and ever growing community formed around code repository code.google.com one should always keep this giant company in mind.
- Our small country lost in the middle of Europe can be a home of a company able to match quality and overall level of its products with global world competition. Although Tovek is not nearly as big and influential as any of the other main leaders, its capabilities are not to be underestimated.

[24] Meaning specifically Google Analytics, Google Trends and described Google Refine.

5 CI and SCM

Social Content Management is a new concept of dealing with corporate (mainly unstructured) data introduced by IBM a few years ago. I personally came in touch with the platform IBM offers as a SCM solution during my internship in early 2012.

The general aim of the SCM was summarized in former chapters as an *effort to create an environment that encourages people in generating social content through interaction and collaboration while providing platform owners (companies) with means of managing this content to be effectively used later.* This concept reflects the shift in paradigm of collaboration and information sharing. With the *Web 2.0* wave the content creation shifted from service owners to service users. Nowadays the main momentum and knowledge base lies not with product vendors and service providers, but with users and communities built around these services and products.

Solid boundaries and formal roles, functions and connections broke down and everything today is a part of a *social* network where communities along with people – their members – form a ground building block. And this does not apply only to public spheres of Internet, but also to corporate segment and organizational intranets, as is shown in Image 6.

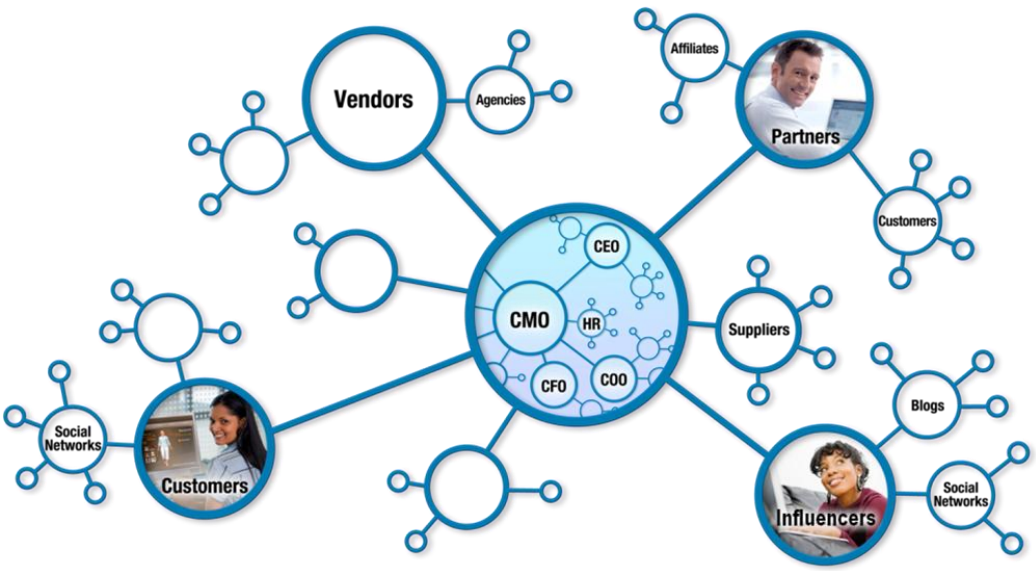


Image 6 - Social Marketplace (source: (IBM, 2012))

This change of paradigm deeply impacts every major business area including Competitive Intelligence, because before every analysis, one must first gather the source data. And to make the content management plausible, that is where SCM comes into play.

5.1 Knowledge Management

Important part tightly related to the concept of Social Content Management is Knowledge Management. As was already pointed out, the main idea behind SCM is to create a business platform where users (people) can generate *social* content and owners (companies) can manage this content to be effectively used later. This goes in line with Nonaka's known *SECI Knowledge-Creating model* and with the whole Japanese business concept in general.

Nonaka and Takeuchi in their publication *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation* discuss the process of creating and sharing knowledge inside (and later even outside) the boundaries of a company. (Nonaka, et al., 1995) They divide the actual knowledge into two distinct groups – *explicit* and *tacit,* and summarize the whole process of acquiring, transforming and sharing the knowledge into a matrix-spiral SECI model shown in Figure 2.

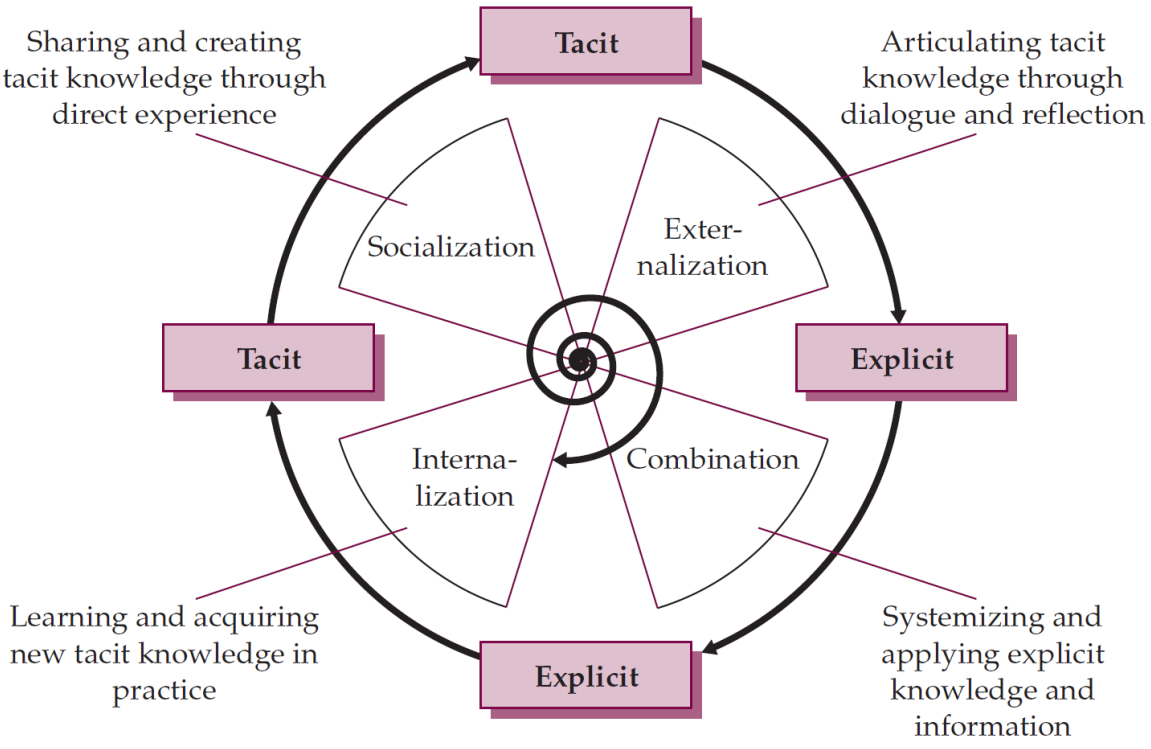


Figure 2 - SECI model (source: (Takeuchi, 2006) originally from (Nonaka, et al., 1995))

Explicit knowledge represents something formal and systematic, that can be easily expressed, communicated and shared. This reflects mainly the *Western* way of dealing with knowledge.

Japanese culture on the other hand works mainly with **Tacit** knowledge, which is looked at as something personal, not easily visible and expressible, difficult to communicate and share.

Social Content Management approach tries to support the whole process of knowledge creation, gathering and sharing on two different fronts, be it the explicit or tacit part – by providing the means of easy information sharing and creating an accessible collaboration platform; while in back-end carefully marking, indexing and categorizing all data as they are created so that they can be used again in the next iteration of knowledge building process.

5.2 Measuring influence

One of the current trends in relation to social networks is the collective effort of companies to exploit the possibilities social networks offer and generate some extra value and profit. And general rule that also applies in this area is that *the bigger influence you have, the better*.

Everlasting problem however is: How to actually measure ones influence – especially in an environment as *loose* and *fuzzy* as social networks?

Mani Karthik discussed several approaches to measuring influence on Facebook – current world largest social network. According to his work one can consider four different approaches on *what* to measure when trying to evaluate ones influence on Facebook (Karthik, 2010):

- **Influence**

  One of the ways is to measure number of comments and stack it up against time. Commenting is one of the pillar functions of Facebook for its users to interact. The more comments a single post has in a short time, the more people is probably *listening* and therefore influenced by it.
- **Reach**

  Somewhat simpler approach is to focus on a number of *Likes*[25]. It may not reveal the details and opinions of previous approach, but on the other hand it shows ones reach.

[25] Facebook way of telling others: „I find something interesting“.

- **Quality**

  Another interesting idea considers overall network growth in terms of *friends*. If one could compare the rate of growth of one's own friend network to something like the *average network growth*, one would get a quick reference on whether one is not *behind*.

  However here it is important to point out, that interesting and somewhat adding to *quality* of one's profile are only external friend requests from others, not one's own bonding initiative.

- **Popularity**

  Lastly in terms of measuring influence and popularity Karthik discusses the possibility of monitoring activity and interaction outside the immediate inner circle of friends.[26] The more influential one is, the more will one's topics be discussed even outside the boundaries of one's network.

There are also several books dedicated solely to Facebook and exploiting its functions to leverage one's business potential – amongst others Brian Cartner's *The Like Economy*. There he, among other things, discusses the possibilities of applying influence principles to Facebook – especially through principle of *reciprocity*. On this account of *give and ye shall receive* Cartner shares his own experience:

"I had demonstrated my expertise in the free course, and in line with the idea of reciprocity, any sense of obligation [my viewers] felt to me would make it more likely they'd select me for that work." (Cartner, 2011 pp. 223-224)

And it is not only the biggest social network of today that draws attention and special interest of various subjects. Also Twitter, a widely used network dedicated to sharing short notes and links, has been a subject of various attempts on how to utilize its hidden *riches*.

Topsy

Stanford University student project *Sentiment140* capable of performing brand sentiment analysis on Twitter was already mentioned in the Prologue. And there are others, such as Topsy – *"a realtime social search engine with more than a half-billion queries processed per month."* (TopsyLabs, 2011)

[26] Facebook currently provides such functionality for Facebook Pages, where it shows graphs and trends of influence under chart: "Talking about".

But Topsy is not just a plain search engine, but also a social analytics tool capable of performing real time social environment monitoring, trends visualization and prediction, or influence and reach analysis. In a limited form it is available publicly for free via web APIs.

Using this publicly available Twitter API[27] for measuring influence and combining it with e.g. Google Refine, one is able even with free open tools to perform basic influence analysis of one's customers, fans and other Twitter *followers*. However unlike Sentiment140, Topsy's internal mechanics of evaluating influence and other metrics are not publicly available and therefore ambiguous. Yet it is still a very useful tool – if not directly for measuring influence, then at least for frequency analysis and trends visualization via the analytics.topsy.com *social* API.

Klout

Californian project launched in 2008 aims to become a Google in the field of social networks. But instead of providing links to web pages most relevant to searched terms, it provides links to people most influential in a given field or area of interest. The official page offers a short explanation:

"In the same way Google analyzes public websites to generate PageRank[28], Klout analyzes publicly-available data to measure a person's influence." (Klout, 2008)

Currently Klout indexes five[29] major social networks including Facebook, Twitter, Google+, LinkedIn and Foursquare, and others like YouTube or Flickr are to come. Based on his or her activity, influence, reach and community membership, each user is given a Klout rank ranging from 0 to 100 (with average at 20). This together with user interaction and giving additional K+[30] creates a living network where everyone has his or her biggest *influencers* always at his or her side.

Interesting idea, however there have been some issues with privacy and internal mechanisms ambiguity in 2011 resulting in some people abandoning Klout. (Moore, 2011) John Scalzi at CNN even called Knout *"a little bit socially evil."* (Scalzi, 2011) How (and whether) the network copes with this wave of criticism is yet to be seen.

[27] See <http://otter.topsy.com>

[28] Google evaluates each web page with a number according to its content, significance and relevance to subject. This number is called PageRank. Search results are then ordered based on it.

[29] Klout works with thirteen social networks, but currently measures influence and other values only based on these five.

[30] An acknowledgement of help / influence similar to +1 on Google+ or Like on Facebook.

5.2.1 Key Account Management

When discussing influence one must not forget one of the most important areas of customer relationship management for every company – so called Key Account Management (KAM), which can be characterized by following definition from Peter Cheverton's book *"How Come You Can't Identify Your Key Customers?"*:

"KAM is the management of those customer relationships that are considered most important to a company. Key accounts are defined as those accounts which are held by customers producing or responsible for the bulk of the profits for a company and/or those which have potential to do so." (Cheverton, 2002)

In conditions of today's e-market one must not underestimate power of social networks. To identify amongst its *fans* and *followers* people with the greatest influence and therefore potential to either help promote or ruin every marketing effort should be one of the *must-dos* of every responsible modern company.

5.3 Summary

This chapter discussed the various aspects of Social Content Management and its relation to Competitive Intelligence. The actual product platform I was introduced to during my internship at IBM (see Image 7) consisted of four individual products that together form the SCM solution.

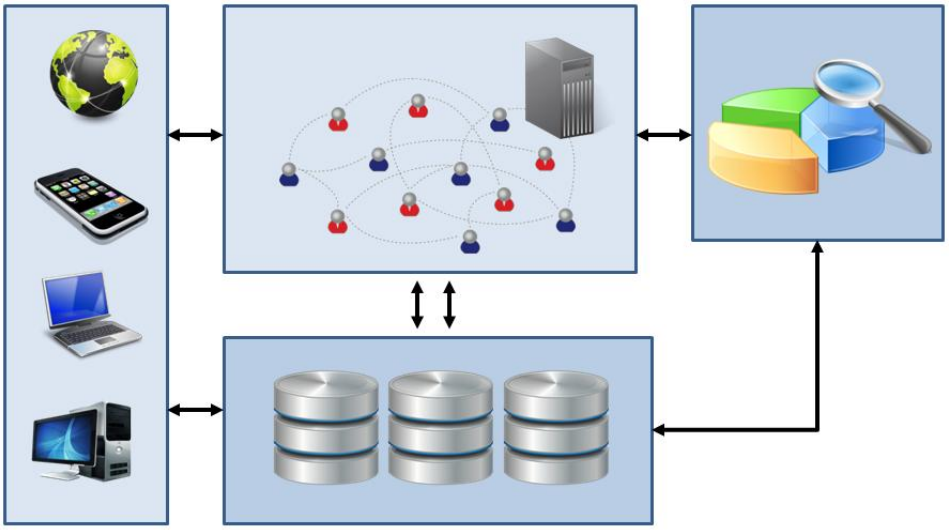


Image 7 - SCM platform model (source: Author)

As a central node of the platform stands **IBM Connections** – a social collaboration tool sometimes referred to as a *"Facebook for companies."* This being backed up by complex

platform for enterprise content management – **Filenet P8 Platform**, which provides the necessary background and allows for advanced management of stored data.

For higher accessibility reasons the platform was also accompanied by **Lotus QuickR** tool capable of integrating standard Windows user interface with FileNet and thus creating another way of reaching, viewing and manipulating stored data.

As a fourth and final pillar then stands the **IBM Content Analytics** with sole purpose of performing complex and detailed analysis over data created via Connections and QuickR and stored in FileNet P8.

This platform as a whole is able to take care of the needs of both Content Management and Competitive Intelligence areas[31].

Somewhat different approach was introduced by Mzinga and its *OmniSocial* service with functionality not much different from the one offered by IBM. The main and key difference is a form of the product – while IBM goes the way of delivering and independent solution working internally inside a company, Mzinga delivers a *"cloud based enterprise-class, social business ecosystem."* (Mzinga, 2011)

Each approach has its pros and cons. How the whole area evolves, and which way will the market go in the future is currently hard to predict. Be it one way or the other, we definitely have some interesting years coming.

[31] Of course speaking about CI in terms of internal data sources, not external competitive analysis.

Chapter 2 – Practical Part

6 Real life project

To test theoretical possibilities and capabilities of current professional products in practice, during my internship at IBM in early 2012 I participated on a real project in cooperation with local team of IBM experts. The project was aimed at competitive intelligence analysis of Slovakian banking market and its results were to be presented during IBM Forum 2012 in Slovakia.

The main purpose of the case study was to test usability of social networks – namely Facebook – for competitive intelligence and sentiment analysis while also testing the limits of IBM analytical software Content Analytics when working in *unknown* (or better *unsupported*[32]) environment.

6.1 IBM Content Analytics

Although many different products and applications were used during the whole project from simple web browsers, company intranet applications and communication tools and other supporting software, the main *core* tool was of course IBM Content Analytics (ICA).

This complex professional analytical software is built around several open source standards, which were fully adopted and further extended to allow interoperability with other tools be it IBM or 3rd party software. ICA itself can be then divided into several components - each specialized in specific area or task, that together support the whole analytical process from data gathering in the beginning to added value presentation and visualization at the end.

As can be seen in Figure 3, Content Analytics stands on two main pillar applications - Text Miner and Administration Console and several components. While Text Miner serves as an end-user client analytical tool, through which the knowledge working and value extraction is done; Administration Console is pure administering platform which allows administrators to configure and monitor each of the key specialized ICA components, that support the whole analytical process.

[32] Although Content Analytics supports many different world languages, Slovak is not amongst them. This led to some interesting problems and limitations, which are discussed further in the text.

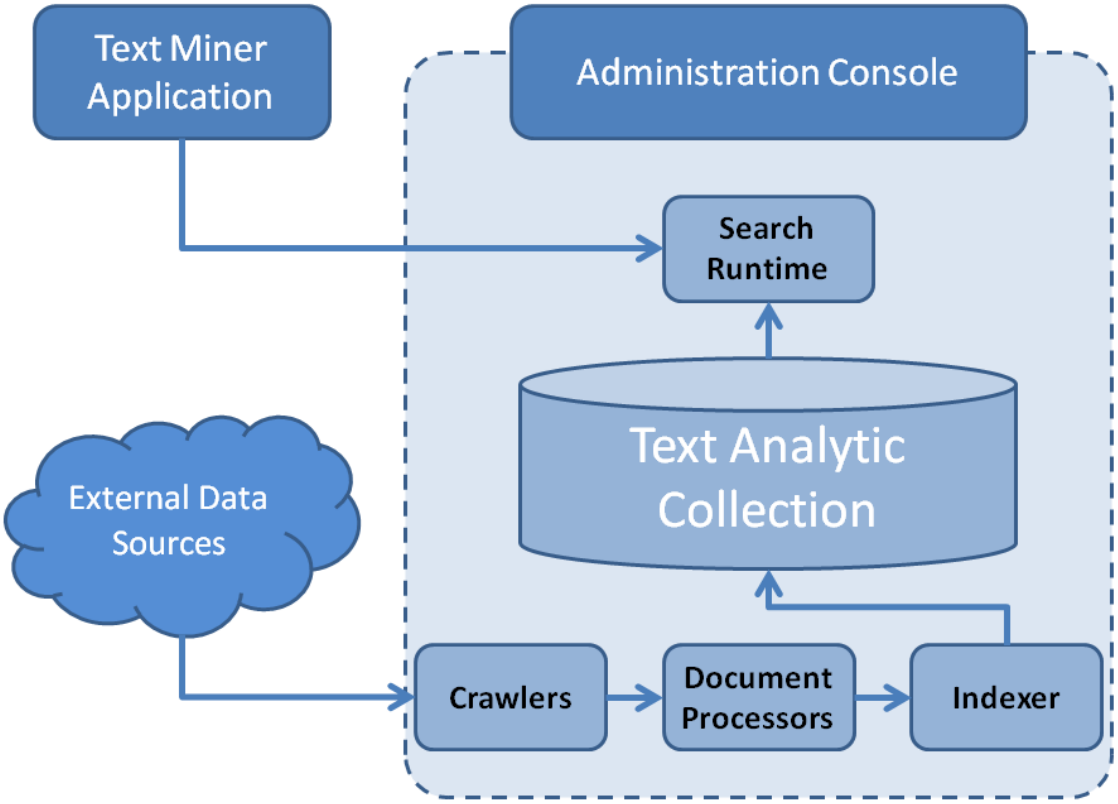


Figure 3 - ICA Component Architecture (source: (IBM, 2004), modified by author)

6.1.1 Content Analytics Components

According to official IBM RedBook documentation, the *core* of the Content Analytics software consists of the following set of components: Crawlers, Document Processors, Indexer, Text Analytic Collection and Search Runtime. (IBM, 2004)

- **Crawlers**

  Crawlers are base components of Content Analytics. The main and only purpose of these pre-programmed, pre-configured *bots* is to gather data from external sources. Based on the nature of the source we can then divide crawlers into several categories such as web-based crawlers, file systems crawlers, relational database crawlers, email crawlers and others.

  The key feature of the crawlers is that they can be run automatically and repeatedly, which allows for automated creation of data collections and their periodical updating.

Gathered raw data is then passed on to next component, which is Document Processor.

- **Document Processors**

  As the name itself implies, the role of the Document Processors is to actually process *raw* data gathered by crawlers and to prepare them for indexing and inclusion into Text Analytic Collection.

  Key part in the processing plays so called *annotators* - various text analytics, which perform individual processing tasks such as document language determination, linguistic analysis, pattern matching and others. Annotators in Content Analytics are based on Unstructured Information Management Architecture (UIMA) open source standard sponsored by Apache organization and use standardized Common Analysis Structure (CAS) to represent each document during the processing, which allow for easy inclusion of other, custom made standard-compliant annotators.

  Generated parsed document enriched of specified annotations and other important metadata is then passed on to the next component for indexing.

- **Indexer**

  Role of the Indexer component is to allow for fast and optimized text mining and analysis of parsed documents which are stored in collection.

  Because Content Analytics indexer is based on *Apache Lucene Indexer*, at the output of UIMA pipeline during parsing phase, documents are converted into special *Lucene document* format. This document is then automatically indexed and stored in text collection so it is available for end-user operations and queries.

- **Text Analytic Collection**

  Text analytic collection is the key entity in the whole Content Analytics system and can be looked at as a complex database. Every single document gathered by crawlers is – after proper parsing and indexing – stored in one of the text analytic collections preconfigured by software administrator.

  Configuration itself is crucial for proper functionality of the whole tool, because it is here where the nature of to-be-gathered (and stored) data and languages to be used is specified.

  When all previously described steps finished and a collection is built, its content is then copied into one or more Search runtime components so it can be accessed by end-user analysts via client Text miner applications. This also allow for

uninterrupted operability of the whole tool even during reindexing of the collection (e.g. when changing parameters to be parsed).

- **Search Runtime**

  Search runtime is a server-based component whose responsibility is to service search and analytic requests of client applications such as Text miner. The whole service runs on SIAPI – IBM's unified Search and Index Application Programming Interface based on java that allows creating one program able to search through different back-end search products (if they use SIAPI, of course) such as the whole IBM product family. (IBM, 2005)

  Client applications such as Text miner can then access SIAPI remotely and communicate with Search runtime via standard protocols such as HTTP or HTTPS.

  As was already mentioned – thanks to the fact that search runtime does not operate directly on *hot* data in collection but has its own copy, it is able to service client applications even during reindexing and reconstruction of the source data collection. This, linked together with the ability of text collections to be associated with more than just one search runtime at a time, creates highly available and largely scalable multiuser environment.

6.1.2 Administration Console

Administration console is a robust administration component of Content Analytics and it is the main tool used by Content Analytics administrator. It allows for easy and fast but complex administration of the product as a whole or its individual parts as they were described above. The most important tasks include:

- Components and collections administration,
- System monitoring,
- Security and user rights configuration,
- Client applications and collections association.

As can be seen in Image 8, graphical user interface (GUI) of the Administration console is browser-based and is therefore platform independent which certainly adds on usability.

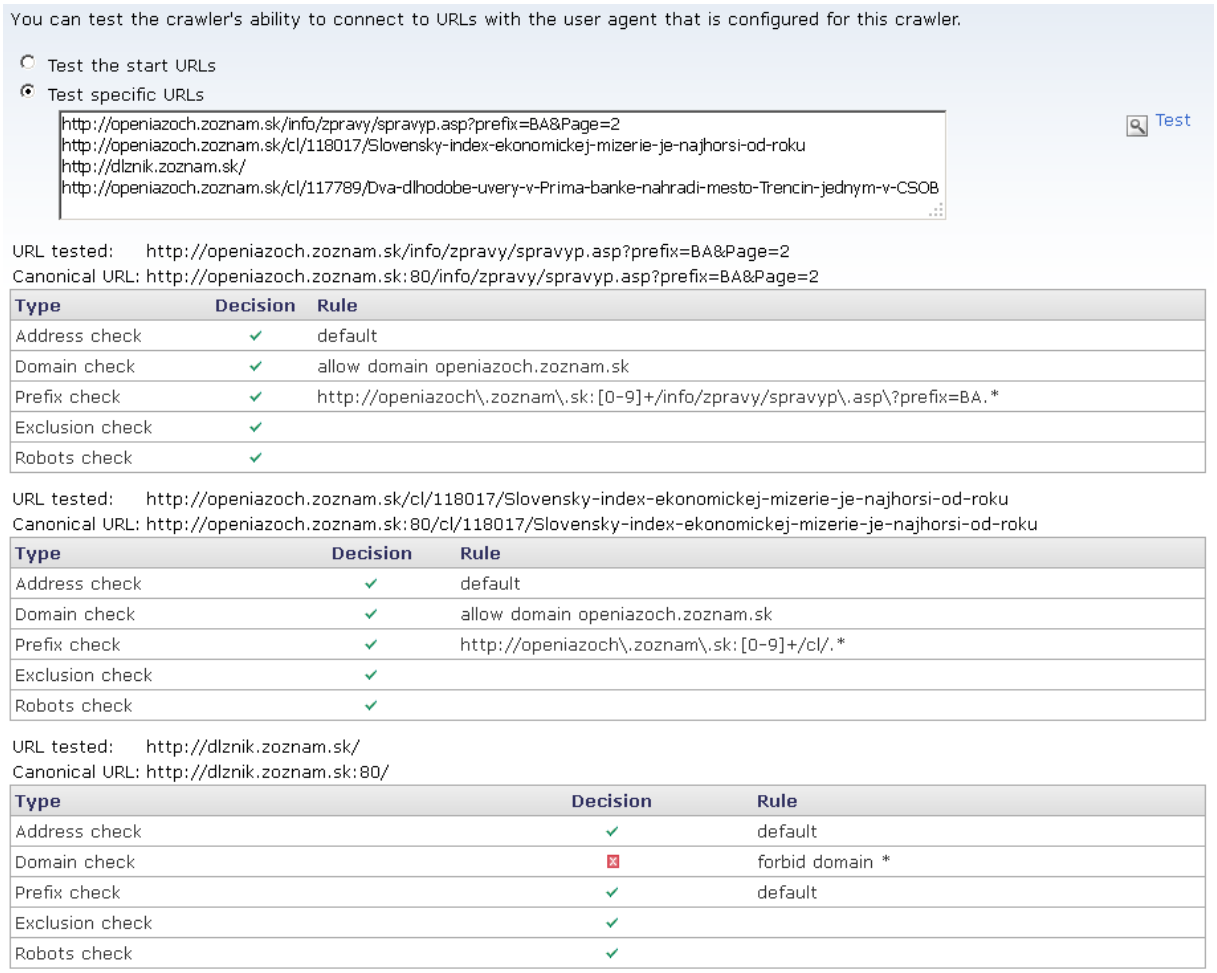


Image 8 - Administration Console – Web Crawlers configuration (source: Author)

6.1.3 Text Miner Application

Text Miner Application is an example of client application using SIAPI. It is a base built-in end-user application component and represents the Content Analytics front-end. After a collection is created and configured by administrator and filled with crawled, parsed and indexed documents, the platform is then prepared for the key *knowledge working* phase. During this phase analysts examine gathered data and through Text miner application perform searching, filtering, grouping, sorting and other operations over individual collections to extract new information and to gain knowledge.

Due to fuzziness of the whole knowledge working process there are no rules, guidelines or advices on how filter, group or sort data in collections to get desired (or sometimes even unexpected) results. Text Miner Applications therefore does not offer any configuration wizards like Administration console but instead focuses on creating multidimensional environment where end-user is allowed to look at the same data from many different views, angles and dimensions. These *tabs* include:

- **Documents view**

  This basic view shows a list of documents based on current search query conditions. Each item in the list contains basic information important to identify the document such as: Document date, title, source, thumbnail of the content and of course link to the whole original document stored in collection (see Image 9).

  This view is very useful in final parts of knowledge working process after the data in collection has been filtered from thousands to just a few relevant documents that need to be manually checked and read (e.g. to eliminate false positive matches).

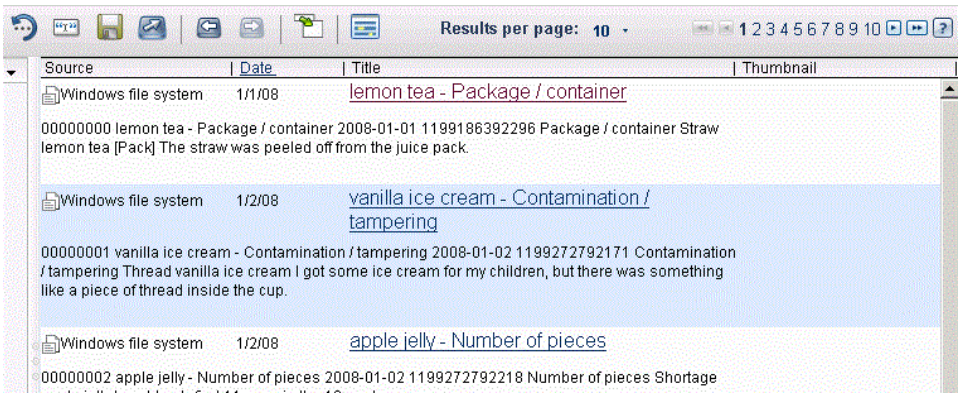
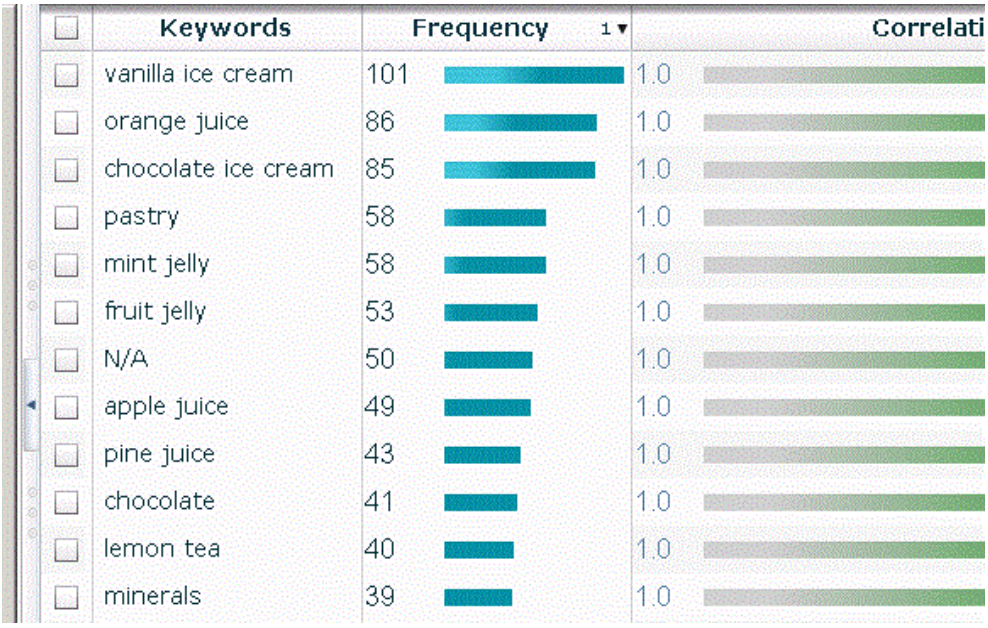


Image 9 - Text Miner: Document view (source: (IBM, 2004))

- **Facets view**

  Facet view shows detailed list of keywords for selected facet. Each item in the list has its *frequency* of occurrence in filtered documents and its *correlation*[33]. As can be seen in Image 10, keywords in the term of facets view do not have to be single word but can also be a whole phrase or a text pattern – depending on what analyst is looking for.

  This is useful to get the overall idea of what is the nature of collected data, what are the most frequent and relevant entities throughout the whole collection or its selected part a where might be the sought knowledge hidden.



Keywords	Frequency	Correlati
vanilla ice cream	101	1.0
orange juice	86	1.0
chocolate ice cream	85	1.0
pastry	58	1.0
mint jelly	58	1.0
fruit jelly	53	1.0
N/A	50	1.0
apple juice	49	1.0
pine juice	43	1.0
chocolate	41	1.0
lemon tea	40	1.0
minerals	39	1.0

Image 10 - Text Miner: Facets view (source: (IBM, 2004))

- **Time Series view**

  Primary purpose of this view is to display distribution of occurrences of relevant documents over time period (see Image 11). Relevance is based on selected facet and created search query.

  This comes useful e.g. in thorough long-time media analysis when inspecting occurrences of company and competitor's products in monitored media.

  Through Time series view one can also easily limit displayed results to specific time period to be able to perform much more detailed analysis on a smaller sample of data.

[33] One might also call it *relevance* or *interrelation* to selected documents.

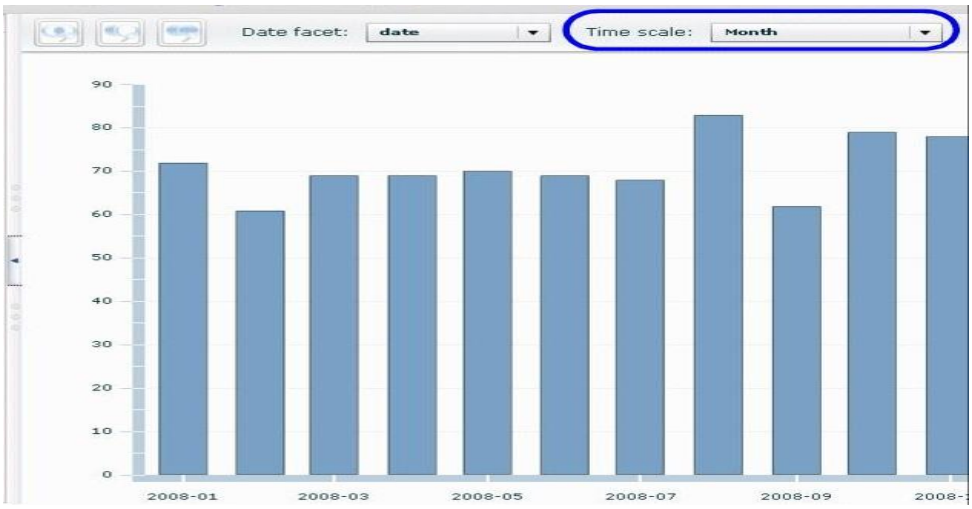


Image 11 - Text Miner: Time Series view (source: (IBM, 2004))

- **Trends view**

  Trends view tab might look the same as Time series view, but only at a first glance. Although Trends view also shows distribution over time, it works solely with facets and not the collection of documents as a whole. Its purpose is to show sharp increases (or decreases) in occurrence of individual keywords and phrases in selected facets and therefore easily reveal some trend-breaking anomalies (see Image 12). These might be caused by seasonal events such as Christmas or tax return deadline, but might also signalize some unexpected events worth investigating in more detail.

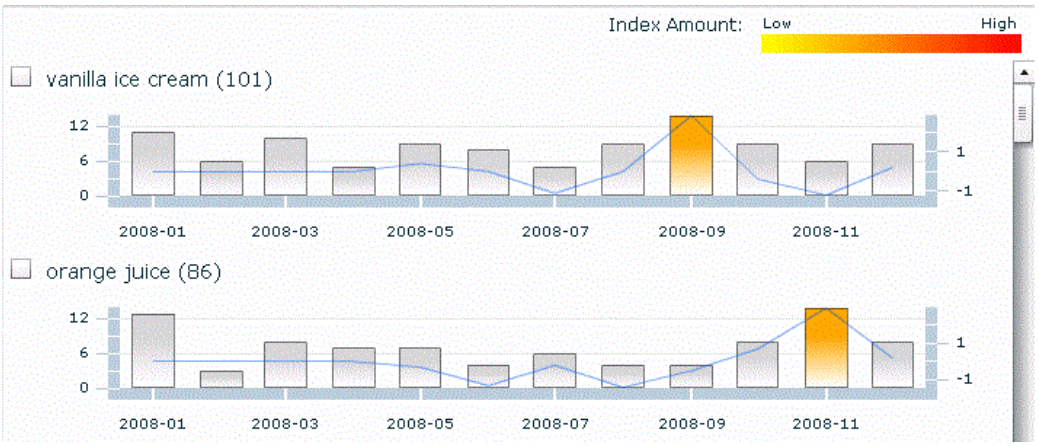


Image 12 - Text Miner: Trends view (source: (IBM, 2004))

This of course has next to no use in one time *ad hoc* analysis, but reveals its full potential after a long time period during which crawlers repeatedly gathered new raw data to be processed, parsed and indexed.

- **Deviations view**

  This tab seems similar to Trends view at a first glance, but its purpose is quite distinct. While Trends view tries to show distribution of occurrences of individual keywords in facets over time and highlight any significant deviations in their frequency based on their previous frequencies, Deviations view focuses on finding deviations amongst facet's keywords for given time period (see Image 13).

  Instead of computing average frequency based on previous time periods and predicting trends, Deviations view compares how the frequency of each individual keyword in facet differs from the average of frequencies of all keywords in facets for a given period. This together with extended filtering options including *Day of week* or *Month of year* allows for identification of interesting patterns in gathered data such as cyclic or seasonal patterns and also alerts analysts in case those patterns are to suddenly change.

Image 13 - Text Miner: Deviations view (source: (IBM, 2004))

- **Facet Pairs view**

  One of the most important tabs in Text miner if not the most important one is this view which allows to display correlation of keywords from two different facets. Resulting set of individual frequencies and correlations can be then viewed either in a simple list form (*table view*), matrix form (*grid view*) or overall view (*bird's eye view*).

- Table view looks similar to single Facet view with the difference that it displays all combinations of keywords in two selected facet (limited to the set of the most frequent for performance purposes).
- Grid view shows the same data transformed into a matrix providing both detailed numerical data for each keyword pair and graphical indicators of important fields (see Image 14).
- Bird's eye view provide overall preview of the whole dataset with the ability to chose a subset and zoom in for more detailed information.

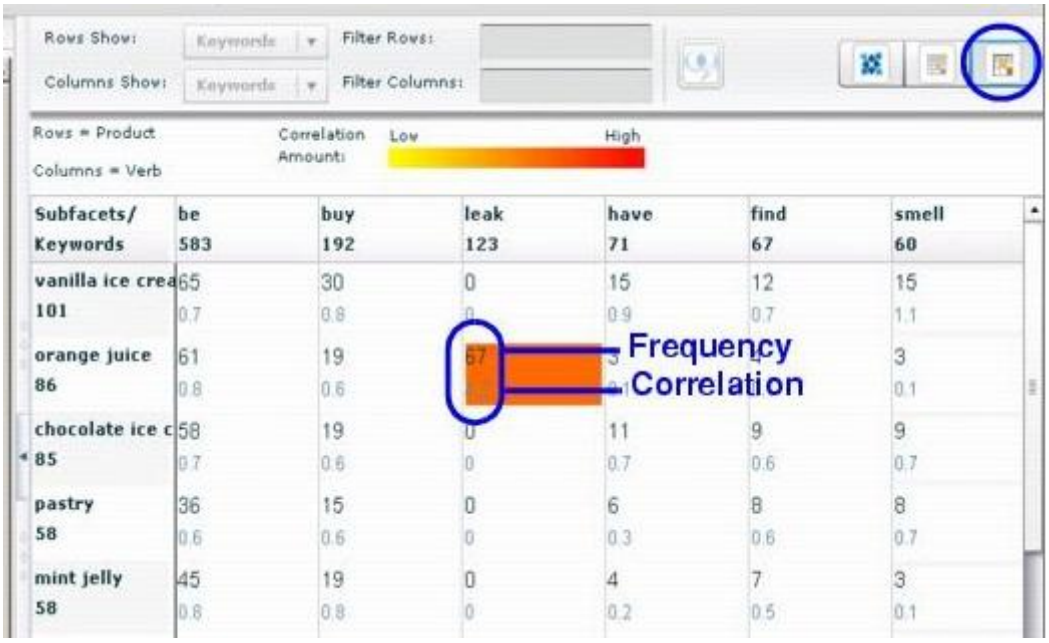


Image 14 - Text Miner: Facet Pairs view (source: (IBM, 2004))

This may not seem as much of an impressive functionality, but it is through facet pairs view where one is able to explore hidden corners of text collection and to e.g. determine overall mood of customers in connection with individual competitors or find a reason for decreasing sales of company's main product.

- **Connections view**

  Connections view represents another way of how to look at individual facet pair, but its functionality is so different from other views described in Facet pairs view that it was given its own tab.

  After selecting two facets the main workplace window displays dynamically computed cloud of facet terms and keywords (represented as bubbles) with connections (represented as lines). Through this simple graphical representation shown in Image 15 one can easily at a first glance recognize the most frequent

terms (biggest bubbles) and the strongest correlations (orange or red lines) between them.

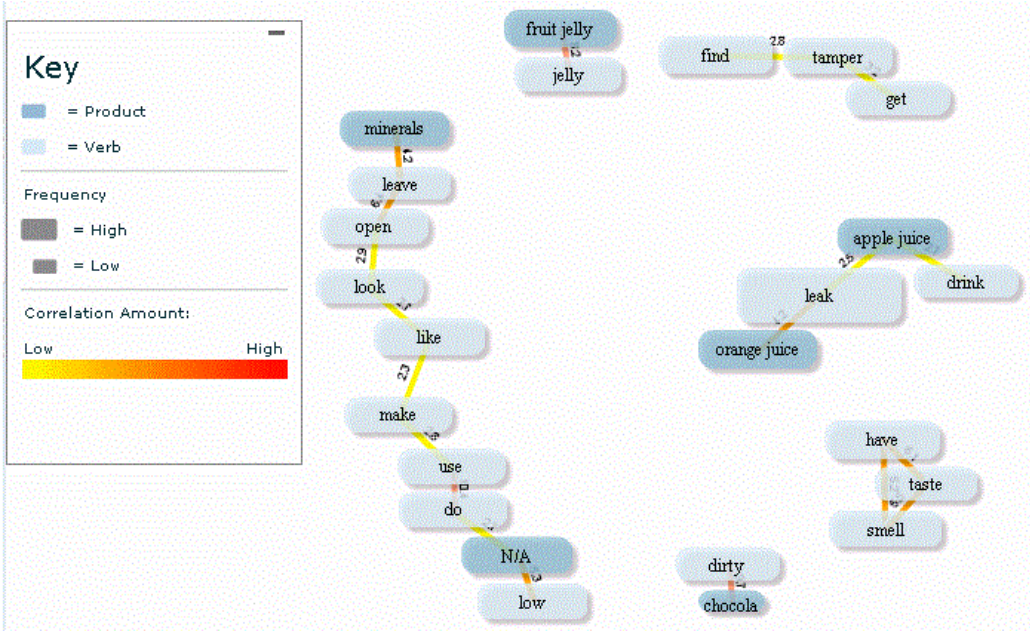


Image 15 - Text Miner: Connections view (source: (IBM, 2004))

Thanks to this visualization of the facet pairs and the ability to manually rearrange the cloud and to further filter it by adjusting the *correlation threshold* level one can reveal some new and unexpected facts and relations e.g. which competitors or products are often mentioned together, what is the key adjective people associate with the investigated product brand, etc.

- **Dashboard view**

  Functionality of the Dashboard view is not meant to help directly during the knowledge working process, although one might be able to see new facts when they are visualized. As can be seen in Image 16, the main purpose of this text miner tab is to visualize the most interesting findings and important facts in graphs, pie charts and tables so they can be understood by business staff and further shared, presented and used in collaboration process. Analysts can choose between several predefined default and custom-made layouts and export visualized data as images to be included into presentations and analysis reports.

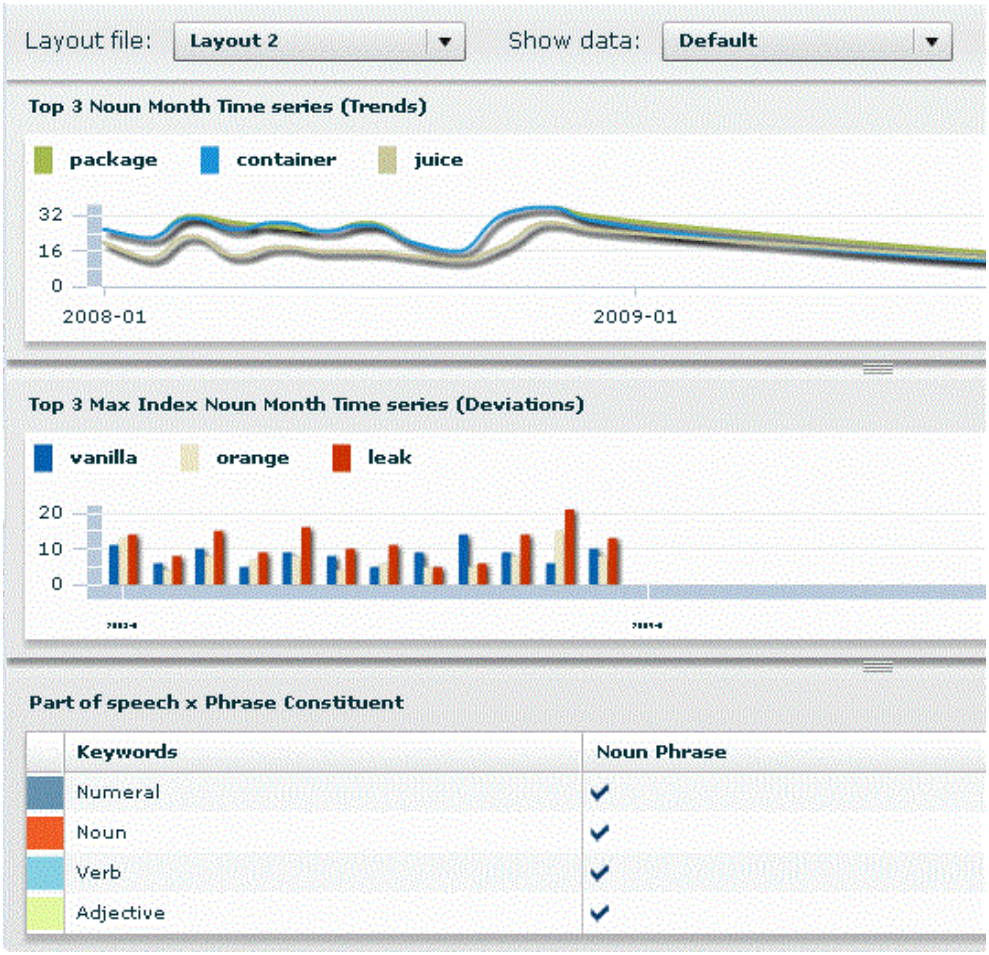


Image 16 - Text Miner: Dashboard view (source: (IBM, 2004))

6.2 Actual Project

The actual real life project took place in March 2012 during winter course of student internships at IBM under the IBM Smart University CZ program. I participated on the project as a member of the realization team together with my student colleagues and IBM mentors and experts. The project itself was preceded by several important smaller projects including programming and configuration of custom *Facebook crawler* plugin[34] for Content Analytics and can be looked at as a culmination of the whole four months internship run.

Because of complexity and variety of different actions requiring usage of several different tools and based on previous experience and expertise of IBM experts, the whole project was divided into multiple phases for better coordination and workflow control:

- *Initial phase* – to define the purpose of the project,
- *Global market analysis* – to gather base source data,
- *Goal definition phase* – to define primary and secondary goals of the analysis,
- *Preparation phase* – to fill and configure ICA dictionaries and crawlers,
- *Data-mining phase* – to gather detailed source data,
- *Knowledge working* – to extract actual knowledge from source data,
- *Presentation phase* – to create scenarios to be presented,
- *Closing phase* – to summarize and close the whole project.

Some phases lasted only a few hours, some lasted days, the core activities of the project, that took place in several repeated iterations, lasted more than a week. To better understand both relative length of each individual phase and mentioned iterations, see Figure 4.

[34] This is described in more detail in Data-mining phase of the project. See chapter 6.2.5.

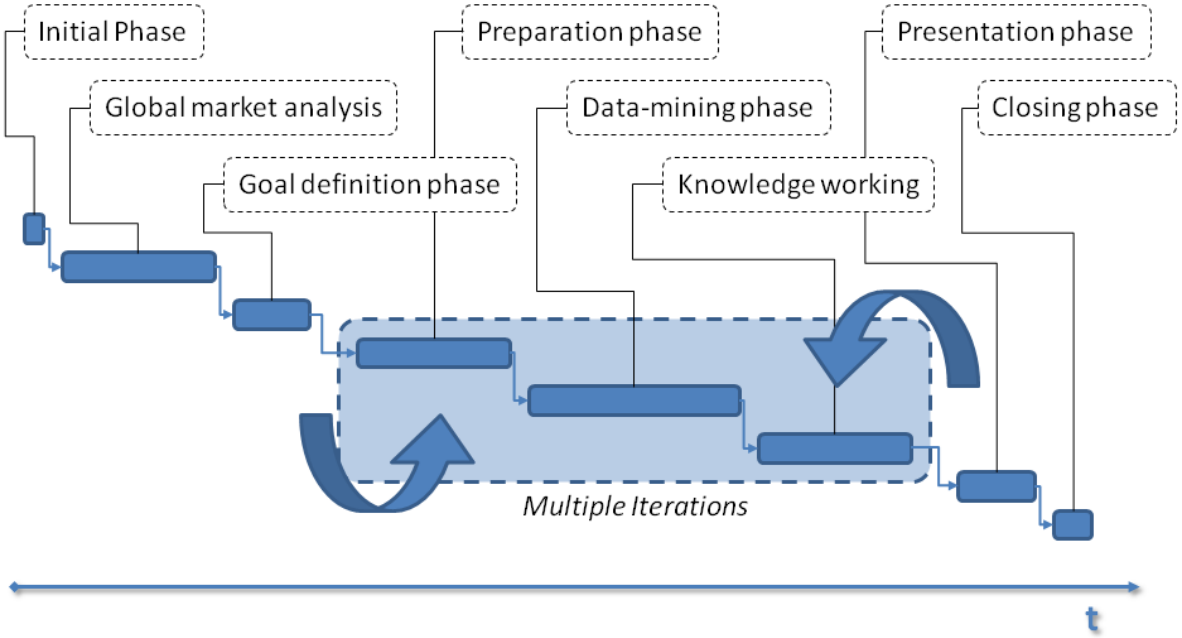


Figure 4 - Project phases diagram (source: Author)

6.2.1 Initial phase

Purpose of the Initial phase was to undergo necessary steps to officially launch the whole project and begin actual work on it. This included especially activities such as determining the general aim and goal of the project and putting together the realization team.

Because the results of the project were to be presented at the IBM Forum 2012 in Slovakia, it was decided that the project would be aimed at Slovak market. Due to small size of the target market[35] there was very limited number of areas suitable for general competitive analysis[36]. Therefore it was decided the project would be aimed at banking sector with emphasis on analyzing its overall state in connection with social networks.

This brought an interesting challenge in fact, that the area of social networks and their usability for competitive intelligence analysis using tools such as IBM Content Analytics was until then untested and unexplored. Created project plan and expected results were therefore based mainly on team's joined knowledge together with *best guess*, because there was no known previous work that could have been used as a reference, guidelines or *best practice*.

The realization team itself in its final form (after thorough discussion about needs and requirements of each project phase) consisted of two IBM expert analysts, ICA language

[35] Population of the whole Slovakia is only 5.4 million people.
[36] Based on size and business potential there were only public sector and financial sector that could have been taken into account.

dictionary specialist, programming developer, consulting native speaker as a language expert and consulting resource analyst.

The individual roles and responsibilities during the whole project were given as follows:

- **Senior IBM expert analyst**

  To oversee the whole project, coordinate all activities from the global perspective and ensure activities comply with the project plan.

  He was a formal project leader responsible for success of the whole project and quality of its final outcome.

- **IBM expert analyst**

  To provide overall support and expertise throughout the whole project. To help coordinate activities during each phase and to communicate with other team members in case of unexpected event or problem.

  He was responsible for quality of individual outcomes of each project phase.

- **ICA dictionary specialist**

  To perform necessary configuration steps for ICA to properly crawl given web sources and parse gathered documents. To fill language dictionaries and set up term facets.

  He was responsible for quality of language dictionaries and term facets created in Preparation phase.

- **Programming developer**

  To program new custom web crawler plugin for Content Analytics able to exploit Facebook web API and to gather publicly available user data for further analysis.

  He was responsible for quality and configuration of custom Facebook crawler for ICA.

- **Consulting language expert**

  To function as a native language support in crucial Preparation project phase, during which Content Analytics' components were configured and filled with language dictionaries tailored for Slovak banking market.

  He was responsible for language quality of dictionaries created in Preparation phase.

- **Consulting resource analyst**

  To perform initial global market analysis and gather available resources suitable for future detailed crawling. And provide support throughout the whole project and each of its individual phases.

  He was responsible for quality of analysis performed in Global market analysis phase.

After it was decided what the goal of the project would be and what were the roles of individual team members, the project advanced into the next phase – Global market analysis.

6.2.2 Global market analysis

The purpose of the Global market analysis was to get general knowledge of the state of the banking market sector, create a comprehensive list of its subjects and their products and gather enough source data to be crawled, parsed and analyzed further in the project.

Although this initial analysis and source searching was done manually and with nothing but a web browser and public search engines, this does not render the results less important. On the contrary – outcomes and findings of this phase served as source data for the rest of the project therefore it was crucial for them to be accurate and truly detailed.

Outcomes can be divided into three parts – list of Slovak banks, list of products and list of sources suitable for further analysis.

- **Slovak banks**

  Bank subjects acting on a Slovak market can be divided into two groups – banks with subsidiary office in Slovakia (*SK Banks*) and foreign banks. Each group has fifteen members as is shown in Table 2.

Table 2 - List of Banks on Slovak market (source: (Banky.sk, 2012))

SK Banks	Foreign Banks
Československá obchodná banka, a.s.	AXA Bank Europe
ČSOB stavebná sporiteľňa, a. s.	Banco Banif Mais, S. A.
EXIMBANKA SR - Exportno-importná banka	BKS Bank AG
OTP Banka Slovensko, a. s.	BRE Bank SA
Poštová banka, a.s.	Citibank Europe plc
Prima banka Slovensko, a. s.	COMMERZBANK Aktiengesellschaft
Privatbanka, a. s.	Crédit Agricole Corporate and Investment Bank S. A.
Prvá stavebná sporiteľňa, a. s.	Fio banka, a.s.
Slovenská sporiteľňa, a. s.	HSBC Bank plc
Slovenská záručná a rozvojová banka, a. s.	ING Bank N. V.
Tatra banka, a. s.	J&T BANKA, a. s.
UniCredit Bank Slovakia, a. s.	Komerční banka, a.s.
VOLKSBANK Slovensko, a. s.	Oberbank AG
Všeobecná úverová banka, a. s.	The Royal Bank of Scotland N. V.
Wüstenrot stavebná sporiteľňa, a. s.	ZUNO BANK AG

More interesting information than from a simple list of subjects on a market can be gained when investigating *online presence*[37] of each bank. This area was subjected to a deeper analysis later during the project and is described in more detail in Appendix. For purposes of showing results of this project phase see following simple summary of the top eight banks with the highest online presence in Table 3.

Table 3 - Banks with the highest online presence (source: Author)

#	Name of Bank
1.	OTP Banka Slovensko, a. s.
2.	Tatra banka, a. s.
3.	ZUNO BANK AG
4.	VOLKSBANK Slovensko, a. s.
5.	AXA Bank Europe
6.	Poštová banka, a.s.
7.	J&T BANKA, a. s.
8.	Wüstenrot stavebná sporiteľňa, a. s.

- **Products and Product types**

  Second part of the market analysis focused on listing all currently available banking products. Data from this report were then used for detailed analysis with the use of IBM Content Analytics.

  After consulting with official sources and financial servers all products were divided into several categories based on their type. These categories were:

  - Current accounts *(bežné účty)*,
  - Mortgages *(hypotéky)*,
  - Internet banking *(internetové bankovnictví)*,
  - Credit cards *(kreditní karty)*,
  - Savings accounts *(sporenie)* and
  - Consumer credits *(spotrební úvery)*.

[37] Online (or web) presence can be defined as *"a collective existence online of a company or individual, where website is one example."* (Cohn, 2010)

- **Sources for Analysis**

  Another important outcome of the General market analysis was a comprehensive list of web sources to be analyzed. Because one of the goals was to inspect in more detail interaction of individual banks with their customers on social networks, the list was to contain next to general sources like information and financial news servers also official banks' profiles and pages on various social networks.

  This part of the analysis turned out to be most problematic of the whole phase, because there weren't simply enough relevant and fresh *e-sources* updated on a regular basis. Whether it is because of mentioned relative small size of the Slovak market or simple lack of public's interest in financial sector, the fact is that quality Slovak financial news servers are rather scarce.

  Even sadder is the situation with social networks. If banks even try to communicate with their clients through social networks, they choose almost exclusively Facebook as their main channel and ignore other options such as Twitter or Google+. And even on Facebook is situation rather bad – from thirty banks on Slovak market only five[38] have functioning and active official Facebook profiles.

  The final list of used sources divided into categories was as follows:

  - General news servers
    - Ekonomika @ SME.sk
    - Ekonomika @ Aktuality.sk
    - O peniazoch @ Zoznam.sk
  - Specialized financial sites
    - Banky.sk
    - Slovenská banková asociácia
    - FINinfo.sk
  - Official Facebook profiles
    - ZUNO
    - Tatra Banka
    - mBank SK
    - UniCredit Bank SK
    - Poštová banka

[38] These are: *ZUNO, Tatra Banka, mBank SK, UniCredit Bank SK* and *Poštová banka*. We also included into analysis *Slovenská spořitelňa* although its profile is marginally less active then any of the "big five", but at least it is somewhat active.

- Slovenská spořitelňa

6.2.3 Goal definition phase

Outcomes of the Global market analysis served as a source data for basic understanding of the state of Slovak banking market. Based on this newly gained knowledge during the Goal definition phase it was possible to specify individual project goals.

Due to somewhat limited range of available sources suitable for detailed analysis via Content Analytics, it was necessary to adjust original goals and expectations accordingly. Current situation didn't allow comprehensive analysis of social networks and therefore it was decided to focus solely on Facebook, specifically on mentioned six banks with active official profiles.

After a thorough discussion, individual goals were set as follows:

- **Frequency analysis**
  What is the position of individual banks in examined news and media in general? How well are they known? How often are they mentioned? Which banks are often mentioned together and why?
- **Product portfolio analysis**
  What are the main products with the most interest from customers? What is a structure of average portfolio? What naming conventions can be found?
- **Customer analysis on Facebook**
  What is a structure of customers – their gender, age, education, etc.? Which are the key accounts *in bank's social network?*
- **Sentiment analysis on Facebook**
  What is the overall mood *in banking market? Which banks have positive / negative reactions to their products and why?*
- **Other findings**
  How active are banks in communication with customers? Are there any problems or questions unanswered? Are there any other interesting things worth pointing out?

After it was agreed on what the actual outcome of the team work would be, the first informal milestone was reached and the whole project moved to the next important part, where the actual IBM Content Analytics was finally involved.

6.2.4 Preparation phase

Preparation phase is one of the three *core* phases of the whole project. Purpose of this phase is to configure each of the main components of IBM Content Analytics software so it can properly crawl given web sources, parse gathered data into documents, index them and then store them in prepared collection (which also required configuration).

Because of fuzziness of the source data and several problems that emerged due to unsupported environment, this whole part of the project – including *preparation*, *data-mining* and *knowledge working* phase – required to be done in several iterations. After each iteration, gathered results were carefully examined and necessary steps and precautions were taken to eliminate all unwanted deviations and exceptions caused by wrong configuration of either crawler or parser. This is also a standard procedure recommended by official ICA RedBook.

Crucial part of this phase was also filling the ICA *dictionaries* and configuration of term *facets*.

As was said earlier, IBM Content Analytics is capable of thorough text analysis using various advanced features such as natural language processing. However these features are available only for supported languages like English. Unfortunately, Slovak is not amongst them. ICA not only does not *speak* Slovak, it also does not understand its grammar rules and sentence-building conventions.

One of the most difficult tasks during the Preparation phase therefore was to try and *teach ICA to understand Slovak*. This is in fact a very complex and lengthy process requiring additional programming. Because there was by no means enough time for such delicate operation, compensatory solution was chosen – ICA's current language dictionaries were extended with the key Slovak terms related to banking area and financial products and also with two sets of *sentiment dictionaries* one for negative and one for positive sentiment. New terms were then grouped into thematic units – facets – which were then used during the parsing and knowledge working phases.

This complex process was done thanks to cooperation of IBM experts, ICA dictionary specialist and native language consultant, who joined forces and were able to make ICA work even in unsupported environment such as Slovak banking market. And although by no near perfect, the results after a few iterations and additional configurations were more than a satisfactory for the purposes of the analysis and set goals.

6.2.5 Data-mining phase

Once project Text Analytic Collection was created and all main components properly set and configured including new dictionaries, project could move to Data-mining phase. During this automated process, ICA crawlers were to be harvesting data from given source web servers and Facebook sites. While to gather data from normal web page one needs nothing but a built-in web crawler, getting relevant information from Facebook requires more sophisticated means.

It was already mentioned that along with several built-in crawlers ICA supports custom made crawlers to be able to satisfy individual needs. In this case it was necessary to use custom Facebook crawler capable of calling Facebook web API. Because there was no such component available, it had to be created from scratch and configured specifically for the needs of this project.

Social data of Facebook users are made available through *Facebook Graph API* which uses *OAuth 2.0*[39] protocol for authentication and authorization. (Facebook, 2012) User data can be divided based on their accessibility (which is subject to individual user privacy setting) to following three groups:

- **Public data**

  Data visible outside the boundaries of Facebook. These data can be viewed and indexed by various search engines and bots without the need for authentication.
- **Facebook-public data**

  Data visible to anyone who is logged in to Facebook. To get this data externally outside of Facebook one must use Graph API and obtain a *user access token*.
- **Private data**[40]

  Private data accessible only with explicit permission of the owner. One must either authenticate as a requested user (i.e. log into his or her account) or as an application (obtain *app access token*) with granted permission to access this user's private data.

For purposes of the project there was no need to access private user data and so the custom crawler was set to work with just publicly available and Facebook-public data from official profile pages of individual banks.

[39] For overview of this protocol see the original draft at <http://tools.ietf.org/pdf/draft-ietf-oauth-v2-12.pdf>

[40] Some basic information such as name and gender cannot be made private and are always visible.

6.2.6 Knowledge working

After all source web servers and Facebook pages were crawled and obtained data parsed and indexed based on the dictionaries and term facets, project could advance to its most interesting phase – knowledge working.

Both beauty and curse of this phase is that there are no direct guidelines, no best practice or anything. Shared proverb amongst knowledge workers is that "*it is as if you are trying to find a black cat inside a black hole*".

Point of the knowledge working is to find some hidden added value in gathered data. And although computers and complex tools might be able to point out some facts and findings that might be useful, highlight some trends and deviations, that might be interesting; in the end it is up to the knowledge worker alone to use his wits and find in this mystery web full of dark corners and dead ends that hidden *treasure* whatever its form would be.

To be able to give solid statement about some fact, one should not limit one's view to just one dimension but should use multidimensional approach and examine source data from various angles and aspects.[41] Truly successful scenario does not suffice with simple *what*, but also tries to cover and explain the *why*. Meaning that although showing some trends and deviations in one's findings might be useful and interesting, the true power and purpose of knowledge working is to help discover *why* such deviations occurred and *why* are the numbers and connections in the findings the way they are. To point a finger at some interesting fact is just one part of knowledge working process, the true Competitive Intelligence comes with the explanation why.

After several iterations during which components were reconfigured; dictionaries were modified; facets reorganized; source data recrawled, reparsed and reindexed; results were finally satisfying enough to be proclaimed sufficient. Then all found patterns and scenarios were put together for final discussion and presentation.

6.2.7 Presentation phase

All successful scenarios found during knowledge working phase together with other interesting findings and facts which emerged throughout the whole project were put together and discussed at the Presentation phase. Each scenario was then subjected to a group analysis

[41] In ICA this can be achieved through different *View tabs* in Text Miner Application.

and evaluation. The most interesting scenarios and outcomes were then chosen to be presented life at the IBM Forum 2012 in Slovakia.

I have chosen some of the final successful scenarios and included them to this thesis. They can be found in Appendix at the end.

6.2.8 Closing phase

Purpose of the Closing phase was to officially end the whole project. During the final meeting each phase was subjected to a detailed discussion. Contribution and work effort of each team member were evaluated and commented by both team leader and co-workers.

As a final outcome of this phase a short report summarizing the findings, benefits and problems of the project together with a list of *best practice* were created to be archived as a *knowledge object* for future projects.

6.3 Findings and Conclusion

The whole project ended a success. We were able to fulfill given goals and prove both usability of social networks for Competitive Intelligence analysis and operability and adaptability of IBM Content Analytics in unsupported environment such as Slovakian banking market.

Although some phases lasted longer than originally planned and emerging issues and barriers due to unsupported language environment at first brought some unexpected results requiring further optimization and re-configuration using Administration Console, the project as a whole finished on time and brought both enough results and business scenarios to be presented at IBM Forum 2012 and new findings and experience in the field of competitive intelligence in general that could be used in future projects and works.

In the following paragraphs are summarized some of the most important and valuable experience that were gathered throughout the whole project.

6.3.1 Benefits

Along with the general success of the whole market analysis the project brought three key beneficial findings:

- **ICA works *everywhere***

  Due to ever changing internet environment with the new options emerging every day it was important to find out whether robust analytical tool such as ICA is flexible enough to adapt to these changes and work even in unsupported and unknown environment.

- **Custom crawler works**

  Another important finding was that custom Facebook crawler was working just fine with public data and that it has been successfully tested on real project. Although originally made to satisfy the needs of this sole project, its general usability rendered him demanded for other projects too and is currently one of the highly downloaded and shared custom components amongst IBM analysts worldwide.

- **Facebook is useful**

  Some of the successful scenarios showed usefulness of social networks (namely Facebook) for Competitive Intelligence analysis. And one does not have to dig for private user data. Even with publicly available data alone can a complex market analysis be done.

6.3.2 Problems

The project didn't avoid some problems and issues. Some were just a minor nuisance but some turned out to be rather serious setbacks we had to deal with.

- **Unsupported environment**

  Obvious problem was already mentioned unsupported environment for ICA. This not only added a lot of extra work during the preparation phase, but also denied us from using advanced language processing mechanisms and features, which significantly lowered quality of performed analysis and relevance of gained results.

- **Lack of relevant sources**

  One of the serious setbacks was a total lack of good relevant sources. This almost meant an end for the whole project, but after an emergency meeting it was decided to give it a go even with the very limited resources there were.

- **Language ambiguities**

  This is an everlasting problem in every language-related computer analysis. All those homonyms, irony, sarcasm and other *advanced language features* computers may never fully understand. However there emerged another specific problem during the sentiment analysis – how to evaluate terms such as *interest rate*? Is high interest rate positive of

negative? With *current accounts* it is certainly positive, but with e.g. *mortgages* it is definitely bad news.

Without advanced text processing features due to unsupported language this was one of the pains we had to deal with during the knowledge working.

- **Banks are not *social***

  Last but not least sad fact is that Slovak banks don't use social networks. Whether it is due to their lack of interest in this internet area, lack of knowledge on how to effectively interact with customers or whether they simply don't find these channels important – that is unknown.

7 Conclusion

This thesis successfully provided a general overview of a current state of Competitive Intelligence market and discussed new opportunities and possibilities along with new approaches to data management in relation to modern social networks.

To be able to discuss those new features, the whole competitive intelligence market area was first analyzed using Porter Model of Five Competitive Forces. The analysis was accompanied with a list of companies that can be considered current major players or even market leaders. Each company was given short characteristics along with the description of its main analytical product. And although the list of CI players is by no means comprehensive, it served its purpose in helping provide general overview of current state of competitive intelligence market and its main subjects and preparing the ground for following parts of the work.

The whole section was closed with discussion about the new areas of interest for CI in relation to modern social networks. Thesis offered a definition and description of a new Social Content Management business area and described several approaches to social (*soft*) data management.

The main part of the thesis then focused on a real life project which took place in early 2012 in cooperation with IBM. Through this detailed case study were described possibilities and limitations of using current professional analytical tool IBM Content Analytics for detailed social network analysis, frequency analysis and currently more and more popular *sentiment analysis*.

Final part of the practical section presented the key benefits and findings of the project[42] and addressed some real issues and problems the realization team encountered. Because of the absence of any related work in this area, the whole case study can be used in the future as a reference model or guidelines on how to approach the fuzzy area of social networks in relation to competitive intelligence and data analysis.

[42] Some of the findings being described in more detail in appendix.

Sources

Autonomy. 2005. AUTONOMY COMPLETES ACQUISITION OF VERITY, INC. *Autonomy.* [Online] 12 29, 2005. [Cited: 4 1, 2012.] http://www.autonomy.com/content/News/Releases/2005/1230.en.html.

—. 2011. HP Acquires Control of Autonomy Corporation plc. *Autonomy.* [Online] 10 3, 2011. [Cited: 1 1, 2012.] http://www.autonomy.com/content/News/Releases/2011/1003.en.html.

—. 2012. IDOL Functionality. *Autonomy.* [Online] 2012. [Cited: 2 12, 2012.] http://www.autonomy.com/content/Technology/idol-functionality-audio-and-speech/index.en.html.

Banky.sk. 2012. Zoznam bánk. *Banky.sk - bavme sa o peniazoch.* [Online] 2012. [Cited: 4 1, 2012.] http://banky.sk/10/zoznam-bnk-bankysk.php.

Berger, Arthur Asa. 2011. *Media Analysis Techniques.* s.l. : SAGE, 2011.

Brody, Roberta. 2008. Issues in Defining Competitive Intelligence: An Exploration. *Journal of Competitive Intelligence and Management, Volume 4, No. 3.* 2008, pp. 3-16. http://scip.cms-plus.com/files/JCIM/02.%20JCIM%204.3%20Brody%20%28WEB%29.pdf.

Business Week. 2007. SAP to buy Business Objects for $6.8B. *Business Week.* [Online] 10 7, 2007. [Cited: 4 1, 2012.] http://web.archive.org/web/20071012175940/http://businessweek.com/ap_working/financialnews/D8S4K2580.htm?chan=top+news_top+news+index_top+story.

Cartner, Brian. 2011. *The Like Economy: How Businesses Make Money With Facebook.* s.l. : Que Publishing, 2011.

Cheverton, Peter. 2002. *How Come You Can't Identify Your Key Customers?* s.l. : Kogan Page, Ltd., 2002.

Cohn, Michael. 2010. What is online presence? *CompuKol Connection.* [Online] 12 25, 2010. [Cited: 4 1, 2012.] http://www.compukol.com/blog/what-is-an-online-presence/.

Competitive Intelligence Foundation. 2008. *Journal of Competitive Intelligence and Management.* 2008.

Computing. 2006. Interview: What is meaning-based computing? *Computing Analysis.* [Online] 8 21, 2006. [Cited: 2 12, 2012.] http://www.computing.co.uk/ctg/analysis/1860116/interview-what-meaning-computing.

Coyne, Kevin P. and Subramaniam, Somu. 1996. 1996, The McKinsey Quarterly #4, pp. 14-25.

Daily Mail. 2012. Watch your wall: New Facebook attack has stolen passwords from 45,000 users - and could be spreading through infected links. *Daily Mail.* [Online] 1 6, 2012. [Cited: 4 15, 2012.] http://www.dailymail.co.uk/sciencetech/article-2083118/Facebook-hacked-Ramnit-worm-stolen-passwords-45-000-users.html.

Facebook. 2012. Authentication. *Facebook Developers.* [Online] 2012. [Cited: 4 1, 2012.] https://developers.facebook.com/docs/authentication/.

Feige, Tomas. 2011. *Cloud Databases - Different approaches to large data management.* Prague : University of Economics, 2011.

Fleisher, Craig S. and Bensoussan, Babette E. 2010. *Business and Competitive Analysis.* s.l. : P.Ed Heg USA, 2010.

Go, Alec. 2009. *Twitter Earth.* [Online] 2009. [Cited: 4 15, 2012.] http://www.twitter-earth.com/#havel.

Go, Alec, Bhayani, Richa and Huang, Lei. 2009. Sentiment140. *Twitter Sentiment Classification using Distant Supervision.* [Online] 2009. [Cited: 4 15, 2012.] http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf.

Google. 2010. Announcing Google Refine 2.0, a power tool for data wranglers. *Google Open Source Blog.* [Online] 11 10, 2010. [Cited: 2 10, 2012.] http://google-opensource.blogspot.com/2010/11/announcing-google-refine-20-power-tool.html.

—. 2010. Deeper understanding with Metaweb. *Official Google Blog.* [Online] 7 16, 2010. [Cited: 2 10, 2012.] http://googleblog.blogspot.com/2010/07/deeper-understanding-with-metaweb.html.

—. 2011. Google Refine. *Google Project Hosting.* [Online] 12 2011. [Cited: 2 10, 2012.] http://code.google.com/p/google-refine/.

—. 2011. Screencasts - Google Refine. *Google Project Hosting.* [Online] 7 11, 2011. [Cited: 2 10, 2012.] http://code.google.com/p/google-refine/wiki/Screencasts.

How Competitive Forces Shape Strategy. **Porter, Michael E. 1979.** 1979, Harvard Business Review.

Howlett, Dennis. 2011. Making sense of HP's Autonomy acquisition. *ZDNet.* [Online] 8 19, 2011. [Cited: 2 12, 2012.] http://www.zdnet.com/blog/howlett/making-sense-of-hps-autonomy-acquisition/3345.

IBM. 2005. DB2 - Search and Index API (SIAPI). *IBM Public Library.* [Online] 2005. [Cited: 3 1, 2012.] http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.db2.ii.of.doc/ap/iiyspsrchind.htm.

—. 2012. *Enterprise Social Network - Future or reality?* 2012.

—. 2012. IBM Content and Predictive Analytics for Healthcare. *IBM.* [Online] 3 16, 2012. [Cited: 4 1, 2012.] http://www-01.ibm.com/software/ecm/content-analytics/predictive/healthcare.html.

—. 2011. IBM LanguageWare. *IBM.* [Online] 2 4, 2011. [Cited: 4 1, 2012.] http://www-01.ibm.com/software/globalization/topics/languageware/.

—. 2004. *RedBook - IBM Content Analytics Version 2.2.* s.l. : IBM, 2004.

Juchelka, Václav. 2012. *Gathering data on the relationship of users to trademarks from social networks.* Brno : Masaryk University, 2012.

Karthik, Mani. 2010. How to measure Facebook Influence? *DailyBloggr.* [Online] 10 14, 2010. [Cited: 4 1, 2012.] http://www.dailybloggr.com/2010/10/how-to-measure-your-facebook-influence/.

Klout. 2008. Klout. *Understand Klout.* [Online] 2008. [Cited: 4 15, 2012.] http://klout.com/understand/privacy.

KM World. 2007. Business Objects gains Inxight. *KM World.* [Online] 7 12, 2007. [Cited: 4 1, 2012.] http://www.kmworld.com/Articles/News/News-Analysis/Business-Objects-gains-Inxight--36890.aspx.

Molnár, Zdeněk. 2009. *Competitive Intelligence.* Praha : Oeconomia, 2009.

Moore, Pam. 2011. Why I Deleted My Klout Profile. *Social Media Today.* [Online] 11 20, 2011. [Cited: 4 15, 2012.] http://socialmediatoday.com/node/389381.

Mzinga. 2011. Create a social business ecosystem. *Mzinga - OmniSocial.* [Online] 2011. [Cited: 4 15, 2012.] http://www.mzinga.com/software/omnisocial.asp.

Nonaka, Ikujiro and Takeuchi, Hirotaka. 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation.* 1995.

Prescott, John E. and Gibbons, Patrick T. 1993. *Global Perspectives on Competitive Intelligence.* s.l. : Society of Competitive Intelligence Professionals, 1993.

SAP. 2011. NETBASE AND SAP JOIN FORCES TO DELIVER SOCIAL MEDIA ANALYTICS TO GLOBAL ENTERPRISES. *SAP.* [Online] 12 12, 2011. [Cited: 12 29, 2011.] http://www.sap.com/corporate-en/press/newsroom/press-releases/press.epx?pressid=17994.

—. 2009. *SAP BusinessObjects Text Analysis.* 2009.

—. 2011. *SAP Text Analysis: Eyes for Text.* 2011.

Scalzi, John. 2011. Why Klout scores are possibly evil. *CNN Money.* [Online] 11 15, 2011. [Cited: 4 15, 2012.] http://money.cnn.com/2011/11/15/technology/klout_scores/index.htm.

SCIP. 2007. SCIP FAQ. *Strategic and Competitive Intelligence Professionals.* [Online] 6 12, 2007. [Cited: 4 15, 2012.] http://www.scip.org/resources/content.cfm?itemnumber=601&navItemNumber=533.

Social Media Research Foundation. 2012. NodeXL: Network Overview, Discovery and Exploration for Excel. *CodePlex: Open Source Community.* [Online] 2012. [Cited: 4 15, 2012.] http://nodexl.codeplex.com/.

Takeuchi, Hirotaka. 2006. *The New Dynamism of the Knowledge-Creating Company.* 2006.

Topsy. 2011. Topsy's Otter API. *Google Project Hosting.* [Online] 2011. [Cited: 2 10, 2012.] http://code.google.com/p/otterapi/.

TopsyLabs. 2011. TopsyLabs. *About TopsyLabs.* [Online] 2011. [Cited: 4 15, 2012.] http://topsylabs.com/company/about/.

Tovek. 2011. Tovek Server. *Tovek Products.* [Online] 2011. [Cited: 4 1, 2012.] http://www.tovek.com/upload/produkty/tovek/tovek-server/ts_info_en.pdf.

Unruly. 2012. Social Ad Effectiveness. *Unruly Media.* [Online] January 2012. [Cited: 3 30, 2012.] http://www.unrulymedia.com/SocialAdEffectiveness.

Wikipedia Contributors. 2006. Porter five forces analysis. *Wikipedia.* [Online] 7 1, 2006. [Cited: 12 30, 2011.] http://en.wikipedia.org/wiki/Porter_five_forces_analysis.

List of Content

List of Images

List of Figures

List of Tables

Appendix – ICA Scenarios

In this appendix are shown the most interesting business scenarios that represent the main outcome of the project described in the practical part of this thesis. Some of these scenarios were presented life at the IBM Forum 2012 in Slovakia.

Frequency analysis of SK Banks

Amongst the *basic* analysis belongs this frequency analysis of Slovak banks. Its purpose is to give reader a simple overview of the *online presence* of all banking subjects on Slovak financial market. As a sign of electronic presence in this case counts any form of electronic data mentioning the company's name – be it official press release, blog post, news article or just a Facebook comment.

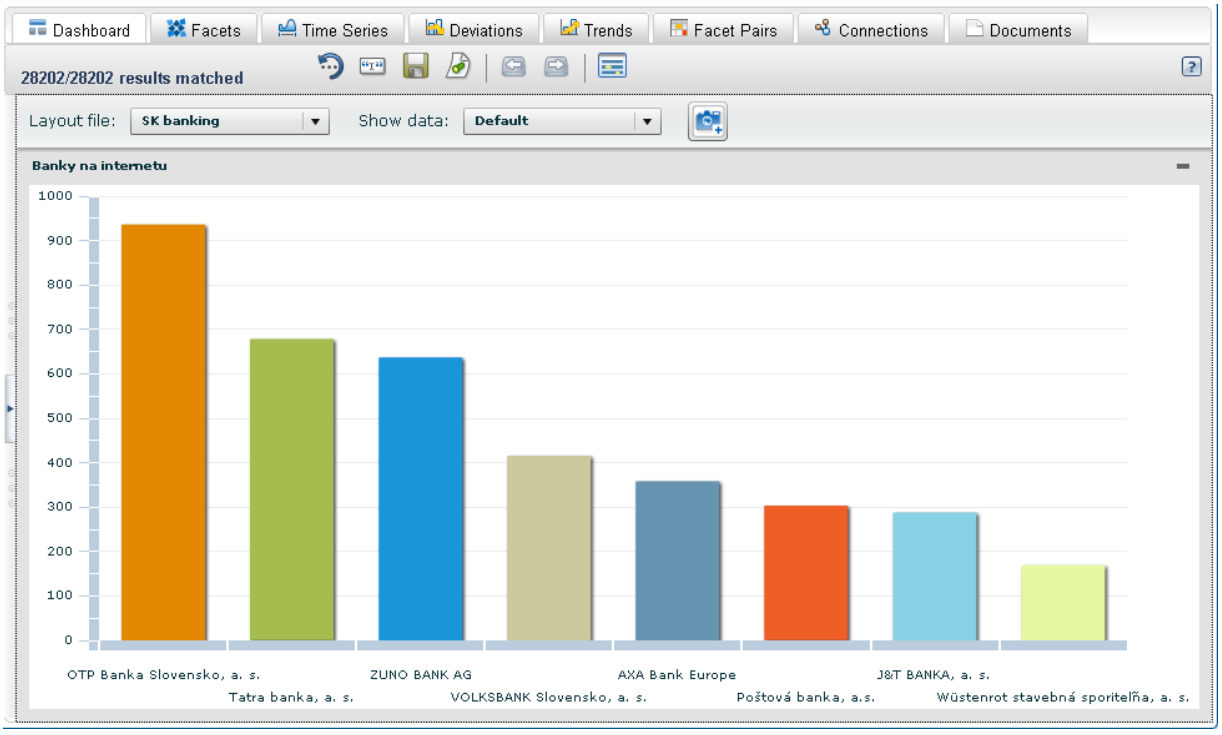


Image 17 - Frequency analysis of SK Banks (source: Author)

This kind of analysis can be broken further down according to form and source of parsed data (and e.g. focus only on presence of companies on facebook, etc.). Results shown in Image 17 represent the most general and simple form, which visualizes the total sum of occurrences of each individual bank no matter the source and form.

Frequency analysis of Product types

This scenario focuses on simple frequency analysis of product types in banking sector. One of the results in *Global market analysis* project phase was a list of main product types used in banking market sector. The product types were:

- Current accounts *(bežné účty)*,
- Mortgages *(hypotéky)*,
- Internet banking *(internetové bankovnictví)*,
- Credit cards *(kreditní karty)*,
- Savings accounts *(sporenie)*,
- Consumer credits *(spotrební úvery)*.

For proper business planning it is necessary to keep oneself informed about the frequency of interest in each individual product type currently on the market. From the Image 18 it is evident, that dominant banking product types are current and savings accounts and internet banking, while consumer credits and credit cards are minority.

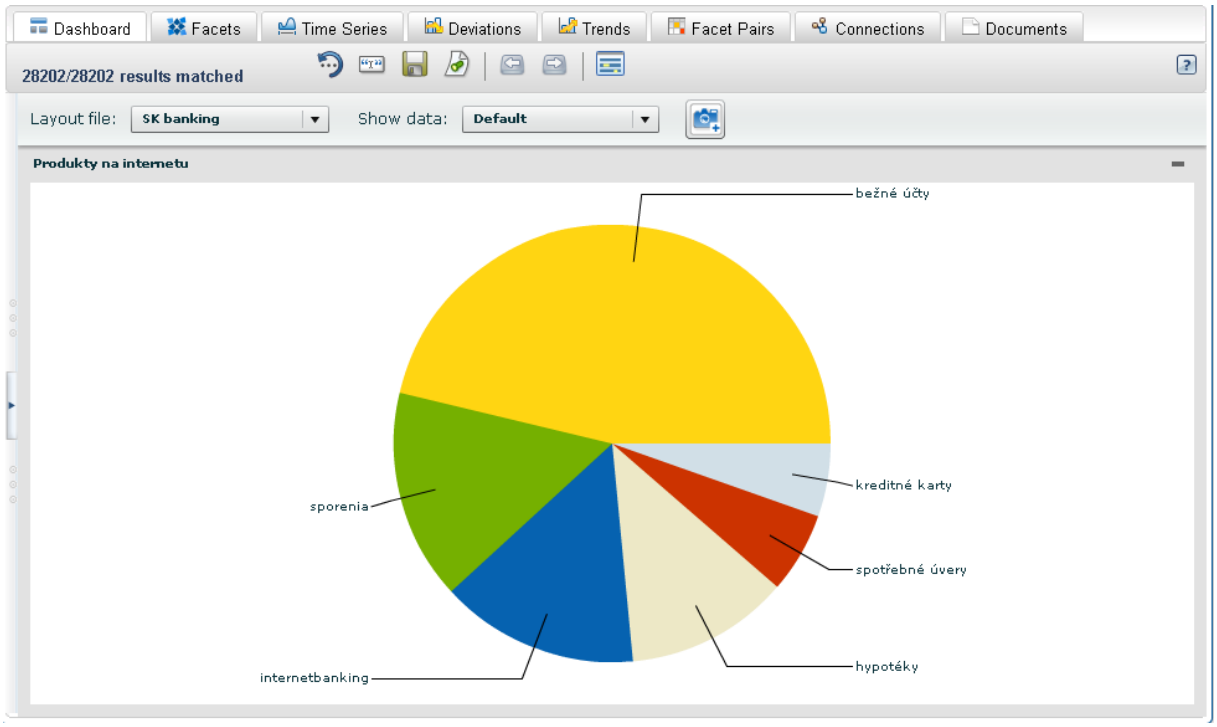


Image 18 - Frequency analysis of Product types (source: Author)

This knowledge helps to make better decisions on what form should company product portfolio be and on what product types should marketing section focus the most.

Communication channels of SK Banks

This scenario belongs to ones focused solely on Facebook. Its purpose is to compare different communication channels each bank uses to keep in touch with its customers (and *fans*) on Facebook. Resources for this kind of specialized analysis were somewhat limited due to the fact that only six Slovak banks have active Facebook profile (or *page*), but the results are still interesting enough to be considered and taken into account.

Rows = Názov stránky
Columns = Typ prispevkov
Correlation Amount: Low High

Subfacets/ Keywords	comment 1565	status 566	link 517	photo 125	video 33
ZUNO 1593	999 10.6	443 12.5	104 2.8	43 4.2	4 0.4
Tatra banka 366	192 8	19 1.4	128 15.5	16 5.3	11 11.9
mBank SK 359	166 6.9	45 4.3	143 17.9	4 0.5	1 0
UniCredit Bank SK 182	77 5.8	22 3.5	41 8.4	38 31.8	4 3.6
Poštová banka 101	85 4.1	9 1.9	41 15.2	7 4.8	9 32.2
Slovenská sporiteľňa 13	1 0	11 20.6	1 0	0 0	0 0

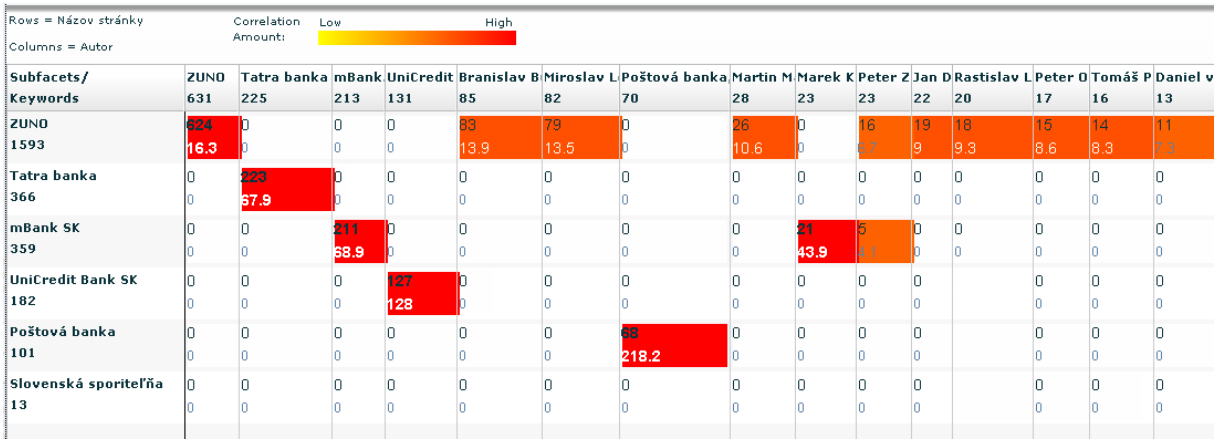
Image 19 - Communication channels of SK Banks (source: Author)

As can be seen in Image 19, except for *Slovenská spořitelna* all banks have active and well commented profile pages with clear dominance of *ZUNO* bank with almost a thousand comments. What is interesting are the different ways they communicate with their customers. While *ZUNO* posts text statuses, *Tatra banka* and *mBank SK* bet on link posting, *UniCredit Bank SK* on photos sharing and *Poštová banka* even communicates through videos. This shows different strategies of each subject and might bring additional interesting results if subjected to a deeper and more detailed analysis. That is however out of scope of this thesis.

Key People in Banks' network

The importance of Key Account Management was discussed in theoretical part of the thesis in chapter 5.2. This scenario shows some real life applications of discussed concepts. Although there is currently no simple way to easily evaluate influence level and therefore potential value of individual Facebook profile as can be done e.g. on Twitter via *Topsy.com*

web service, one is at least able to identify the most active customers in one's network and aim one's attention to them and their needs.



Subfacets/ Keywords	ZUNO 631	Tatra banka 225	mBank 213	UniCredit 131	Branislav B 85	Miroslav L 82	Poštová banka 70	Martin M 28	Marek K 23	Peter Z 23	Jan D 22	Rastislav L 20	Peter O 17	Tomáš P 16	Daniel v 13
ZUNO 1593	624 16.3	0 0	0 0	0 0	33 13.9	79 13.5	0 0	26 10.6	0 0	16 6.7	19 9	18 9.3	15 8.6	14 8.3	11 7.3
Tatra banka 366	0 0	223 67.9	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
mBank SK 359	0 0	0 0	211 68.9	0 0	0 0	0 0	0 0	0 0	21 43.9	5 4.1	0 0	0 0	0 0	0 0	0 0
UniCredit Bank SK 182	0 0	0 0	0 0	127 128	0 0	0 0	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0
Poštová banka 101	0 0	0 0	0 0	0 0	0 0	0 0	68 218.2	0 0	0 0	0 0	0 0		0 0	0 0	0 0
Slovenská sporiteľňa 13	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0

Image 20 - Key People in Banks' network (source: Author)

From Image 20 is apparent that (not surprisingly) each bank is actively posting to its facebook profile, but we can also see that there are several individuals not directly connected to the banks that are also actively posting on their profiles. Specifically on *mBank SK* profile these are *Marek K.*[43] and *Peter Z.*, on *ZUNO* these are *Branislav B.*, *Miroslav L.*, *Martin M.* and others.

When inspected in more detail both *Branislav B.* and *Miroslav L.* proved to be valuable customers and fans of ZUNO, because they both post either neutral or positive comments and reply to other people's problems[44]. This may not seem much of a deal, but people like this are cornerstones of each community or *fan base* and should be taken good care of. After all – they help build good company's name and they do it for free just because they believe in the brand.

Sentiment Analysis

Because of its *fuzzy* nature, sentiment analysis may be the most delicate and difficult scenario of the whole project. While other scenarios worked mainly with structured and clearly defined data and its metadata, during sentiment analysis one must dive into a sea of unstructured text and natural language processing filled with homonyms, sarcasm and irony.

Results of such analysis can be seen in Image 21 and Image 22 which show matrix distribution of Facebook comments, statuses, links and photos amongst individual Slovak banks with positive sentiment or negative sentiment respectively.

[43] All names were for purposes of this thesis anonymized, data from analysis however include direct links to each individual's Facebook profile.

[44] This detailed analysis is not present in the thesis due to its total length and high level of detail.

Rows = Názov stránky
Columns = Typ prispevkov
Correlation Amount: Low High

Subfacets/ Keywords	comment 126	status 82	link 32	photo 11	video 4
ZUNO 121	59 13.4	57 20.1	2 0.2	3 1.4	0 0
mBank SK 56	30 12.9	18 10.3	8 7.9	0 0	0 0
Tatra banka 29	17 12	0 0	9 18.7	1 0.2	2 5.9
UniCredit Bank SK 23	8 4.9	5 3	4 4.9	5 23.1	1 0.6
Poštová banka 13	6 4.9	0 0	7 25.9	0 0	0 0
Slovenská sporiteľňa 1	0 0	1 0.6	0 0	0 0	0 0

Image 21 - Sentiment Analysis - positive posts on FaceBook (source: Author)

Rows = Názov stránky
Columns = Typ prispevkov
Correlation Amount: Low High

Subfacets/ Keywords	comment 42	status 22	link 7	photo 1
ZUNO 42	23 30.4	17 39.7	2 1.9	0 0
mBank SK 13	6 12.2	4 10.6	3 17.9	0 0
UniCredit Bank SK 4	3 9.9	0 0	0 0	1 10.2
Tatra banka 3	2 4.5	0 0	1 2	0 0
Poštová banka 2	1 0.5	1 0.9	0 0	0 0

Image 22 - Sentiment Analysis - negative posts on FaceBook (source: Author)

Maybe the most interesting findings can be achieved by comparing not absolute values but relative shares of both positive and negative comments on each banks profile. Results of simple division of above values with total sums of comments from Image 19 are visible in Image 23.

As can be seen, banks keep an average share of negative comments at mere 2.742% with slight lead of *UniCredit Bank SK* (with 3.9%). This implies a surprisingly very peaceful and friendly atmosphere one wouldn't expect based on experience from various general internet discussions and forums. With the hypothesis of a friendly environment goes also a much higher share of positive comments with the average of 12.072% and the highest value at *mBank SK's* 18.07%.

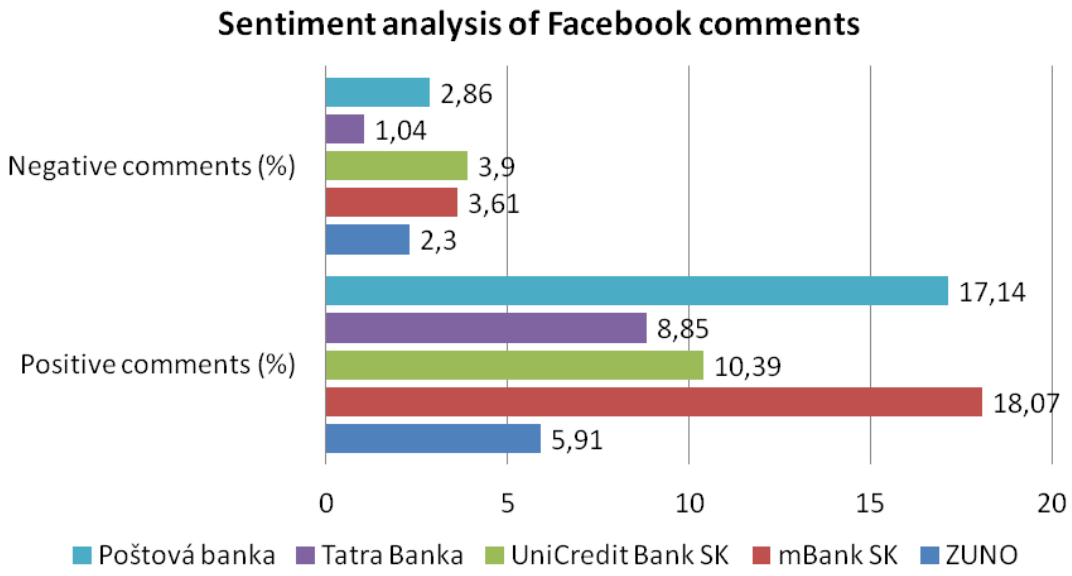


Image 23 - Sentiment Analysis - share of total FB comments (source: Author)

However, when evaluating such analysis one must be extra cautious not to get deluded by the results and not to come to wrong assumptions. This e.g. tells us nothing about state of the banking market and customers satisfaction, only that IF customers have problems with current products of Slovak banks, they don't complain about them on their official Facebook pages.