

Web Mining for Strategic Intelligence in South Africa

Lynnda Wagner & Jean-Paul Van Belle

lynndaw@mweb.co.za & jvbelle@commerce.uct.ac.za

Information Systems Department, University of Cape Town

Correspondence to:

Professor Jean-Paul Van Belle

Department of Information Systems

University of Cape Town

Private Bag

7701 Rondebosch

South Africa

Email: jvbelle@commerce.uct.ac.za

Phone: +27-21-6504256

Fax: +27-21-6502280

Web Mining for Strategic Intelligence in South Africa

Lynnda Wagner & Jean-Paul Van Belle

lynndaw@mweb.co.za & jybelle@commerce.uct.ac.za

Information Systems Department, University of Cape Town

Abstract

This paper explains how web mining can be used to gather strategic business intelligence and describes the various associated concepts. The results of an exploratory study on how web mining for strategic intelligence is actually used in South African organizations, is then presented. The research methodology is based on semi-structured interviews with competitive intelligence professionals. The analysis was done using a qualitative research approach and follows a thematic analysis methodology. The findings are discussed under the three major themes which were identified, namely general intelligence practices; Web-based intelligence; and skills development and education.

Introduction

Corporations are increasingly relying on Web information for their knowledge on products and services, market trends, companies, laws, etc. Collecting this type of information is gathering strategic intelligence. Since a primary concern of senior management is the organization's capability to align its internal operations with that of the changing external environment, the Web offers a rich repository of immediately accessible, prolific information. Mining the Web for intelligence can lead to the realization of competitive advantage, and should form an integral part of strategic planning.

The amount of information on the Web is staggering and growing exponentially. As such, it is not practical to manually cover the full extent of the Web. Although search engines are powerful tools, they are still incapable of providing tailored information, which means that users must still sift through copious search results before finding something within the right scope. Technologies are being developed to enable online data analysis with the ultimate goal being to gain insight from online data. Thus, Web mining has emerged as an active area of research and development.

The objective of this exploratory research is to survey how Web data is collected and assimilated into intelligence by South African business intelligence professionals. The analysis is by means of qualitative research methods.

Strategic Intelligence

Strategic management is primarily concerned with the long-term policy of a company, as opposed to day-to-day operations. Strategic intelligence is derived primarily from the theory and practice of national security and military intelligence. However, more recently, the terminology has been applied in a commercial sense.

Business intelligence (BI) supports the information needs of an organization's internal operations (day-to-day processes, research and development (R&D), material and labor efficiencies) and external circumstances (customer needs, threats from competition, supplier reliability, domestic and international regulation, technology). By contrast, *strategic intelligence* involves the collection of information about the external factors that have the potential to impact the business' mission and strategic direction. Since companies operate in a dynamic, sometimes volatile, business environment, executives rely far more on external than internal information. Therefore, the focus of this research is on strategic intelligence which addresses an organization's external environment.

Strategic analysis deals with the evaluation of strategic intelligence to form an overall picture of the factors influencing the organization's environment. The information is collected, disseminated, digested and incorporated into policy decisions. When conducting an external environmental scan, enterprises should look at two facets of the environment: societal environment and competitive, or

industry, environment. The societal environment relates to macro-level forces that have an indirect impact on the company whereas the competitive environment has a direct impact on the organization. The societal environment deals with broader forces that do not necessarily impact upon an organization's operations but greatly influence its future direction. These include economic trends, developments in technology, social pressures and changes in the government or political arena. The framework under the acronym of PEST – political, economic, social and technical – provides a good structure for a societal environmental scan.

Much of the information required for scanning the political and legal environment is available on the Internet. Government websites host a plethora of information on macro-level environmental forces such as public policy, business regulation, product safety, government contracts, census data about demographics and so forth. The technological environment can be assessed through an assortment of online publications. Business wires and press releases can be good sources of intelligence. Most companies have websites which can provide information about their products and services, mission and vision, prices, distribution, management, financial position, etc. Private companies without a Web presence can be researched through government sites that provide online access to public record information. Subscriber-based online forums such as Listservs and Mailing Lists offer yet more resources. One can learn about an industry and its key players through these roundtables. Similarly, newsgroups can draw interest groups within a niche market.

Because a strong focus of BI is on gathering competitive information (CI), the terms BI and CI are often used interchangeably and this practice will also be adopted in this paper, although it is hereby recognized that BI actually covers a larger domain.

Web Mining

Ineffectiveness of Standard Search Engines

When conducting research on the Internet, most people use standard search engines such as Google, Yahoo, Infoseek and so on. Unfortunately, Internet search engines are proving themselves insufficient as users must sift through profuse search results before finding something within the right scope. Even *meta-search engines*, such as MetaCrawler and Dogpile, which connects to multiple search engines and consolidates the search results from all do not solve these problems. Kosala & Blockheel (2000) highlight a range of problems associated with today's search engines when retrieving pertinent information from the Web. These problems include: low precision (due to irrelevant search results); low recall (due to the inability to index the whole Web); inability to create new knowledge from Internet data; lack of personalization of the information. The large portion of the Web that is not covered by search engines is referred to as the Hidden Web. Most of this refers to pages that are dynamically created from databases. (Adams 2001)

Key Concepts of Web Mining

Knowledge discovery is the concept of gleaning previously unknown information from data. Knowledge discovery in databases (*KDD*) is the method of gaining insight from data that is stored in an organization's databases. Web mining is essentially the application of KDD processes to the Internet. Therefore, Web mining is the automated process of finding, analyzing, retrieving and storing meaningful, applicable information from the Internet.

Research activities in Web content mining include information retrieval (IR), categorization and clustering of Web documents and information extraction (IE) from Web pages (Chen 2003). Once data has been extracted, the remaining process is text mining. These concepts are explained below.

Web crawlers, also referred to as spiders, agents, bots, ants, etc., are programs that retrieve information from Web resources. They are widely used by search engines to crawl through the Web and index Web pages. They can also be used to personalize searching. For example, they can be used to collect data for Web mining.

Since it is possible for search results to return hundreds, or even thousands, of pages, a *clustering* process can be used to organize the retrieved information into groupings. Unfortunately, effective

clustering is reliant upon the parameters and metrics used to evaluate the similarity among Web pages (Silvestri *et al.* 2004). Matters for consideration include: identification of relevant attributes; weighting; method selection and similarity measure; limitations on computational/memory resources; speed and reliability of retrieved results; ability to make changes in the database and selection of ranking algorithms (Kobayashi & Takeda 2000).

Information extraction is the process of pulling out data from retrieved documents. Its main objective is to extract data and transform free text into structured data, such as XML, to be stored in a database. Encoding Web data into database entries will allow better retrieval, organization and analysis of Web data (Adams 2001; Flesca *et al.* 2004). The computational linguistics community applied the same concept of discovering trends and patterns through statistical computation to *real text data mining* (Hearst 1999). By using text category assignments to find new patterns or trends and the discovery of new themes within those text collections, *real* trends may be discernable. Grimes (2004:31) describes text mining as a ‘cousin’ of data mining and as using statistical analysis to extract concepts, detect relationships, and classify unstructured documents into categories.

Web Intelligence

A new field of study has emerged around the Web. “*Web Intelligence (WI)* is a new direction for scientific research and development that explores the fundamental roles as well as practical impacts of Artificial Intelligence and advanced Information Technology on the next generation of Web-empowered products, systems, services, and activities,” (Zhong & Yao 2003:1). WI-related topics include: Web agents; Web mining and farming; Web information retrieval; Web knowledge management; the infrastructure for Web intelligent systems; and social network intelligence.

Web Intelligence must be linked to a process. The intelligence professionals that were interviewed consistently referred to the SCIP model (see Figure 1) based on (Herring 1999).

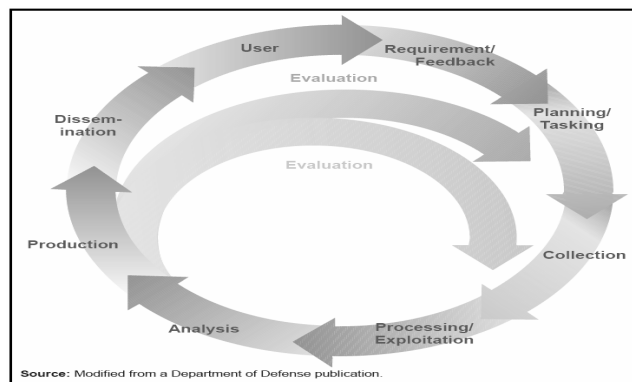


Figure 1: Intelligence Cycle

Strategic Intelligence and Web Mining

With the expansion of Internet usage, along with the growth in user population, more and more people are relying on Web information to increase or share their knowledge regarding companies, products and services, market trends, social interests, laws, etc. Therefore, the Internet has become an effective source of strategic intelligence. With its ease of access to information that transcends conventional boundaries and its revolutionary means of data acquisition and use, the Web can change the way businesses conduct an environmental scan (Tan *et al.* 1998).

Currently, Web-based technologies are at the forefront of IT. Information exchange on the Internet is rapidly changing the way environmental scanning for strategic intelligence is conducted. With the accelerated development of IT, companies must continually develop their IS capabilities in order to quickly assess and exploit opportunities as well as manage threats within the business environment. The challenge to build a system for scanning hordes of information on the Web for strategic information calls for Web mining.

Research Approach

This empirical study on the use of web mining for strategic information in South Africa is exploratory and qualitative of nature. After a literature survey had been conducted to determine how Web mining can be used *in principle* to collect intelligence, an assessment of how the Web is actually used for intelligence purposes in South Africa was then carried out by means of semi-structured interviews.

Research Design

A semi-structured, or focused, interviewing technique was chosen to facilitate open-ended discussions through organized questioning. A completely structured approach can cause rich data to be overlooked. Open-ended questions that probe, clarify and draw out details can get a participant to expound upon his/her experience. This is key, in order to gain an understanding of intelligence gathering practices in South African companies.

A semi-structured interview questionnaire (not included here due to space restrictions but available from the authors) was used as a guideline and consisted of a combination of structured, Likert scale and open-ended questions. A few of the questions were derived from a study on CI practices in South Africa conducted by Viviers *et al.* (2002) which, itself, was based on a Canadian survey.

Interviewees were inter-dispersed throughout South Africa. Therefore, it was more practical to conduct telephone interviews since they are less expensive and less time-consuming than face-to-face interviews. Plus, telephone interviews are generally conducive to studies with a small set of questions and a short timeframe.

Sample Frame and Data Collection

As CI in South Africa is in its infancy (Calof & Viviers 2001:63), only a small number of companies can be studied. For the focused interviews, the target audience was South African companies that have a CI discipline. The rationale behind this was to find individuals who already have some foundational knowledge of the intelligence process. Since associations already exist around the field of CI, it was sensible to target members of CI circles or participants in CI workshops. Therefore, the sample frame included members of the South African chapter of the Society of Competitive Intelligence Professionals (SCIP) and attendees of the Competitive Intelligence for Exporters Workshop (13 November 2003).

An e-mail was sent to each targeted respondent introducing the scope of this study and inviting that person to participate in a telephonic interview. At the end, there were 13 interview participants out of a total of 36 targeted. The interviews ranged from 15 minutes to 1 ½ hours in duration.

Five of the interview participants were CI consultants who compile intelligence for clients as a core business. It was very constructive to speak to these field experts. Eight of the respondents, including the five mentioned, were full-time intelligence employees. The participants operated in a multitude of industries ranging from the wine industry to human capital management. Although a majority of the interviewees worked for a medium-sized company, small and large companies were also represented. On the whole, the sample frame yielded quality respondents from a broad base of industries.

Furthermore, an ideal strategic intelligence methodology is one that incorporates a comprehensive intelligence cycle. As previously mentioned, data analysis must still be conducted. Results of the analysis must then be assimilated into a report and distributed to the relevant parties to be used, or considered, in strategy formulation. Feedback from strategic managers can then lead to the next cycle.

Data Analysis

Thematic Analysis

A standard approach for analyzing qualitative data obtained from interviews is a thematic analysis. Accordingly, a thematic analysis was used to extract meaning from the semi-structured interviews. In effect, the thematic process described by Ganesh *et al.* (2003) was employed to garner themes from

the interview data. As such, the basic structure of the questionnaire was used to draw up categories for the collected data. These categories were intended to organize the data for comparisons. The raw data was summarized via paraphrases and assigned to the categories. For each participant's data, major and minor themes were noted. Concepts with dominating patterns were distinguished as the major themes. Thereafter, themes from the sub-samples were compared across the entire sample to identify the common themes. The identified themes were coded. The thematic codes were then assigned across the sub-samples to identify all manifest themes. Finally, the data collection was then sorted by category.

The major themes can be classified into three major areas: general intelligence practices; Web-based intelligence; and skills development and education.

General Intelligence Practices

Intelligence Process

All of the participants have some mechanism for collecting intelligence. Seven out of the 13 respondents have a systematized intelligence process as a whole. These individuals follow the SCIP model of planning, collection, analysis and communication of intelligence. A CI consultant described the process used for all their clients as, "The firm's competitive environment is systematically collected, analyzed and disseminated to policymakers who can act on it." Although the techniques may vary, the process remains the same. One of the CI consultants explained that the methodologies used depend on the clients' needs but they are conducted in the "standard SCIP way".

The remaining six participants practice intelligence in an ad hoc way. Although an intelligence policy is non-existent, they do make an effort to gather intelligence as it is perceived to be important in their businesses.

Intelligence Sources

Secondary sources provide a majority of the information and can be in hardcopy or electronic formats. These sources include industry/market authorities such as trade magazines, industry reports, publications by professional associations, research reports and so on. Fee-based commercial databases such as LexisNexis, Factiva and Inetbridge are also used to search for industry- or company-specific information. Government publications and databases were also mentioned as being used for the collection of intelligence. Information is being gathered from government agencies such as DTI, Registrar of Companies, ITC, Justice Bureau, CSS, WTO, UN, etc.

Sources are generally documented especially by those with a structured intelligence process. Either a library of hardcopies is kept or electronic copies are stored in a PC. Some of the interviewees even have a more sophisticated system such as an in-house database and an indexing system.

Data Analysis

Among interviewees that follow a structured intelligence cycle, a qualitative data analysis is generally conducted. Various qualitative analysis models are applied depending on customer needs. Techniques that were mentioned include industry analysis, market analysis, Porter's Five Forces model, competitor analysis, risk analysis, customer profiling, management profiling, patent analysis, experience curve analysis, growth strategy analysis and stakeholder analysis.

Qualitative data analysis is conducted manually. Statistical or data mining software can be used to conduct quantitative data analysis as an adjunct to the qualitative analysis. However, qualitative data analysis is the primary method used for analyzing data. This is performed manually since computer programs are not sophisticated enough to carry out this task. One of the consultants elaborates:

"Computers still cannot answer the question 'so what?' That is an intellectual exercise. Computers cannot think and AI is too expensive. IT can be used for collection, collation and organizing data. Analysis is putting the puzzles together. A computer cannot put all the pieces of multiple puzzles together."

One consultant mentioned that available BI programs are not well-suited for analyzing external data:

“BI tools are not really great for external information but they are good for internal information. For external data, you need to apply your mind to it and it cannot be done with a machine. BI tools cannot come up with conclusions.”

Web-based Intelligence

Web Usage

Collectively, the Web is used by all of the respondents in the intelligence gathering process. They all use search engines to retrieve data from the Web. While some used basic search functions by entering rudimentary queries and clicking on the resultant links, others used advanced search terminology to achieve more precise results. Using semantically fine-tuned search queries was a sub-theme of advanced searches. One consultant pointed out that business libraries are checked for industry-specific lexicons which are then used in query semantics. Although the same types of search engines were used, these advanced searches yield much more relevant search results. Another CI consultant stated that their aim is to retrieve no more than 20 documents at a time through focused queries.

With exception to one interviewee, the Web is checked regularly, often on a daily basis. This is in order to stay current on events. For example, news alerts are used to ensure that the intelligence worker is notified of up-to-date information. Another reason for daily usage is to conduct searches on different facets of the external environment. A third reason is that intelligence professionals conduct regular searches for different clients.

Opinion of the Web as a Source

An overwhelming majority of the interviewees were of the opinion that the Web is a good source of global information. Some of the respondents use the Web quite extensively. One person claims that 90% of their company intelligence is obtained from the Web and likened the Web to a big brother who “is always there when you need him and who knows everything.” A consultant’s statement about the Web as a source of information was, “It’s an effective resource. There’s added value from the Internet.” One analyst used to believe that market research always required hands-on experience. Now, he has realized the amount and type of information that can be found on the Web. His comment about the Web was, “It is an invaluable source with an awful lot of information!” Others maintain that the Web is a good source but the information available is limited. One respondent said, “It’s excellent for first-level information but poor for finer details.” He cites that company websites, as an example, only provide general information.

A sub-theme that emerged is that the Internet is underutilized in South Africa. This point was conveyed by the comment:

“South Africa lacks information on the Internet. [...] We don’t use the Internet to the full extent. There is very limited or older information on the Web compared to other international players. The Internet should be leveraged better.”

Consequently, the respondents use multiple sources to supplement Web information. For example, proprietary databases are consulted for details on South African entities.

Web Data Validation

Since data quality is important in strategic intelligence, the need for Web data validation emerged. One consultant stressed that, even though Web data is usually found to be valid, it still needs to be tested. This was reiterated by one senior analyst who said, “It [the Web] is a good source but you must validate the information.”

Information retrieved from the Web is typically validated through cross-referencing against reliable sources. Reliable sources include specialized databases, personal contacts and other niche publications. Another identified technique for testing data validity was a critical evaluation of the document source or author. The reputation of the source publishing the information is evaluated. The author’s reputation can be another factor. One consultant explained that if the author is known from past experience or historical data, then the author’s interest in the subject matter is evaluated. The consultant explained that more confidence is held in an article that covers an author’s personal interests rather than his/her company’s interests.

Systems and Tools

Various systems are used in the intelligence process. Some of the respondents have an in-house system for storing and organizing extracted information. An analyst shared that his company has a full-fledged in-house system to support the intelligence process. A few have an actual data warehouse with years of compiled history. The data warehouse is then mined and used in conjunction with current information. While others have basic electronic files of reference lists of URLs that have been organized into different industry categories.

Some software packages are used to organize information. Two respondents spoke positively of Brimstone. One consultant found Brimstone to be the “best product for CI”. He explained that Brimstone has two components: a CRM component and a capability of supporting intelligence. Brimstone allows a user to search the Web using search engines, extract the Web pages and organize the documents. However, there is no package that can effectively analyze data. For this reason, data analysis is conducted manually.

Skills Development and Education

Advanced Web searching is more effective and efficient. One respondent shared that the team of analysts at his company make a point of retrieving no more than 20 Web pages. This requires highly fine-tuned search queries. Another individual asserts that an experienced analyst does not need exhaustive Web data since he/she will know how to retrieve just the right amount of information to get to an answer. Obtaining precise search results is predicated upon skilled searching.

Proficient searching stems from the searcher’s knowledge of the appropriate terminology to use, sites to explore and Web search tools. Some of the companies have addressed the importance of skills development by developing their own training program or sponsoring in-house training by intelligence experts. Some of the respondents recognize the value of skilled searching and have expressed a desire to learn better searching techniques and about search tools that are available.

Education empowers intelligence workers with knowledge. Those who have a structured intelligence approach regularly attend conferences, workshops and courses on CI, knowledge management, Internet searching, etc. One consultant is doing a PhD on the skills required for effective CI.

Knowledge also delivers better intelligence products. Intelligence gatherers should be mindful of the intelligence process because it is how data is collated and presented that transforms it into intelligence. This was supported by a consultant who said that, “Internet information is not intelligence. It’s what is done with the information that makes it intelligence.”

Minor Themes

Some of the minor themes are worth mentioning.

Need for an Ethical Intelligence Practices

A few participants have highlighted the need for an ethical means of collecting data. Investigations must be done in a legal and proper manner. There is a lawful and ethical way of deploying an intelligence strategy. The Society for Competitive Intelligence Professionals has a code of ethics that can be used in formulating an intelligence policy. One of the consultants emphasized the need for an awareness of ethical intelligence practices. He said that he first educates his clients on what CI is and is not. He states that awareness must be created because some people have the wrong idea about intelligence. He explained that intelligence must be collected in a legal manner. For example, privacy rights must not be infringed. He further clarified that CI is not industrial espionage. Publicly available information can be used effectively for intelligence.

Counterintelligence

Another minor theme of interest that emerged had to do with counterintelligence. Counterintelligence is essentially defensive intelligence. This includes counter measures taken when intelligence surveillance is detected. For example, one of the respondents expressed a need to discern between *real* intelligence from measures taken by the competitor to mislead him. A proactive counterintelligence approach can be taken as well. One consultant mentioned that he and his

intelligence team ensure that they cover their virtual track as a counterintelligence tactic. This makes sure that their intelligence gathering trail cannot be traced back to them.

Conclusions and Implications

The study on intelligence collection in South Africa revealed that Web mining can be used to assist business decision-making. This is advocated further by the fact that professionals who operate an intelligence consulting company use the Web to retrieve strategic information for their clients. However, it is important to distinguish between Web data that is free and data that is available through online information services. Although the two are publicly available on the Web, a subscription fee is required to access the latter. If the Internet is underutilized in a given region, such as the case in South Africa, then limiting data extraction to free Web data will negatively impact the quality of the Web mining system. It is recommended that data from proprietary databases are incorporated, especially when there is limited information on the standard Web.

One of the main implications of the research is the opportunity which presents itself for small and medium-sized organizations. Given the amount of strategic information that can be collected from the Web and the availability of tools that facilitate the Web mining process, small to medium-sized organizations can leverage Web data to gain a competitive advantage. By using a Web mining system to retrieve, extract and process Web data, companies can conduct their own environmental analysis without having to spend a lot of money. The cost of market analysis publications or consultation can be circumvented. Obviously, if an organization is price-sensitive then the Web mining tools should be selected based on budgetary restraints.

References

- Adams, K.C. (2001). "The Web as a database: new extraction technologies & content management", *Online*, 25 (2), 27-32.
- Calof, J.L. & Viviers, W. (2001). "Adding competitive intelligence to South Africa's knowledge management mix", *Africa Insight*, 31 (2), 61-67.
- Chen, H. (2003). "Introduction to the JASIST special topic section on Web retrieval and mining: A Machine Learning Perspective", *Journal of the American Society for Information Science and Technology*, 54 (7), 621-624.
- Flesca, S., Manco, G., Masciari, E., Rende, E. & Tagarelli, A. (2004). "Web wrapper induction: a brief survey", *AI Communications*, 17 (2), 57-61.
- Ganesh, U.; Miree, C. & Prescott, J. (2003). "Competitive intelligence field research: moving the field forward by setting a research agenda", *Competitive Intelligence & Management*, 1(1), 1-12.
- Grimes, S. (2004). "Consumer and enterprise search: not an exact match", *Intelligent Enterprise*, 7 (9), 22-31.
- Hearst, M.A. (1999). "Untangling text data mining." *Proceedings of ACL '99: 37th annual meeting of the association for computational linguistics*, University of Maryland, June 20-26, pp. 1-8.
- Herring, J.P. (1999). "Key Intelligence Topics: A Procedure to Identify and Define Intelligence Needs". *Competitive Intelligence Review*, 10 (2), 4-14.
- Kobayashi, M. & Takeda, K. (2000). "Information retrieval on the Web." *ACM Computing Surveys*, 32 (2), 144-173.
- Kosala, R. & Blockheer, H. (2000). "Web mining research: a survey." *Proceedings of SIGKDD*, 2 (1), 1-15.
- Silvestri, F., Perego, R. & Orlando, S. (2004). "Assigning document identifiers to enhance compressibility of Web search engines indexes", *Proceedings of SAC '04*, March 14-17, 2004, Nicosia, pp. 600-605.
- Tan, S.S.L., Teo, H., Tan, B.C.Y. *et al.* (1998), "Environmental scanning on the Internet", *Proceedings of the International Conference on Information Systems*, Helsinki, 76-87.
- Viviers, W., Saayman A., Muller, M-L. & Calof, J. (2002). "Competitive intelligence practices: A South African study", *South African Journal of Business Management*, 33 (3), 27-37.
- Zhong, N., Liu, J. & Yao, Y. (2003). *Web intelligence*, Berlin: Springer.