2007

Web Mining for Strategic Intelligence: South African Experiences and a Practical Methodology

Lynnda Wagner
University of Cape Town, Rondebosch

Jean-Paul Van Belle
University of Cape Town, Rondebosch, jvbelle@commerce.uct.ac.za

Web Mining for Strategic Intelligence: South African Experiences and a Practical Methodology

Lynnda Wagner & Jean-Paul Van Belle[1]
Department of Information Systems, University of Cape Town, Rondebosch, ZA-7701, South Africa
[1]jvbelle@commerce.uct.ac.za

Abstract. This paper aims to provide a practical methodology for using web mining to gather strategic business intelligence. It first describes some concepts relating to web mining and reports some exploratory findings on how web mining for strategic intelligence is actually used in South African organizations. The paper then suggests some practical steps for organizations wishing to establish their own web mining procedures.

Keywords: Web Mining; Strategic Intelligence; Competitive Intelligence; SCIP Process; South Africa.

1 Introduction

In a global and rapidly changing global business environment, it is imperative for companies to systematically conduct local and global environmental scans to equip executive management with the necessary knowledge to plan, lead, organize and control their business. Ideally, organizations employ mechanisms for efficient strategic intelligence collection, which empowers management to forecast and influence the companies' future direction. This is especially important for smaller businesses in emerging countries which increasingly have to compete globally but are faced with infrastructural limitations such as high bandwidth costs and more difficult access to information about global technology and market trends.

The Web is an excellent source of current, insightful and easily accessible information. Indeed, 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. Enterprises now have the opportunity to devise a web mining system to rapidly assess its external environment. This will enable companies to be aware of trends and/or changes outside of their operations that can negatively or positively impact their business. This empowers businesses to become proactive towards emerging market issues. Also, it allows management to detect opportunities and threats, and evaluate how to adapt their operations accordingly. Perhaps most of all, it facilitates decision-making with regards to the future direction of the company. Overall, the information gained can be used to formulate corporate strategies, the ultimate goal of strategic management.

Unfortunately, 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 suggest a practical way to implement a process for supplying strategic managers with effective strategic intelligence.

2 Competitive and Strategic Intelligence

Currently, there is a lot of interest in and around *competitive intelligence* (CI), which forms part of the function known as business intelligence. As defined by [1:109], CI is public and private information regarding competitors' present and future activities as well as the behaviors of suppliers, customers, technologies, acquisitions, markets and the general industry environment. *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). Although BI covers a larger area than CI, a strong focus of BI is on gathering competitive information and thus the terms BI and CI are often used interchangeably in practice.

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. Environmental scanning enables a business to determine trends in the macro-environment that are important to the company, detect potential threats and opportunities or changes for the business as a result of those trends, be future-thinking and alert management to converging trends. Generally, when organizations conduct an environmental scan, they look at two facets of the environment: the societal environment and the 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.

Thus, the Internet has become one of the top sources of intelligence for CI professionals. For example, the Internet Intelligence Index (www.fuld.com/i3/index.html), compiled by Fuld & Company [2], is an index of about 600 intelligence-related websites. It is divided into three areas: 1) general business Internet resources; 2) industry-specific Internet resources; and 3) international Internet resources.

3 Web Mining

Web mining can be classified into three areas: content mining, structure mining and usage mining. Web content mining relates to discovering meaningful information from the contents of web documents. Web structure mining deals with the categorization of web pages based on the link structures. And web usage mining draws conclusions about how users interact with Web sites [3]. This paper focuses on the use of web content mining in gathering strategic intelligence.

Generally, internet search engines are inadequate for web content mining. Common problems include: low precision (due to irrelevant search results), low recall (due to the inability to index the hidden web), inability to create new knowledge from Internet data and lack of personalization of the information [3]. The hidden web refers to a significant portion of the web not indexed by engine spiders – typically pages that are dynamically created from databases [4]. Even the use of meta-search engines does not remedy most of the above problems.

The following explains some of the key concepts relating to web content 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. Web content mining activities include information

retrieval (IR), categorization and clustering of Web documents and information extraction (IE) from Web pages [5].

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 e.g. to collect data for Web mining [6]. During the crawling process, there are two ways that Web content can be analyzed: content-based and link-based [7]. In the content-based approach, the body text of the web page is scrutinized to determine whether it is relevant to the search criteria. Link analysis uses the hyperlinks and the anchor text describing those links to evaluate page content. Link analysis is widely used and is the basis for the most effective page ranking algorithm [8].

Because search results usually return thousands of pages, a clustering process is normally used to organize and group the retrieved information. Unfortunately, effective clustering is reliant upon the parameters and metrics used to evaluate the similarity among Web pages [8]. 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 [9].

Finally, 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, usually in XML format, to be stored in a database. Encoding Web data into database entries will allow better retrieval, organization and analysis of Web data [4; 10]. The computational linguistics community applied the same concept of discovering trends and patterns through statistical computation to real text data mining [11]. 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. [12:31] 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 (WI) is a new sub-discipline that explores the fundamental roles and practical impacts of Artificial Intelligence and advanced Information Technology on the next generation of Web-empowered products, systems, services, and activities," [13:1]. 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. Some of the techniques proposed in the literature include MDR - "Mining Data Records in Web pages [14], superpage classification and pruning [15] and a number of approaches related to the semantic web, including IBM's WebFountain project [16].

The challenge to build a system for scanning hordes of information on the Web for strategic information calls for a practical web mining methodology.

4 CI in South Africa: Some Findings

This section reports some crucial findings from a small-scale exploratory survey conducted among CI practitioners in South Africa. A semi-structured interview with

open-ended questions was conducted by telephone due to the dispersed nature of the sample. The sampling frame included members of the South African chapter of the Society of Competitive Intelligence Professionals (SCIP). Because CI in South Africa is in its infancy [17:63], only a small sample of thirteen individuals could be obtained, but 8 of these were full-time intelligence employees or CI consultants. The interviews ranged from 15 minutes to 1 ½ hours in duration. The following highlight some of the more prominent findings, the more detailed results are reported in [18].

4.1 Web Use

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.

4.2 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 following 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.

4.3 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.

4.4 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.

4.5 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."

5 A Practical Web Mining Process

This section proposes some practical steps which a smaller but global-thinking organization can undertake to create their own web mining process or methodology. These steps – loosely organized in a sequential methodology – can and indeed should be customized and/or supplemented by other activities relevant to the experience of the organization. It is informed by and maps directly onto the initial phases of the proprietary SCIP Intelligence Cycle which is used by most of the Intelligence Officers which were surveyed.

5.1 Building a Company-Specific Taxonomy

As a starting point for Web mining, search queries must be fine-tuned in order to help alleviate low recall precision. A key strategy for improving precision in IR is to create a company-specific taxonomy. This should include all possible terminology and relevant concepts used in the industry, sector, niche market and company. Not only is it industry-specific, it is also company-specific.

There are online information services that provide industry-specific taxonomies. These fee-based services can be consulted when constructing the company-specific taxonomy. An example of this is the Taxonomy Warehouse (http://www.taxonomywarehouse.com/). It can be beneficial to use an already-compiled domain-specific taxonomy to save time and effort. However, company-specific terms, such as the list of actual competitors, major customers and suppliers, etc., must still be added.

In addition, inputs from the PEST (political, economic, sociological and technological) and Porter's Five Forces model (new entrants, buyers, suppliers,

rivalry and substitutes) should also be included in the company-specific taxonomy. A list of terms within each contributory factor must be compiled. For example, when considering the technological environment within a PEST analysis, a list of specific technologies, related components or tools, companies heavily involved in R&D within the technology sector, etc., must be generated first. This is done for each of the four PEST components as well as Porter's five competitive factors. Relevant website addresses should also be included. The accumulated lexicon will be the basis of the Web search parameters for each aspect of the environmental analysis.

5.2 Meta-searching

Meta-search engines combine the results from multiple search engines. This allows for greater coverage of the Web since limitations of one search engine may be accommodated by another search engine. Furthermore, meta-search engines incorporate algorithms to eliminate duplicate Web pages. It makes economical sense to exploit this technology as a part of a Web mining system.

Terms from the company-specific taxonomy will be used in the meta-search engine query fields. The meta-searcher will retrieve URLs that relate to the query terms. This can range from customers to niche-market product categories. This is solely dependent on what terms are queried. When researching a particular area (i.e. technology), the terms from the related section of the PEST and/or Porter's Competitive Forces taxonomy in addition to the general company-specific terms will be used. It is important to set the meta-searcher to retrieve the maximum number of pages from each search engine in order to have a large base of Web data to be processed further.

Another search strategy is to use reverse link analysis. This type of search is possible in most meta-search engines, as well as standalone search engines, by typing "link:" followed by the URL. A reverse link analysis retrieves the URLs of pages with an in-link, or hyperlink, to a specific website (i.e. competitor or customer site). This type of search can reveal some interesting information. For example, a potential customer's website which cites a competitor's website may disclose some form integrated supply chain arrangement. An alternative use of reverse link analysis is merely to ensure inclusion in public Web directories, or wherever else, the competitor is listed.

Meta-search results require a lot of fine-tuning in order to establish relevancy within an intelligence strategy. "Enterprise search tools must negotiate access to content stores, apply techniques that classify and organize results, and embed the search function behind the scenes in operational applications," [12:24]. Essentially, the meta-search engine is the 'content store' and mining techniques will be applied to organize and process the data in order to transform it into intelligence.

5.3 Crawling Web Pages

Running personal spiders on a client machine allows more CPU time and memory to be allocated to the search process and provides more options [7]. Most search engines

re-crawl based on an average of once a month and, therefore, possibly retain 'dead pages' at any given time. An intelligence analyst needs more up-to-date data. For this reason, a personal crawler that can refresh data without having to rely on search engine updates is more appropriate.

Personalized Web crawling can also customize search further and improve Web coverage. Since meta-search engines automatically rank pages based on the PageRank algorithm, certain Web pages may be omitted because they were not within the top ranked pages. A company's ranking order will not always be identical to the search engines' since the relevancy of Web pages can be subjective. In order to circumvent the possibility of omissions, a personalized Web crawling process is ideal.

The Web crawler is fed with all the URLs obtained from the meta-search results. Likewise, the website addresses listed in the company-specific taxonomy (i.e. customer sites, competitor sites) are also fed into the crawler. The user must designate how many levels of linked Web pages the crawler must investigate. This will be dependent on how comprehensive the company-specific taxonomy is and how deep the desired search must be. Cache memory, processing speed and storage capacity are other determinants. The user must use discretion in deciding the level of scope (how wide) and depth of the crawl. If the objective is to obtain an overview of the whole competitive environment, the scope is wide but may only be two levels of link analysis deep. On the other hand, in-depth competitor profiling would require a very narrow search scope and very deep link analysis.

Many commercial spidering and crawling tools are available. In addition, it is relatively easily to build customized tools for specific needs or purposes [6].

5.4 Searching Online Proprietary Databases

It is necessary to supplement Web data with data from other credible sources such as commercial, governmental or private databases. This will augment Web data in case limited information is available on the Web. However, another reason for doing this is to validate Web data by cross-referencing it against reliable sources. Therefore, data extracted from proprietary databases must be clustered and summarized separately from the bulk of the retrieved data. This will likely result in smaller data sets used merely to test the reliability of the retrieved data. These standalone data sets will be compared to the results of Web clustering. Once data validation is satisfied, these data sets can also be added to the extracted data.

5.5 Searching the 'Hidden Web'

Larger companies have their own searchable websites. These sites make up the 'hidden Web'. Since conventional search engines do not index dynamic Web pages, Web crawlers do not have access to the database from which the dynamic pages are created. Because an off-the-shelf solution is not known at this time, the only option is to perform a manual search within the websites or use custom spider to automate some of this [6] Again, relevant terms from the company-specific taxonomy can be used as the search criteria.

5.6 Clustering Web Pages

Once the Web crawler has retrieved potentially relevant pages, clustering will be conducted to organize the pages into categories. Since it is probable to be bombarded with excessive data, it is important to perform clustering before actual extraction in order to avoid downloading irrelevant data. This is especially important in the context of economies where bandwidth is relatively expensive (such as South Africa). Consequently, user interaction is important at this stage of the Web mining process. The user must select which clusters of data will go to the next stage of data processing.

It is advisable to use an on-the-fly clustering technique. This would save the time and effort required to manually construct a categorical hierarchy. Several researchers [5; 4; 19] recommend that automation is used to cope with Web information overload. So, if and when automation is reasonably possible, it is sensible to use it. In addition, ad hoc clustering would allow for groupings to emerge organically from retrieved data. Allowing information to surface is usually inherent in knowledge discovery.

Using a clustering tool which takes stopwords into account is important. Elimination of stopwords can reduce the level of noise in data. Precision and efficiency can be significantly improved when noise is filtered out from the clustering process. Furthermore, comparisons between text documents can be conducted more accurately when non-discriminative words are purged.

The resultant catalog of clusters is often referred to as a taxonomy. This is not to be confused with the original company-specific taxonomy although the cluster-inspired taxonomy is likely to inform and refine the company-specific taxonomy which can then be used in future searches. This cycle will repeat as and when new clusters are generated.

5.7 Extracting Web Data

In the previous stage, clustering was used to help filter out key concepts from the hordes of retrieved data. In the next phase, IE aims to identify and remove pertinent data from the selected clusters. As recommended by [20] this is the final stage of preprocessing unstructured data prior to text mining. Again, the objective of IE is to transform unstructured data into structured database entries. From there, text mining techniques are used to analyze the text.

An IE system needs to know how to locate information for extraction. A wrapper induction system that is trained to learn the extraction rules is ideal. The user then provides the system with sample documents from which extraction rules are to be generated automatically. From the training documents, the system is able to decipher how to locate the start and finish of data to be removed. A Web induction system that can formulate extraction rules from a small set of examples is ideal.

The wrapper also needs to be designed to convert data into XML format. This is the transformation of unstructured data into structured data. Once the XML format has been created, the XML document is then stored in a database.

5.8 Mining Text for Information

The final stage of the proposed Web mining system is text mining. This process uses linguistic analysis, statistical and computational techniques, to identify entities, entity relationships and interrelated concepts. This is the point where data analysis leads to knowledge discovery, or intelligence.

Once data compiled from multiple information sources is aggregated, text mining can reveal unexpected associations or confirm hunches about environmental trends. Associations between organizations and/or concepts may also emerge. The user should be equipped with as much of a synopsis as possible. A summarization and/or visual representation of the text mining results is valuable. A system which allows for the user to interact with the results would be a bonus.

It is important to note that text mining software is not meant to conduct a thorough analysis since text mining technologies are not well-developed enough to do that. Again, summarization and visual representation can be helpful in presenting patterns and trends. However, the user must have some analytical skills and be familiar enough with the data set to really understand what the result means [21].

5.9 Repeating the Cycle

The above process is likely to be repeated at regular or ad-hoc intervals. In this case, a baseline historical database of original web information is kept and special rules can be introduced to weigh incremental and new information against static baseline and historical information. Indeed later cycles are believed to benefit greatly from historically collected information (as incorporated in the company specific taxonomy) and differential data items will typically highlight significant evolutions and trends.

Also, any of the phases can feedback into earlier stages for future iterations e.g. the company-specific taxonomy will be updated and refined with results from the cluster analysis.

5.10 Commercial Web Mining Tools

The appendix lists a number of commercially available Web mining tools. Software packages that include multiple components are available. Seemingly, knowledge management systems tend to incorporate multiple Web mining components.

5.11 Integration of Web Mining Components

The proposed Web mining architecture consist of the following components: 1) company-specific taxonomy; 2) personalized Web crawling with meta-searching; 3) manual search and extraction of data in proprietary databases, including the Hidden Web; 4) Web IE and 5) text mining. With the exception of the third component, Web mining tools are commercially available that address the other four components.

A detailed technical architecture is beyond the scope of this paper. However, a key consideration is the ability to integrate all of the components into one Web mining

system. Some software packages have incorporated multiple components into their suites. Nevertheless, an all-encompassing program is unavailable. Therefore, the issue of integration must be addressed.

XML makes integration of standalone or disparate components possible. Wrappers can be written to translate data into XML format. Thereafter, XML schema can be used to enable the flow of XML data from one component to the other.

6 Conclusions and Implications

The study on intelligence collection in South Africa revealed that Web mining can be used to assist business decision-making, especially in the context of having to compete in an increasing global arena [22].

It is important to distinguish between Web data that is free and data that is available through online information services, the latter often at a cost. Where the Internet is underutilized in a given region, such as the case in South Africa as in many other emerging economies, 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.

The proposed Web mining process only fulfills the phases between planning and processing. As highlighted previously, data analysis must still be conducted. Results of the analysis must be assimilated into a report. The report must then be distributed to the relevant parties to be used, or considered, in strategy formulation. Feedback from strategic managers can lead to the next cycle. In addition, ongoing evaluation of the specific processes should be conducted. The proposed Web mining architecture does deal with the evaluation of data validity. This is done by cross-referencing the retrieved data with information from authoritative sources. Nevertheless, an ongoing appraisal of the strategic intelligence system, as a whole, is a good discipline.

Computer systems continue to reduce manual data processing. This is demonstrated by the Web mining tools that have been developed to automate Web mining. However, systems cannot make the actual business decisions. A comprehensive strategic intelligence system can only equip executive management with information about the environment. Executives must still analyze how changes in the environment affect the organization, determine the company's strengths and weaknesses, formulate strategies that incorporate the nature of its operations and industry and then implement those strategies.

One of the main implications of the research is the opportunity which presents itself for small and medium-sized organizations that have to contend increasingly with global competitions but, equally if not more importantly, are given the opportunity to compete in this global market. 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 local and global 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.

The ever-increasing computing power of PCs brings the possibility of using (clusters of) them to crawl relevant but sizeable areas of the Web. Although the initial crawling cycle may take significant time and resources, user-designated websites can be re-crawled at regular intervals to ensure that regular updates are made from key websites.

Finally, it must be conceded that the proposed Web mining process is mainly conceptual at this point. Testing the Web mining architecture can only be done through the real-world implementations.

References

[1] Velder, R.G., Vanecek, M.T., Guynes, C.S., & Cappel, J.J. (1999). 'CEO and CIO perspectives on competitive intelligence', Communications of the ACM, 42 (8), 109-116.

[2] Kennedy, S. D. 1998. "Competitive intelligence on the Web 101", Information Today, 15 (9), 40-41.

[3] Kosala, R. & Blockheel, H. (2000). "Web mining research: a survey." Proceedings of SIGKDD, 2 (1), 1-15.

[4] Adams, K.C. (2001). "The Web as a database: new extraction technologies & content management", Online, 25 (2), 27-32.

[5] 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.

[6] Hemenway, K. & Calishain, T. (2004). "Spidering Hacks: 100 Industrial-Strength Tips & Tools". Cambridge: O'Reilly.

[7] Chen, H., Chau, M. & Zeng D. 2002. 'CI Spider: a tool for competitive intelligence on the Web'. [Electronic] Available: http://ai.bpa.arizona.edu/~mchau/papers/CISpider.pdf

[8] Silvestri, F., Perego, R. & Orlando, S. (2004). "Assigning document identifiers to enhance compressibility of Web search engines indexes", Proceedings of SAC '04, Nicosia, 600-605.

[9] Kobayashi, M. & Takeda, K. (2000). "Information retrieval on the Web." ACM Computing Surveys, 32 (2), 144-173.

[10] Flesca, S., Manco, G., Masciari, E., Rende, E. & Tagarelli, A. (2004). "Web wrapper induction: a brief survey", AI Communications, 17 (2), 57-61.

[11] 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.

[12] Grimes, S. (2004). "Consumer and enterprise search: not an exact match", Intelligent Enterprise, 7 (9), 22-31.

[13] Zhong, N., Liu, J. & Yao, Y. (2003). Web intelligence, Berlin: Sprinter

[14] Liu, B., Chin C.W. & Ng, H.T. (2003). 'Mining topic-specific concepts and definitions on the Web', Proceedings of WWW 2003, May 20-24, Budapest, 251-260

[15] Ester, M., Kriegel, H. & Schubert, M. (2002). 'Web site mining: a new way to spot competitors, customers and suppliers in the world wide Web', Proceedings of SIGKDD, Edmonton, 249-258.

[16] Gruhl, D., Chavet, L., Gibson D., Meyer, J., Pattanayak, P., Tomkins A. & Zien, J. (2004). 'How to build a WebFountain: an architecture for very large-scale text analytics', IBM Systems Journal, 43 (1), 64-77.
[17] Calof, J.L. & Viviers, W. (2001). "Adding competitive intelligence to South Africa's knowledge management mix", Africa Insight, 31 (2), 61-67.
[18] Wagner, L. & Van Belle, J.P. (2005). "Web Mining for Strategic Intelligence in South Africa", Proceedings of 6th IAABD Conference, Dar es Salaam [Electronic].
[19] Song, M., Song, I-Y. & Hu, X. (2003). 'KPSpotter: a flexible information gain-based keyphrase extraction system', Proceedings of WIDM'03, New Orleans, pp. 50-53.
[20] Feldman, R., Aumann, Y., Liberzon, Y., Ankori, K., Schler, J., & Rosenfeld, B. (2001). 'Information retrieval and text mining: a domain independent environment for creating information extraction modules', Proceedings of the CIKM'01, Atlanta, 586-588.
[21] Robb, D. 2004. 'Taming text', Computerworld, 38 (25), 40-41.
[22] 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.

Appendix: Some Commercial Web Mining Tools

Table 1. Mapping of some commercial web mining tools on the proposed process steps.

Tool	Taxonomy	Meta-searcher	Web Crawling	Web Clustering	Information Extraction	Text Miner	Scalability	XML
WebSphinx			X					
FOCI		X		X				
KartOO		X		X				
CI Spider			X	X				
TetraFusion		X	X	X	X	X		
Predictive Text Analytics				X	X	X		
Autonomy		X	X	X		X		X
Verity		X	X	X		X		X
Convera RetrievalWare			X	X	X	X	X	
ClearResearch					X	X		
InXight				X		X		
Dataware II KM Suite						X		
SemioMap		X				X	X	
Enterprise Intelligence Center						X		

Tool	Taxonomy	Meta-searcher	Web Crawling	Web Clustering	Information Extraction	Text Miner	Scalability	XML
SAS TextMiner			X	X	X	X	X	
Brimstone		X		X	X			
Copernic		X	X	X	X			