December 2003

Competitive Intelligence and the Web

Robert Boncella
Washburn University

COMPETITIVE INTELLIGENCE AND THE WEB

Robert J. Boncella
Washburn University
bob.boncella@washburn.edu

Abstract

Competitive intelligence (CI) is the selection, collection, interpretation and distribution of publicly-held information that is strategically important to a firm. A substantial amount of this public information is accessible via the World Wide Web. This tutorial describes some of the difficulties in using this information resource for CI purposes and some of the solutions.

Keywords: Competitive intelligence Internet searching and browsing Intelligence monitoring Information verification Web Mining

Introduction

The purpose of this paper is to provide an overview of how the Web can be used to carry out competitive intelligence. This paper is structured as follows: Following a definition of Competitive Intelligence, the logical structure of the World Wide Web is presented. The sections that follow present the techniques that can be used to carry out CI projects and some of the problems associated these techniques. In particular, information gathering, information analysis, information verification, and information security as related to CI are discussed.

Competitive Intelligence

The Society of Competitive Intelligence Professionals defines Competitive Intelligence as *"the process of ethically collecting, analyzing and disseminating accurate, relevant, specific, timely, foresighted and actionable intelligence regarding the implications of the business environment, competitors and the organization itself"(*SCIP, 2003).

This process involves a number of distinct activities undertaken by a firm engaged in a CI project. An effective CI project is a continuous cycle, whose steps include (Herring, 1998):

1. Planning and direction (working with decision makers to discover and hone their intelligence needs);
2. Collection (conducted legally and ethically);
3. Analysis (interpreting data and compiling recommended actions)
4. Dissemination (presenting findings to decision makers)
5. Feedback (taking into account the response of decision makers and their needs for continued intelligence).

After step 1 is completed steps 2 and 3 are the keys to a successful and efficient CI process. Any number of information resources are consulted to carry out steps 2 and 3. A comprehensive list of these information resources may be found in (Fuld, 1995).

In addition Internet information resources are being used more frequently in the CI process. The reasons for this trend. include:

1. A business Web site will contain a variety of useful information, usually including company history, corporate overviews, business visions, product overviews, financial data, sales figures, annual reports, press releases, biographies of top executives, locations of offices, and hiring ads. An example of this information is **http://www.google.com/about.html.**
2. The cost of this information is, for the most part, free.
3. Access to open sources does not require proprietary software such as a number of commercial databases.

The Web Structure

The HTTP protocol and the use of Uniform Resource Locators (URL) determine the logical structure of the web. This logical structure provides a natural retrieval technique for the contents of the Web. The logical structure of the Web can be understood as a mathematical network of nodes and arcs. The nodes represent the web documents and the arcs are the URLs located within a document. A simple retrieval technique is one that starts from a particular HTML document and follows the links (arcs) from document to document (node to node). The process of "following the links" means document retrieval. This process is also referred to as Information Retrieval (IR). The content of the retrieved documents can be evaluated and a new set of URL becomes available to follow.

The retrieval techniques are graph search algorithms adapted to use a document's links to implement and control the search. An example of a graph search algorithm is a breadth first search on links contained in the initial document. A modification would a best first search based algorithm.

Information Gathering on the Web

The most common method for gathering information from the Web is the use of "search engines." Examples of these are: AltaVista (**http://www.altavista.com**), Infoseek (**http://www.infoseek.com**), Yahoo! (**http://www.yahoo.com**) and Google (**http://www.google.com**). These search engines accept a user's query, an expression of keywords, and return a set of web pages or documents that satisfy the query to some degree. Further, this set of pages and documents are organized in some fashion.

A Web search engine usually consists of the following components.

1. Web Crawlers or Spiders are used to collect Web pages using graph search techniques.
2. An indexing method is used to index collected Web pages and store the indices into a database.
3. Retrieval and ranking methods are used to retrieve search results from the database and present ranked results to users.
4. A user interface allows users to query the database and customize their searches. For more details on Web Crawlers see Chen, et al. (2002).

In addition to the general search engine types, a number of domain specific search engines are available. Examples of these are

- Northern Light, a search engine for commercial publications, in the domains of business and general interest.
- EDGAR is the United States Securities and Exchange Commission clearinghouse of publicly available information on company information and filings.
- Westlaw is a search engine for legal materials.
- OVID Technologies provides a user interface that unifies searching across many subfields and databases of medical information.

A third type of search engine is the *meta-search* engine. Two examples are MetaCrawler (www.metacrawler.com) and Dogpile (www.dogpile.com). When a meta-search engine receives a query it connects to several popular search engines and integrates the results returned by those search engines. Meta-search engines do not keep their own indexes but in effect use the indices created by the search engines it used to respond to the query.

Given the size of the Web, using a graph search algorithm approach, it takes a long time to crawl and index all the relevant Web pages associated with a query, even for a domain-specific search engines. Many Web pages may be "crawled" but not indexed. The results is outdated or incorrect information. This "static" type of informational retrieval will not take in to account continuous updating of dynamic content Web pages. The result is information that is not current. In addition to time and currency of information, the number of pages that satisfy the user's query is a problem.

It is estimated that the Internet is composed of over 552.5 billion web pages or documents, and is growing by 7.3 million pages a day (HMI). These pages or documents can be classified into two basic types, the "surface Web," those pages or documents that are freely available to any user. The number of these types of pages and documents is estimated to be approximately 2.5 billion; and "deep Web" pages and documents which consists of dynamic pages, intranet sites, and the content of Web-connected proprietary databases. The number of deep Web documents is estimated to be 550 billion. Deep Web documents are generally accessible only to y members of organizations that produce them or purchase them, such as businesses, professional associations,

libraries, or universities. Internet search engines such as Google, AltaVista and Lycos usually do not index and retrieve deep Web pages. This distinction is important to keep in mind when doing a CI project. It means that some of the most valuable information, such as full-text scholarly journals, books still in copyright, business market information, and proprietary databases can only be retrieved by users with subscriptions, searching with specialized software. For a review of tools for searching the deep Web see Aaron and Naylor (2003).

Another difficulty with gathering information using the surface Web is that a number of sites are starting to charge a fee for access to information. (Murray and Narayanaswamy, 2003).

Information Analysis

Given the large number of pages an uncontrolled search might generate it becomes necessary to control the search. This can be done through control of the graph search techniques. Controlling the search is, in effect, a rudimentary analysis of the information being retrieved. In effect, the search should only return those Web pages that a relevant to the query. Sophisticated Web search engines are able to work in this way to some extent. This initial form of analysis is referred to as Web Mining. For a more complete discussion of this see Dunham (2003).

Web Mining

Web mining can be categorized into three classes: Web Content Mining, Web Structure Mining, and Web Usage Mining.

Web Content Mining

Web Content Mining refines the basic search technique. Web Content Mining can be subdivided into Web Page Content Mining and Search Result Mining. Web Page Content mining can be viewed as text mining. In Web Page Content Mining, the graph search algorithm is controlled by the contents of the page. An unsophisticated search engine will use keywords to control the graph search algorithm. This techniques returns a set of pages that can either be searched again using a refinement of the initial search or this set of returned pages can be "text mined" using more complex text mining techniques.

Web Structure Mining

Web Structure Mining uses the logical network model of the Web to determine the importance a Web page. One method is the *PageRank* technique (Page and Brin, 1998). This technique determines the importance of Web information on the basis of the number of links that point to that Web page. The idea is that the more Web pages that reference a given Web page the greater the importance of the page. This technique combined with keyword search is the foundation of the Google search engine. Another technique is the Hyperlink-Induced Topic Search (HITS) (Kleinberg, 1999). HITS finds Web pages that are hubs and authoritative pages. A hub is a page that contains links to authoritative pages. An authoritative page is a Web page that best responds to a user's query.

Web Usage Mining

Web Usage Mining performs data mining on Web logs. A Web log contains "clickstream" data. A clicksteam is a sequence of page references associated with either a Web server or Web client (a web browser being used by a person). This data can be analyzed to provide information about the use of the web server or the behavior of the client depending upon what clickstream is being analyzed.

Text Mining

In some instances even the most carefully designed query sent to a well-designed Web search engine results in hundreds if not thousands of "hits." The third step in a CI project is to perform analysis on the information collected. It is a daunting task to organize and summarize hundreds of Web page or documents. Automating this process is quite useful.

The goal of text mining is to perform automated analysis of natural language texts. This analysis leads to the creation of summaries of documents, determining to what degree a document is relevant to a user's query, and clusters document. Text mining applications are available commercially.

Regardless of how efficient and/or effective the information analysis task is performed, its usefulness is determined by the quality of the information retrieved. Because of the unsupervised development of Web sites and the ease of referencing other Web pages, the user has no easy method of determining if the information contained on a Web page is accurate. The possible inaccuracies may be accidental or intentional. Inaccuracies are a significant problem when the Web is used as an information source for a CI project. The issue is information verification.

Information Verification

Web search engines perform an evaluation of the information resources. The HITS and PageRank techniques evaluate and order the retrieved pages as to their relevance to the user's query. This evaluation does not address the accuracy of the information retrieved.

Confidence in the accuracy of the information retrieved depends on whether the information was retrieved from the surface web or the deep web. The deep web sources will be more reliable than the surface web sources and will require less verification than the information retrieved from surface web sources. In either case one should always question the source and if possible confirm with a non-Web source for validation. In assessing the accuracy of the information retrieved it is useful to ask the following questions. Who is the author? Who maintains (publishes) the Web site? How current is the Web page? Further suggestions and more detail on methods of verifying information retrieved from the Web, either deep web or surface web, can be found at the following Web sites:

- **http://www.uflib.ufl.edu/hss/ref/tips.html** (date of access April 18, 2003).
- **http://www.vuw.ac.nz/~agsmith/evaln/index.htm** (date of access April 18, 2003).
- **http://www.science.widener.edu/~withers/webeval.htm** (date of access April 18, 2003).
- **http://www.ithaca.edu/library/Training/hott.html** (date of access April 18, 2003).
- **http://servercc.oakton.edu/~wittman/find/eval.htm** (date of access April 22, 2003).

Information Security

Recognizing the possibility of your firm being the focus of someone else's CI project, information security becomes a concern. These concerns include:

- assuring the privacy and integrity of private information,
- assuring the accuracy of its public information
- avoiding unintentionally revealing information that ought to be private.

The first of the concerns can be managed through the usual computer and network security methods.

The second concern requires some use of Internet security methods. In general a firm must guard against the exploits that can be carried out against Web sites. Some of these exploits are Web Defacing, Web page hijacking and Cognitive Hacking.

Web Defacing involves modifying the content of a Web page. This modification can be done in a dramatic and detectable fashion. However, and perhaps more dangerous, the content can be modified in subtle ways that contribute to the inaccuracy of the information.

Web Page Hijacking occurs when a user is directed to a web page other than the one that is associated with the URL. The page to which the user is redirected may contain information that is inaccurate.

Cognitive hacking or semantic attack is used to create a misperception about a firm's image. For an example of this type of attack and counter measures, see Cybenko, Giani, and Thompson (2002) .

Another form of this attack is to build a Website that is a repository for negative information about a particular firm. There are number of Websites whose URL contains the word "sucks" as part of the URL. The countermeasure to this type of attack is for the firm to monitor those sites that are trying to create a negative image of the firm and respond appropriately.

The issue of unintentionally revealing sensitive information is a difficult one to address. In the course of doing business in public, a firm my reveal facts about itself that individually don't compromise that firm but, when taken collectively, reveal information that is confidential. For example listing position openings on a public Website may reveal details about that firm that ought to be held private. For an example of this see Krasnow (2000). A countermeasure to this security breach is for the firm to carryout a CI project against itself.

Conclusion

The foregoing has provide an overview of the issues associated with implementing a CI project using the "open sources" of the Web. The methods and techniques associated with information gathering and information analysis are to a great degree automated by using Web Spider and text mining. The assurance the validity of results based on these actives is not well automated. In particular, information verification, at this stage, requires human intervention. Perhaps a stream of research can be started that will lead to methods that automate the process of information verification of Web sources in general and surface Web sources in particular. With regard to information security and CI, the issue of assuring the accuracy of a firm's public information and providing countermeasures to cognitive hacking may require the firm to monitor its "information presence" on the Web.

References

Aaron, R. D., and Naylor, E. "Tools for Searching the 'Deep Web'," *Competitive Intelligence Magazine* (4:4), Online at **http://www.scip.org/news/cimagazine_article.asp?id=156** (date of access April 18, 2003).

Calishain, T., and Dornfest, R. *Google Hacks: 100 Industrial-Strength Tips & Tools*, Sebastopool, CA: O'Reilly & Associates, 2003.

Chakrabarti, S. *Mining the Web: Discovering Knowledge from Hypertext Data,* San Francisco, CA: Morgan Kaufmann, 2003.

Chen, H., Chau, M. I., and Zebg, D. "CI Spider: A Tool for Competitive Intelligence on the Web," *Decision Support Systems* (34:1) 2002, pp. 1-17.

Cybenko, G., Giani, A., and Thompson, P. "Cognitive Hacking: A Battle for the Mind," *IEEE Computer* (35:8) August, 2002, pp. 50–56.

Dunham. M. H. *Data Mining: Introductory and Advanced Topics,* Upper Saddle River, NJ: Prentice Hall, 2003.

Fleisher, C. S., and Bensoussan, B. E. *Strategic and Competitive Analysis,* Upper Saddle River, NJ: Prentice Hall, 2000.

Fuld, L. *The New Competitor Intelligence*, New York: Wiley, 1995.

Herring, J. P. "What Is Intelligence Analysis?" *Competitive Intelligence Magazine* (1:2), 1998, pp., 13-16 (available online at **http://www.scip.org/news/cimagazine_article.asp?id=196**)

Kleinberg, J. M. "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM* (46:5), September, 1999, pp. 604-632.

Krasnow, J. D. "The Competitive Intelligence and National Security Threat from Website Job Listings," 2000, **http://csrc.nist.gov/nissc/2000/proceedings/papers/600.pdf** (date of access April 18, 2003).

Lyman, P., and Varian, H. R. "Internet Summary," Berkeley, CA: How Much Information Project, University of California, Berkeley, 2000 (available online at **http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html**; date of access April 18, 2003).

Murray, M., and Narayanaswamy, R. "The Development of a Taxonomy of Pricing Structures to Support the Emerging E-business Model of 'Some Free, Some Fee'," *Proceedings of SAIS 2003*, 2003, pp. 51-54.

Page, L., and Brin, S. "The Anatomy of a Large-Scale Hypertextual Web Search Engine," 1998, **http://www-db.stanford.edu/~backrub/google.html** (date of access April 22, 2003).

Schneier, B. "Semantic Attacks: The Third Wave of Network Attacks," *Crypto-gram Newsletter,* October 15, 2000, **http://www.counterpane.com/crypto-gram-0010.html** (date of access April 18, 2003).

SCIP (Society of Competitive Intelligence Professionals), **http://www.scip.org/** (date of access April 18, 2003).