Mining the Internet for Competitive Technical Intelligence

By Alan L. Porter, David J. Schoeneck, and Paul R. Frey, Search Technology, Inc.
Diana M. Hicks and Dirk P. Libaers, Georgia Tech

In a 2005 issue of CIM we described a "tech mining" approach to generate competitive technical intelligence (CTI) (Porter, 2005). Tech mining addresses managerial issues by deriving empirical knowledge, primarily from patent and research publication abstract databases. This article extends the resource base to be mined to the internet.

Technology intelligence detects opportunities based on early identification of emerging technologies pertinent to a company's business interests. Moreover, it identifies areas in the competitive landscape with limited or no competition, where corporate strategy can exploit these areas. Competitive intelligence tracks competitor activities to spot threats early. CTI blends elements of both.

CTI RESOURCES

Software can aid in extracting intelligence from database searches — for example, by retrieving research publication abstract records on "fuel cells" or patents assigned to "International Fuel Cells" (now a unit of United Technologies). Extracting knowledge to meet strategic intelligence needs is well and good, but companies want more!

Table 1 presents a larger picture of competitive technical intelligence resources. These resources exploit technological content from publicly accessible and clients' confidential databases, and also extract information from business and general databases such as LexisNexis and Factiva. This kind of empirically derived knowledge from databases and the internet should be complemented by suitable tacit knowledge from individuals. For instance, first map the hot spots of fuel cell research and development (R&D) activity, then have technical experts refine and interpret the prospects (Table 1 E). Additionally, tap business experts to explore the ramifications of enhanced technical capabilities (Table 1 F).

Users of CTI information want answers to their questions rather than nicely defined puzzle pieces. That's a tall order, but there are practical ways to extend the information compilation to include the internet (cells C and D in Table 1). We first draw upon search engines such as Google to augment our database-derived results from the internet, then look at specific sites. For instance, our fuel cells search identifies an active research center at Georgia Tech. We would then look up their web site to check whether key researchers are still located there, see their most recent research efforts, and obtain contact information. But we need more.

A typical searcher is looking for ONE result. Sometimes this is recovering a previously known source; other times it is discovering a new one (Battelle, 2005). For CTI purposes, we often want to capture an entire body of information. Taking the fuel cell illustration, we identified a set of active R&D center web sites. We then probed further by profiling what fuel cell types those active centers emphasize to spot trends as key centers shift toward emerging technologies, or to discern

TABLE 1. COMPETITIVE TECHNICAL INTELLIGENCE RESOURCES

Source	Technological Content	Contextual Content
Databases (empirical)	**A.** Compiled, filtered, organized R&D (publication, patent, etc.) information	**B.** Compiled, filtered, organized business and socioeconomic information
Internet (empirical)	**C.** Diffuse, up-to-the-minute, ill-structured technical information	**D.** Diffuse, up-to-the-minute, ill-structured business and socioeconomic information
Human (tacit)	**E.** Technical expertise (tacit knowledge)	**F.** Business and context expertise (tacit knowledge)

the range of applications. Here's how we developed this type of internet-derived intelligence.

INNOVATION STRATEGIES OF SMALL FIRMS

A recent study completed for the Small Business Administration investigated the innovation strategies of long-lived, highly innovative small firms (Hicks et al., 2006). The focus was on learning the technology commercialization strategies used by small companies that patent heavily. The traditional modes of studying such topics are surveys or longitudinal studies. Surveying suffers from many flaws — limited extent of feasible queries (not too many questions), self-reporting biases, nonresponse biases (how different are those who don't respond?), high cost, and so on. Longitudinal data — information on attributes from the same objects (firms, individuals, agencies, web sites, etc.) over a specified time period — are notoriously hard to come by. So, we mined the company web sites by building an "Innovative Firms Application Wizard" that uses Google's Application Programming Interface (API) search capabilities.

Previous work identified a growing cadre of highly innovative small firms (Hicks, 2002). These firms have fewer than 500 employees, are independent and long-lived, are not bankrupt, and have at least 15 U.S. utility patents assigned to them in a five-year period. They are "serial innovators." We sought to gain insight into these innovative firms' technology commercialization strategies. Having a hefty number of patents, what did they do with them?

- Do they create positional advantages based on this patent estate?
- Do they engage in strategic patenting to close off areas to competitors?
- Is technology licensing a core business activity of the firm?

CONTENT ANALYSIS OF WEB SITES

A content analysis of the firms' web sites helped answer the question of patent utilization. Scholars who review web site content analyses note serious issues (McMillan, 2000; Opoku, 2005). Web site designs vary and so do their communication objectives. Ellinger and others (2003) found that the mission statement or "about" section of a web site was almost universally present. Perry and Bodkin's (2000) examination of corporate web sites led them to conclude that the sites focused on institutional advertising. Sullivan (1999) concluded that image creation is the most important function of corporate web sites. These considerations suggested focusing on select pages or sections of web sites. The Google API enables such selectivity, but the content analysis mined the full web sites. This reflects the fact that small firm sites vary widely in style and depth.

The firm sample began with 516 small businesses with 15 or more patents issued from 1998 to 2002. In 2006, 407 remained independent and solvent, and had viable web sites. The small firms represent many high-tech sectors such as biotechnology, medical equipment, and software. A substantial number were in highly innovative sectors such as semiconductors and pharmaceuticals. These firms can be called the "usual suspects." The data set also includes firms working in imaging and display, optical components, tissue engineering, plastics, material handling, batteries, consumer goods, and many other specialties. Overall, the sheer variety of firms is striking. We also built a control sample of a like number of firms named as direct competitors of the chosen serial innovators by Hoover's Company and Capsules Database.

To investigate which business strategies such firms use to commercialize their innovations, we sought the frequency of keywords relevant to potential technology commercialization activities, as clusters of keywords could conceivably frame distinct technology commercialization strategies. First, a

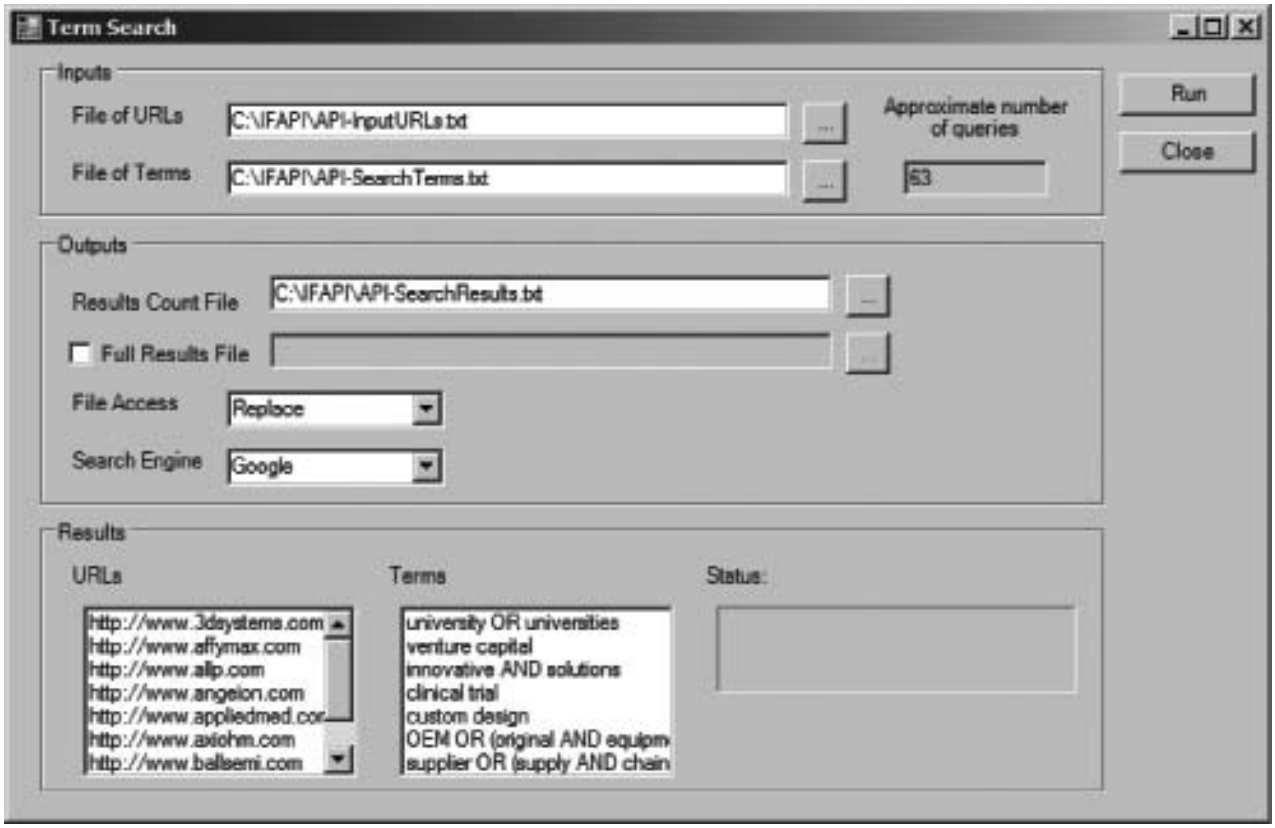

Figure 1. The "Innovative Firms Application Wizard" Search Interface

literature review developed a list of relevant terms, then we checked and revised them based on their prevalence and distribution over a sample of about 80 web sites. Web site scanning identified the functional activities important to the firm, such as research and development, licensing, production and sale of products, provision of a service, and so on.

The cornerstone of this endeavor to get at internet content is an algorithm that used Google's SOAP (Simple Object Access Protocol) API. This allowed us to create software that automates targeted searches. We wrote an "Innovative Firms Application Wizard" to exercise this capability (the interface is shown in figure 1), which enabled us to search selected web sites for selected terms.

PATTERNS OF OCCURRENCE

The main experiment's internet search looked for the occurrence frequency of 89 terms or phrases across the 814 company web sites (407 serial innovator firms and 407 control firms) — roughly 72,000 searches. (Table 2 shows sample information returned.) Google's count of pages for the target site (sometimes this is exact; sometimes it is estimated, especially for larger sites) was used to normalize the frequencies. For instance, 3dsystems mentions "university" 90 times, whereas Affymax does so only 11 times, but Affymax's hit rate is 11/78 versus 3dsystems' 90/12,500. So university emphasis is relatively higher for Affymax.

Table 2 shows patterns of occurrence in sample results. For instance, Ballsemi emphasizes original equipment (potential interest in serving as supplier?). Affymax and Angeion, among others, repeatedly mention "clinical trial," but Appliedmed does not — suggesting certain medical interests. Custom design is relatively most emphasized by Allp.

Internet search and retrieval issues successfully addressed included the following:

- Searching sites where the opening page is a flash file
- Handling redirected web sites
- Enabling Boolean operators (note the inclusion of "OR" and "AND" in the search terms illustrated in table 2)
- Using term tallies — avoiding too generic terms such as "solution" by using suitable multiword phrases, such as "integrated solutions," via a certain degree of trial and error

STRATEGIC APPROACHES

Factor analyzing the resulting keyword by web site patterns discerned six different strategic approaches to

TABLE 2: SAMPLE RESULTS RETURNED BY THE INNOVATIVE FIRMS APPLICATION WIZARD'S GOOGLE SEARCH

	Web site pages	University OR universities	Venture capital	Innovative AND solutions	Clinical trial	Custom design	OEM OR (original AND equipment)	Supplier OR (supply AND chain)
3dsystems	12,500	90	2	119	0	5	78	39
Affymax	78	11	6	0	32	0	0	0
Allp	78	16	0	1	4	3	6	3
Angeion	398	19	0	6	16	0	4	2
Appliedmed	183	1	0	7	0	0	10	0
Axiohm	59	8	0	0	0	0	1	0
Ballsemi	9,960	2	0	0	0	0	289	1
Bunnomatic	339	45	0	31	11	1	16	0

commercializing technologies (Hicks et al., 2006). For instance, one set of web site terms suggests a company seeking to contract its research capabilities. This factor shows heavy usage of phrases such as:

- R&D AND testing
- Research AND contract
- Technology development

Another factor suggests firms striving to provide solutions, with web sites using terminology such as:

- Total AND solution
- System AND (integration OR solution)
- Product AND performance

Further analyses showed that firm factor fits varied by their industrial sector.

GAINING KNOWLEDGE ABOUT A SET OF SITES

Thanks to Google, we implemented a simple way to scan many web sites relating to an intelligence interest. This might be a set of sites addressing a target technology or competitor organizations active in one's domain. To our knowledge, this is a novel way to gain knowledge about a set of internet sources collectively. By contrast, most search applications help a human locate that one "nugget." Here we wanted to learn about how a sizable group of small serial innovator firms use their technologies (patents) to go after business.

Another approach is to apply a web crawler such as WebQL to bring back information from an internet search set (Watts and Porter, forthcoming). That works nicely when you have a well-defined search focus. In that paper we retrieved conference proceedings (sets of papers) from particular Institute of Electrical and Electronics Engineers conferences relating to a theme, then analyzed them further. As table 2 implies, the present analytical needs could not be met by such an approach. Here the large set of web sites to be scoured lack a common search algorithm to identify them. Conversely, if we searched for occurrences of the 89 search phrases, we would retrieve a huge and useless set of sites. The focused search approach is vital to completing the CTI picture (see table 1).

NEXT STEPS

As with most research efforts, this one opens up many possible extensions. Here are some possible next steps:

- Retrieve Google summaries. This is possible using the Google API. For those pursuing web content analyses, these could be quite valuable.

- Determine ways to refine "within-site" search. We investigated whether one could limit searching to key pages, those telling "about" the company or describing its "mission." We could not identify a reliable nomenclature to do so.
- Deepen the searching. The Google search engine does not find (or count) any pages not publicly linked. Unless a page appears in a site tree from the home page, it will not be searched.
- Combine capabilities for interactive internet text mining. You could use within-site search functionality as a form of "detail window" to query sites and display results.
- Pursue more subtle analyses. Refine content analyses based on classifying web site emphases, such as marketing or research bravado. This involves a two-stage analysis: classify the web site based on certain characteristics, and differentiate the next analytical steps. For example, with sites classed as research focused, look for certain patterns to discern emphasis on performing contract research for others versus licensing of the firm's intellectual property.
- Tapping human expertise (cells E and F of table 1) remains expensive and time-consuming. Can we extend empirically based analyses to reduce the need for topical expertise?

ACKNOWLEDGMENTS

The research reported in this article was supported by the Small Business Administration, Office of Advocacy under contract SBAHQ05M0292. The special Google application wizard was written by Chuck Howard of Personalized Software Development. We appreciate Google's provision of the SOAP Search API. We note that it is limited to noncommercial use and 1,000 queries per day, and availability appears curtailed — see http://code.google.com/apis/soapsearh/reference.html. Google provides a license key for such use. See Google's Advanced Search at www.google.com/advanced_search for a sense of capabilities.

REFERENCES

Battelle, J. (2005). *The search*. Penguin Books, New York.

Ellinger, A., Lynch, D., and Hansen, J. (2003). "Firm size, web site content, and financial performance in the transportation industry," *Industrial Marketing Management*, v32(3), pp177-185.

Hicks, D. (2002). Serial innovators: the small firm contribution to technical change, Report prepared for the U.S. Small Business Administration under contract number SBAHQ-01-C-0149.

Hicks, D. M., Libaers, D. P., Porter, A. L., and Schoeneck, D. J. (2006). A taxonomy of serial innovators, Final Report to the U.S. Small Business Administration, Office of Advocacy (Order No. SBAHQ05M0292).

McMillan, S. J. (2000). "The microscope and the moving target: the challenge of applying content analysis to the world wide web," *Journalism and Mass Communication Quarterly*, v77(1), pp80-98.

Opoku, R. (2005). "Communication of brand personality by some top business schools online," Licenciate thesis, Lulea University of Technology, ISSN 1402-1757.

Perry, M., and Bodkin, C. (2000). Content analysis of 100 company web sites, *Corporate Communications: An International Journal*, v5(2), pp87-97.

Porter, A. L. (2005). "Tech mining," *Competitive Intelligence Magazine*, v8(1), pp30-36.

Porter, A. L., and Cunningham, S. W. (2005). *Tech mining: Exploiting new technologies for competitive advantage*. Wiley, New York.

Sullivan, J. (1999). "What are the functions of corporate home pages?" *Journal of World Business*, v34(2), pp193-210.

Watts, R. J., and Porter, A. L., (forthcoming). "Mining conference proceedings for corporate technology knowledge management," *International Journal of Technology Management*.

Alan Porter is Director of R&D for Search Technology and co-directs the Technology Policy and Assessment Center at Georgia Tech, where he is Professor Emeritus of Industrial and Systems Engineering, and Public Policy. His current interests focus on improving the tools for technical intelligence. Contact Alan at aporter@searchtech.com.

David Schoeneck is Research Analyst with Search Technology. Contact David at daves@searchtech.com.

Paul Frey is President of Search Technology. Contact Paul at paulf@searchtech.com.

Diana Hicks is Professor and Chair of the School of Public Policy, Georgia Institute of Technology. Before this she was the Senior Policy Analyst at CHI Research, Inc. where she conducted numerous policy analyses for government agencies based on empirical information in patent and paper databases. Contact Diana at dhicks@gatech.edu.

Dirk Libaers is an Assistant Professor of Entrepreneurship and Innovation at the Henry W. Bloch School of Business and Public Administration, University of Missouri-Kansas City, and was previously affiliated with Georgia Tech. Contact Dirk at libaersd@umkc.edu.