Communications of the Association for Information Systems

Volume 5 Article 8

January 2001

The Present and Future of Internet Search

Wendy Lucas
Bentley College, wlucas@bentley.edu

William Schiano
Bentley College, wschiano@bentley.edu

Katherine Crosett
Kalex Enterprises, Inc, kcrosett@kalexenterprises.com

Communications of the Association for Information Systems

Volume 5, Article 8
April 2001

THE PRESENT AND FUTURE OF INTERNET SEARCH

Wendy Lucas
William Schiano
Computer Information Systems Department
Bentley College

Katherine Crosett
Kalex Enterprises, Inc.

wschiano@bentley.edu

ELECTRONIC COMMERCE

THE PRESENT AND FUTURE OF INTERNET SEARCH

Wendy Lucas
William Schiano
Computer Information Systems Department
Bentley College

Katherine Crosett
Kalex Enterprises, Inc.

wschiano@bentley.edu

ABSTRACT

Search engines were crucial in the development of the World Wide Web. Web-based information retrieval progressed from simple word matching to sophisticated algorithms for maximizing the relevance of search results. Statistical and graph-based approaches for indexing and ranking pages, natural language processing techniques for improving query results, and intelligent agents for personalizing the search process all show great promise for enhanced performance.

The evolution in search technology was accompanied by growing economic pressures on search engine companies. Unable to sustain long-term viability from advertising revenues, many of the original search engines diversified into portals that farm out their search and directory operations. Vertical portals that serve focused user communities also outsource their search services, and even directory providers began to integrate search engine technologies from outside vendors.

This article brings order to the chaos resulting from the variety of search tools being offered under various marketing guises. While growing reliance on a small set of search providers is leading to less diversity among search services,

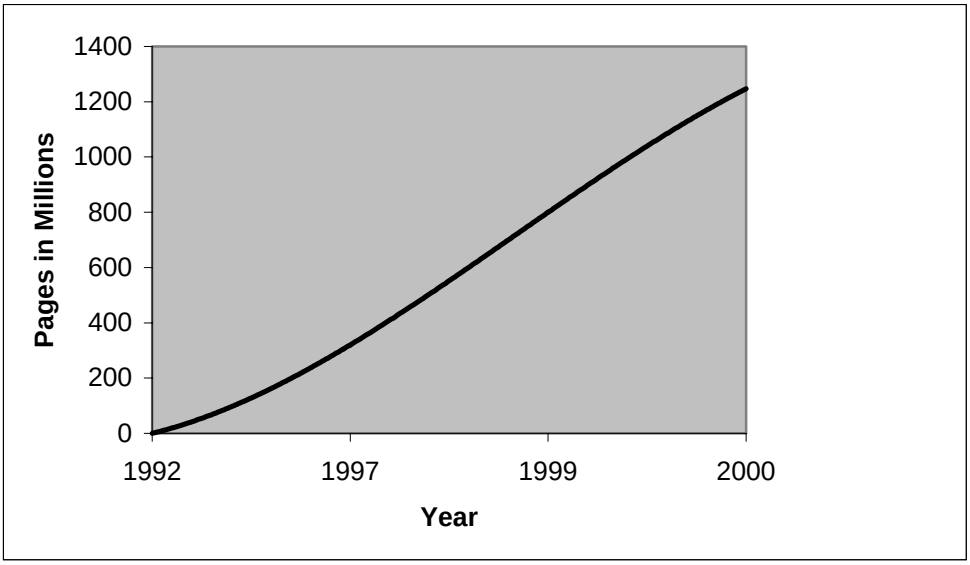
users can expect individualized searching experiences that factor in personal information. The convergence of technology and business models also results in more narrowly defined search spaces, which will lessen the quantity of search results while improving their quality.

Keywords: search engines, ranking algorithms, relevancy, personalization, portals, vortals

I. INTRODUCTION

The Internet was heralded as a free market and search engines praised as facilitators. Faced with myriad challenges inherent in the open structure of the Web, however, search engines saw their coverage decline while the number of pages continues to grow rapidly. Figure 1, based on estimates in Sullivan [2000b] and Lawrence [1999]), shows the growth achieved. In addition to uncertainty surrounding the size of the Web, the uneven quality of its contents greatly affects the tasks search engines must perform to provide relevant responses to users' queries. Efforts by page authors to outsmart indexing and ranking software to achieve top placements in search engine listings further exacerbate this problem.

Search engine companies pursued a variety of strategies to increase the number of people of who visit their sites and to widen the array of services available to these visitors. Some established search providers, such as Yahoo!, Excite, and Lycos, evolved into full-service portals. Vortals, or vertical portals, sprung up to address the growing number of user groups with targeted search and directory services. Still others differentiated themselves by focusing on a unique technology or marketing concept. Even with these steps, few companies are profitable and most face formidable economic challenges.



Based on estimates in Sullivan [2000b] and Lawrence [1999]

Figure 1. Pages on the World Wide Web

This article first identifies the set of technologies required for Web searching. The degree of sophistication needed in each of these areas and the directions search providers are pursuing are examined. Then, changes in the search engine industry and in the priorities of search services are examined in light of the economic issues of scale and scope [Chandler, 1990] fueling them. It is this combination of technology and economic forces that is shaping the future of Internet search.

II. TECHNICAL CHALLENGES OF SEARCH

Information retrieval (IR) originally focused on indexing and retrieving information from textual databases with fixed structures that reflect their content. By contrast, Web pages are of widely varying quality, their internal structural integrity is not enforced, and their numbers are constantly changing. Web size estimates are confounded by the absence of a mechanism for measuring the number of password-protected pages, those with dynamically updated content, pages with specialized formats, pages to which access by search engines is

prohibited, and peer to peer servers that may not be consistently online. META and other subject-related tags in HTML documents are not required and cannot be relied upon as accurate indicators of content. The profile of search engine users is also quite different from that of traditional information retrieval systems users, who are typically trained professionals. Most users of the Web are novice searchers, with little understanding of optimal query formulation techniques.

The uncertainty surrounding the size and quality of Web contents coupled with search engine users' lack of training greatly affects the difficulties associated with providing relevant results. The primary tasks that search engines perform in this pursuit include traversing and indexing the contents of the Web, applying relevancy-ranking algorithms to determine matches from their index to a user's query, and providing users with an interface for specifying their queries and viewing their results [Gudivada et al., 1997]. A search service provider's ability to satisfy the needs of its users rests on how effectively these tasks are performed. Figure 2 summarizes these tasks.

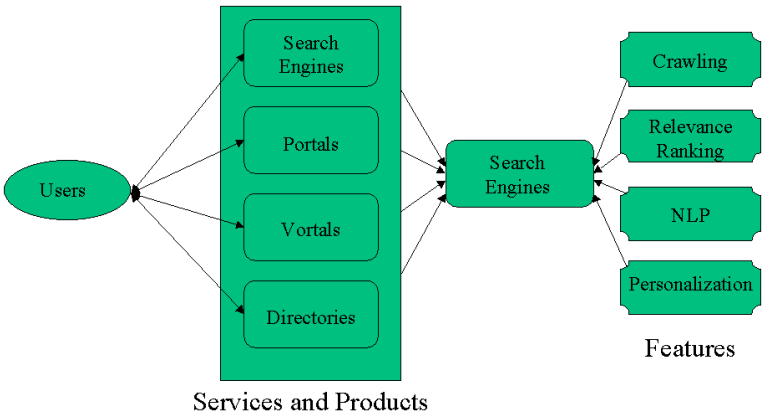


Figure 2. Internet Search Model

The following paragraphs describe current industry practices in each of these areas, identify technologies that will move search providers closer to meeting the needs of their users, and offer examples of companies engaged in those technologies.

CRAWLING AND INDEXING THE WEB

An up-to-date, accurate index is crucial to the success of Internet search. Search engines use software referred to as “robots,” “spiders,” “crawlers,” or “wanderers” to traverse the Web and gather up pages. The contents of those pages are passed to software for automatic indexing, which associates each word in the index with the pages in which it occurs [Gudivada et al., 1997]. A robot may traverse and index all links encountered without regard to the quality of the pages found. This approach is most likely to be taken by a large-scale search engine that seeks to maximize the breadth of its coverage, such as AltaVista, Fast Search & Transfer ASA (FAST), or Northern Light. The major search engines (see the Appendix), however, proved to be far from adequate at maintaining comprehensive, accurate indices of the entire Web. Lawrence and Giles [1999] found that the combined coverage of eleven major full-text search engines was 42% of the indexable Web. Overlap in coverage among the search engines was relatively low, with no individual search engine indexing more than 16% of the Web. They also found that the indexing of new or modified pages could take several months or more. Peer-to-peer networks also rely on search engines to find content on distributed end-user machines acting as servers. Because these machines frequently go offline or change locations, search indices for peer-to-peer networks such as Napster and Scour are refreshed each time a user logs in.

Inaccurate and incomplete indices contribute to the low retrieval effectiveness of today’s leading search engines [Gordon and Pathak, 1999, Leighton and Srivastava, 1999]. While metasearchers can compensate for poor coverage by submitting queries to multiple search engines [Selberg and Etzioni, 1997], they typically share the weaknesses of those they utilize. In addition, as more and more sites turn to search outsourcing companies like Inktomi, FAST, and Google, the overlap in coverage should increase, lessening the value added by a metasearcher.

An alternative to indiscriminate Web-wide indexing is to impose a crawling order that seeks to visit more important pages first [Cho et al., 1998]. Quality, or

importance, of a page is calculated as an independent measure from relevance to a user-specified query and is stored in the index for use in ranking pages, as described in the next section. A common ordering metric is termed “link popularity,” and is based on the premise that the number of links leading to a page is an indicator of that page’s importance. Inktomi applies link popularity metrics to its crawling order.

The contents of a page’s URL can also be used for determining crawling order. SearchEdu.Com, for example, includes only pages with “.edu” extensions in their domain names. A selective search engine may decide to put off visiting a page on the basis of its ordering metric until other more promising pages have been indexed, or may decide to exclude the page from its index entirely. Selective indexing is a viable means for providing users with focused databases that are also more manageable. Metasearch-like interfaces can then be used for identifying the appropriate sources.

The appeal of specialized indices that meet the needs of particular segments of the population should only increase with the continued growth of the Web, as they filter out many of the irrelevancies found when conducting Web-wide searches. Even for focused indices, however, reliable information about page content is hard to come by without human intervention. Standard metadata classifications are being developed that will provide structured information about page content, such as the Dublin Core metadata element set for describing Web resources, and the Resource Description Framework [Brickley and Guha, 2000], which provides an architecture for metadata. The World Wide Web Consortium’s advocacy of XML-based XHTML as the standard for all pages will facilitate the adoption of metadata classifications because XHTML requires far more structure in documents than HTML.

Metadata standards can only be useful if used correctly. Some page authors engage in deliberately deceptive practices, referred to as “spamming the index,” that attempt to mislead search engines about the content of their pages to achieve higher rankings. This competition is driven by the fact that most users only look at the first page of search results [Silverstein et al., 1998]. While indexing

algorithms attempt to weed out spammed pages, effective retrieval algorithms for discerning truly relevant, high-quality pages are also needed.

RELEVANCE RANKING

The earliest search engines based their retrieval algorithms on the similarity of query terms to Web page content. This measurement remains a key component in many of today's search algorithms. Search engines list links to pages matching a user's query in decreasing order of relevance, which can most simply be defined in terms of a page's similarity to a query. Each page in a search engine's index as well as each query entered by the user can be represented by a vector of the form (t_1, t_2, t_3, …, t_n), in which *n* is the number of unique terms [Harman, 1992]. If t_i is present in the page or query being represented, its value is 1. Otherwise, its value is 0. A Boolean match between the page and query can then be calculated as the dot product of the two vectors, with weighted matches calculated by weighting terms in the page vectors. While the specifics of the ranking algorithms used by search engines are proprietary, most claim to give higher weightings to terms appearing near the top of the page, particularly if they are within title tags. Terms in header tags and META tags or in close proximity to one another may also boost a page's rank.

Search engine ranking algorithms are often based on standard information retrieval models, including the vector space model and probabilistic models [Harman, 1992, Gudivada, 1997 #7]. The former, and more commonly used, rewards query terms that occur more frequently in a document than in the collection as a whole. Probabilistic models, which give higher weights to terms that previously appeared in relevant documents, are harder to implement because they depend on relevance judgments from users and the need for accurate estimates of conditional probabilities that a term occurs in a relevant document.

Algorithms based solely on the commonality of terms between a Web page and a query are limited in their effectiveness because of the uneven quality of Web pages. Statistical techniques, such as Latent Semantic Indexing (LSI) [Deerwester

et al., 1990], go beyond the concept of term matching to derive the true meaning of a document. LSI is a method in which the latent semantic structure of a document is estimated. First, a matrix that correlates terms to documents is constructed, from which factors representing common-meaning components are extracted. Each document is then represented by a vector of uncorrelated indexing terms, which may or may not have appeared in the document but are close to its meaning based on an overall pattern of term usage. A query is also represented as the weighted sum of its component term vectors, which are compared to the document vectors to find those coming closest to it, as measured in terms of highest cosines. Excite uses a proprietary statistical method called Intelligent Concept Extraction™ that is similar in concept to LSI for identifying terms related to a user's query and searching for concept-related pages.

These types of approaches help foil a common spamming technique, which is the repetition of a popular search term throughout a Web page, even though that term may have little or nothing to do with the page's content. Terms may be hidden using a variety of approaches, including matching their text color to the page's background color, or creating transparent images and placing the terms within the alternate text fields of their HTML tags. If a word is unrelated to the true content of a page, then it should not be included in the vectors used in statistical methods like LSI for representing key document concepts.

Retrieval algorithms that judge the quality of a Web page as an independent measure that is then used in ranking documents also demonstrated superior performance over those based on standard information retrieval models. One such measure of perceived quality is provided by the Direct Hit search engine, which determines "page popularity" based on of the number of people who visit a page, the amount of time they spend there, and other related metrics. Direct Hit, a subsidiary of Ask Jeeves, monitors these behaviors and makes their results available to several popular search engines for use in their ranking algorithms.

The graphical structure of the Web also provides valuable information about a page's importance. The Web can be represented as a set of nodes joined together by directed links, where each node corresponds to a page and the *anchor tags*

within HTML documents define the links between pages. The *ancestors* of a page are defined as those pages containing links to it, while a page's *descendents* are reached by following a page's outgoing links. Many search engines factor the number of ancestors of a page into the calculation of that page's quality. Google uses a variation of the link popularity measurement applied by Inktomi for ranking pages within its index. Called PageRank™ [Brin and Page, 1998], this metric bases a page's rank on both the number of ancestors to a page and the importance of each of those ancestors, as measured by the number of pages linked to them. The number of links each ancestor contains is used to normalize the measurement, so that ancestors with fewer outgoing links will contribute more weight. This calculation is combined with various page parameters, including term proximity, font size, and the text found in anchor tags, the latter of which is associated with both the page in which it appears and the page to which it is providing the link.

The CLEVER search engine [Chakrabarti et al., 1998] is built upon a link-based algorithm that classifies pages as being either *authorities* or *hubs* [Kleinberg, 1999]. Authorities are the best sources of information on a topic, while *hubs* provide collections of links to authority pages. Pages are assigned initial numerical hub and authority scores. Each hub score is then updated as the sum of the authority scores of its descendants, and each authority score is recalculated as the sum of the hub scores of a page's ancestors.

Rankdex, an experimental search engine, employs a method called Hyperlink Vector Voting [Li, 1998] that makes use of the label field within anchor tags for ranking pages. These labels, which are provided by outsiders rather than page creators, are expected to present a less biased representation of page content. Label fields of links pointing to a page therefore serve as that page's descriptors, and are used in determining similarity to a query.

While link-based quality measurements should reduce spammed pages in an index, as they are less likely to be linked to, they are not impervious to manipulation. Page creators can engage in "reciprocal linking," in which each provides links to the others' Web pages. Link popularity ranking methods are also biased toward

more established pages that built up a network of incoming links. Newer pages are therefore more likely to be overlooked due to their lack of connections, regardless of how innovative their content may be. Despite these caveats, both statistical and graph-based approaches increase the usefulness of search results. The final ingredient that must be added to the mix is effective communication between the user and the search engine.

USER INTERACTION

The importance of involving the user in the interface design process must be brought to bear on search engine development for these systems to realize their full potential [Shneiderman et al., 1998]. A prerequisite for retrieving relevant results is the correct interpretation of the needs of the user, which is facilitated by an understanding of the relationship between how a user interacts with the system and what the user is attempting to accomplish [Stary, 1999].

Casual users are often unmotivated or unwilling to express their information needs as queries. A study of approximately one billion queries contained in an AltaVista query log found that 72.4% contained two or fewer query terms and 79.6% contained no Boolean operators [Silverstein et al., 1998]. Users who do use Boolean operators often do so incorrectly. One reason for these findings is that query syntax varies between engines, requiring users to remember a different set of rules for each engine they visit. AltaVista's standard search, for example, supports (+) and (-) operators for specifying the mandatory inclusion and exclusion, respectively, of search terms. Only the advanced search feature, however, supports the standard Boolean operators. If a user enters `recipes AND fruit` to the main search form rather than the advanced search page, AltaVista will search for three terms: `recipes`, `AND`, and `fruit`. Yet most users avoid the advanced search pages and seldom read the "hints and tips," as they believe they are intended for more experienced users than they are [Pollock and Hockley, 1997].

Web interface designers were, and continue to be, slow to facilitate search for casual users. When a user enters a few terms into a search box, the likely

expectation is that the search engine will seek all documents containing the terms entered. In reality, most search engines perform a disjunctive comparison in which documents containing any of the terms are retrieved. An exception is Google, which defaults to performing a conjunctive comparison. This simple design decision can have a major impact on the effectiveness of a user's query.

To understand the intent of a user's query, search engines are making use of Natural Language Processing (NLP) techniques. These include automatically truncating, or stemming, search terms so that both plural and singular forms are included in a search, automatically identifying proper nouns based on the use of upper case letters, and recognizing phrases based on word proximity [Liddy, 1998]. Some search engines, including GO.com, Lycos, and Northern Light, go beyond basic stemming by searching for other forms of a word, so that a search on "assumption" will also find documents containing "assume," "assumes," and "assuming." Others, such as HotBot, include stemming as an option in their advanced search page.

The intended meaning behind a query is complicated by two factors termed *synonymy* and *polysemy*. The first refers to the fact that many synonyms exist for the same word. For example, a search on "user interface" may ignore documents about "human computer interaction," although they are likely to be relevant to the user. *Polysemy* refers to the problem of words having more than one meaning. If a user enters a query on "java," is she interested in the programming language or in where to find a good cup of coffee? Techniques such as LSI and the use of an online thesaurus for expanding a query help in dealing with the synonymy problem. Relevance feedback, which refers to the modification of queries by adding new terms and re-weighting existing ones based on user feedback, helps with the polysemy problem and has been shown to yield more relevant search results [Salton and Buckley, 1990]. It has limited appeal to most Web searchers, however. A study of Excite's query log found that only about 5% of users' queries took advantage of the relevance feedback mechanism provided [Jansen et al., 1998].

The SimpliFind™ search engine forces user involvement in identifying the meaning of a search term. It uses a semantic network built on the WordNet® online lexical reference system [Fellbaum, 1998] for finding documents that contain not only the terms specified by the user, but related concepts as well. After entering a search term, users select its meaning from a pull-down menu or, if the term is not in the database, are prompted to enter a meaning that is added to the database for future use. The original query is then expanded using associated words.

Personalizing a search by factoring in the interests of a search's initiator is another means for discerning the concept of a query and shows great potential for improved Web searches. Intelligent agents that acquire knowledge of a user's interests and preferences through interaction and monitoring can focus searches toward results that are more likely to be relevant to that user. The personalization of search results should lead to continually better performance as the agents learn from each interaction.

This concept is demonstrated by a user-adapted intelligent interface to AltaVista that filters information on the basis of a user model [Ambrosini et al., 1996]. The user modeling subsystem draws on stereotypes to represent the typical user. Artificial intelligence techniques discern the stereotype that best fits the user during the modeling phase. The information filtering module filters retrieved documents on the basis of user characteristics. It also employs a semantic network for factoring in the occurrence of semantic links and terms. Preliminary experiments found that this system improved on the performance capabilities of AltaVista by about 20% [Ambrosini et al., 1996].

Letizia is an autonomous interface agent for Web browsing that recommends pages to users in real-time [Lieberman, 1997]. It records the URLs chosen by a user and compiles a user profile based on their page content. Page analysis is performed using a standard information retrieval measurement in which the match between a term and a document is calculated as the product of term frequency times the inverse document frequency, or TFIDF [Salton and Buckley, 1988]. While the user is searching, Letizia searches the Web space that

is near the user's current position and presents results thought to be of interest in an independent window.

Glance [2000] describes a search assistant that combines agent technology with the graphical structure of the Web. A community of users is able to engage in a collaborative search through the use of a software agent called the community search assistant. All of the queries submitted by the community are stored in the form of a graph in which related queries are linked together. Users can then follow these links to a set of search results.

The visualization of both the query formation process and the results of a search can enhance a user's understanding and aid in query reformulation [Rao et al., 1995]. The SketchTrieve prototype [Hendry and Harper, 1997], in which the emphasis is on providing a "secondary notation," or visual cues, allows users to represent and organize search activities. Users can customize a menu containing a list of service categories and submenus featuring kinds of services. Several visualization techniques also exist for presenting different views of search results. The DropJaw prototype system [Karlgren et al., 1998] clusters search results over two dimensions: user-defined genre-based document categorizations, such as informal and private vs. public and commercial, and dynamically generated content-based clusters. Search results are presented using a multi-dimensional visualization that allows users to drag and drop subsets of a document set for regrouping.

Understanding and anticipating information needs and effectively communicating search results are critical to effective user interaction. Strategies using natural language processing, personalization, and customization should have profound effects on the ways people interact with search engines and they interact with us. Applying these strategies to search engines that selectively build and rank pages in their indices using statistical and graph-based techniques should lead to the next generation of Web searching tools.

III. ECONOMIC CHALLENGES OF SEARCH

In the early 1990's, consumers of online services frequented the closed networks of CompuServe, AOL, and Prodigy, which charged fixed monthly and variable connect time fees for access to public and proprietary content. These companies, which dominated the market, were able to generate both revenues and profits using this business model. In the mid-1990's, competition arrived as a result of Congress's 1993 vote to legalize the commercial use of Internet technology that had been developed with federal money. Research projects and student hobbies focused on organizing the Web were transformed into professional search engine companies. Casual users also began to access the Web for research purposes. The first search engine companies, namely Yahoo!, Lycos, and AltaVista, appeared. Within a few months, scores of other search engines began operations. As Chandler [Chandler, 1990] showed, such competition and diffusion of market share is common in the early stages of industries.

The competition for site visitors soon intensified and extended beyond the provision of search engines as companies tried to leverage their customer bases through economies of scale and scope. Many search engine companies diversified into full-service portals, offering free e-mail, news, home page services, and even free Internet access to enhance "stickiness," the amount of time users spend at the site. Although searching does send visitors elsewhere to satisfy their information needs, search tools became a competitive necessity for portals to attract and retain customers.

However, as Table 1 reflects, profits failed to materialize, and are still out of reach for the majority of today's search companies. The expansion into portals, while increasing stickiness, also increased costs. This chapter examines the financial pressures faced by these companies, the paths chosen in their quest for profits, and the effects of these choices on the future of Internet search.

Table 1. Performance Data for Publicly Traded Search Engine Companies ($000)

(continued on next page)

		About(1)	Ask Jeeves	Excite(2)	goto	Infospace
Ticker Symbol		BOUT	ASKJ	ATHM	GOTO	INSP
# unique visitors(4)		20,637	10,931	26,958	8,841	NA
2000Q3	NetPPE	17,423	16,682	336,494	26,914	47,569
	Revenue	20,129	29,029	160,533	25,050	57,695
	Gross Profit	9,951	18,706	80,168	21,740	47,331
	% revenue	49%	64%	50%	87%	82%
	Prod Dev	4,808	6,343	23,818	3,534	10,152
	% revenue	24%	22%	15%	14%	18%
	Sales/Mktg	11,348	20,896	79,244	21,185	34,408
	% revenue	56%	72%	49%	85%	60%
	Net Profit	(18,869)	(38,460)	(668,710)	(46,103)	(48,699)
	% revenue	-94%	-132%	-417%	-184%	-84%
FY1999	NetPPE	9,401	7,416	176,077	12,703	4,503
	Revenue	26,962	22,026	336,955	26,809	36,907
	Gross Profit	9,351	7,943	193,899	20,596	31,648
	% revenue	35%	36%	58%	77%	86%
	Prod Dev	8,386	8,610	54,805	3,689	3,189
	% revenue	31%	39%	16%	14%	9%
	Sales/Mktg	48,597	35,305	130,725	34,459	23,695
	% revenue	180%	160%	39%	129%	64%
	Net Profit	(55,096)	(52,929)	(1,457,638)	(29,262)	(21,694)
	% revenue	-204%	-240%	-433%	-109%	-59%
FY1998	NetPPE	3,302	879	35,937	1,336	1,239
	Revenue	3,722	800	155,360	822	9,623
	Gross Profit	(494)	(599)	125,874	(607)	7,989
	% revenue	-13%	-75%	81%	-74%	83%
	Prod Dev	3,114	1,712	29,557	1,232	1,245
	% revenue	84%	214%	19%	150%	13%
	Sales/Mktg	7,890	2,301	63,074	9,645	6,286
	% revenue	212%	288%	41%	1173%	65%
	Net Profit	(15,578)	(6,806)	(37,559)	(14,023)	(9,057)
	% revenue	-419%	-851%	-24%	-1706%	-94%
(1)About quarterly data is 2000Q2						
(2) In 1999, Excite merged with broadband access provider @Home.						
(3) Lycos quarterly data is 2000Q1;annual is ended July 31, 1999						
(4) Source: Media Metrix, September 2000.						

Table 1. Performance Data for Publicly Traded Search Engine Companies ($000)

(continued from previous page)

		Inktomi	Looksmart	Lycos(3)	Yahoo!
Ticker Symbol		INKT	LOOK	TRLY	YHOO
# unique visitors(4)		NA	13,518	30,780	52,679
2000Q3	NetPPE	83,580	11,595	10,759	98,098
	Revenue	78,588	33,364	78,603	295,548
	Gross Profit	67,838	10,933	64,839	254,688
	% revenue	86%	33%	82%	86%
	Prod Dev	17,293	8,921	12,570	30,060
	% revenue	22%	27%	16%	10%
	Sales/Mktg	39,470	23,335	38,921	109,171
	% revenue	50%	70%	50%	37%
	Net Profit	(8,544)	(12,915)	122,410	47,665
	% revenue	-11%	-39%	156%	16%
FY1999	NetPPE	83,580	11,595	7,471	58,111
	Revenue	223,484	48,865	135,521	588,608
	Gross Profit	191,600	41,947	106,794	486,809
	% revenue	86%	86%	79%	83%
	Prod Dev	55,961	26,593	26,279	67,511
	% revenue	25%	54%	19%	11%
	Sales/Mktg	122,182	59,082	78,807	214,887
	% revenue	55%	121%	58%	37%
	Net Profit	(9,441)	(64,663)	(52,044)	61,133
	% revenue	-4%	-132%	-38%	10%
FY1998	NetPPE	17,362	1,979	3,960	31,007
	Revenue	20,426	8,785	56,060	245,100
	Gross Profit	15,610	7,199	43,547	192,946
	% revenue	76%	82%	78%	79%
	Prod Dev	12,173	4,765	26,758	33,917
	% revenue	60%	54%	48%	14%
	Sales/Mktg	21,452	10,975	35,036	124,734
	% revenue	105%	125%	62%	51%
	Net Profit	(22,355)	(12,858)	(28,440)	(12,674)
	% revenue	-109%	-146%	-51%	-5%

(1)About quarterly data is 2000Q2

(2) In 1999, Excite merged with broadband access provider @Home.

(3) Lycos quarterly data is 2000Q1;annual is ended July 31, 1999

(4) Source: Media Metrix, September 2000.

STRATEGY

Among companies still operating their own search engine infrastructure, three primary strategies emerged:

- Infrastructure
- Directory-Based
- Niche Focus

Each of these models, and its impact on Internet search, is explored in the following subsections.

Infrastructure

Several portals either outsourced their entire search operation or negotiated business arrangements that give them access to efficient technologies offered by infrastructure specialists. In July 2000, Yahoo! replaced Inktomi with Google for handling its Web search operations. Two years earlier, Inktomi succeeded AltaVista, which had taken over for Open Text in mid-1996. Google also powers the search capability of Netscape, while Inktomi powers or provides supplementary results to several search services, including HotBot, MSN Search, Snap, GoTo.com, and LookSmart.

In June 2000, Lycos changed its business model by adding technology from FAST and Inktomi to its own engine to improve search results. Lycos stated that “We’re outsourcing the spidering and cataloging of the big search engine. With the number of people it requires, we can’t make a business out of spidering the entire Web” [Bray, 2000].

Such outsourcing offers opportunities for “back end” players to sell their services, but the willingness of Yahoo! and others to switch vendors frequently implies low switching costs and therefore heavy price pressure on the vendors.

While users can expect to see fewer variations in search technologies and results across portals as consolidation continues, they can expect a more personalized searching experience. Yahoo!, for example, stores user-provided zip codes so it can offer location-specific search results. Direct Hit also offers

personalized search results at its site by factoring in the user's year of birth, gender, and zip code. The more information a portal stores on a user, the more personalized its searches become and, the portal hopes, the less likely the user will be to go elsewhere for conducting a search.

Directory-Based Alternatives

Directories offer a valuable approach for enhanced searching and are integrated into most of today's search tools. Several companies fill this need with either a Web-wide or industry-specific offering. About.com (formerly The Mining Company) uses paid human guides to hyperlink sites on specific topics and now contains a significant directory. The company consistently appears in the top ten on Media Metrix's Top 50 list by attracting casual users who explore topics of interest [Media Metrix].

Other organizations derive revenues from selling their searchable directories to individual or corporate users and tailoring them for their use. InfoSpace, Switchboard, and LookSmart, all businesses with successfully IPO's, employ this tactic, which grants consumers ease of use but relinquishes a Web-comprehensive search.

Companies like VerticalNet provide searchable networks that play the role of "vortal metasearchers" by assisting users in identifying the appropriate community. Vortals, like portals, outsource most of their searching operations. EoExchange, which provides the search infrastructure for VerticalNet, combines current search technologies with industry-specific catalogs. At its Web site, the company notes the importance of including metadata and popularity-based measurements like Google's within search algorithms. More and more sites like VerticalNet can be expected to offer personalized, focused search and directory services based on a common set of search technologies.

Niche Players

New firms are exploring alternative business models that rely on a unique technology or marketing twist for generating revenues. GoTo.com sells search

engine keyword positions to advertisers and reveals the price paid for each entry. Its paid listings currently appear at Netscape and will soon be at AOL. Other companies that ventured into pay-for-position search services include FindWhat.com, Kanoodle.com, and RocketLinks. RocketLinks displays results from Google after its paid listings, while the other three all list results from Inktomi.

Ask Jeeves, whose paid link system is also used by Go2Net, allows users to enter questions using natural language. Human editors build its knowledgebase of answers. Recently, Ask Jeeves expanded into the corporate marketplace by providing company-specific knowledge bases that can be used for customer targeting, e-commerce, and e-support applications.

RealNames developed a navigation system that is integrated into Internet Explorer and other search services. This system lets users type a brand name into the browser address box for finding the appropriate Web site. Entering "Ford," for example, will lead to Ford Motor Company's site. Vendors are charged a yearly fee for each keyword that is assigned to them.

The marketplace for firms with alternative business models is growing, but it is too soon to know what strategies will ultimately prove successful. While paying-for-position schemes offer alternative sources of revenue, they are not always well received by users - witness AltaVista's ill-fated attempt at this venture [Sprenger, 1999]. Proven technology innovations, however, are likely to be integrated into existing search services, as evidenced by Google's quick ascent to search outsourcing provider.

REVENUE SOURCES

Three primary revenue sources are available to search engines:

- advertising,
- ancillary income derived from site visitors, and
- the provision of service to other sites.

Advertising

Advertising is the major source of revenue for most search engines. Lycos, for example, reported $93.44 million in 1999 advertising revenues and Yahoo! reported $529.9 million for a similar period. Such heavy reliance on advertising poses a threat to search engine companies. Advertising as a percent of total revenues dropped from 74.5% to 68.9% for Lycos between 1998 and 1999, and Yahoo! warned in its 1999 annual report that continued growth in advertising revenues is doubtful. While total spending for online advertising is growing [Cohen, 2000] and is projected to continue to do so [Lawrence, 2000], rates for online advertising have dropped precipitously since mid-2000 [Dvorak, 2000].

Ancillary Income

With advertising revenue in doubt, many search engines companies extended their offerings into electronic commerce in hopes of leveraging their base of users by selling them other products and services. Yahoo, Lycos and Excite all began to offer shopping. However, given the difficulties faced by companies trying to sell to consumers profitably over the Web, it seems unlikely that these e-commerce endeavors will be a major profit source in the near future.

Provision of services

Searching is one of the most resource-intensive of all Web site operations, with indexing, cataloging, and retrieval processes being expensive to develop, operate, and maintain. Portals therefore established a variety of teaming agreements with outside vendors who offer ways to streamline search operations. Providing these services is a source of revenue for search engine companies and offsets their fixed development costs. In addition, services allow providers to maintain a focus on search. Some search engines, such as at Northern Light, may also charge for premium content found in a search.

COSTS

Because of negligible duplication costs, information businesses are often assumed to have nearly zero marginal costs. Search engines, however, require processing power and infrastructure to deliver their product. Table 1 shows the net property plant and equipment of the publicly traded search engine companies. The cost of processing power and infrastructure maintenance is reflected in the gross margin numbers. A shortcoming of the directory model is that it does not scale well. The exponential growth of the Web shown in Figure 1 translates into an exponential growth in the number of people required for indexing and periodically re-indexing sites.

Quiver, an infrastructure provider to vortals, anonymously collects bookmarks and Internet usage behaviors from Web site communities for use in the automatic generation of directories. The hub and authority-based algorithms used in the CLEVER search engine described earlier are applied for ranking sites within Quiver's directories. In the spring of 1999, Yahoo! instituted a service that guarantees site evaluations within seven business days in exchange for a fee. Those who do not choose this fee-based service can expect to wait weeks or even months to be reviewed.

To overcome staffing costs, Netscape's Open Directory (formerly NewHoo) uses a volunteer-compiled effort that attempts to cover the entire Web. Originally applauded as consistent with the Zeitgeist of the Internet, the shortcomings of this loosely controlled operation show up in long queues, inconsistent quality, and bias. Google, which formed a partnership with Open Directory in the spring of 2000, expects the infusion of its technology to ease the cumbersome browsing of the directory's data [Google, 2000].

With the least profitable search engine showing over 75% gross margins on a virtual business with fairly low fixed capital requirements, why do most of these companies lose money? The answer lies in two dominant costs, product development and sales and marketing. As Table 1 shows, product development spending ranges from 9 to 65% of revenue, reflecting the difficulty of keeping up with algorithmic research and developing new services for maintaining

competitive portal sites. With intense competition for consumer spending and a desire to remain standing, expenditures on sales and marketing are seen as essential for driving traffic. The resulting range of expenditures that are 37 to 180% of revenue ultimately suppresses profitability.

IV. CONCLUSIONS

Search engines bring order to a chaotic Web and are indispensable to many of us. Advances in statistical, popularity, and graph-based algorithms are improving the accuracy of indices. A better understanding of the context of indexed pages should help foil the attempts of spammers. Metadata standards, when implemented, will aid in the automatic classification of document contents. Customizable interfaces that utilize natural language programming, personalization, and visualization techniques hold great promise for enhancing both user interaction and the relevancy of search results.

Technology alone, however, will not ensure the success of a search engine. The alternative business models of portals, vortals, and directories offer users a variety of choices for meeting their searching needs, as was shown in Figure 2 in Section I. Each relies on a small set of search technology and content providers, which leads to less diversity in search tools and in results from searches across common domains. If the history of other industries is any indicator, economies of scale and scope will continue to support consolidation in the industry, contributing to a reduction in available search services.

Search services are becoming more personalized in order to improve customer retention. The site where we invest most of our time should come to know us best, lessening our need and desire to go elsewhere. Focused search spaces based on communities of users are gaining prominence, and should become more vital as the Web continues to expand. Intelligent agents that help users navigate this increasingly complex space and guide users to sites of interest to them will become our constant companions. We are only at the beginning of this evolutionary process that will soon make Web searching, as we know it today, a thing of the past.

EDITOR'S NOTE: This article was received on February 20, 2001. It was with the authors 1 month for one revision. The article was published on April 17, 2001.

REFERENCES

EDITOR'S NOTE: The following reference list contains hyperlinks to World Wide Web pages. Readers who have the ability to access the Web directly from their word processor or are reading the paper on the Web, can gain direct access to these linked references. Readers are warned, however, that
1. these links existed as of the date of publication but are not guaranteed to be working thereafter.
2. the contents of Web pages may change over time. Where version information is provided in the References, different versions may not contain the information or the conclusions referenced.
3. the author(s) of the Web pages, not AIS, is (are) responsible for the accuracy of their content.
4. the authors of this article, not AIS, are responsible for the accuracy of the URL and version information.

Ambrosini, L., V. Cirillo, and A. Micarelli. (1996) "A User-Adapted Interface for a Search Engine on the World Wide Web." *WebNet, Toronto, Canada, 1996*.

Bray, H. (June 19, 2000) "Lycos to Hand Off Net-search Business," in *The Boston Globe*, pp. C1. Boston, MA.

Brickley, D. and R. V. Guha (2000) "W3C Resource Description Framework (RDF) Schema Specification, Proposed Recommendation," http://www.w3.org/TR/PR-rdf-schema/ (January 8, 2001).

Brin, S. and L. Page. (1998) "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in *Proceedings of 7th International World Wide Web Conference, Brisbane, Australia, 1998*.

Chakrabarti, S., B. Dom, D. Gibson, S. R. Kumar et al. (1998) "Experiments in Topic Distillation." *ACM SIGIR Workshop on Hypertext Information Retrieval on the Web, Melbourne, Australia, 1998*.

Chandler, A. D. (1990) *Scale and Scope: The Dynamics of Industrial Capitalism*. Cambridge: The Belknap Press of Harvard University Press.

Cho, J., H. Garcia-Molina, and L. Page. (1998) "Efficient Crawling through URL Ordering," in *Proceedings of 7th International World Wide Web Conference, Brisbane, Australia, 1998*.

Cohen, A. (2000) "Why Online Advertising Is Failing," *Sales and Marketing Management* (152) 11, pp. 13.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Laudauer et al. (1990) "Indexing by Latent Semantic Analysis," *Journal of the Society of Information Science* (41) 6, pp. 391-407.

Dvorak, J. C. (2000) "State of the Banner," *Forbes*, December 25, pp. 273.

Fellbaum, C., ed. (1998) *WordNet: An Electronic Lexical Database*: MIT Press.

Glance, N. S. (2000) “Community Search Assistant.” *AAAI-2000 Workshop on AI for Web Search, Austin, TX, 2000*, pp. 29-34.

Google (2000) "Google Extends Award-Winning Search Service with Addition of Netscape's Open Directory Project," http://www.google.com/press/pressrel/pressrelease15.html (April 10, 2001).

Gordon, M. and P. Pathak (1999) "Finding information on the World Wide Web: the retrieval effectiveness of search engines," *Information Processing & Management* (35) 2, pp. 141-180.

Gudivada, V. N., V. V. Raghavan, W. I. Grosky, and R. Kasanagottu (1997) "Information Retrieval on the World Wide Web," *IEEE Internet Computing* (1) 5, pp. 58-68.

Harman, D. (1992) “Ranking Algorithms,” in W. B. Frakes and R. Baeza-Yates (Eds.) *Information Retrieval: Data Structures & Algorithms*, Upper Saddle River, NJ: Prentice Hall, pp. 363-392.

Hendry, D. G. and D. J. Harper (1997) "An Informal Information-Seeking Environment," *Journal of the American Society for Information Science* (48) 11, pp. 1036 - 1048.

Jansen, B. J., A. Spink, J. Bateman, and T. Saracevic (1998) "Real Life Information Retrieval: A Study of User Queries on the Web," *SIGIR Forum* (32) 1, pp. 5-17.

Karlgren, J., I. Bretan, J. Dewe, A. Hallberg et al. (1998) “Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres.” *Eighth Delos Workshop: User Interfaces for Digital Libraries, Stockholm, Sweden, 1998*.

Kleinberg, J. M. (1999) "Authoritative sources in a hyperlinked environment," *Journal of the ACM* (46) 5, pp. 604-632.

Lawrence, S. (2000) "Net Ads Keep On Ticking," *The Industry Standard*) September 4, 2000.

Lawrence, S. and C. L. Giles (1999) "Accessibility of Information on the Web," *Nature Magazine* (400) 6740, pp. 107-109.

Leighton, H. V. and J. Srivastava (1999) "First 20 Precision among World Wide Web Search Services (Search Engines)," *Journal of the American Society for Information Science* (50) 10, pp. 870-881.

Li, Y. (1998) "Toward a Qualitative Search Engine," *IEEE Internet Computing* (2) 4, pp. 24-29.

Liddy, E. D. (1998) "Enhanced Text Retrieval Using Natural Language Processing," http://www.asis.org/Bulletin/Apr-98/liddy.html (April 11, 2001).

Lieberman, H. (1997) “Autonomous Interface Agents.” *ACM Conference on Computers and Human Interface, Atlanta, Georgia, 1997*, pp. 67-74.

Media Metrix (2001) http://www.mediametrix.com/usa/data/metrixcentral.jsp (March 15, 2001).

Pollock, A. and A. Hockley (1997) “What's Wrong with Internet Searching,” in *D-lib magazine*, vol. March.

Rao, R., J. Pedersen, M. Hearst, J. Mackinlay et al. (1995) "Rich interaction in the digital library," *Communications of the ACM* (38) 4, pp. 29-39.

Salton, G. and C. Buckley (1988) "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management* (24) 5, pp. 513-523.

Salton, G. and C. Buckley (1990) "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science* (41) 4, pp. 288-297.

Selberg, E. and O. Etzioni (1997) "The Metacrawler Architecture for Response Aggregation on the Web," *The IEEE Expert* (12) 1, pp. 8-14.

Shneiderman, B., D. Byrd, and W. B. Croft (1998) "Sorting out searching: A user-interface framework for text searches," *Communications of the ACM* (41) 4, pp. 95-98.

Silverstein, C., M. Henzinger, J. Marais, and M. Moricz. (1998) *Analysis of a Very Large Alta Vista Query Log*. Compaq Systems Research Center 1998-014.

Sprenger, P. (1999) "AltaVista Nixes Paid Search," Wired News, http://www.wired.com/news/business/0,1367,20906,00.html (April 11, 2001).

Stary, C. (1999) "Toward the Task-Complete Development of Activity-Oriented User Interfaces," *International Journal of Human-Computer Interaction* (11) 2, pp. 153-182.

Sullivan, D. (2000a) "Major Search Engines…The Major Search Engines," http://searchenginewatch.com/links/Major_Search_Engines/The_Major_Search_Engines/ (April 3, 2001).

Sullivan, D. (2000b) "Search Engine Sizes," http://www.searchenginewatch.com/reports/sizes.html (April 3, 2001).

APPENDIX

The following are considered to be the major search engines based on either the amount of their usage or on how well known they are [Sullivan, 2000a]:

AOL Search	HotBot	Northern Light
AltaVista	iWon	Open Directory
Ask Jeeves	Inktomi	Raging Search
Direct Hit	LookSmart	RealNames
Excite	Lycos	Yahoo!
FAST Search	MSN Search	WebTop
GoTo	NBCi	
Google	Netscape Search	

ABOUT THE AUTHORS

Wendy Lucas is assistant professor in the Department of Computer Information Systems at Bentley College in Waltham, MA. She received her M.S. in Management from the Sloan School of Management at M.I.T., and her Ph.D. from the Electrical Engineering and Computer Science Department at Tufts University. Her research interests include information retrieval and presentation, information visualization, and visual query languages.

William T. Schiano is assistant professor at Bentley College and senior consultant at thoughtbubble productions. His research focuses on the technical and strategic management of electronic commerce initiatives. He is the author of several case studies and articles and a co-author of Cyberlaw (West Publishing, 2000).

Katherine Crosett is vice president at Kalex Enterprises, Inc., a New Albany, Ohio-based Internet consulting firm. Her research involves business model design for start-ups. She holds an MBA from the University of Vermont.

Communications of the Association for Information Systems

ISSN: 1529-3181