International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Issue No. 03, March 2016
www.ijates.com
ijates
ISSN 2348 - 7550

WEB MINING-THE DEEP WEB CONTENT IS HIGHLY RELEVANT TO QUALITY, AND RELEVANCE TO INFORMATION SEEKERS.

Swetha Narayan

Department of Computer Application, Administrative Management College-Bangalore, India

ABSTRACT

Searching on the Internet today can be compared to dragging a net across ocean, most of the web's Information is buried far down on dynamically generated sites and search Engine never finds it. Internet content is considerably more diverse and the volume certainly much larger than commonly understood. he deep Web in a study based on- Public information on the deep Web is currently 400 to 550 times larger than the commonly defined World Wide Web, The deep Web is the largest growing category of new information on the Internet. On average, deep Web sites receive fifty percent greater monthly traffic than surface sites and are more highly linked to than surface sites; however, the typical (median) deep Web site is not well known to the Internet-searching public. More than half of the deep Web content resides in topic-specific databases. more than 200,000 deep Web sites presently exist. our key find include all the deep web content of every information need to market and domain.

Keywords- Analysis of Deep websites, Page views, search engines, surface web , web quality.

II. INTRODUCTION

Most of the Web's information is buried far down on dynamically generated sites, and standard search engines never find it.Traditional search engines create their indices by crawling surface Web pages. To be discovered, the page must be static and linked to other pages. Traditional search engines can not "see" or retrieve content in the deep Web — those pages do not exist until they are created dynamically as the result of a specific search. Because traditional search engine crawlers cannot probe beneath the surface, the deep Web has heretofore been hidden.

To put these findings in perspective, published in *Nature* estimated that the search engines with the largest number of Web pages indexed (such as Google or Northern Light) each index no more than sixteen percent of the surface Web. Since they are missing the deep Web when they use such search engines, Internet searchers are therefore searching only 0.03% — or one in 3,000 — of the pages available to them today. Clearly, simultaneous searching of multiple surface and deep Web sources is necessary when comprehensive information retrieval is needed. this conclude with requests for additional insights and information that will enable us to continue to better understand the deep Web.

II. THE DEEP WEB

1.1 Introduction

Within the strict context of the Web, most users are aware only of the content presented to them via search engines such as Excite, Google, AltaVista, or Northern Light, or search directories such as Yahoo!, About.com, or Look Smart. Eighty-five percent of Web users use search engines to find needed information, but nearly as high a percentage cite the inability to find desired information as one of their biggest frustrations. According to a recent survey of search-engine satisfaction by market-researcher NPD, search failure rates have increased steadily. The genesis of the Deep Web content study was to look afresh at the nature of information on the Web and how it is being identified and organized.(figure 2)

III. BRIEF ON WORKING OF SEARCH ENGINE

Search engines obtain their listings in two ways: Authors may submit their own Web pages, or the search engines "crawl" or "spider" documents by following one hypertext link to another. The latter returns the bulk of the listings. Crawlers work by recording every hypertext link in every page they index crawling. figure 1.

3.1 Surface Web

The surface Web contains an estimated 2.5 billion documents, growing at a rate of 7.5 million documents per day. Today, the three largest search engines in terms of internally reported documents indexed are Google with 1.35 billion documents (500 million available to most searches).see figure 1.

3.2. Hidden Value on the Web-searchable databases

Sites that were required to manage tens to hundreds of documents could easily do so by posting fixed HTML pages within a static directory structure. However, beginning about 1996, three phenomena took place. First, database technology was introduced to the Internet through such vendors as Bluestone's Sapphire/Web (Bluestone has since been bought by HP) and later Oracle. Second, the Web became commercialized initially via directories and search engines, but rapidly evolved to include e-commerce. And, third, Web servers were adapted to allow the "dynamic" serving of Web pages (for example, Microsoft's ASP and the Unix PHP technologies).

It has been said that what cannot be seen cannot be defined, and what is not defined cannot be understood. Such has been the case with the importance of databases to the information content of the Web. And such has been the case with a lack of appreciation for how the older model of crawling static Web pages today's paradigm for conventional search engines no longer applies to the information content of the Internet.

3.3 Analysis of Standard Deep Web Sites

Analysis and characterization of the entire deep Web involved a number of discrete tasks:

1. Qualification as a deep Web site.
2. Estimation of total number of deep Web sites.
3. Size analysis.
4. Content and coverage analysis.
5. Site page views and link references.

6. Growth analysis.
7. Quality analysis.

3.3.1 Deep Web Site Qualification:

An initial pool of 53,220 possible deep Web candidate URLs was identified from existing compilations at seven major sites and three minor ones. After harvesting, this pool resulted in 45,732 actual unique listings after tests for duplicates. Cursory inspection indicated that in some cases the subject page was one link removed from the actual search form. Criteria were developed to predict when this might be the case. The Bright Planet technology was used to retrieve the complete pages and fully index them for both the initial unique sources and the one-link removed sources. A total of 43,348 resulting URLs were actually retrieved.

3.3.2. Estimation of Total Number of Sites.

The basic technique for estimating total deep Web sites uses "overlap" analysis, the accepted technique chosen for two of the more prominent surface Web size analyses. We used overlap analysis based on search engine coverage and the deep Web compilation sites. The technique is illustrated in the figure.

3.3.3. Content Coverage and Type Analysis

- Content coverage was analyzed across all 17,000 search sites in the qualified deep Web pool (results shown in Table 1); the type of deep Web site was determined from the 700 hand-characterized sites (results shown in Figure 6).
- Broad content coverage for the entire pool was determined by issuing queries for twenty top-level domains against the entire pool. Because of topic overlaps, total occurrences exceeded the number of sites in the pool; this total was used to adjust all categories back to a 100% basis.
- Hand characterization by search-database type resulted in assigning each site to one of twelve arbitrary categories that captured the diversity of database types. These twelve categories are:
- Topic Databases subject-specific aggregations of information, such as SEC corporate filings, medical databases, patent records, etc.
- Internal site — searchable databases for the internal pages of large sites that are dynamically created, such as the knowledge base on the Microsoft site.
- Publications — searchable databases for current and archived articles.
- Shopping/Auction.
- Classifieds.
- Portals — broader sites that included more than one of these other categories in searchable databases.
- Library — searchable internal holdings, mostly for university libraries.
- Yellow and White Pages — people and business finders.
- Calculators — while not strictly databases, many do include an internal data component for calculating results. Mortgage calculators, dictionary look-ups, and translators between languages are examples.
- Jobs — job and resume postings.
- Message or Chat .
- General Search — searchable databases most often relevant to Internet search topics and information.

  These 700 sites were also characterized as to whether they were public or subject to subscription or fee access.

3.3.4. Growth Analysis

The best method for measuring growth is with time-series analysis. However, since the discovery of the deep Web is so new, a different gauge was necessary. who searches associated with domain-registration services return records listing domain owner, as well as the date the domain was first obtained (and other information). Using a random sample of 100 deep Web sites and another sample of 100 surface Web sites we issued the domain names to who is searching and retrieved the date the site was first established. These results were then combined and plotted for the deep vs. surface Web samples.

3.3.5. Quality Analysis

Quality comparisons between the deep and surface Web content were based on five diverse, subject-specific queries issued via the bright planet technology to three search engines (AltaVista, Fast, Northern Light) and three deep sites specific to that topic and included in the 600 sites presently configured for our technology. The five subject areas were agriculture, medicine, finance/business, science, and law.

The queries were specifically designed to limit total results returned from any of the six sources to a maximum of 200 to ensure complete retrieval from each source. The specific technology configuration settings are documented in the endnotes.

IV. RESULTS AND DISCUSSION

This study is the first known quantification and characterization of the deep Web. Very little has been written or known of the deep Web. Estimates of size and importance have been anecdotal at best and certainly underestimate scale. For example, Intelliseek's "invisible Web" says that, "In our best estimates today, the valuable content housed within these databases and searchable sources is far bigger than the 800 million plus pages of the 'Visible Web.'" They also estimate total deep Web sources at about 50,000 or so.

60 Deep Sites Already Exceed the Surface Web by 40 Times(few observed websites are listed in Table).

V. FIGURES AND TABLES

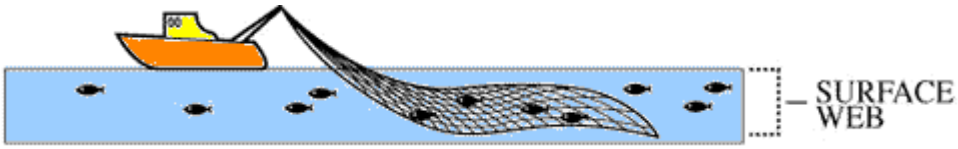

Figure 1: Web Surface

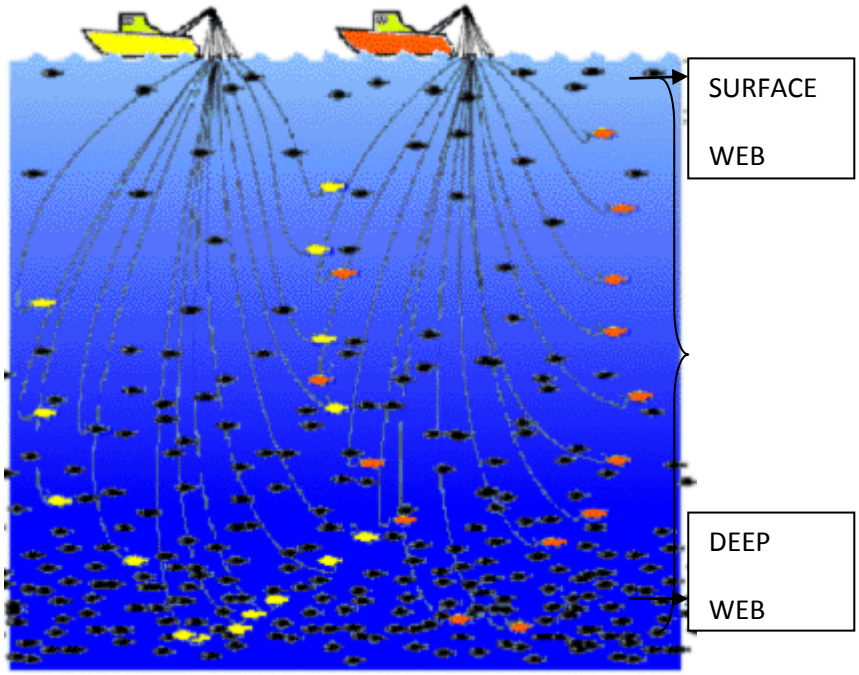

Figure 2: Deep Web

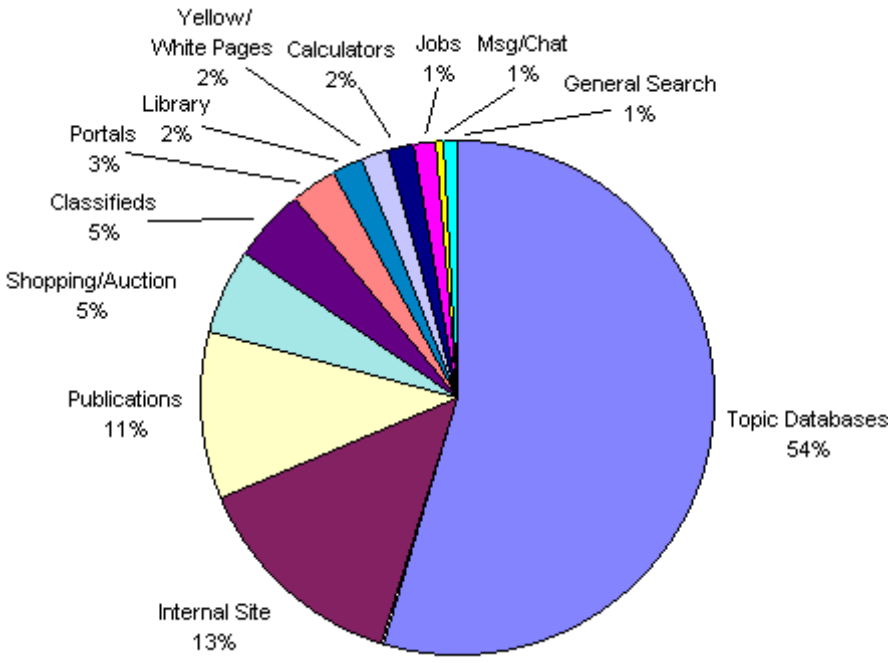

Figure 3: Distribution of Deep Web Sites by Content Type.

Table: Few Deep Sites Already Exceed the Surface Web by 40 Times:

Name	Type	URL	Web size(GBs)
National Climatic Data Centre (NOAA)	Public	http://www.ncdc.noaa.gov/ol/satellite/satelliteresources.html	366,000
Alexa	Public (partial)	http://www.alexa.com/	15,860
MP3.com	Public	http://www.mp3.com/	4,300
Amazon.com	Public	http://www.amazon.com/	461
Adobe PDF Search	Public	http://searchpdf.adobe.com/	143

VI. CONCLUSION

Serious information seekers can no longer avoid the importance or quality of deep Web information. But deep Web information is only a component of total information available. Searching must evolve to encompass the complete Web.

Directed query technology is the only means to integrate deep and surface Web information. The information retrieval answer has to involve both "mega" searching of appropriate deep Web sites and "meta" searching of surface Web search engines to overcome their coverage problem. Client-side tools are not universally acceptable because of the need to download the tool and issue effective queries to it. Pre-assembled storehouses for selected content are also possible, but will not be satisfactory for all information requests and needs. Specific vertical market services are already evolving to partially address these challenges. These will likely need to be supplemented with a persistent query system customizable by the user that would set the queries, search sites, filters, and schedules for repeated queries.

These observations suggest a splitting within the Internet information search market: search directories that offer hand-picked information chosen from the surface Web to meet popular search needs; search engines for more robust surface-level searches; and server-side content-aggregation vertical "info hubs" for deep Web information to provide answers where comprehensiveness and quality are imperative.

REFERENCES

[1]. http://wdvl.com/Internet/Protocols/ and http://www.webopedia.com/Internet_and_Online_Services/Internet/Internet_Protocols/.return to text.
[2]. http://www.google.com.
[3]. http://www.alltheweb.com and quoted numbers on entry page.
[4]. http://www.northernlight.com .

[5]. G Notess, "Searching the Hidden Internet," in Database, June 1997 (http://www.onlineinc.com/database/JunDB97/nets6.html).

[6]. Alexa Corp., "Internet Trends Report 4Q 99".

[7]. B.A. Huberman and L.A. Adamic, "Evolutionary Dynamics of the World Wide Web," 1999; see http://www.hpl.hp.com/research/idl/papers/webgrowth/.

[8]. C. Sherman, "The Invisible Web," [formerly http://websearch.about.com/library/weekly/aa061199.htm].

BOOKS

- K. Bharat and A. Broder, "A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines," paper presented at the Seventh International World Wide Web Conference, Brisbane, Australia
- C. Sherman, "The Invisible Web".

Proceeding Papers

- The technology acceptance model and the World Wide Web.