Fibber McGhee's Internet[1] – *dealing with internet contents chaos*

Carl A. Singer, Ph.D.

Introduction

The full capability of the web as an information repository is nowhere near being realized. There are two fundamental reasons: not properly identifying with the user and drawing the incorrect system boundaries on this opportunity. This is as a result of the following two classic issues:

> A daunting challenge continually facing computer professionals is building systems for "just plain folks." Computer professionals not only tend to live, eat and breathe computers, but they also are heavily engrossed in the systems that they are designing and building. Contrast their system intimacy with that of the casual user and there is a potential disaster.
>
> The search engine is not "the system." The system boundary must include both the web contents, itself, and the search engine. Considering only the search engine misses the mark.

Such is the case today with the web. Although the web is somewhat ubiquitous and today's search engines provide general users with remarkable capability undreamt of even a decade ago, in reality search engines unleash only a small portion of the web's potential. Even as many users are becoming more agile in their ability to search the web there is a fundamental shortcoming – the web contents, itself.

The logical structure of the web contents has largely been ignored. A laissez-faire approach to contents has put the entire burden on search engines, rather than a possibly more effective approach that cooperatively weds explicit contents structure with receptive search engines. Briefly, I believe that there are two reasons for this situation: Technically, it's a tough challenge to design any "best" contents structure. Socio-politically, the web is a "free spirit" and there is something sinister about any attempt to shackle this spirit.

This note informally discusses a contents oriented approach to addressing the current situation.

Search problems

In lieu of a formal taxonomy of search related problems – here's a short list of common problems.

1. Not finding what you're looking for.
2. Getting too much contents upon retrieval.
3. Wading through irrelevant contents.
4. (Purposely provided) Unwanted contents: False contents, advertisements, objectionable material.
5. Time and energy (and aggravation) spent searching.

[1] Fibber McGhee's Closet was a U.S. cultural icon. Fibber McGhee was the protagonist of an "old time" radio comedy show that ran from 1935 to 1959. Whenever he'd go to his hall closet there would be a crash as a cluttered mess would tumble out. Fibber would then proclaim "gotta clean out that closet one of these days."

Examples – Here are several illustrative examples from my recent search attempts. The reader may relate to their own experiences in searching the web.

1. Type in my name, *Carl Singer*, and you'll get 447,000 "hits" using Google; 760,000 via Yahoo. You'll also get many singers (vocalists) named Carl; several people (living and / or dead) who share my name, you'll get hits for all sorts of miscellaneous entries containing "Singer" and "Carl" in close proximity. You may also get a paid advertisement for a book about John *Singer* Sargent written by someone whose first name is Carl. Put quotes around my name, *"Carl Singer"* and you'll get about 1000 hits. Add my middle initial, *"Carl A. Singer",* and you'll get a more manageable 67 hits, but lose some "Carl Singer" entries. Remarkably, in all cases the first "hit" is a German University's (Universitat Trier) Science Bibliography citing an ACM Journal article that I co-authored in 1976[2]. To further complicate matters, there is a Computer Science Professor, Carl ***P.*** Singer, (not me) at DePauw University. Understandably, some databases confuse his "collaborative colleagues" (his co-authors) with mine.
2. I've made a product selection decision and I now wish to execute a purchase (based on net price, vendor reputation, location etc.) and I'm deluged with product reviews.
3. I'm looking for parts for a product that I currently own – I get inundated with sites that want to sell me this product, but few that provide parts or related information.
4. Searching for technical articles on, *"multivariate analysis"* I got some articles, but had to wade through resumes and course descriptions that contain this term. I needed to refine my search by process of elimination, using awkward, contents-oriented constructs to eliminate contents containing "biography" "resume" or "vita" – a time consuming and inexact effort at best. There is no "smart" facility to specify that I want journal articles, only "dumb" word searches.

   Compare this to when I worked at Bellcore (Bell Communications Research – now Telcordia) where we had a full strength research library. I was fortunate to have the assistance of an outstanding research librarian. I could pose a problem and interactively discuss my needs with this librarian. I would then receive a refined selection of articles. If necessary, additional refinement would result from further dialogue. I would then get my articles. Additionally, proper fees had been paid for the reprints, etc. Thus, the research librarian greatly enhanced my intellectual, search-oriented capabilities and also dispensed with administrative issues for me. In contrast today many, if not most, of us are on our own when it comes to locating source materials. We get little feedback beyond the number of "hits" or suggested alternate spellings.

   Again, the search engine has little problem finding contents with the desired phrase, *multivariate analysis*, but unlike my discussions with the research librarian, I currently have no direct way to specify that I want a technical article. There is apparently nothing in the web contents to distinguish between technical articles and, say, course descriptions or resumes.

In each of the above examples there is a shortfall of both context and contents: the inability for the user to clearly communicate the context of their web search activities and the lack of descriptive information to better identify the contents.

[2] Jay F. Nunamaker, Benn R. Konsynski, Thomas Ho, Carl Singer: Computer-Aided Analysis and Design of Information Systems. Communications of the ACM, Volume 19, Issue 12, December, 1976.

Solution Approaches

The purpose of this brief paper is to call attention to an opportunity, not to design a system, thus "solution approaches" are meant only as suggestions and conversation starters.

Context

The user's context might be communicated via a menu choice with categories such as: general information, research, purchase, product information, recreation, etc. A user might build several default context profiles and select from among these. Filtering or modifying the search based on an understanding of the user's context is complex and requires continued exploration.

A conversational "engine" might solicit more context-related information from the user. More ambitiously, "smart" tools could communicate with the user to better learn the context of the search.

Contents

There are at least two solution approaches to addressing web contents shortcoming. The conventional wisdom seems to point towards more powerful, intelligent search engines that, like a human research librarian, can distinguish, for example, between a journal article and a course description. A second approach focuses on contents organization and self-identification. Here is a skeletal suggestion towards this second approach:

Facilitate a header section where contents contributors could freely (voluntarily) describe the context-related characteristics of their submission. Determining what header information might be useful could be done in one of two ways: (1) by committee, or (2) by usage – that is let individual contents contributors use their ingenuity and see what settles out. I recommend this approach.

Another, more sophisticated approach to providing useful header-like information is to have an intelligent agent generate this information upon analyzing (parsing) new contents as it is initially posted. The contents author / submitter might, optionally, have the ability to verify or edit this header-like information. A hybrid of contents provider-entered and intelligent agent generated header information is certainly a possibility.

To illustrate this concept here are examples of voluntary, context oriented contents header information that would synchronize with context enhanced searches:

Contents Type:

Journal, Book, Publication, News, Opinion, Chat, Biography, Resume, Course Description, Commercial, Product Information, Sales, Hobby, HowTo, Games, Musings, Entertainment.

Universality:

Universal, Organization (internal use only) – identify organization, Private / limited

Timeliness / duration:

Creation Date, Timeless, Expires on (*date*), 90 days, 1 year, 3 years.

Contents Provider:

Author Name, Organization

Note: There is some header-like information that is routinely associated with contents in today's web. For example, *creation date* and *language* seem to be ubiquitously associated with contents.

I have informally identified four categories of header information, I am sure that other categories will surface as ideas come forth from the contents providers. A search engine could exploit this additional, voluntarily provided information with compound queries that focus on both header (contents characteristics) and contents. For example: (**Technical Article** *on*) multivariate analysis.

Conclusion

This note is a challenge to the various involved communities to identify and measure the problems and opportunities relating to web contents and retrieval. I do not pretend to have a workable solution – only a partially identified problem, and an SOP (seat of pants) approach to addressing it.