



Read the following material and then answer the questions at the end. Much is discussed on the Internet regarding its growth and user-driven, decentralized nature. This part overviews the current state of searching and search services on the Internet. The essential arguments are that your time is well spent learning how to issue more effective queries and to understand the basic operations of the search services you employ.

Topic 1: Status of the Internet and Searcher's Frustrations

Many have likened the Internet to a huge, global library. While true in some aspects, it has some unique differences. There is no central "card catalog"; the Internet's growth is outpacing the ability of humans or technology to keep up with it; its sheer size is unknown and perhaps unknowable; and content is (to say the least) of uneven quality.

Here's some of what we know (or think we know) about information on the Internet:

- There are from 200 million to more than 320 million documents publicly available on the Internet
- Document growth is, at minimum, doubling each year
- Two-thirds to three-quarters of all users cite finding information as one of their primary uses of the Internet
- Two-thirds to three-quarters of all users cite the inability to find the information they seek as one of their primary frustrations (second only in frustration to slowness of response)
- Of the major search engines, estimated coverage of the documents on the Internet ranges from a high of 34% to a low of 3%
- Combining multiple search engines in a given search can increase the likelihood of finding the information desired by a factor of 3.5 or more
- Different search formats and conventions make it difficult to search multiple engines at one time
- Use of structured, or 'Boolean' queries, while known to help obtain better search results, can be difficult and frustrating for some users to learn.

One of the challenges of the Internet is to make its value available to the millions of new users who have had no formal training or experience in query formulation or search strategies.

Topic 2: Search Engine and Directory Basics

The major search services on the Internet are essential starting points for users seeking information. As such, they routinely are some of the most visited locations on the Web.

Search services can be divided into two groups, commercial and non-commercial. Commercial search services go to the effort to catalog information

RHP Computer Applications Class

5-3 Internet Search Problems

Name: _____



on the Internet to attract attention and advertising revenues. Non-commercial services exist for many different reasons.

Search services on the Internet come in two main flavors: 1) 'search engines' that index words or terms in Internet documents; and 2) 'directories' that classify Web documents or locations into an arbitrary subject classification scheme or taxonomy. Most of the above are examples of the former; Yahoo, Mining Company and LookSmart are examples of the latter.

Search engines use 'spiders' or 'robots' to go out and retrieve individual Web pages or documents, either because they've found them themselves, or because the Web site has asked to be listed. Search engines tend to "index" (record by word) all of the terms on a given Web document. Or they may index all of the terms within the first few sentences, the Web site title, or the document's metatags. Due to the ever-changing nature of the internet, the services must re-sample their sites on a periodic basis. Some of these services re-sample their sites on a weekly or less-frequent basis.

Precision, recall and coverage are limiting factors for most search engines. Precision measures how well the retrieved documents match the query; recall measures what fraction of relevant documents are retrieved. Coverage refers to what percentage of the potential universe of relevant documents is cataloged by the engine. For example, consider a search engine with 10 documents, five of which mention eagles, out of a total universe of 50 potential documents mentioning eagle (45 of which are not indexed by that engine). A query on eagle that returned four documents and two others from this engine would have a precision of 0.66, a recall of 0.80 and coverage of 0.10.

Precision is a problem because of the high incidence of false positives. (That is why you get so many seemingly irrelevant documents in your searches.) This is due to imprecision in the query (searching on eagle and missing the mention of eagles), indexing mistakes by the engine, and keywords entered by the Web document developer that do not actually appear in the document. Coverage is a problem for all engines, with the largest ones only covering at most one third to one half of publicly-available documents .

Search directories operate on a different principle. They require people to view the individual Web site and determine its placement into a subject classification scheme or taxonomy. Once done, certain keywords associated with those sites can be used for searching the directory's data banks to find Web sites of interest.

These distinctions by search service are not clean in all cases. The Excite search engine, for example, uses 'morphological analysis' for determining its keyword matches . While construction of the index is more akin to a search engine, in operation Excite can work like a directory. As other search engines

RHP Computer Applications Class

5-3 Internet Search Problems

Name: _____



begin classifying information into directory-like clusters, these distinctions are likely to continue to get fuzzier.

For searches that are easily classified, such as vendors of sunglasses, the search directories tend to provide the most consistent and well-clustered results. This advantage is generally limited solely to those classification areas already used in the taxonomy by that service. Yahoo, for example, has about 1,400 classifications (excluding what it calls 'Regional' ones, which are a duplication of the major classification areas by geographic region) in its current taxonomy. When a given classification level reaches 1,000 site listings or so, the Yahoo staff split the category into one or more subcategories. If a given topic area has not been specifically classified by the search directories, finding related information on that topic is made more difficult. Another disadvantage of directories is their lack of coverage because of the cost and time in individually assigning sites to categories.

Most searches of a research or cross-cutting nature tend to be better served by the search engines. That is because there is no classification structure behind the listings; only whether the keywords requested appear in that search engine's index database or not.

The flexibility of indexing every word to give users complete search control, such as provided by AltaVista or OpenText, is now creating a different kind of problem: too many results. In the worst cases, submitting broad query terms to such engines can result in literally millions of potential documents identified. Since the user is limited to viewing potential sites one-by-one, clearly too many results can be a greater problem than too few.

Increasingly, the growth of the Internet is causing the specialization or balkanization of search services. Lawyers, astronomers or investors, for examples, may want information specifically focused on their interest topics. By cataloging information in only those areas, users interested in those topics are better able to keep their search results bounded. Such specialization can also lead to more targeted advertising on those search service sites. Again, though, like the directories, such specialization can limit search results to the boundaries chosen by the service, which may or may not conform to the boundaries sought by the user.

The ultimate challenges to any of these centralized search services, therefore, are to: 1) keep pace with explosive document growth; 2) understand the "boundary" needs of their user communities; 3) provide sufficient "intelligence" to infer what users are really asking for even when their queries don't specify it; and 4) ensure sufficient coverage to provide one-stop searching. In the race for eyeballs, user retention and repeat visits are key.



Topic 3: How Search Services Rank Documents

A Web page, or document, can contain various kinds of content (as opposed to display or presentation options like sound, animation or frames), some of which is not shown when you view the document in your browser:

- **Title** – an embedded description provided by the document designer; viewable in the titlebar (it is also used as the description of a newly created bookmark by most browsers)
- **Description** – a type of metatag which provides a short, summary description provided by the document designer; not viewable on the actual page; this is frequently the description of the document shown on the documents listings by the search engines that use metatags
- **Keywords** – another type of metatag consisting of a listing of keywords that the document designer wants search engines to use to identify the document. These too, are not viewable on the actual page
- **Body** – the actual, viewable content of the document.

Search engines may index all or some of these content fields when storing a document on their databases. (Over time, engines have tended to index fewer words and fields.) Then, using proprietary algorithms that differ substantially from engine to engine, when a search query is evaluated by that engine its listing of document results is presented in order of ‘relevance.’ Because of these differences in degree of indexing and algorithms used, the same document listed on different search engines can appear at a much higher or lower ranking (order of presentation) than on other engines.

Though not hard and fast, and highly variable from engine to engine, four factors tend to influence greatly the ranking of a document in a given query:

1. **Order a keyword term appears** – keyword terms that appear sooner in the document’s listing or index tend to be ranked higher
2. **Frequency of keyword term** – keywords that appear multiple times in a document’s index tend to be ranked higher
3. **Occurrence of keyword in the title** – keywords that appear in the document’s title, or perhaps metatag description or keyword description fields, can be given higher weight than terms only in the document body
4. **Rare, or less frequent, keywords** – rare or unusual keywords that do not appear as frequently in the engine’s index database are often ranked more highly than common terms or keywords.

Some engines, notably Excite, attempt to “infer” what you mean in a query based on its context. Thus, the meaning of **heart** can differ if the context of your search is cardiac disease as opposed to Valentine’s Day. The methods by which these inferences are made are statistically based on the occurrence of some words in conjunction with others. Though useful for simpler queries, such inference techniques tend to break down when the subject of the query or its modifiers do



not fit expected query relationships. For commonly-searched topics, this is generally not a problem; for difficult queries, it is a disadvantage to standard full-text indexing.

Cottage industries have emerged to help Web site developers place themselves higher in the search engines' listings (it is clearly more valuable to be within the first few listings sent to a user than be buried hundreds, or thousands, of documents lower). A constant battle is being waged between the engines and those desiring high listings from jimmying the system to "unfair" advantage.

Crude, early attempts to "spam" search engines to get higher listings included adding hidden terms like "sports" that were searched frequently but not the real subject of the document. Other techniques were to use certain keywords repeatedly, such as "cars cars cars cars cars" to get a higher frequency rating. Another was to cram the page with high-interest terms using the same color as the overall Web page, thus "hiding" the added keywords. The leading search engines have caught on to these and now have automated ways to prevent the worst of these spamming techniques.

More subtle techniques, however, are hard to prevent. For example, a listing for ski resorts in Utah could also add hidden tags for "Caribbean" or "beach resort" knowing that wealthy Caribbean travelers may also be looking to take ski vacations. If you as the searcher asked for Caribbean vacations you may logically wonder why you've gotten a listing for Utah ski resorts. It is because of such techniques (among others) that you can sometimes get document listings from a search that seemingly have nothing to do with your query.

So, differences in how search services rank documents, how developer's themselves choose to characterize their Web documents, and just simple errors in how computers process and index these pages can all lead to highly variable ranking results from different search services.

Topic 4: Characteristics of Searchers and What Takes Search Time

Professional information searchers do not have a single style. There is no "correct" way to search on the Internet. Search styles have been described as ranging from 'ants' – the carefully planned, methodical search hoping to get exact results on the first try – to 'grasshoppers' – intuitively jumping from topic to topic, refining results as more is learned . Only you can determine what your style is.

There is only one meaningful measure for a successful search: getting the results you desire. And within that context, there is only one meaningful basis for judging whether one search strategy or another is superior: whether those results are obtained faster.

Surfing and browsing on the Internet are seductive. One begins with an objective in mind, finds new tidbits of interest, and hours later can wonder where the time

RHP Computer Applications Class

5-3 Internet Search Problems

Name: _____



has gone. It is often difficult to apply metrics against whether the original search interest was obtained, or whether the whole process was productive or not. So, let's look at some aspects of a typical search. The example assumes a 56.6 KB modem and a relative "fast" time for the Internet. This is perhaps an optimistic mid-range for current users of the Internet. The example is only meant to be illustrative:

Search Step	Est. Process Time (sec.)	No. Repeats	Total Time (min.)	Cumulative Time (min.)
Formulate Query	120	3	6.0	6.0
Issue Search	10	3	0.5	6.5
Get Search Listings from Service (30/query)	10	9	1.5	8.0
Review Documents; Select for Download	12	50	10.0	18.0
Download Document	15	50	12.5	30.5
Review Document	18	50	15.0	45.5
Average Time per Document (90 document example)				0.5

These estimates are likely an underestimate. Recently, information professionals using the Web to do searches in comparison with traditional online search services like Dialog found it took on average 2.4 minutes per document to get acceptable results.

Whatever the actual "average" search time is, it will not apply to your circumstances in any case. However, what is the case is that certain aspects of searching can add delays to getting desired results and increase frustration:

- No matter how precise or accurate the query, a large percentage of results returned by search services will not be what you're looking for
- Actual search time in getting candidate listings from services is relatively fast; the one-by-one document download and review is the most time consuming part of the process
- Larger listings of candidate documents from the services require more evaluation time
- Often too little time is spent on search and query formulation; any improvements you can make toward more precise and accurate queries will lead to fewer documents to review and faster overall times to the results you want.

RHP Computer Applications Class

5-3 Internet Search Problems

Name:



The essential conclusion is that time is well-spent in understanding how to pose a proper query and how to take advantage of the way that search services work.

RHP Computer Applications Class

5-3 Internet Search Problems

Name:



Assignment

Email me the answers to the following questions at <mailto:rhp@denovodental.com>. In the subject line, type **Compapp 5-3 lastname**.

1. According to the table, what is the average length of time it takes to search for a document per document.
2. Which search engine uses morphological analysis?
3. Name three search engines that are organized by directories that classify Web documents or locations into an arbitrary subject classification.
4. There are hundreds of millions of documents available on the Internet. What is the maximum percentage of documents that are covered by the search engines.
5. Name a search engine that indexes every word of an Internet document.