ranks results are often unavailable to users. Depending on the search service, users may constrain their searches in "Chron job" several ways including: (Unix process) - by using Boolean operators to add or exclude terms - by limiting the range: collection, domain, or language - by specifying date, author, region of origin, or other metadata - by specifying media type Search page "Chron job" (and enters terms) (Unix process) Results page selects a result from The results page normally contains a list of pointers (e.g. URLs) and descriptions of each. Web page The requested page normally

formats the results and returns contains (at least) one of the

words which the user originally entered. But whether or not the page helps the user reach his or her goals depends on a great many factors including: - user's experience with search

- user's experience with the domain
- clarity of the user's goals
- appropriateness of the collections
- how well the search service is maintained

by setting limits on 1) the number of links (or hops) it can travel away from a seed URL 2) the depth of path (or number of sub-directories

in a URL) it will accept

Administrators will restrict the activities of a spider

Index builder List of URLs

Administrators will restrict the activities of an index builder by setting limits on how often it indexes 1) pages that change frequently

(indexed every day) 2) pages that change rarely

Search Engine

Front-end

page generation

system

Administrators configure page

generation systems to define the format of results pages.

Ranked results are

then passed to

Administrators configure a search

engine by defining what indices

it searches, what rules it uses for

matching, and how it ranks results.

(indexed every few days) 3) pages that seem to be static (indexed less frequently)

The expanded list of URLs found by the spider defines a collection of pages which the index builder will index.

re each visited by the indexe

A seed list of URLs define starting points for a spider.

List of URLs

Expanded

Log Files

The spider, the index builder, and the search engine have no notion of a site. Rather they operate on collections which are generally parts of a site (for example, a bug list or forum). Searches on most sites query the indices of multiple collections.

attempt to find search terms in

(may also stem or expand the term

or look it up in a list of synonyms

and then attempts to find the related words too)

Set of all objects in a collection

Set of all objects on servers crawled by service

Set of all objects on the web

Beginning from a seed URL, the spider "crawls" the web collecting links and following to other pages which in turn

(may or may not include metatags)

Object

for each page and thus builds Object record (metadata - kevwords associated with the object)

> Index of collection (a table of keywords, pointers to related object locations - e.g. URLs and other information)

Group of indices of collections

If a search term matches nore than one object (e.g. URL), the results must be ranked

The lexicon is a list of terms (frequently searched) for which the editors have provided preferred URLs that they want "forced" to be first on results pages. Forced results are sometimes sold.

Many algorithms can be used to rank results:

- explicit forcing (as with a lexicon)
- number of mentions of a term on a page
- mention in a title or metatag
- proximity of mentions
- number of links in or out (popularity)
- number of hits received (traffic)

which is sent back to the requester's computer

sends a request to the hosting server for

sends search terms to