Volume 5, Number 1, © JSSE 2006 ISSN 1618-5293

Bert Verstappen (with the assistance of Nadja Ulrich)

www.HuriSearch.org - A Search Engine for Human Rights Information

This article provides a brief overview of search engines and describes how they work. It highlights HuriSearch, a human rights search engine, initiated by HURIDOCS - Human Rights Information and Documentation Systems, International. The article describes the background and aims of HuriSearch and the way in which the project has been implemented. This is followed by a comparison between HuriSearch and the general search engine Google, concerning the relevancy of searches and the depth of crawling. Finally, the future perspectives of HuriSearch are outlined.

Keywords:
Search engine, human rights, HURIDOCS, HuriSearch, human rights education, Internet, Meta tags, Human Rights Information and Documentation System, NGOs, websites of human rights

1 Search Engines[1]

1.1 Introduction

Over the past few years, the Internet has become an important tool to foster Human Rights Education. Its ability to provide cheap and far-reaching information on basically every topic, accessible from every point of the world, makes it powerful. But within its strength lies its weakness: in order to quickly find the right information when you need it, one has to rely on search engines. But the selection criteria of the large search engines do not always correspond to the needs of a person searching for human rights information.

An ever growing number of human rights organisations have turned to the World Wide Web as a powerful and cost-effective medium for informing the world about human rights and their efforts at the grass-roots level to protect and promote them.

Search engines make use of robots, spiders, crawlers, and various other computer programmes that trace hyperlinks across the Web. Robots follow hyperlinks from one document on the Web to the next and they index Web documents and send the results back to the database.

When you search for a specific term in a search engine, an enormous database is checked and the results are presented in a list. Because there is so much information available on the Web, these results may amount to several thousand of so-called hits. This is no problem if what you are looking for is presented among the first 30 or so hits. If this is not the case, most users give up.

Some search engines index more web pages than others, or index web pages more often than others. Some index every word on every page; others only part of the document. Each search engine also has:

- its own system for collecting sites and adding to the database;
- its own system for organizing sites in the database;
- its own way of searching the database;
- its own way of establishing relevance and presenting information.

A small difference in one of these elements leads to differences in the results for the same search on different engines.

1.2 Ranking and Relevancy

Relevancy is difficult to determine on an automated basis, because the concept itself is subjective. Main search engines use panels of human editors to help measure and fine-tune their search results, with different methods used by each engine. Measuring relevance is expensive and slow because of the need for human intervention.

One of the main rules in a ranking algorithm involves the location of keywords on a Web page. Search engines will also check to see if the search keywords appear near the top of a Web page, such as in the headline or in the first few paragraphs of text. They assume that any page relevant to the topic will mention those words right from the beginning.

Frequency is another factor in how search engines determine relevancy. A search engine will analyse how often keywords appear in relation to other words in a web page. Those with a higher frequency are often considered more relevant than other web pages.

Link analysis is used by several engines as part of their ranking algorithm, most notably by Google. By analysing how pages link to each other, a search engine can determine what a page is about and also whether that page is considered to be "important" and thus deserves a ranking boost. In addition, sophisticated techniques are used to screen out attempts by webmasters to build "artificial" links designed to boost their image. Link analysis is not the same as link popularity - it is the quality and not the quantity that counts. In other words: it is more important to get links from good Web pages that are related to the topics you want.

With competitive considerations in mind and to make it more difficult for website owners to manipulate their rankings, developers of search engines do not supply vital information on how search engines search and rank results. However, this information might aid users in determining their search strategies. Information for users displayed on search screens is therefore nothing more than search tips and tricks.

Another obstacle to finding out more about search engines is the fast and frequent changes these engines themselves undergo: they are continuously being upgraded to make them ever faster, ever more precise, and even more advanced than those of their close competitors. The documentation that comes with them frequently lags behind these developments. The users' interests are often secondary.

Meta tags are HTML tags that are written into the head section of an HTML page and provide information about the page, such as the title, the author, a description and keywords. They are intended for computers rather than humans. Meta tags provide web page authors with some influence over which keywords are used by some search engines to index their documents. The "title" meta tag is usually displayed in the description of the document that appears when it comes up as a search engine hit. The "title" and "description" meta tags are used by several major search engines. Many less serious Web developers abused in particular the "keyword" meta-tag, by adding many irrelevant keywords. Meta tags also offer the ability to prevent pages from being indexed at all.

1.3 Types of Search Engines

While previously there used to be a strict distinction between web directories (hierarchically ordered by subject categories) and search engines, nowadays most large, general search engines have built-in subject lists.

Meta search engines offer the possibility to do a search with several search engines simultaneously without having to consult each search engine separately. They function as an intermediary; they pass on the query to the search engines and afterwards order the results.

Traditionally, search engines are thought to be general information resources from which a wide range of information could be extracted based on general keywords. As the Internet became more populated with both users and content, a migration from general information sources to specific information sources is only natural. Specialised or "vertical" search engines are not new but have become more popular, in particular for commercial purposes such as buying and selling products. Trying to find a used car using a general search engine is like trying to find a needle in a haystack. A search engine dedicated to used cars would likely guide the user faster to more accurate information.

2 HuriSearch

2.1 About HURIDOCS

HURIDOCS - Human Rights Information and Documentation Systems, International - established in 1982, is a decentralised global network of organisations concerned with human rights information.

Vision: A world where the power of information and communication is harnessed in the service of human rights

Mission: To strengthen the effectiveness and credibility of human rights organisations and national human rights institutions by enhancing their capacity to manage and communicate information

HURIDOCS facilitates human rights documentation work by:

- developing tools and techniques for human rights monitoring and information handling (e.g. standard formats for information recording and exchange);
- co-operatively organising training courses and workshops on human rights information handling (including collection, organisation, reporting and dissemination); and
- providing advice and support on the management of documentation centres and information systems.

2.2 Background and Aims of HuriSearch

The users of human rights information find themselves in the same situation as other Web user communities, namely one of having to sort through hundreds or thousands of Web pages in order to find those of genuine relevance to human rights work.

Queries for human rights information that are submitted to general search engines provide a considerable degree of less relevant or irrelevant information. Also, general search engines index only certain parts of each site, or have particular selection criteria that exclude several smaller sites, or rank them lower. This implies that in particular sites from non-governmental organisations based in developing countries and societies in transition do not appear frequently in the result lists of such searches - even though these sites often contain the most relevant and up-to-date information.

Several human rights groups have created more or less extensive lists of links to sites related to their own. Some organisations have developed directories of relevant links, searchable by geographical focus and theme. In the framework of the HuriSearch project, HURIDOCS staff compared and accumulated such sources of information on human rights sites. It found that these lists are far from complete and often do not have a clear scope. In addition, many lists are not updated frequently and therefore contain many dead links.

The main limitation of these link lists and directories is that they do not allow searching several sites at once. Users looking for particular information are obliged to browse the World Wide Web by jumping from one site to the other. While this is a useful way to discover what sites are available and what information they contain, it is cumbersome and time-consuming when one searches for particular pieces of information.

The large majority of Web sites of non-governmental organisations are of modest size (from a few pages to 20 Mb.) and often do not contain tools that allow for searching the site.

With regard to searching relevant information on the Web: while larger organisations often have well-trained and full-time staff responsible for this, in smaller NGOs this work is usually done by persons who also have other tasks to fulfill and may not have the necessary expertise and training. HuriSearch seeks to facilitate their retrieval work.

HuriSearch provides a solution to the various problems outlined above by allowing users to search one stop all sites that are included in the project. It is public and free of charge, and available to all persons interested in human rights information: human rights activists, policy makers, researchers and students, journalists and the public in general. The project reflects the increasing interest of the human rights community to use the Internet as a tool for the protection and promotion of human rights. By providing access to information in dozens of languages, it will also be an important tool to enhance human rights education.

HuriSearch aims to:

- Improve the availability of human rights information for all users of human rights information: the NGO community itself, international and national bodies, the media and research community and the public at large.
- Enhance the visibility of all human rights sites, big or small, well-known or not, in whatever language they may be and in whatever region they may be produced or hosted. The project will in particular enhance the visibility of sites of smaller organisations, which are less likely to be retrieved through other search tools.
- Improve the quality of human rights Web sites by providing human rights organisations with advice and training on how to boost the semantic quality of their sites by appropriate usage of existing and emerging Web standards (Dublin core meta tagging, semantic Web, etc).

English is the dominant language of the Web and many organisations in non-English speaking countries provide information in English to serve an international audience. It is important that persons who are not familiar with this language have access to human rights information - this can be considered as a right. HuriSearch is able to handle 77 different languages and can also handle documents and user queries in various non-Latin scripts, including Arabic, Chinese and Cyrillic.

2.3 Selection Criteria

During its first phase of development, HuriSearch included only sites of non-governmental human rights organisations, with the purpose of making their material more visible. The large majority of the almost 1900 sites which were included by May 2005 were found through active searching. Sources included references in e-mails received, existing directories of links, links from individual sites, and other search techniques. Users of HuriSearch could also make suggestions for sites to be included.

The working definition used during this phase for including sites is that they should be sites of non-governmental organisations which list "human rights" among their principal areas of work, or which are undertaking main activities in the field of human rights. The basis was a self-definition of organisations, as expressed on their Websites. Sites of organisations focusing on development issues were included when they declared to work from a human rights approach. Of course, human rights were defined broadly, including civil, political, economic, social and cultural rights. Sites which focus on social justice were also included.

While trying to identify relevant sites, HURIDOCS identified some sites which may require discussion, for example sites with a clear political bias. HURIDOCS intends to establish an Advisory Board for HuriSearch, with recognised human rights experts. This body is to look into issues of inclusion and exclusion as well as other relevant topics.

2.4 Achievements

HURIDOCS explored the possibility of establishing a human rights search engine beginning in 2000. Rather than itself maintaining and hosting HuriSearch, HURIDOCS opted for collaboration with the company Fast Search & Transfer (FAST), the world leader in enterprise search solutions. FAST provides businesses and government organisations with the ability to intelligently and dynamically access, retrieve and analyse information in real time.

HuriSearch has various search options including "use word variants", "exact phrase" and "disable dynamic abstract". Searches can be narrowed down according to size of document, and the results include a "list of related topics" which is useful for refining the search results and also for users who are not English mother tongue and may have difficulties expressing terms precisely.

HuriSearch was publicly launched in July 2003, through announcements on several mailing lists and Websites. It has been tested by legal professionals, NGO staff, researchers and students from various parts of the world. Feedback from users has been encouraging and generally confirmed our ideas about the usefulness of a tool like this.

By July 2005, HuriSearch crawled and indexed the sites of ca. 1,900 non-governmental human rights organisations selected by HURIDOCS from various Web directories and other resources. HuriSearch included over 750,000 documents and has been visited by almost 14,000 users from all over the world. The feedback from these users has been quite promising. A Content Management Board for HuriSearch is being established, which is to provide advice on the criteria for inclusion and exclusion of particular sites. The Board is to contain human rights experts with insight into IT issues and different language backgrounds. The final decision-making authority with regard to HuriSearch is with the Board of HURIDOCS.

2.5 Analysis of the performance of HuriSearch

In September 2004, Mr. Patrick MÃ¼ller, in charge of Documentation and information at the European Committee for the Prevention of Torture in Strasbourg, wrote an (unpublished) article Searching for human right reports - or: What has the turkey to do with Turkey? in which he compared the performance of HuriSearch with that of the general search engine Google. The tests he carried out were repeated on 26 May 2005 under http://www.HuriSearch.org/search/ and http://www.google.com/.

Table 1. Are the hits more relevant in HuriSearch than in Google?

Search	Number of hits in HuriSearch	Comment	Number of hits in Google	Comment
Polizei Deutschland (police Germany)	883	*Very relevant documents, however only from few Websites.	821,000	Some relevant left (e.g. right at the top the "official site of the German police"), in addition, completely irrelevant ones (e.g. a Website "under construction", without contents and indication of author. More relevant hits for a search on "Menschenrech

				te Polizei Deutschland" (human rights police Germany)
"corporal punishment" women	1,413	Above all documents about concrete actions/campaigns on the topic.	186,000	Under the first 20 hits, there are some clearly "unwanted" results. Various other hits are semi-relevant - it concerns particularly trivial to semi-scientific definitions.
Torture	67,226	Very relevant, and large variety of sources (and countries)	12,600,000	The hits are a black-and-white mixture of relevant and unwanted results. The first hit is the Website of the World Organisation Against Torture, the second is the adult site torturegarden, followed by Wikipedia. The first page of results also includes What Torture Method Would You Be?
ÐŸÑ‹Ñ‚Ðº Ð¸ (= "torture"	1,497	Very relevantly, of some	220,000	Some relevant hits from news sites and

in Russian)		selected NGOs		Human Rights Watch. On the first page, there are also non-existing pages and www.deadhouse.ru/gallery/tortures
"right to information"	2,499	Some relevant texts, however also unexpected texts of laws (included in NGO Websites).	210,000	The first page contains a variety of mostly relevant hits, from governments, newspapers, intergovernmental organisations and NGOs. An odd one is the Karnataka State Police Housing Corporation Limited.
racism in Sweden	3,994	To a large extent relevantly, with some less relevant hits (e.g. a site on racism in South Africa, which mentions once "Sweden".	661,000	Mostly relevant NGO sites, research articles and messages.

Conclusion: The "recall" is always substantially higher in Google. However, the "precision" is smaller in Google than in HuriSearch. The relevance of the hits in Google depends largely on the precision of the search terms. General terms (e.g. "right to information") or potentially ambiguous terms (e.g. "degrading treatment") lead to loss of relevance in Google. Google cannot be held responsible for the content of websites it indexes but persons looking for human rights information are often presented with inappropriate material. A striking example would be for persons working on the issue of ill-treatment of women. A search on Google for "women raped in prison" consistently lists pornographic sites on the first results page.

Similarly when carrying out searches on specific topics, e.g. discrimination against specific groups, generalist search engines mix bona fide sites that deal with the issue from a news standpoint or one of respect and tolerance with sites that condone violence or discrimination against the group. HuriSearch indexes peer reviewed sites which results in significantly increased precision.

Two weaknesses are noticeable with HuriSearch. Firstly, different hits of the same Website are given as single hits, while Google is in principle limited to two hits per site (with a link "More results fromâ€¦" wherever relevant). This leads to a higher variety of the results. This will be modified within HuriSearch in the near future. Secondly, at the time of writing HuriSearch (as a matter of choice) is limited to NGOs. Nonetheless other hits found by Google (e.g. News, governmental sites, etc.) also refer to relevant material.

Table 2. Does HuriSearch index more deeply than Google?

Website	**Hits in HuriSearch**	**Hits in Google**	**Deviation Google/HuriSearch**
www.bannet.org	159	140	- 14%
www.ahc.org.al	430	467	+9%
www.dir-info.de	413	402	- 3%
www.sosf.ch	733	31	- 2365%
www.ihd.org.tr	993	955	- 4%
www.amnesty.org	3,941	6,830	+ 58%
www.gisti.org	1,294	6,150	+ 475 %

www.savetibet.org	2,360	17,600	+746%

In considering these results, it should be noted that HuriSearch does not include certain types of files, in particular images, which are included in Google.

Result: The assumed thesis that HuriSearch indexes more deeply, could not be proven. It is interesting that HuriSearch and Google have equal numbers of hits for small Websites, but that Google indexes more pages for large Websites.

2.6 The Future of HuriSearch

With regard to the further development of HuriSearch, HURIDOCS is now planning to create a HuriSearch "one-stop shop" for human rights information.

Firstly, in addition to the sites of NGOs, additional collections will be added for human rights sites of intergovernmental organisations (IGOs), national human rights institutions (NHRIs) and academic institutions. The number of sites of IGOs is over 30 and includes sites of organisations dealing particularly with human rights as well as the human rights sections of sites of organisations with a more general focus. For NHRIs there will be approximately 70 sites of institutions which follow the Paris Principles. The number of sites of academic institutions is about 100 at the moment and will be further expanded. Users will be able to choose in which (combinations of) the four categories they would like to search.

Since October 2005, HuriSearch runs under a new version of FAST Data Search which has additional features for obtaining results by source, country and language. In addition, for each search result a list of most frequently occurring keywords will be compiled, which makes it easy to refine the search. The total number of documents included will be over two million.

HURIDOCS will also set up an Advisory Board for HuriSearch, consisting of members representing the various stakeholders. This Board will oversee the evolution of HuriSearch and give advice when relevant issues arise in relation to its contents, such as the criteria for inclusion and exclusion.

The next stage could be entitled Semantic HuriSearchAchievements: Building the semantic human rights web. In line with recent developments within the Web community, the human rights community needs to adopt semantic web standards for their websites if the web is not to become a bottomless pit of unmanageable and unreliable data. HURIDOCS has already been sensitising the human rights community to the need to adopt meta-data and other semantic web standards for their websites. The "Dublin Core" set of meta-tags have been recommended by HURIDOCS as far back as 1998 and is gradually being adopted by the human rights web community. Meta-tags are used in emerging standards such as the Resource Description Framework (RDF), which is a language for representing information about resources in the World Wide Web and the RSS (Really Simple Syndication) format for sharing news and the content of news-like sites.

Notes

[1] This section is mainly based upon various articles in SearchEngineWatch: www.searchenginewatch.com.