Journal of Universal Computer Science, vol. 1, no. 4 (1995), 247-250
submitted: 22/9/94, accepted: 23/10/94, appeared: 28/4/95, ©Springer Pub. Co.

WAIS and Information Retrieval on the Internet

Helmut Mülner
JOANNEUM Research - IHM, Graz, Austria

1 Abstract

WAIS (Wide Area Information Servers), a development of Thinking Machine Corporation, turned out to be one of the main search engines in connection with the World Wide Web (WWW). This article gives a short overview of WAIS, its history, its basics and some connected developments.

Category: H.5.1

2 Searching on the Internet

The World Wide Web has no inherent facilities to search for informations. All you can do is following links.
But if a beginner browses the WWW (s)he soon will discover that there are lots of pages that mention or offer tools for searching the net.
On close inspection these tools fall into two categories:
Tools to collect information which are usually used to build indexes and tools to query the collected information.
These tools are not integrated in WWW-servers but use a gateway to some other service or program.
Several methods are in use. To name a few of them: grep, perl, archie, netfind, wais, veronica, X.500, whois, finger, ftp to Usenet FAQs and other archives, telnet (hytelnet).
The most important of these services is the interface to WAIS, the Wide Area Information Servers.

3 WAIS

WAIS is is an architecture for a distributed information retrieval system. WAIS is based on the client server model of computation, and allows users of computers to share information using a common computer-to-computer protocol.
It started as a joint effort of Dow Jones News Services ("contents"), Thinking Machine Corporation ("computing power"), Apple Computer ("user interface experience") and KPMG Peat Marwick ("users").
The concept was created in 1989, 1990 the first prototype was ready. 1993 the development leader of WAIS, Brewster Kahle, founded WAIS Inc. to provide commercial WAIS software and services. In Sept. 1994 526 servers were installed worldwide.
WAIS consists of several components: It defines a protocol for communicating queries between clients and servers. It contains an index builder (waisindex) to

collect information. It has a server that ansers queries using the index(es). And there are clients for different platforms.
The WAIS server came in two forms: A commercial server maintained by WAIS Inc., and a free server (freeWAIS) which is now supported by the CNIDR.

4 CNIDR

The Clearinghouse for Networked Information Discovery and Retrieval
Its goals are to

- Promote and Support the implementation and use of networked information discovery and retrieval software applications such as the Wide Area Information Server (WAIS), World Wide Web, the Internet Gopher, freeWAIS, and archie.
- Coordinate to Create Consensus among NIDR applications developers to ensure compatibility and interoperability.
- Disseminate Information about NIDR applications to the network community as well as those active with NIDR applications development.
- Collect or Create Documentation and manuals, Project information, Binaries and source code, Bibliographies and General information.
- Classify Protocol standards and compliance; Identify, classify and integrate noteworthy projects and Identify and cross-reference provider and consumer communities
- Distribute Collected materials and information, Classified materials and information and Educational and research materials

One of the achievements of the CNIDR was the implementation and support of the freeWAIS package. While developing freeWAIS the people at CNIDR concentrated on standard aspects of data exchange protocols which led to better support for the Z39.50 standard. One consequence of this was the renaming of freeWAIS to zdist.

5 Z39.50

Z39.50 - "Information Retrieval Service Definitions and Protocol Specification for Library Applications" - is an American National Standard that was approved in 1988 by the National Information Standards Organization (NISO), an American National Standards Institute- (ANSI) accredited standards writing body that serves the library, information, and publishing communities.
Several companies implemented this standard or variants of this; but it did not develop large scale acceptance.
The WAIS protocol is an approximate implementation of this standard; it includes several extensions and 5 omissions.
Z39.50 is an applications-layer protocol within the OSI reference model developed by the International Standards Organization (ISO). Its purpose is to allow one computer operating in a client mode to perform information retrieval queries against another computer acting as an information server especially in the field

of online library catalogs.
The standard was significantly rewritten for its next version in 1992. One important step in this version of the standard was alignment with ISO 10162/10163, the Search and Retrieval (SR) Service Definition and Protocol Definition. It also incorporated some features of the WAIS protocol.
The next version (Version 3) of the standard was balloted in December 1994.

6 freeWAIS-sf

As the CNIDR concentrated on Z39.50, a group at the University of Dortmund (U. Pfeifer, T. Huynhz) took over the further development of freeWAIS .
They started out in Summer 1993 with bugfixes for freeWAIS-0.202. As they got no feedback from the original developers, they published their own version in September 1994 and name it freeWAIS-sf (sf is for structured fields).
The enhancements included

- field structures (text, date, numeric)
- complex Boolean searches
- stemming
- phonetic coding
- document format specification language
- better installation
- locales
- bug fixes

The package includes detailed instructions for linking to WWW and gopher.

7 Information Retrieval - Basics

One of the achievements of WAIS was that the "general" public of Internet users learned about modern concepts of Information Retrieval (IR).
The classic problem of IR is the balance between recall (defined as number of relevant document that are retrieved by a query divided by the total number of relevant documents), precision (number of retrieved documents that are relevant divided by the total number of retrieved documents) and ease of query formulation.
Boolean queries have many problems, so modern IR very often uses ranked queries with different methods, from simple coordinate matching to vector space and statistical models.
The general idea is that a query is just a (simple) document and the retrieval works by computing the "similarity" of the query document to the database documents resulting in a ranked list of similar documents.

8 Conclusion and Resources

The spreading of the ideas that were the basis of the WAIS retrieval engine can improve the world of the WWW by delivering the means to incorporate sophisticated search engines.
Some resources that should be considered in future developments are:

Managing Gigabytes (mg) a book and freely available software by I.H. Witten, A. Moffat and T.C. Bell.

SMART a system developed by G. Salton and documented in several books and articles.

PAT the commercial system by R.A. Baeza-Yates and G.H. Gonnet that was used in the Oxford English Dictionary project.