Search Facilities for Internet Relay Chat

Taher H. Haveliwala*
Stanford University
Computer Science Department
taherh@cs.stanford.edu

ABSTRACT

The Internet encompasses a diverse array of information sources that have been indexed for efficient search, including the Web, Usenet, and email (both personal mail and specialized mailing lists). One information source, publicly accessible over the Internet, yet unarchived and unindexed, is the Internet Relay Chat (IRC) system. We are archiving several of the more useful technical-support-oriented IRC channels, with the goal of extracting, archiving, and indexing information that would help satisfy users' information needs.

1. INTRODUCTION

Information sources on the Internet can be categorized as depicted in Table 1. Of the major information sources on the Web, most have been archived and indexed.[1] The glaring exception is the Internet Relay Chat (IRC) system. Although a small number of transcripts for special events are occasionally published on the Web by individuals, there is no continuous, large-scale archive of IRC channels available. Techniques were presented in [2] for assisting users in *finding* relevant channels, although the problem of archiving and indexing channels was not considered. We discuss below some of the key issues involved in indexing IRC channels.

Table 1: Taxonomy of Internet Information Sources

	Unarchived	**Archived**
Unindexed	IRC	Web (*Internet Archive [5]*)
Indexed	Web (*Google [4]*, *AltaVista [1]*)	Usenet (*Google [4]*) mailing lists (*Geocrawler [3]*)

2. INTERNET RELAY CHAT

An IRC network consists of a set of independent servers running daemons that exchange messages using the IRC protocol [6]. A user can connect to IRC by using a client program to connect to one of hundreds of IRC servers. Note that there are many independent IRC networks; users can only communicate with other users connected to a server that is a part of the *same* network. Popular IRC networks include Dalnet, EFNet, Undernet, and OpenNet, each consisting of tens or hundreds of individually operated servers. Upon logging on, the user chooses a name, known as a *nick*. The user can view a flat list of available *channels* currently in existence on the IRC network they have joined. The client can join one of the channels (or create a new channel), view the current participants, and take part in the discussion.

*Supported by NSF Grant IIS-0085896
[1]Note that in the case of the Web, current snapshots have been indexed, although the historical archives have not.

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.
ACM 1-58113-513-0/02/0007.

3. INDEXING CHALLENGES

Using an automated script (known as a *bot*) that acts as an IRC client, archiving a set of channels is not difficult. However, various properties of IRC make the archives difficult to index effectively. A few are listed below:

1. **Dynamic channels:** Channels are created on demand. In general, there is no ownership of channel names.
2. **Flat channel organization:** Channels do not exist as part of some categorical hierarchy.
3. **Highly informal:** Because of its real-time nature, communication tends to be highly informal, making it hard to winnow useless chatter from contentful messages.
4. **Multiplexed threads of discussion:** There can be several simultaneously occurring threads of conversation at one time within any particular channel.

As part of our initial investigation, we are archiving some of the more useful technical-support-oriented IRC channels on Dalnet, including `#apache`, `#freebsd`, `#linux`, `#perl`, `#php`, and `#python`. Our goal is to generate and index useful extracts that could provide a valuable resource to users seeking support information.

4. REFERENCES

[1] AltaVista
http://www.altavista.com/.

[2] Neil W. Van Dyke, Henry Lieberman, and Pattie Maes. Butterfly: A conversation-finding agent for internet relay chat. In *Proceedings of the 1999 International Conference on Intelligent User Interfaces*, January 1999.

[3] Geocrawler
http://www.geocrawler.com/.

[4] Google
http://www.google.com/.

[5] The Internet Archive
http://www.archive.org/.

[6] J. Oikarinen and D. Reed. Internet relay chat protocol. *Internet Network Working Group RFC 1459*, May 1993.