Looking for technical reports

Robert Meolic

Abstract— **This paper is about technical reports, a special type of research papers, that are getting more and more popular source of scientific knowledge. Five most important search engines for locating technical reports are described. Some other sources of research papers on the Internet are mentioned, too. Reading the paper you can improve your skills to search for scientific knowledge significantly. This is also the main purpose of the paper.**

Keywords— **computer science, technical report, search engine, digital library.**

I. Introduction

A good research in computer science is usually based on acquiring information from other authors, who are working on similar topics. To enable exchanging information between scientists, many conferences, colloquia, and symposia are organised. Additionaly, technical and scientific journals publish numerous papers every day. One may think that this enormous quantity of information satisfy all the needs of today's researchers. However, in the age of Internet this is not completely true.

Internet enables people from different places in the world to communicate with each other and exchange data quickly and very easy. Therefore, it seems to be an ideal medium for exchanging scientific papers. Some advantages of electronic publishing over classic publishing are:

- low costs,
- easy distribution to very large population,
- enables complex search engines, which help users locating paper with interesting contents,
- better evidence about which papers are the most interesting and who reads them, etc.

However, there also arise some problems with publishing papers on the Internet. Scientific papers need to be reviewed to be trustworthy. In fact, the reviewing process determines the limit of how many papers can be published every year. On the other side, nobody can prevent one to publish her/his paper, even unreviewed, on the Internet. The inspiration for such act can be very similar to that why beta versions and sharewares version of software are giving away. If you publish your ideas often and quickly, then:

- you also get response from others more quickly,
- your ideas have more chances to get considered,
- you have more chances to attract others to your work.

Robert Meolic is with the Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia. E-mail: meolic@uni-mb.si .

In the last decade, many institutions, especially the academic ones, recognize the importance of publishing the knowledge obtained by their researchers. The weaknesses of publishing in journals become very evident. For example, it is not unusual that two or more years pass before a submitted paper is published. Also, presenting papers at the conferences has some disadvantages. Conference fees are becoming very high. Moreover, many conference proceedings have only small circulation and therefore published papers do not reach many other authors. To overcome these difficulties, internal reports, abstracts, extended abstracts, preprints, unreviewed papers, and similar materials are increasingly used for exchanging information between scientists within institution and also worldwide. Nowadays, we refer to this documents usually as *technical reports*.

Most of technical reports available on the Internet have the following common properties:

- they are collected and published by institutions where the authors are employed,
- they have an identification label, which enables one to cite them,
- they are free for download and use.

Although technical reports are usually only locally reviewed, they are an important source of knowledge, especially if you are looking for new topics. Many good ideas appear first as a technical report and after that as a paper in a journal. The main advantage is that technical reports enable one to publish her/his ideas when they occur to her/him and not only at the end of the research. Technical reports also act as an important part of information exchanges when working in a team. They can even be a kind of milestones during the project.

II. Using search engines to locate technical reports

In this section we give an overview of some popular search engines, which can be used for locating technical reports. They can be divided into two groups:

1. general search engines, which look for information on the Internet,
2. special search engines, which look for documents in the collections of bibliographies of scientific literature.

Fig. 1. The Collection of Computer Science Bibliographies

In the group of general search engines there are all popular Internet searchers. Let us list just few of them:

- AltaVista: `http://www.altavista.com/`,
- Google: `http://www.google.com/`,
- GoTo: `http://www.goto.com/`,
- Yahoo: `http://www.yahoo.com/`.

These search engines produce very good results when looking for a particular technical report. They are especially convenient when searching for a paper with a given identification label or for a paper from an unknown institution. You can also find a copy of document, for which the original was removed from Internet.

However, in this paper we are more interested in the group of special search engines. We have inspected the following search engines:

- The Collection of Computer Science Bibliographies (CCSB): `http://liinwww.ira.uka.de/bibliography/index.html`
- CORA - Computer Science Research Paper Search Engine: `http://cora.whizbang.com/`
- The Computing Research Repository (CoRR): `http://www.acm.org/repository/`
- Unified Computer Science TR Index (UCSTRI): `http://www.cs.indiana.edu:800/cstr/`
- Networked Comp. Sci. Technical Reference Library (NCSTRL): `http://cs-tr.cs.cornell.edu/`

Unified Computer Science TR Index (UCSTRI) [8] is a WWW service which provides a searchable index over thousands of existing technical reports, theses, preprints, and other documents broadly related to computer science. This service has been in operation since May 1993 and has enjoyed significant attention [2]. It was an attempt to unify a wide variety of technical documents broadly related to computer science as a searchable index. The entire index currently consists of 14111 items found at 185 different sites. UCSTRI was one of the first attempts to collect indices of contents from technical reports sources. Although UCSTRI is still usable and helpful, it seems not to be maintained very much since 1994 and therefore it is usefull only if looking for a paper in an archive which already existed in that year.

The Collection of Computer Science Bibliographies (CCSB) is located at Lehrstuhl Informatik für Ingenieure und Naturwissenschaftler in Karlsruhe, Germany (Figure 1). The authors currently report the following statistical numbers about the referenced publications: 498487 journal articles, 320440 conference papers, and 110521 technical reports. Each item in the collection is a BibTeX record. In the case of technical reports, BibTeX record contains a link to the full paper, where available. The bibliographies are collected using various Internet search tools and by contributions from individuals. They are automatically converted to BibTeX format. The local BibTeX copies of the bibliographies are updated with every new release of the bibliography collection (about every month). The statistic about accesses to the CCSB shows that this search engine is pretty popular. They noted 2165837 completed requests from Jan 12, 2000 to Sep 28, 2000 (259 days). This is an average of 8859 requests per day. The statistic also shows that in the same period about 50 MB of data was transferred from CCSB per day and that it was used by scientist all over the world.

CORA - Computer Science Research Paper Search Engine is the result of a continuing research project at Just Research (a company located in Pittsburgh, near the campus of CMU) with Carnegie Mellon University graduate and undergraduate students (Figure 2) [7]. CORA is a special-purpose search engine covering computer science research papers. It allows keyword searches over the partial text of Postscript-formatted papers it has found by spidering the Web. Currently, it provides access to over 50,000 research papers on all computer science subjects. The results are displayed by extracted title, author, and abstract. The extraction results are also used

Fig. 2. CORA - Comp. Sci. Research Paper Search Engine

to provide automatically generated BibTeX entries. Citation references are processed to provide forward and backward crosslinks — showing both, papers referenced by the current paper and papers that reference the current paper. The papers are automatically categorized into a "Yahoo-like" topic hierarchy with 75 leaves. The citation link structure is analyzed in order to identify seminal and survey articles in each category.

The Computing Research Repository (CoRR, see Figure 3) started in September 1998 through a partnership of ACM, the Los Alamos e-Print archive (LANL), and NCSTRL (Networked Computer Science Technical Reference Library). It is available to all members of the community at no charge. Everyone has to submit her/his paper manually by email, by FTP, or by using Web interface provided by LANL. It is interesting that they do not accept submissions with omitted figures, tables or sections, nor they accept 'abstract only' submissions. From their viewpoint, such submissions are unhelpful to readers and of very limited archival value. Authors submitting a paper classify their papers in two ways: the first is by choosing a subject area from a list of subject areas and the second is by choosing a primary classification from among the roughly 100 third-level headings in the 1998 ACM Computing Classification System. CoRR is a part of NCSTRL collection and therefore their material can be searched through the NCSTRL form, too.

Fig. 3. The Computing Research Repository

Networked Computer Science Technical Reference Library (NCSTRL) is a common interface to the technical report collections of its (currently over 100) member institutions (Figure 4) [1], [4], [6]. It has been funded by DARPA and the National Science Foundations, with most of the technical work recently being carried out at Cornell University. For the most part, NCSTRL institutions are universities that grant PhDs in Computer Science or Engineering, with some industrial or government research laboratories. NCSTRL has its own viewer, which enables one to view documents without downloading them. NCSTRL is running the very succesful Dienst protocol and software [5].

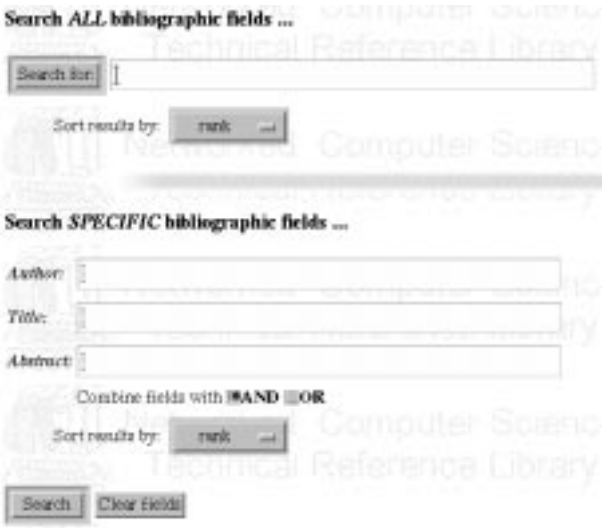

Fig. 4. Networked Comp. Sci. Technical Reference Library

III. A COMPARISON OF SEARCH ENGINES

We made a simple test of search engines. Results of the test are shown in Figure 5. There, the number of documents found is presented for each search engine. Because the reported papers not always exist, we also tested the reliabilty of given links. We checked first 10 hits within each search result. Only the links that directly led to the full paper in any format were considered as the good ones.

The results show that CCSB, CORA, and NCSTRL find much more papers than UCSTRI and CoRR. On the other hand, CoRR and NCSTRL report only papers which are really accessible, while not all links reported by UCSTRI, CCSB, and CORA are usable. We did not check the content of papers. Because of manually submitting and classifying papers we believe that CoRR is somehow superior in comparison to other repositories.

This simple test cannot be a criterion of which search engine is better. It is an individual decision of which search engine to use in a particular situation. Maybe, the best way to get interesting papers is to try all of them.

Search string	UCSTRI	CCSB	CORA	CoRR	NCSTRL
“speech recognition”	6	109	546	37	64
“temporal logic”	2	>170	252	10	167
“database architecture”	21	13	7	0	11
“digital signature”	0	74	35	3	28
“branch prediction”	1	57	89	0	56
“distributed multimedia”	0	60	132	1	63
“text categorization”	0	76	75	7	7
“mobile computing”	3	159	238	0	44
“real-time systems”	44	>170	422	0	233
“parallel computing”	3	>170	449	2	133
$\sum$	80	>1058	2245	60	806
reliability	45%	50%	75%	100%	100%

Fig. 5. A comparison of search engines

IV. Conclusions

Technical reports are an important resource of knowledge in computer science. They enable quick insight and also a broad overview of existing topics. Moreover, they can serve as a good source of references to other scientific publications. In this paper we presented all important search engines for locating technical papers. We excluded only the WATERS project (*Wide Area Technical Report Service*) which appeared in 1992 and seems not to be alive anymore [3].

If you are looking for a technical report from a particular institution, you can also search or browse their index if they have one. For example, a very significant index of technical reports in computer science is that from School of Computer Science at Carnegie Mellon University (`http://reports-archive.adm.cs.cmu.edu/`).

Although technical papers can be very helpful, they cannot replace journal papers. Therefore, you must never forget two important electronic sources of knowledge in computer science: IEEE Computer Society Digital Library (`http://www.computer.org/publications/dlib/`) and ACM Digital Library (`http://www.acm.org/dl/`). In the future, we can expect that huge libraries of informations and knowledge will appear, which will change the way we are looking for data on the Internet. For example, some rudiments of such libraries are LIBERATION Electronic Library (`http://www.iicm.edu/liberation`) and The New Zealand Digital Library (`http://www.nzdl.org/fast-cgi-bin/library`).

References

[1] James R. Davis and Carl Lagoze. The Networked Computer Science Technical Report Library. Technical report, Cornell University, Computer Science, 1996. TR96-1595.

[2] Edward A. Fox. World-wide web and computer science reports. *Comm. ACM*, 38(4):43–44, April 1995.

[3] James C. French, Edward A. Fox, Kurt Maly, and Alan L. Selman. Wide Area Technical Report Service: Technical Reports Online. *Communications of the ACM*, 38(4):45–45, April 1995.

[4] Carl Lagoze. NCSTRL: Experience with a Global Digital Library. In *DL'97: Proceedings of the 2nd ACM International Conference on Digital Libraries*, page 269, 1997.

[5] Carl Lagoze and James R. Davis. Dienst: An Architecture for Distributed Document Libraries. *Communications of the ACM*, 38(4):47–47, April 1995.

[6] Barry M. Leiner. The NCSTRL Approach to Open Architecture for the Confederated Digital Library. *D-Lib Magazine*, December 1998. ISSN 1082-9873, `http://www.dlib.org/`.

[7] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the Construction of Internet Portals with Machine Learning, 2000. Draft accepted for journal publication. Kluwer Academic Publishers, `http://www.cs.cmu.edu/~knigam/papers/cora-jnl.pdf`.

[8] Marc D. VanHeyningen. The Unified Computer Science Technical Report Index: Lessons in indexing diverse resources. In *Proceedings of the Second International WWW Conference*, Chicago, October 1994.

Robert Meolic was born in 1972. He received his B.Sc. and M.Sc. degree in computer science from the University of Maribor, Faculty of Electrical Engineering and Computer Science in 1995 and 1999, respectively. Currently he is working towards a Ph.D. in computer science. His research interest is in the domain of formal verification of concurrent systems.