==
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 18:2 February 2018
India's Higher Education Authority UGC Approved List of Journals Serial Number 49042
==

Towards Creating Virtual Library

Rukhsana Shawl, M.Phil. (LIS). M.LISC.

==

Present Developments – References Search

Whereas a decade ago, the state of the art in a research area could be found out by reading conference proceedings and journals in the local library, nowadays it is additionally necessary to find these electronic publications on the web. Traditional search engines do not help for this task, because they do not index e.g. postscript documents, which is the electronic format of many preprints appearing on the web.

The few existing searchable indices for postscript documents either cover too large fields—all of computer science, for example—to be really helpful, or they depend on some submission procedure which delays the appearance of the documents on the web. A searchable index for scientific papers which is specialised in a relatively small research area and also allows to find the latest new documents.

Three Steps for Efficient Search

This proceeds in three steps. In the first step, a list of names of people who are active in the research under consideration is constructed. This information is obtained from electronic computer science bibliographies, and therefore the names found can be seen as "certified." In the second step, the Home Pages of these people are found. In the third step, these Home Pages are used as starting points for a search engine, which collects scientific papers in the area close to these Home Pages. The documents are used to create a searchable index which is accessible on a web server.

HPSearch

HPSearch is a web-based, topic-oriented information system for finding and watching scientifically relevant personal Home Pages. Personal Home Pages have obtained an important position in scientific communication Preprints and long versions of published papers, project descriptions, course material and contact information can be found there.

HPSearch shows two advantages in comparison to standard search engines:

- System has collect a large amount of topic-specific data. The Home Pages of 75,000 scientists were searched. Their names are mainly provided by the DBLP.
- HPSearch has built up a domain-specific knowledge.

HPSearch works as follows: A set of candidate Home Pages is formed with the help of usual search engines. Each page is rated and a ranking is built. The highest rated candidate Home Pages are visited by an agent and are rated again, a final ranking is built. The result is stored in a database.

The user queries HPSearch by a Web-interface (Figure 2) and receives a hypertext-view of the results. He has also the possibility to start a Web search if the name is not in the database. HPSearch is implemented in Java. In order to maintain the information content the results are reconsidered regularly.

Now, we can see personal Home Pages, rating function and ranking, maintenance of information, architecture of the system, working plan of the search and experiences.

Personal Home Pages

A scientifically personal Home Page is a hypertext page (actually in HTML) which is created by the scientist himself or by his order. Its main function is to present users an overview over his work. The classification of hypertext pages and especially the identification of personal Home Pages is a difficult task. There are not any fixed rules which describe scientists' personal Home Pages.

We determine characteristics with the help of a set of training documents (reference set). These hypertext pages are known as personal Home Pages of computer scientists. A second set of pages (which is determined by the results of the search engines) is used for control (control set). Both sets have a size of 1,000 elements. A characteristic C has to be both

- relevant, a certain percentage R of documents in the reference set have it, and
- significant, the ratio
- percentage of documents with the characteristic C in the reference set
- percentage of documents with the characteristic C in the control set
- has a certain value.

Experiences shows that expedient values are R = 2.5 % and S = 2. Here obtain about 500 such characteristics and highlight the most important facts:

- The probability that certain words appear in different sections on the page is very different and find out that it is reasonable to distinguish the sections title, first header-tag on a page, all other header-tags, labels from links, and other text within a HTML-page. For example the word publications appears in 0.2 % (2.3 %) in the title, 0.7 % (1.0 %) in the first header, 12.2 % (1.4 %) in other headers, 25.4 % (3.4 %) in labels and 16.6 % (5.5 %) in other text in the reference set.
- The name of the person is mostly found in title (89.8 % versus
20.0 % in control set), first header (51.0 % vs. 4.8 %) or URL (44.0 % vs. 3.0 %).
- The -character (65.5 %) or strings like “people” appear (in combination with the name) in the URL.
- There are references to his publications.
- The size of a personal Home Pages is small. The median size in the reference set is 3KB, in the control set it is 9KB.

The domain-specific knowledge itself is changing permanently too. A comparable test from May 5, 1997 shows for example that 46.0 % of the documents have a “-” -character in the URL. Therefore, we repeat the above procedure every month automatically.

Rating Function

From the characteristics worked out, a 2-step rating function is formed. Rating function instead of classical text classification algorithms like naive Bayes, nearest neighbour or decision trees because of the following reasons:

- The structure of the document is hard to handle with pure text Classification.
- The number of characteristics is too high.
- We rate some off-the-page criteria (e.g. URL-structure, search engine provided information, links from other Web-pages).
- In the first step of the search, we do not have the entire document.
- Everything in the system including the rating criteria is highly dynamic. With the automatic update-able rating function the system keeps implementable.
- Since there is a query term, the name of the person searched for, our application is no pure classification problem.

In step 1, HPSearch evaluates the entries received by the search engines: Title, URL, description and position made by search engines ranking algorithms are made available.
After creating a first ranking (presorting), HPSearch reads the best candidate Home Pages, whereby header, links, meta-tags and normal text are evaluated (step 2). A final ranking is built.

Maintaining of Information

Maintaining of information is a primary task of HPSearch. High actuality must be kept. Therefore, HPSearch has two parametrically controllable mechanisms which work periodically and event-oriented. Evaluation of DBLP access-logs: Daily, HPSearch evaluates the access- logs of the person pages in DBLP. Names unknown to HPSearch or names for which the search took place a long time ago are searched for. In a middle-sized time period, a function which depends on the date of search and the score of the best page decides whether a search is started or not. This direct interconnection to the access-log of DBLP guarantees that the data collected by HPSearch is up-to-date.

IF

```
parameter:    SEARCHDAYS1,    SEARCHDAYS2    with
SEARCHDAYS1 < SEARCHDAYS2
input:   name of a person
output:  start Search (search will be started)
IF       (name not in HPSearch ) start Search = true
-ELSEIF       (days since last search > SEARCHDAYS2) start Search = true
ELSEIF        (days since last search < SEARCHDAYS1) start Search = false
ELSE     start Search = f (days since last search, best score)
```

Validation of URLs

Weekly HPSearch checks for each name the URL of the best rated page on existence. Each URL is checked at least twice before it is removed from the HPSearch -database. Then HPSearch starts a new search for that name.

```
Validation o
Parameter:    CHECKDAYS1, CHECKDAYS2 with CHECKDAYS2 = CHECKDAYS1 + 7
Input:   URL
Output:       start Search (search will be started)
IF       (days since last visit of the URL < CHECKDAYS1)
Start Search = false
ELSEIF        (URL is valid) enter new date of visit;
Start Search = false
ELSEIF        (days since last visit of the URL > CHECKDAYS2)
Start Search = true
ELSE     start Search = false
```

It is important to notice that the above algorithms and the determination of characteristics work completely automated.

Architecture

The architecture of HPSearch is shown in figure 3. The core of HPSearch is connected with all components. The arrows depict the data-flow. HPSearch asks queries to search engines which return result pages. Candidate Home Pages are read from the Web.

Results are stored persistently in the database which runs with DB2. A hypertext view is generated to show the results. An export-file for DBLP is generated. With the user-interface HPSearch can be queried.

Working Plan of the Search

The search is started by a user or by the mechanisms.

- The name is ascertained.
- Queries for different search engines are generated.
- Responded pages are parsed, entries are extracted.
- The entries are rated, a ranking is determined (Presorting).
- The best hits are visited and rerated, a final ranking is determined.
- The result is stored in the database.
- A hypertext page is generated.

Beyond the above working plan, there are some minor but important extensions:

- We perform two forms of URL-unification independently:
- A string “index.” or a final “/” -char is not significant.
- Host names are mapped to their IP-number. are unified:
- Expand the set of candidate Home Pages,
- by shorten a URL to a user starting page and
- by adding pages which are linked from pages in the process and the label of that link contains the name of the searched person.

HPSearch provides about 38,000 names for which a probably good Home Page was found. In an evaluation test, HPSearch searches for persons whose Home Page was entered manually in the DBLP. HPSearch found in 84 % of them a correct personal Home Page. In DBLP, only 79 %of manually entered URLs are correct, other URLs were broken or out of date. The best search engine shows the correct personal Home Page in 60 % on position 1. Some users want to send correct or remove wrong personal Home Pages. We have built a form for entering that pages. The manually entered Home Pages from DBLP are included in the HPSearch database, after URL checking.

Mops

Mops is a relatively simple search engine. Its goal is to provide an index of scientific papers found on or close to a given list of web addresses. From these documents, an index is created. It can be searched through a web interface. It answers questions by giving links to the documents containing the search items, their initial lines, and the choice to browse their contents in ASCII format.

Finding and Sorting Documents

The start point of Mops is a list of web addresses. Given a web address, it starts a' search from the given address following the links down to a given depth. It turned out to be sufficient to set the depth to 1 or 2. Therefore, only a small area around each given web address is searched, what makes the search quite fast. Mops collects all postscript, dvi and pdf documents (also in compressed form) found, because scientific papers in computer science are available mostly in these formats.

For each document, the date when it was found, the sequence of links which led to it, and the "name" of the link - i.e. the text <a href=URL> name of the link </a> used by the web page's author for the description of the link - is stored. When a document was found already earlier, it is not collected again, if its last collection date is not too long ago. Anyway, its sequence of links and name is updated accordingly.

Dependent on the web address of the document, it is decided whether it is included into the collection of scientific documents, or into the collection of documents which have to do with lectures, classes, or other, non- scientific matters. Moreover, there are documents which appear under different addresses, e.g. because the web server is accessible with different names, or because a scientist belongs to different research institutes. It is tried to find these “duplicates” and to prevent that they appear more than once in the index.

Creating Index

Each document is stored in ASCII format, which is generated Using pstotext This allows to use glimpse index for creation of an index for each collection. Search through this index is available by a web interface which uses glimpse. Besides searching for documents in which certain terms appear, it is also possible to search directly for links, or to search through the names of the documents. For each document found, its original URL, its description, its initial lines and the first matching lines are shown. It is possible, to browse through the complete text of each document, and to view all the lines matching the query. Each search result can be refined by further queries.

Quality of the Mops Index

Mops consists of a bunch of Perl scripts. It runs on an ordinary PC under a Linux operating system, and it uses lots of freely available software like glimpse, peri, pstotext, cgiwrap, apache. The search engine usually runs for 30 hours during the weekend. Afterwards, the index is updated. Here produced a CD-ROM which contains all the data used for the index search of Mops and the scripts for the Web interface. To install the Web interface on a standard Linux PC takes copying one file from the CDROM to the disk. One is then able to search the complete index from CD-ROM without HPsearch Joins Mops

The quality of the Mops index depends on the list of starting points for the search. Mops searches only a small area around each starting point. If the starting points are too far away from the documents, they either will not be found, or the search takes too much time. Usually, scientists provide links to their scientific papers either on their Home Page or on a neighboured web page. Therefore, Home Pages are good starting points for the search.

The first Mops index was created for scientific papers in complexity theory. A “hand written” list of about 60 web addresses and a list automatically created using the interface to DBLP and HPSearch is used as starting points for the search. This list consists of Home Pages of complexity theorists, and of technical report servers of several universities. Currently, there are about 7,000 documents in the collection. On average, there are 40 new documents found each week.

Another index was created for scientific papers in the BDD area. The first step was to get the names of the scientists in this field. This was managed by querying the DBLP server for publications which have “BDD” or “Binary Decision Diagram” in its titles. The bibliographic data obtained in this way yields names of scientists of the considered field. With these names, HPSearch produced a list of personal Home Pages.

This is used as starting points for the search of Mops. Within two runs of the search engine, about 1000 scientific papers and about 250 other papers were collected. Meanwhile, few “hand written” addresses were added. The index now contains 2,500 papers. Its usefulness can be seen by the fact that the BDD portal has a link to it.

A Case Study: Jewish Virtual Library

The Jewish Virtual Library (JVL) is an online encyclopedia published by the American-Israeli Cooperative Enterprise (AICE), one of whose “principal objectives is to enhance Israel’s image by publicizing novel Israeli approaches to problems common to both our nations and illustrating how Americans can learn from these innovations.” Launched in 1998, it is a comprehensive website covering topics about US-Israel relations, Israel, the Jewish people, and more.

The JVL website was originally created in the late 1990s under the name The Jewish Student Online Research Center (JSOURCE). Since then, the JVL gradually grew in popularity over the years, and in 2012 reported 9.6 million unique visits, and 1.1 million unique visitors in April 2014. Foreign policy expert Mitchell G. Bard is the encyclopedia’s Executive Director.

The JVL relies on hundreds of history books, scientific studies, various encyclopedias, articles, archives, maps, and material from museums for its bibliography, and "takes a scholarly, independent approach" - as companies, individuals and foundations may become sponsors of wings of the Virtual Library. According to the JVL, the Library covers material that cannot be found anywhere else in the world, such as information about joint U.S.-Israel projects, and the treatment of Americans during the Holocaust. It explains that it received permission to use materials from the Library of Congress, from the American Jewish Historical Society, the Anti-Defamation League, the Simon Wiesenthal Center, the Ministry of Foreign Affairs (Israel), and Prime Minister's Office, Rabbi Joseph Telushkin (author of Jewish Literacy), and dozens of other resources.

The Library has 13 wings: History, Women, The Holocaust, Travel, Israel & The States, Maps, Politics, Biography, Israel, Religion, Judaic Treasures of the Library of Congress and Vital Statistics and Reference.

The JVL is constantly updating, changing and expanding, and includes more than 60,000 articles and nearly 10,000 photographs and maps related to Jewish history, Israel, Israel–United States relations, the Holocaust, antisemitism, and Judaism, as well as various statistics, information about politics, biographies, travel guides, and Jewish women throughout history. The website includes the complete text of the Tanakh and most of the Babylonian Talmud.
The JVL contains many articles and studies conducted by AICE, principally involving American-Israeli cooperation. In addition, it has information about Israel education in America, including information about Israel Studies and course materials on Israel-related subjects. It also provides book and movie reviews, a "latest News" page, many publications, and a "Virtual Israel Experience" online project.

Reception

In fact, a PBS web page for the film The Jewish Americans lists the JVL as a resource "For Statistics and Analysis About Jews in America Today", with the description, "A division of the American-Israeli Cooperative Enterprise, the Jewish Virtual Library is a comprehensive online Jewish encyclopedia, covering everything from Antisemitism to Zionism. More than 13,000 articles and 6,000 photographs and maps have been integrated into the site. Their Vital Statistics section has an exhaustive list of current statistics and comparative data." The Jewish Virtual Library has been cited by CNN, New York Times, BBC, CBS News, Fox News, The Los Angeles Times, USA Today, Bloomberg, among others.

It is listed as reference by academic libraries at Pennsylvania State University, Michigan State University, University of Washington, King's College, London, and the University of Delaware. JVL states that it has received awards from Britannica Internet Guide Selection, USA Today Hot Site, and the Best of the Jewish Web from the Jewish Agency for Israel, the Academic Excellence Award from Study Web and others. John Jaeger, in an article published by the Association of College and Research Libraries, said of the JVL: "This library, once it is entered, is more like a living encyclopedia than it is anything else.

One has options to click on, such as history, women, biography, politics, Israel, maps, and Judaic Treasures at the Library of Congress, with each launching a person into a different realm. The site is extremely well put together."[Karen Evans of Indiana State University wrote that the site is comprehensive, with "easily accessible, balanced information".

==

References

1. Bard, Mitchell (June 15, 2001). "Empty slogans don't help Israel – The Wisconsin Jewish Chronicle". Jewishchronicle.org
2. Karen Evans, Jewish Virtual Library at the Internet Reviews Archive, College and Research Libraries News, a division of the American Library Association at Bowdoin College, Oct 2002.
3. Kirkpatrick, David D. "Benjamin Netanyahu News – The New York Times". Topics.nytimes.com.
4. Montopoli, Brian (December 11, 2009). "White House Hanukkah Party Spawns Anger". CBS News.
5. Roberts, Sophie (January 2005). "To be Jewish is to question". BBC.

6. Waters, Rob (March 4, 2008). “For European Jews, Living to 100 Is Partly a Tale of Two Genes”. Bloomberg.

===

Rukhsana Shawl, M.Phil. (LIS). M.LISC.
Librarian
Government Degree College
Ganderbal 191201
Jammu and Kashmir
India
rukhsana54@yahoo.com