

Salticus: Guided Crawling for Personal Digital Libraries

Robin Burke
California State University, Fullerton
Dept. of Information Systems and Decision Sciences
Fullerton, CA 92834
rburke@fullerton.edu

ABSTRACT

In this paper, we describe Salticus, a web crawler that learns from users' web browsing activity. Salticus enables users to build a personal digital library by collecting documents and generalizing over the user's choices.

Keywords

personal digital library, business intelligence, web crawling, document acquisition

1. INTRODUCTION

One solution to the problem of information stability on the World-Wide Web is the creation of a "personal digital library," [1] a personal collection of documents from on-line sources, stored in a local cache and organized for easy access.

The Document Organization and Navigation Agent (DONNA) is a project that seeks to build personal digital libraries to support competitive business intelligence. A personal library, by its nature, does not have the problems of scale presented by public libraries. On the other hand, all aspects of a document's life-cycle must be managed by a single individual.

This paper describes our work in the area of document acquisition. The business intelligence professionals with whom we work use the Web for a large proportion of their information gathering activity, and have a set of strategies for seeking out information of different types. Many of these strategies are not accessible to standard link-based web crawlers. For example, the analysts we studied made heavy use of site-based search engines.

Salticus is a document collection agent that observes a user's browsing and document collection behavior and makes predictions about other useful documents. The user can use these predictions to broaden the collection process and automate it for future visits to the same pages.

2. WEB CRAWLING

The standard approach to document collection and indexing on the web is the use of a web crawler. [2] If the Web is viewed as a graph with the nodes as documents and the edges as hyperlinks, a crawler typically performs some type of best-first

search through the graph, indexing or collecting all of the pages it finds.

This approach is suitable for building a comprehensive index, as found in search engines such as Google or AltaVista. For a personal digital library, we must be more selective. One approach is to focus the crawler on a particular site and mirror its complete contents. Several commercial tools have this capability, for example, WebReaper.¹

This is also unsatisfactory for several reasons. First, many of the documents returned in a complete mirror are irrelevant to the analyst's work. Second, many documents available to a human user are not accessible with link-based crawling at all, because they are accessible only through login pages or other form-based access controls. It should be noted that such controls are often put in place precisely so that documents can be made available to human users and not to web crawlers.

3. SALTICUS

3.1 Document collection

Salticus² is a part of the DONNA system that acts as a proxy between the user and the web. It observes all of a user's browsing behavior including form submission, authentication interactions and cookie creation. To record browsing behavior, the user initiates an "excursion" onto the web with Salticus observing. The system builds a list of the interactions that occur between the user and the various web servers visited, and caches the documents that are accessed. When the user decides to collect a document, it is transferred to the appropriate collection within DONNA.

With this capability, the system operates much like document collecting systems like iHarvest³ that also build personal collections of web documents. Where Salticus differs is in its ability to generalize over document collection actions and to predict other documents to gather.

3.2 Structural generalization

There are several ways that one might generalize about a user's document gathering activity. One possibility is to generalize over the content of links or documents downloaded, building a representation of the user's interests. [3, 4] For this approach to work, the crawler must either rely entirely on the text of the hyperlink to the document as evidence of its content, or it must download every document and analyze it to determine its value.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...\$5.00.

¹ <URL: <http://www.otway.com/webreaper/> >

² Salticus = Search Augmented with Learning Techniques In a Customizable User Spider. It is also the name of a genus of spiders, known for its jumping abilities and advanced eyes.

³ <URL: <http://www.iharvest.com/> >

With the text associated with links not typically very informative, and the prospect of processing every document not practical in an interactive system, we chose a third way, which is to focus on the structure of documents. Salticus has a “Collect” mode, which enables the user to select items to collect on a page without navigating to them. As the user selects items to collect, Salticus builds an XPath [5] representation of each selected link, and it generalizes over these structural descriptions to predict what else on the page might be of interest.

We have found that very simple pre-fix/post-fix generalization is sufficient for the tasks we are supporting. For example, suppose the user has selected three links with the following XPath representations:

```
HTML:BODY:TABLE[1]:TR[1]:TD[1]:A[1]
HTML:BODY:TABLE[1]:TR[2]:TD[1]:A[1]
HTML:BODY:TABLE[1]:TR[3]:TD[1]:A[1]
```

Salticus predicts that the links the user is interested in will be those that exhibit the same variation at the user’s choices. It selects the largest common prefix from these paths and the largest common postfix, and replaces the varying part with “*” or “don’t care” pattern. In this case, the pattern becomes

```
HTML:BODY:TABLE[1]:TR[*]:TD[1]:A[1]
```

which is the first link in the first cell of every row of table 1. If the user instead had selected links in table 1 and in table 2, then the generalization would include both the table and the row.

3.3 Example

Suppose a user initiates a Salticus excursion and enters the address of KMWorld, an on-line publication in the area of knowledge management. Once at the site, the user enters the term “workflow” in the site’s search box. The next page shows a list of 25 documents containing this term. The user selects “Collect” mode in Salticus and clicks on the first several links. At each click, the associated document is retrieved and collected by the proxy, but the original page is still displayed to the user. Then the user clicks on Salticus’s “Predict” function, the system predicts that all of the documents should be collected, and the user accepts the predictions, copying all of the files into the collection.

3.4 Automated operation

Once the user has performed this excursion, Salticus has a record of the steps required to return to the same “place” in the future. Note that with the widespread adoption of database-backed web sites and session-based technologies such as ASP and ColdFusion, the URL has ceased to become a useful identifier for internal pages within a site. A user trying to reissue a URL from a previous session will typically get redirected to the outermost part of the site to login and walk through the application once more.

It is therefore not sufficient to record and replay the URLs that a user has visited. Instead, Salticus makes use of the XPaths recorded for each interaction. When replaying an excursion autonomously, Salticus starts by issuing a URL for the first interaction, but for every subsequent interaction, it follows the XPath to the same location on the page where the user clicked in the previous session. This method is robust in the face of session-based URLs. It is therefore possible for the user to send

Salticus to collect “workflow” related documents from KMWorld in the future, as long as the site is not redesigned.

4. FUTURE WORK

Salticus pattern-based prediction of useful documents has worked well in our informal evaluation of the collecting patterns of intelligence analysts. However, it has some significant limitations. It cannot capture more complex patterns of collection, such as “every third link” or “all rows of only tables 2 and 4”. We are investigating where our current model fails and how more complex predictions might be made.

The problem of automated revisitation brings to the fore the problem of identity: what constitutes a new page? Or a new version of an already-collected page? In the world of session-based URLs, no page will have the same URL that it did when previously visited, even if its contents are the same. We believe that we will be able to use our path-based representation as additional evidence in determining the novelty of documents.

We are also seeking to identify high-level search behaviors, such as the site search engine strategy in the workflow example above, higher-level collection patterns for Salticus to recognize. This would enable the system to make predictions that go across sites and pages, and provide more possibilities for automation.

5. CONCLUSION

Salticus is a document acquisition agent that assists a user building a personal digital library. Salticus tracks user browsing and document collection and generalizes over the user’s collection actions. Salticus’s path-based representation enables it to avoid the problems associated with URLs as identifiers.

6. ACKNOWLEDGMENTS

The DONNA project is sponsored by the FileNet Corporation, a leading provider of document management software, and by the University of California’s DiMI program. Christie Delzeit and Melissa Hanson implemented Salticus’s user interface.

7. REFERENCES

- [1] Rasmusson, A & Olsson, T & Hansen, P. 1998. A Virtual Community Library: SICS Digital Library Infrastructure Project. Research and Advanced Technology for Digital Libraries, CDL’98. Lecture Notes in Computer Science, Vol. 1513. pp 677-678. Springer Verlag.
- [2] Heydon, A, and Najork, M. A. 1999. A scalable, extensible web crawler. World Wide Web, 2(4):219-229, December 1999.
- [3] Chakrabarti, S., van der Berg, M., & Dom, B. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. In Proceedings of WWW8.
- [4] Miller, R. C. and Bharat, K. 1998. “SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers.” Proceedings of WWW7, pp. 119-130, Brisbane, Australia, April 1998.
- [5] World-Wide Web Consortium, 1999. XML Path Language (XPath) Version 1.0. <URL: <http://www.w3.org/TR/1999/REC-xpath-19991116>>