ParaSite: Mining the Structural Information on the World-Wide Web

by

Ellen Spertus

S.B., Computer Science and Engineering, MIT (1990)
S.M., Electrical Engineering and Computer Science, MIT (1992)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1998

Author .....
Department of Electrical Engineering and Computer Science
February 6, 1998

Certified by .....
Lynn Andrea Stein
Class Of 1957 Career Development Associate Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

ParaSite: Mining the Structural Information on the World-Wide Web

by
Ellen Spertus

Submitted to the Department of Electrical Engineering and Computer Science
on February 6, 1998, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

The World-Wide Web is potentially the world's largest knowledge base but only if new information retrieval techniques are developed to take advantage of its unique characteristics, particularly the semi-structured information within pages, across pages, and in page names. Because these types of structure are represented in such different ways, a large number of specialized tools have been required to gather structural information. I provide a relational database interface to the Web called Squeal, which encapsulates these different types of structure in a uniform manner, allowing the user to query the Web in Structured Query Language (SQL) as if it were a database. A novel "just-in-time" interpreter automatically retrieves information from the Web as implicitly demanded by user queries, a technique which could be applied not just to the Internet but to other sources of data too large to be precomputed into a database. The level of abstraction provided by Squeal allows the user to easily create agents that make full use of the previously-untapped information on the Web. One such "ParaSite" is a simple structure-based recommender system that compares favorably to the best text-based system.

Thesis Supervisor: Lynn Andrea Stein
Title: Class Of 1957 Career Development Associate Professor

Acknowledgments

The University of Washington (UW) was my academic home away from home for the last two years. I am grateful to Oren Etzioni and Dan Weld for taking me into their Internet Softbot group. Other UW folk who made me feel at home were Alan Borning, Lauren Bricker, Steve Hanks, Frankye Jones, David (Pardo) Keppel, Ed Lazowska, Sean Sandys, my office-mates Marc Fiuczynski, Jack Lo, Brendan Mumey, Kurt Partridge, and Xiaohan Qin, and of course Keith Golden.

My work benefited greatly from interaction at UW with Oren Etzioni, Marc Friedman, Keith Golden, Nick Kushmerick, Tessa Lau, Marc Langheinrich, Greg Lauckhart, Alon Levy, Neal Lesh, Greg Linden, Kurt Partridge, Mike Perkowitz, Rich Segal, Erik Selberg, Jonathan Shakes, and Stephen Soderland of the University of Washington. I'm particularly grateful to Rich for badgering me into realizing that the user need not explicitly request transfers from the Web to the SQL database.

Alberto Mendelzon, Gustavo Arocena, and George Mihaila of the University of Toronto have generously shared their time, code, and expertise. Their WebSQL system was a major influence on this work.

I also benefitted from code made freely available by Arthur Do (HtmlStreamTokenizer [8]), Santeri Paavolainen (MD5 [30]), Original Reusable Objects, Inc. (OROMatcher [29]), and Sriram Sankar, Sreenivasa Viswanadha, and Rob Duncan (JavaCC [36]).

This thesis is the culmination of many years as an MIT student. Faculty and staff who were especially helpful and encouraging over the years include Bill Dally, who introduced me to research and supervised my bachelor's and master's theses; John Guttag, for repeatedly helping me beyond the call of duty; Tom Knight, who always encouraged me with even my most random ideas; Marilyn Pierce, for her helpfulness and warmth; Jerry Sussman, who is MIT incarnate; and Bill Weihl, whom I could always go to for excellent advice. Other people who encouraged me over the years were Judy Goldsmith of the University of Kentucky and David Lewis of AT&T Research.

From the beginning of freshman year through the completion of three theses, Nate Osgood has provided me with invaluable emotional and intellectual support. I am privileged to have worked with such a wonderful person. I always enjoyed staying with him, Carol Collura and Norm Margolus, Kathy Knobe, my sister Andrea Spertus, and Lisa and Greg Tucker-Kellogg on my visits to MIT.

Philip Greenspun provided me with a much-needed education about databases. Other MIT folk with whom I had valuable discussions about my thesis include Michael Ernst (now at UW), David Karger, Brian LaMacchia (now at Microsoft), Richard Lethin (now at Equator Technologies Consulting), Henry Lieberman and Jim Miller.

I was generously funded by the National Science Foundation for three years of my graduate study and by the Intel Foundation for one particularly crucial year.

I am very grateful to my committee: Lynn Andrea Stein, Ken Haase, Tom Knight, and Pattie Maes. I was privileged to have such a strong and diverse committee, with Lynn's strength in AI, Ken's and Tom's broad backgrounds, Pattie's expertise with agents and collaborative filtering, and all of their interests in novel approaches to information retrieval. I also appreciate help from their assistants: Marie Lamb, Agnieszka Meyro, and Annika Pfluger.

I could not have done this without the support of my family. I am particularly grateful to my father and brother for stimulating my interest in computers and math.

I am not sure I could have done this without the tremendous support and assistance from my advisor Lynn Stein and my fiancé Keith Golden. Lynn took me on as a student under non-optimal circumstances and stuck with me despite the geographic obstacles and the many other demands on her time. She challenged me at just the right level, treating my ideas and work with respect but demanding that I then go and fully develop them. It is a rare teacher who manges to be both rigorous and compassionate. These meant even more to me than her vital intellectual contribution.

Keith Golden helped me in more ways than I could list, including discussing research with me, teaching me about AI, and showing me by example that a thesis could be completed gracefully and thoroughly, but most of all by loving me and believing in me. I feel that I'll be able to do anything I want with him at my side and have a lot of fun too.

Contents

List of Figures

List of Tables