Abstract

Ever increasing amounts of information are available in digital form for use in existing and emerging applications. Data sources, formats and descriptions are accessible in a diversity unimaginable a few years ago. This information seldom comes with a complete specification or schema, even though much of it contains some regular structure. Existing specifications or ontologies, developed separately from the data, are of no direct benefit in organizing such volumes of data. Tools are needed to assist domain experts linking information from diverse and changing sources.

This dissertation presents the SKEIN system which is designed around an algebraic framework. SKEIN is a suite of tools for managing semantic heterogeneity between information sources. The presentation focuses on one large scale repository developed using the algebra. This repository, or nexus, is a graph of dictionary terms related by their definitions as extracted from an on-line Oxford English Dictionary resource. Two algorithms over the nexus provide assistance to experts in domain interoperation. ArcRank computes the most relevant arcs between terms, building on an extension of PageRank. All Pairs Similarity uses ArcRank values to compute which terms have the most similar link structure.

The nexus is a directed labeled graph, four times the size of two other lexical repositories, WordNet from Princeton U. and MindNet from Microsoft Research, but required orders of magnitude less development and maintenance effort. The operators used to build the repository are generic and apply equally well to thesauri, encyclopedias, and other dictionaries. The use of the nexus reduces the effort expended by the expert in matching terms between other sources. Given the task of pairing up English language pages of NATO government websites, SKEIN achieved 70% of the matches obtained by a human expert, without generating any false matches. The nexus and assorted algorithms, when used in the context of the SKEIN system, constitute the first steps towards the systematic interoperation of heterogeneous data sources.

Acknowledgments

More than most endeavors of this type, this work is the result of many people's efforts and patience. First and foremost, I thank my wife Dorothy, without whom none of it would have been possible. There is no doubt in my mind that her love was the fuel that rekindled my desire to complete this work. I owe a debt to everyone who believed in me when I wasn't sure I should believe in myself. Carolyn Tajnai and Suzanne Bentley at the Stanford Computer Forum helped me early on. Nice experiences, like the design of the T-shirt commemorating the 32nd anniversary of the Computer Science Department, are a result of those contacts. Laura Haas at IBM Almaden Research lab gave me new confidence in my ability. The Garlic team provided a great environment to learn to enjoy coding again.

I thank Gio Wiederhold, my advisor, who gave me a great problem to study, and remarkable freedom to pursue various approaches to attack it. Professors Erich Neuhold and Rudi Studer, on sabbatical from Germany, contributed to the creative process. My fellow research group members Prasenjit Mitra, Vasan Pichai and Danladi Verheijen forced me to stop taking shortcuts when explaining ideas. Conversations with Mark Musen, Harold Boley and Martin Kersten sharpened the technical arguments. Interaction with Stefan Decker helped close gaps in my work. True friends, such as Narayanan Shivakumar, Luca de Alfaro, Sudarshan Chawathe, have helped me in both good and bad times. My friends at Gigabeat Inc. allowed me to develop my ideas in a commercial setting.

It is only fitting to remember the example set by my managers and coworkers at the Toshiba R & D center in Kawasaki, Japan, in particular, T. Kodama, T. Kamitake, and Y. Shobatake. It was their thorough knowledge of the networking field that first inspired me to consider graduate studies. I recognize my brother Jean-Luc, who just completed his own Ph.D., and from whom I learned the virtues of healthy competition. Last but not least, I thank my parents, who have always shown me love, and still provide moral support.