

ITR/IM+SOC+HCI GEOWEB SUMMARY

Our proposal is based on the premise that the whereabouts of individuals, organizations, objects, and activities that are referenced on the Web will remain of prime interest to those who seek to learn about them. We intend to develop methods to make this valuable but often "buried" location information easily accessible. As more of our routine activities are coupled to the Web, a growing number of queries and searches will include a geographic component.

Geographic context on the Web takes on many forms, from simple postal addresses to the venue of some event (2002 Winter Olympics: *Salt Lake City*) or a focus of interest (Elian Gonzalez: *Florida* and *Cuba*). It can be discovered using templates, by constrained text analysis coupled with standard geographic sources, by tracing in and out links, by analyzing the pattern of accesses, and occasionally even from the location of the server.

To determine the prevalence and nature of geo-references, we are already testing a geographically-unbiased sampling scheme. Examination of the resulting sample will guide the development of algorithms for extracting geographic information. We will construct an indexing scheme that reflects geographic focus and extent. Although functional mark-up languages like XML allow authors to index their own pages, our goal is to accomplish this automatically, post facto, and without reliance on the author.

Currently we envision indexing using ever-active crawlers or agents, but the index will have to be combined with existing subject hierarchies at search time. The index will be incorporated in a prototype search engine that will constrain searches by both topical and geographic scope of the query. Efficient implementation of the joint constraint is itself of interest. We will also initiate research on evaluating the utility of hidden geo-information for education, social-science research, business, and mundane personal queries.

Our team has complementary expertise in pattern recognition and machine learning, natural language analysis and information retrieval, database theory and data mining, network mapping, and in geographic analysis. We have collaborated in various combinations for many years.

Almost all of the requested budget is for supporting students to make the transition from undergraduate to graduate research. The nature of the project is attractive to students of all ages and origins. They will be exposed to the research approach of four different departments at three universities. We will ensure that every student will spend several weeks in at least two of the partner schools. We will maintain a Web page on ethical issues related to the project, including privacy, security, intellectual property, and Web authors' and reviewers' responsibilities.