Intelligent Agents for Organizing the Web

Susan Gauch
University of Kansas

Introduction

- Information overload on the Web
- Many possible search engines
- Need intelligent help to
  - select best search engine
  - customize results for the user
  - browse the Web

Searching Overview

- Many search engines
  - different spiders
  - different retrieval algorithms
  - different results
- Which to use?
  - differs depending on query
  - generally want information from more than one

Distributed Agent Approach

- ProFusion is an Agent-based meta-search engine which communicates with multiple, distributed search engines
  - http://www.designlab.ukans.edu/profusion
- Routes user queries to most appropriate search engines
- Communicates in parallel
- Fuses results returned

Architecture

- Knowledge Sources
  - no private index
  - meta-knowledge about strengths of search engines with respect to a collection of categories
  - lexicon which associates word with the same collection of categories

Architecture (cont.)

- Agents
  - one Master Agent which controls search
    - » routes query to most appropriate search agents
    - » fuses information returned
  - one Faciliator Agent per search engine which communicates with it
  - one Information Filtering Agent which identifies new information for registered users

Query Routing

- for each word in query
  - use lexicon to map from word -> categories
  - use meta-knowledge to map from categories -> top three search engines
- if no query word are in dictionary, use default best three

Fusing Search Results

- Rank order results
  - normalize scores for all retrieved urls
    - » search engines report match values differently
  - multiply score by confidence factor for each search engine
    - » average value of performance over 13 categories
  - rank order based on result
- Remove duplicates and broken links

Evaluation

- 12 graduate students in Computer Science
- Select own queries (mostly C.S.)
- Run on each of 6 basic + 2 meta-search engines + ProFusion (manual and auto)
- Give relevance judgments of top 20 urls retrieved by each (240 urls judged * 10 search engines = 2,400 urls judged)

Evaluation (cont.)

Search Engine	#rel	#broken	#rel. dup.	Precision
Alta Vista	108	31	9	0.41
Excite	129	7	7	0.51
InfoSeek	99	16	12	0.36
Lycos	119	17	26	0.39
Open Text	72	32	18	0.23
WebCrawler	92	18	19	0.30
MetaCrawler	98	24	13	0.35
SavvySearch	127	29	15	0.47
Manual Pick 6	142	13	8	0.56
Auto Pick 3	170	19	4	0.69

Evaluation (cont.)

- ProFusion
  - Highest precision by far (69% vs 51% Excite)
  - Automatic Pick better than Manual Pick
    - » 69% versus 56%
  - comparable broken links to most (19) except Excite (7)
  - Fewest duplicates (4/170 = 2.3% relevant duplicates versus 7/129 = 5.4% for Excite)

Future Work

- Need to provide user-tailored tools to
  - browse personal collections
  - browse the Web
- Need to actually distribute the search
- Approach is centered around a global taxonomy (for the Web) which is tailored for individual sites and users

User Agents

- index own collection to support personal searching
- characterize personal collection with respect to personal taxonomy
  - meta-knowledge for personal information agent
    - » find and filter new information
  - basis of browsing contents of personal collection

Site Agents

- index own site to support local searches
- characterize site with respect to global taxonomy
  - meta-knowledge for routing queries to this site
  - basis of browsing contents of a specific site

Web Agents

- collect meta-information from Site Agents
- route queries to most appropriate sites for distributed processing
- browse Web via meta-knowledge (taxonomy of sites/pages automatically collated from collected meta-information)