Data Discovery on the Information Highway

Susan Gauch
University of Kansas
Introduction

• Information overload on the Web
• Many possible search engines
• Need intelligent help to
  – select best information sources
  – customize results
  – browse the Web
  – handle non-textual information

University of Kansas
ProFusion: Searching the Web

• 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 Broker Agent which controls search
    • routes query to most appropriate search agents
    • fuses information returned
  – one Facilitator Agent per search engine which communicates with it
  – one User Information Filtering Agent which identifies new information for registered users
Figure 1. Agent Intercommunication and Control Flow Diagram

University of Kansas
Dispatch Agent: 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
Dispatch Agent: Fusing 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
Search Agent

- encapsulate knowledge for each underlying search engine in a “competence module”
- map from standard query representation to specific syntax for each search engine
- connect to, and receive results from, search engines
- parse result page and extract contents into standard format (URL, weight, title, summary...)
- normalize weights
University of Kansas
Results from your search: "Pentium memory prices"

Altavista contributed 10 items.
Excite contributed 10 items.
HotBot contributed 9 items.
Retrieved 27 unique item(s).

To automatically receive updates on this subject, click here.

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Title</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>Prices: Pentium 133Mhz</td>
<td><a href="http://www.colemic.com/prices/p5-133.html">http://www.colemic.com/prices/p5-133.html</a></td>
</tr>
<tr>
<td></td>
<td>Summary: CII Sterling Pentium 133Mhz Mini-Tower or Desktop Case Intel Motherboard w/512K Pipeline CACHE. Intel Pentium 133Mhz CPU. (Upgradeable to a P5-200 Intel.</td>
<td></td>
</tr>
<tr>
<td>0.9500</td>
<td>Denny's Prices : Pentium Systems</td>
<td><a href="http://www.nightowl.net/~dfry/prices1.html">http://www.nightowl.net/~dfry/prices1.html</a></td>
</tr>
<tr>
<td></td>
<td>Summary: Denny's Prices : Pentium Systems. To place an order call at Phone (314-285-4434) or send me Email at : Denny Fry. These Pages are designed and maintained.</td>
<td></td>
</tr>
</tbody>
</table>
Learning Agent: Adaptation

• adapt to network load
  – monitor and set individual time-out values

• adapt to broken search engines
  – identify down search engines
  – prevent them from being selected
  – invoke guarding agent to periodically check status
Adaptation (cont.)

• adapt to changing search engine protocol
  – generic pattern matching grammar for parsing search engine results

• adapt to changing search engine performance
  – automatically calibrate quality of search engine results in each category
  – adjust confidence factors based observations of user behavior (which item in ranked list they select first)
User Agents: Personalized Search

• Users may register personal queries with ProFusion to be automatically re-run on a periodic basis

• Query results are presented in three categories
  – new
  – relevant
  – possibly relevant
ProFusion: Current Thrusts

• index own collection to support searching personal collection
• characterize personal collection with respect to personal taxonomy
  – basis of browsing contents of personal collection
• incorporate user’s feedback to filter out and prioritize new results
Extension: Distributed Search

• currently, spiders collect all information centrally
  – lots of traffic, disk space, overloaded sites
  – “supermarket” approach
• dispatch queries to “best” sites
  – “specialty store” approach
• challenges
  – identify the best sites for each query
Distributed Search: 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
Distributed Search: Brokers

- 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)
Discovering Video Information

• VISION: Video Indexing for Searching Over Networks
  – create a database of video clips indexed by their associated closed captions
  – locate related information via Web searching to augment video clips

• Goals: entirely automatic, real time
Figure 1. The architecture of the VISION Digital Video Library
Summary

- many sources of information
- need a consistent interface to locate information regardless of
  - where it is
  - what format it is in
- one source is not enough
  - locate and fuse information from multiple sources