From search engine to hypertext: a spectrum of possibilities

Peter Brown

Computer Science
Univ. of Exeter, UK

Background to work
Links with Southampton University. These are early ideas. Extremes of retrieval
Unconstrained: search engine Constrained: a hypertext page including, say, 5 links to other pages that supply further information. We often want retrieval between these extremes .. .. especially when using a small screen. Points on the spectrum
1: Full IR, e.g. search engine. 2: Information Filtering, which is like IR but: user's query is supplied in advance. delivery of documents is proactive. 3: Site search on the web; more generally searching a sub-hierarchy. Points on the spectrum (cont.)
4: Context-aware retrieval: may be proactive, e.g. delivery of information as tourist nears a site. may be interactive, but query is wholely or partly supplied automatically. 5: User-tailored

retrieval: similar in principle to previous, but with tailoring of information to a user (e.g. preferences, past history, bookmarks). 6: Dynamic hypertext links: source and/or destination of links are not fixed, but are calculated by a function. Points on the spectrum (cont.)

7: Composite retrieval, e.g.: (1) do an ordinary retrieval. (2) apply one or more 'filters' to the results of (1). Perhaps carry this further and have several autonomous retrieval agents + a consolidator. 8: simple web links: the retrieval has been done in advance by the author, who has provided links to the most relevant documents. author has a piece of context: the web page the reader is looking at. Underlying mechanism: a three-stage retrieval process

(1) specification of (1a) document collection and (1b) query. (2) issue of

query to retrieval engine; delivery of relevant documents. (3) selection by user of a retrieved document. ... plus possible further step of marking relevant document fragments, and explaining *why* **each is relevant. There may be iterations round these stages, e.g. building and using a cache involves iterating over (1) and (2) above. Timing** *For large amounts of information, some elements of (1) to (3) above must be done in advance in order to get acceptable performance.* **.. but subsequent change then causes problems. Dimensions of the spectrum**

Queries: permitted forms, who specifies, who issues. Document collection: who specifies. Iterations. Timing: which stages done in advance. Delivery: proactive/interactive. Most dimensions involve a choice between freedom and constraint. Current applications typically pick certain points on the

spectra, but provide some mechanisms to ease/impose constraints. Example: easing of constraints within hypertext system
Use of associated glossary or dictionary. 'Find me a web page like this one'. Just-in-time retrieval agents: annotating the current document with automatically-created links to relevant material; also Autonomy products. More generally, using *intensional* rather than *extensional* links. Proposal: the read/write interface
Do not distinguish between browsers (for reading) and editors/word-processors (for writing). Thus you have a browser that allows annotation, deletion, highlighting, etc., by the user. Current read/write document is called the *document of interest*, and is one part of the user's context. The document of interest can be used as a basis for retrieving further information the user

wants. Proposed future application: SUPERIF

Information Filtering, but documents are found by a resource-discovery agent, which knows about user interests. Uses read/write interface. Queries are adapted automatically according to how the user reacts to the documents delivered; thus queries evolve slowly over time. Caters for different user 'hats'. Aims to provide proactive delivery of all the new information the user needs. Proposed future application: SUPERIR

SUPERIR is an IR system that only does context-aware retrieval (but document of interest is part of this context). Uses read/write interface. Has a set of predefined document collections prepared in advance (like set of web pages used by search engines). Difference between SUPERIF and SUPERIR

Query known in advance vs. document collection known in advance. Conclusions

We have tried to find automatic processes that help find relevant documents without user intervention. One key to this is the use of context. Another key is the read/write interface. Some operations must be done in advance. The more automation there is, the more a user needs a facility to find out why a particular document has been delivered.