Selecting Task-Relevant Sources for Just-in-Time Retrieval*

David B. Leake and Ryan Scherle
Computer Science Department
Indiana University
150 S. Woodlawn Avenue
Bloomington, IN 47405
{leake,rscherle}@cs.indiana.edu

Jay Budzik and Kristian Hammond
The Institute for the Learning Sciences
Northwestern University
1890 Maple Avenue
Evanston, IL 60201, U.S.A.
{budzik,hammond}@ils.nwu.edu

Abstract

"Just-in-time" information systems monitor their users' tasks, anticipate task-based information needs, and proactively provide their users with relevant information. The effectiveness of such systems depends both on their capability to track user tasks and on their ability to retrieve information that satisfies task-based needs. The Watson system (Budzik *et al.* 1998; Budzik & Hammond 1999) provides a framework for monitoring user tasks and identifying relevant content areas, and uses this information to generate focused queries for general-purpose search engines and for specialized search engines integrated into the system. The proliferation of specialized search engines and information repositories on the Web provides a rich source of additional information pre-focused for a wide range information needs, potentially enabling just-in-time systems to exploit that focus by querying the most relevant sources. However, putting this into practice depends on having general scalable methods for selecting the best sources to satisfy the user's needs. This paper describes early research on augmenting Watson with a general-purpose capability for automatic information source selection. It presents a source selection method that has been integrated into Watson and discusses general issues and research directions for task-relevant source selection.

Introduction

As the volume of available information grows, the burden of information access grows as well. "Just-in-time" (JIT) information systems address this problem by shielding the user from the information access task. Instead of requiring a user to recognize the need for information and initiate queries to satisfy it, these systems observe the user's actions in a task context, anticipate the user's information needs, gather the needed information and present it to the user before the user requests it. Such systems require methods for (1) determining the type of information the user requires, and (2) focusing retrieval on information that satisfies the user's needs.

There are now large numbers of focused information sources on the web, providing a rich range of specialized information aimed at satisfying particular information needs.[1] Finding the right sources itself requires expertise, limiting the usefulness of these systems for non-expert users. However, if their information can be provided automatically, this drawback is nullified. This paper describes ongoing research on enabling just-in-time information systems to automatically select information sources that are appropriate to the user's needs.

We are investigating this problem with SourceSelect, a source selection system integrated with the Watson system (Budzik *et al.* 1998; Budzik & Hammond 1999). Watson automatically fulfills users' information needs by monitoring their interactions with everyday applications, anticipating their information needs, and querying Internet information sources for that information. The initial version of Watson focuses on identifying task-relevant content areas and automatically generating content-relevant queries for general-purpose search engines and a small set of specific search engines associated to particular query types by hand-made strategies. SourceSelect provides an initial approach to adding a general-purpose capability for identifying and accessing content-relevant search engines. Given a query from Watson, the system does a two-step retrieval, first using vector-space retrieval methods to associate queries to relevant sources, and then using automatically-generated queries to guide search within those sources. In the combined system, Watson monitors user activities, identifies relevant content areas, and provides SourceSelect with context information. The SourceSelect system determines appropriate information sources, formulates queries to those

*David Leake gratefully acknowledges the support of the InfoLab and Computer Science Department of Northwestern University during his sabbatical leave. His research is supported in part by NASA under award No NCC 2-1035. Ryan Scherle's research is supported in part by the Department of Education under award P200A80301-98. The research of Kristian Hammond and Jay Budzik is funded by a grant from the National Science Foundation, McKinsey and Company, and gifts from Microsoft Research.

[1] For a sampling of some of these, see The Scout Report (http://wwwscout.cs.wisc.edu/scout/report).

sources, sends off those queries, and collates their results for Watson to pass them on to the user. No user intervention is required to target candidate sources.

The paper begins by sketching the Watson framework and discussing the value of specialized information source selection. It next describes the system and the source selection methods it implements. It then discusses central issues for intelligent source selection and how the approach relates to other current approaches.

Just-in-Time Information Access: The Watson Framework

The Intelligent Information Laboratory (InfoLab) at Northwestern University is developing a class of systems called Information Management Assistants (IMAs). These systems observe users as they go about completing tasks in everyday software applications and uses its observations to anticipate the user's information needs. They then automatically fulfill these needs by querying traditional information sources such as Internet search engines, filtering the results and presenting them to the user. IMAs embody a just-in-time information infrastructure in which information is brought to users as they need it, without requiring explicit requests. Essentially, they allow these applications to serve as interfaces for information systems, paving the way for removing the notion of query from information systems altogether.

The first IMA developed at the InfoLab is Watson, an IMA that observes user interaction with applications such as Netscape Navigator, Microsoft Internet Explorer, and Microsoft Word. From its observations and a basic knowledge of *information scripts*—standard information-seeking behaviors in routine situations—Watson anticipates a user's information needs. It then attempts to automatically fulfill them using common Internet information resources.

The conceptual architecture for IMAs has four components (Budzik *et al.* 1998):

- The ANTICIPATOR uses an explicit task model to interpret user actions and anticipate a user's information needs.
- The CONTENT ANALYZER employs a model of the content of a document in a given application in order to produce a content representation of the document the user is currently manipulating.
- The RESOURCE SELECTOR receives the representation produced by the CONTENT ANALYZER and selects information sources on the basis of the perceived information need and the content of the document at hand, using a description of the available information sources. In most cases, this results in an information request being sent to external sources. A result list is returned in the form of an HTML page.
- The RESULT PROCESSOR interprets and filters the result list. Results are gathered and clustered using several heuristic result similarity metrics, effectively eliminating redundant results (due to mirrors, multiple equivalent DNS host names, etc.). The resulting list is presented to the user in a separate window.

The above mechanism allows Watson to suggest related information to a user as she writes or browses the Web. Watson observes user interaction with Microsoft Word and Internet Explorer, and uses information sources ranging from general-purpose information repositories such as newspaper archives or AltaVista, to special-purpose information sources such as image search engines and automatic map generators.

When a user navigates to a new Web page, Watson suggests pages related to the topic of the page at hand. Similarly, as a user composes a document in Microsoft Word, Watson suggests Web pages on the topic of the document she is composing. This is illustrated in Figure 1.

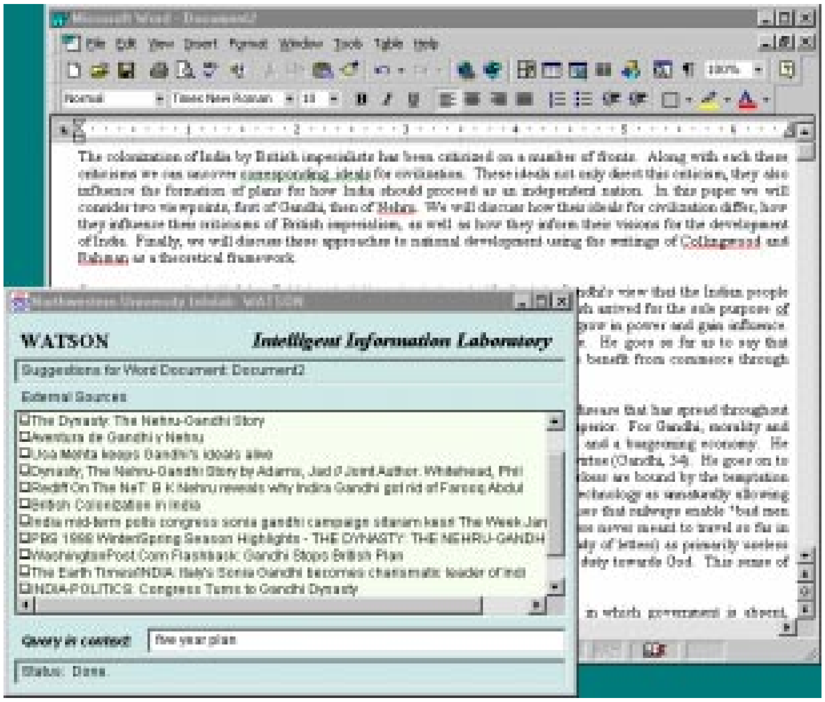
Motivations for Automatic Source Selection

A well-known problem in generating Internet searches is that queries usually return a wide range of information that may not be relevant to user tasks. For the query "home sales," for example, the first page of results for a recent query to AltaVista contained pointers to information on real estate, realtors and mortgages. This is useful information if the user is interested in the mechanics of selling a home. However, if the user is an economist interested in economic indicators, these references are of little use.

If the context for the "home sales query" is known to be that the user is working on a document on economics, it is possible to anticipate the type of result that will be useful. One way to do this is to add additional search terms. This can be useful, but it is sometimes difficult even for an expert to select the right query terms for the desired subset of information to be retrieved.

Sending queries to specialized search engines makes it possible to delineate context in advance of the query itself. A search engine such as CNN financial, for example, provides a focus towards financial news, and sending the "home sales" query there yields the information an economist might want: information on changes in aggregate sales trends.

The number of specialized search engines and repositories is large and rapidly increasing, providing the opportunity to select task-relevant sources to improve search results—if the right sources can be found. Unfortunately, finding the right sources can itself require considerable expertise. However, if a system such as Watson could automatically provide information from the right sources, the usefulness of its results could po-


Microsoft Word - Document2
File Edit View Insert Format Window Tools Table Help
Normal
Times New Roman
12
100%

The colonization of India by British imperialists has been criticized on a number of fronts. Along with each these criticisms we can uncover corresponding ideals for civilization. These ideals not only direct this criticism, they also influence the formation of plans for how India should proceed as an independent nation. In this paper we will consider two viewpoints, first of Gandhi, then of Nehru. We will discuss how their ideals for civilization differ, how they influence their criticisms of British imperialism, as well as how they inform their visions for the development of India. Finally, we will discuss these approaches to national development using the writings of Collingwood and Rahman as a theoretical framework.

...ndhi's view that the Indian people
...sh arrived for the sole purpose of
...grow in power and gain influence
...e. He goes so far as to say that
...benefit from commerce through

...disease that has spread throughout
...perior. For Gandhi, morality and
...and a burgeoning economy. He
...virtue (Gandhi, 34). He goes on to
...those are bound by the temptation
...technology as unnaturally allowing
...ow that railways enable "bad men
...re never meant to travel as far in
...ity of letters) as primarily useless
...duty towards God. This sense of

...in which government is absent,

Northwestern University InfoLab: WATSON
WATSON
Intelligent Information Laboratory

Suggestions for Word Document: Document2

External Sources

- □The Dynasty: The Nehru-Gandhi Story
- □Aventura de Gandhi y Nehru
- □Usa Mehta keeps Gandhi's ideals alive
- □Dynasty, The Nehru-Gandhi Story by Adams, Jad // Joint Author: Whitehead, Phil
- □Rediff On The NeT: B K Nehru reveals why Indira Gandhi got rid of Farooq Abdul
- □British Colonization in India
- □India mid-term polls congress sonia gandhi campaign sitaram kesri The Week Jan
- □PBS 1998 Winter/Spring Season Highlights - THE DYNASTY: THE NEHRU-GANDH
- □WashingtonPost.Com Flashback: Gandhi Stops British Plan
- □The Earth Times/INDIA: Italy's Sonia Gandhi becomes charismatic leader of Indi
- □INDIA-POLITICS: Congress Turns to Gandhi Dynasty

Query in context: five year plan

Status: Done.

Figure 1: Watson suggesting information sources to assist in a research paper.

tentially be increased without burden for the user. The goal of the SourceSelect project is to develop methods for automatically identifying relevant information and satisfying the information needs.

SourceSelect

SourceSelect bridges the gap between a representation of the type of information relevant to the user's task, as generated by Watson, and information sources on the Internet. The aim is a scalable approach that can improve focus while requiring minimal knowledge to be coded. Consequently, we have begun by investigating the use of IR methods to form the association between queries and sources. The choice of sources is based whenever possible on easily accessible information that does not require representing the focuses of the search engines by hand.

Our method divides the search engines used by Watson into two groups, general and specific. Every focused search engine has a list of keywords associated with it, keywords gathered from the META tags on the search engine's main page. A small percentage of search engines do not have keywords in META tags; their keyword lists are constructed manually. The system can currently access six specialized search engines for various topics: CNN, CNNfn, Indiana University, the India engine Khoj, HumorSearch, and ESPN.

When a query is generated by the Watson engine, the SourceSelect uses a vector-space retrieval algorithm (Salton & McGill 1983) to go against the keywords for each search engine, to find specialized search engines relevant to the query (recall that Watson's queries are processed to include terms associated with the task context). This identifies a set of search engines whose focuses are believed relevant to the query, based on a pre-set threshold for sufficient relevance. (This threshold has been set arbitrarily, but we plan to investigate the effects of tuning.) The query is then sent to the selected specialized search engines in addition to the general search engines. For some search engines, the length of the query is reduced to the first few terms to improve retrieval performance. When results are returned from these search engines, they are sent back to the Watson engine for clustering and display

When the selected search engines are especially appropriate, this method can markedly improve the quality of the results generated for a query. For example, while browsing a page on `www.cbs.com` concerning the Dow Jones industrial average crossing the 10,000 mark, the suggestions in Table 1 were generated by standard Watson and Watson with SourceSelect (page titles are shown). The original version of Watson found some sites that relate to financial news, but the results were not very useful for someone with an interest in the Dow. With SourceSelect, the keywords Watson generated for the paper matched with keywords for the CNN financial search engine, and better results were produced.

Two key questions for this approach are whether the selection of specialized search engines will improve re-

Standard Watson	Watson with SourceSelect
WDBJ 7 news at 6 for 08/11/96	Technology Stocks Slip In Lackluster Trading
The 6 O'Clock Report, Wednesday, 8/5/98	When a fund company is publicly traded - Mar. 19, 1999
85 Documents about 'Dog bites & Stats'	Dow manages slight gain in early morning trading
http://www.io.com/ nuka/Text/kpnuka981029.txt	Toon Inn
Log from the hatching of . . . Kereneth's Clutch Ista...	Dow slides 31.13 in jittery trading
Factors Influencing Media Coverage of Business Crises	CNNfn - Dow Squeezes out gain - Nov. 4, 1997
	CNNfn - Dow breaks . . . losing streak - June 17, 1996
	Dow closes up 337.17 in record gain on busiest . . . day ever
	Tech Stocks Solid As Dow And Nasdaq Gain

Table 1: Example of Watson results with and without source selection.

sults for queries in their context area, and whether possible erroneous selection of specialized search engines will degrade performance for queries that are not in their content area. Informal trials are encouraging and we are now designing experiments to test these two questions.

Issues

Issues for automatic source selection include how to identify the user's information needs, how to select sources relevant to those needs, and how to access and exploit the information they provide. We discuss each of these in turn.

Identifying needed information

A key goal of IMAs is to automatically provide users with the right information, rather than forcing them to interrupt their tasks as they notice needs for information and try to satisfy those needs through manual searching. Achieving this goal depends on the system being able to determine what information is relevant to the current goals, without directly querying the user. In principle, abductive plan recognition could be used to explain the user's actions and anticipate information needs. In practice, however, there are many reasons this is not possible: it is too difficult to generate high-level explanations for user behavior, processing cost is too high, too much background knowledge is required, and too many explanations are possible for the observed behaviors.

The Watson approach is to use limited task knowledge, at the level of how particular applications are used and how to infer content information likely to be relevant, to guide its description of relevant content. For example, Watson's knowledge includes that headings in documents are likely to be important. Based on this knowledge, it describes the important content of a document by generating a term vector that gives greater weights to terms in headings. Thus content-relevance is used as an easier-to-compute proxy for task-relevance.

An issue to explore is whether it is worthwhile to preserve the context independently of a query describing information needs within that context. In this approach, the context alone would be used to select specialized search engines to then be presented with the query that assumes that context.

Source characterization

Our initial method for describing the focuses of specific search engines relies on the keywords selected by search engine developers to describe them. These tags provide a reasonable first pass to characterizations, but there is no guarantee that these tags will be accurate. (In some cases the inaccuracies are intentional, as search engines add popular tags merely to increase the chance that the tags for their search engines will match queries presented to other search engines, to increase their traffic.) We plan to explore other methods for characterizing information sources, such as generating term vectors directly from crawling site contents for accessible repositories. We also plan to investigate methods for more flexible matching of page descriptions, such as using a hierarchy to provide more flexible matching for related terms.

Engine-Specific Query Generation

Being able to select specialized information sources raises interesting questions about how to transform general queries into queries that exploit the contextual focus provided by a specialized search engine. When generating a query for a general-purpose search engine such as AltaVista, much of the query content is needed to disambiguate the required context. Once a context is established by the specialized source, that information is no longer necessary. Some search engines automatically AND the terms in queries as their default processing mode (e.g., ESPN), making it possible that the additional terms included for disambiguation will prevent useful information from being retrieved. (For www.humorsearch.com, which has a very small database, queries with more than two terms appear to seldom retrieve *any* results.) In general, being able to access specialized information sources raises interesting questions of how to tailor queries to those sources, in light of both the information needed and the characteristics of the sources themselves.

Parser Selection and Wrapper Generation

Accessing specialized search engines requires having mechanisms for extracting the information that they return and making it available in a useful form. This corresponds to the well-known problem of wrapper generation. SourceSelect relies on hand-coded wrappers to access its information sources, but ideally would exploit either a standard set of wrappers to allow semi-automatic selection or wrapper learning methods (e.g., (Kushmerick, Doorenbos, & Weld 1997)) to facilitate the addition of new sources. Effective methods for wrapper generation are one precondition for automatic addition of new information sources.

Collating Results

A final issue is how to merge the results of multiple specialized sources. SourceSelect currently relies on heuristic clustering algorithms in Watson to group results. These algorithms use information such as the titles of pages and the structure of URLs to decide when two pages are similar. For specialized information sources, these heuristics could be augmented with heuristics that also consider the implicit context provided by the sources of the information themselves.

Perspective

The basic Watson system addresses task-relevant focusing by automatically generating queries relevant to content areas associated with the task. The addition of SourceSelect adds task-based focusing for selecting where the query is sent. The premises of this approach contrast dramatically with those of a search engine such as Google (http://www.google.com), in which the goal of a search is to find a "consensus" answer. In our approach, the goal of a search is to find the answer most relevant to a specific information-seeking context, and the use of specialized resources helps assure the relevance of the result to that context.

Surprisingly little work has been done on source selection. The most notable example, a previous version of SavvySearch (Dreilinger & Howe 1997) kept track of how well search engines handled past queries, and used vector-space retrieval to match the current query to a search engine that has previously done well with similar queries. ProFusion (Gauch & Wang 1996) used a handbuilt knowledge hierarchy to categorize queries and select relevant search engines. More recently, an agent-based learning system was added to ProFusion to manipulate each engine's place in the hierarchy based on past searches (Fan & Gauch 1999).

Older systems, like Metacrawler (Selberg & Etzioni 1995) use only general search engines and send the query to all of them. Bandwidth constraints limit the number of search engines that can be queried. The current incarnation of SavvySearch (http://www.savvysearch.com) now appears to use this approach as well.

The Internet Sleuth (http://www.isleuth.com) is a search engine that indexes other specialized search engines. It allows the user to effectively perform a source-selection algorithm by hand.

Apple's Sherlock (http://www.apple.com/sherlock) allows the user to select the search engines that will be queried. This approach puts the burden of source selection entirely on the user. The user is forced to remember which search engines give the most relevant results for each type of query he may want to use.

The GlOSS (Gravano, Garcia-Molina, & Tomasic 1994) system obtains the index from each of its information sources, and combines these indices to form a meta-index, which is used for source selection. The drawback of this approach is that all of the information sources must cooperate by providing their indices in order for the meta-index to be built.

EMIR (Kulyukin 1999) maintains positive and negative keyword vectors for each of its information sources. Like GlOSS, it needs the cooperation of the information sources to maintain an accurate representation of their contents.

Most of these systems, use general-purpose search engines for their information sources. While general-purpose search engines provide the broadest coverage, focused search engines can have a much greater concentration of relevant links within their subject area. When Watson's contextual information is added to basic source selection, focused search engines appear to provide better results than general search engines.

Conclusion

SourceSelect augments Watson's just-in-time retrieval framework with the capability to choose specialized information sources related to the current context. The goal is to leverage off existing information resources to automatically provide the user with task-relevant information. The current version of SourceSelect matches term vector descriptions of the content area of interest to descriptions from the tags of specialized search engines to select sources expected to be relevant to those content areas, queries those sources, and forwards those results to Watson for presentation to the user. Initial tests have been encouraging; next steps include addition of other specialized search engines, formal evaluation, and exploration of alternative methods for describing task-relevant content and selecting information sources.

References

Budzik, J., and Hammond, K. 1999. Watson: a just-in-time information environment. In *AAAI Workshop on Intelligent Information Systems*. In Press.

Budzik, J.; Hammond, K.; Marlow, C.; and Scheinkman, A. 1998. Anticipating information needs: Everyday applications as interfaces to internet information resources. In *Proceedings of the 1998*

World Conference on the WWW, Internet, and Intranet.

Dreilinger, D., and Howe, A. 1997. Experiences with selecting search engines using meta-search. *ACM Transactions on Information Systems* 15(3).

Fan, Y., and Gauch, S. 1999. Adaptive agents for information gathering from multiple distributed information sources. In *Proceedings of the 1999 AAAI Spring Symposium on Intelligent Agents in Cyberspace.* AAAI Press.

Gauch, S., and Wang, G. 1996. Information fusion with ProFusion. In *WebNet '96: The First World Conference of the Web Society*, 174–179.

Gravano, L.; Garcia-Molina, H.; and Tomasic, A. 1994. Precision and recall of gloss estimators for database discovery. In *Proceedings of the third international Conference on Parllel and Distributed Information Systems (PDIS '94).*

Kulyukin, V. 1999. Application-embedded retrieval from distributed free-text collections. In *Proceedings of the Sixteenth National Conference on Artifical Intelligence.* AAAI Press. In press.

Kushmerick, N.; Doorenbos, R.; and Weld, D. 1997. Wrapper induction for information extraction. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* Morgan Kaufmann.

Salton, G., and McGill, M. 1983. *Introduction to modern information retrieval.* New York: McGraw-Hill.

Selberg, E., and Etzioni, O. 1995. Multi-service search and comparison using the metacrawler. In *Proceedings of the Fourth World Wide Web Conference*, 195–208.