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

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) 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. The proliferation of specialized search engines and information repositories on the Web provides a rich source of focused information, but exploiting them depends on having methods for selecting the best sources to satisfy the user's needs. This paper describes research on augmenting Watson with the capability for automatic information source selection. It presents two automatic methods for source selection that have been integrated into Watson, compares their performance, and points to general problems and opportunities for task-relevant source selection.

Introduction

As the availability of information grows, the burden of finding the right information increases as well. "Just-in-time" information systems shield the user from information access tasks by observing the user's actions in a task context, anticipating the user's information, gathering the needed information and presenting 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 information that satisfies task-based needs. There are large numbers of focused information sources on the web, providing a rich range of specialized information aimed at satisfying particular information needs.[1] However, finding the right sources requires its own expertise and is a considerable burden on the user. This paper describes ongoing research on how to automatically select information sources that are appropriate to the user's needs.

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

Our system is integrated with Watson (Budzik *et al.* 1998), a system that 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 focused on identifying task-relevant content areas and automatically generating content-relevant queries for general-purpose search engines. The proliferation of specialized search engines and information repositories on the Web provides a rich source of more focused information, but exploiting them depends on having methods for selecting the most appropriate sources to satisfy the user's needs. This paper describes research on how to select and access those souces. A system has been developed to perform a two-step retrieval, using vector-space content matching from information retrieval to associate queries to sources, and then using automatically-generated queries to guide search within those sources. The system has been integrated into Watson to provide the system with additional capabilities for focussed retrieval. In the combined system, Watson monitors user activities, identifies relevant content areas, and provides area information to the system. The system determinines appropriate information sources, formulates queries to those sources, sends off those queries, and collates their results for Watson to pass them on to the user. No user intervention is required.

As background, 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 embodies and presents results from initial tests. It then discusses central issues for the approach and how the approach relates to other methods for source selection and improving the quality of search results.

Just-in-Time Information Access: The Watson Framework

The Intelligent Information Laboratory (InfoLab) at Northwesern University is developing a class of systems called Information Management Assistants (IMAs).

These systems observe users as they go about completing a task in everyday software applications such as Microsoft Word and Internet Explorer, and use these 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 everyday applications (e.g., Netscape Navigator, Microsoft Internet Explorer, and Microsoft Word). Watson uses a basic knowledge of information scripts—standard information-seeking behaviors in routine situations—to anticipate 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. 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. This representation is fed to the RESOURCE SELECTOR, which 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, which is interpreted and filtered by the RESULT PROCESSOR using a set of result similarity metrics. The resulting list is presented to the user in a separate window.

When an information need is anticipated in Watson, Watson selects appropriate sources and transforms the original document representation into a query. This query takes the form of an internal query representation, which is then sent to selected information adapters. Each information adapter translates the query into the source- specific query language, and executes a search. Information adapters are also responsible for collecting the results, which are gathered and clustered using several heuristic result similarity metrics, effectively eliminating redundant results (due to mirrors, multiple equivalent DNS host names, etc.).

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 process is illustrated in Figure 1.

The Importance of 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 may be useful information, if the user's interest is in the mechanics of home selling. However, if the user is an economist interested in economic indicators, these references are of little use. If the context is known to be a financial document, it is possible to anticipate the type of result that will be useful and constrain the search. However, even adding search terms to constrain context may not be sufficient to result in the desired subset of information being retrieved.

Sending queries to specialized search engines makes it possible to 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 sources to improve search results—if the right sources can be found. Unfortunately, finding the right these sources can itself require considerable expertise. However, if a system such as Watson could automatically provide information from the right sources, the usefulness of results could potentially be increased without burdening the user. The goal of the SourceSelect project is to develop methods for automatically identifying relevant information and satisfying the information needs.

SourceSelect

The Source Selector bridges the gap between the representation of the user's task content, as generated by Watson, and information sources on the Internet. More specifically, it:

- Receives information about the content of an area of interest, provided by Watson in the form of a term vector.

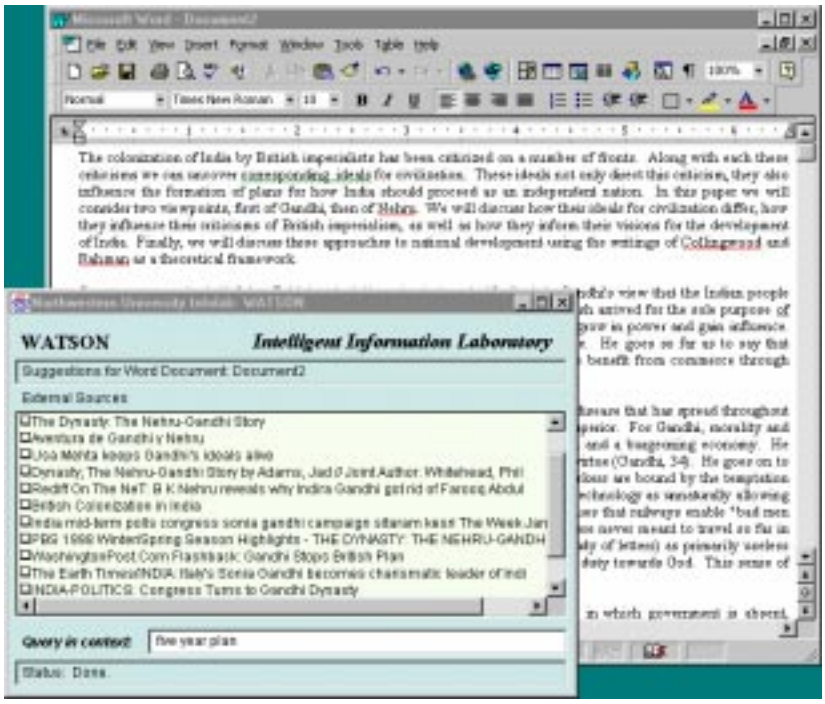

Figure 1: Watson suggesting information sources to assist in a research paper.

An Initial Test

Issues

Content-Relevance as Proxy for Task-Relevance

Wrapper Generation

Engine-Specific Query Generation

Perspective

Compare to Waston's current method, saavysearch, etc.

Conclusion

References

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.*