CORE
Metadata, citation and similar papers at core.ac.uk
Provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 1998 Proceedings

Americas Conference on Information Systems (AMCIS)

December 1998

Extracting Maximum Benefits from Web-Based Searching

Matthew Montebello
Cardiff University

Follow this and additional works at: http://aisel.aisnet.org/amcis1998

Recommended Citation
Montebello, Matthew, "Extracting Maximum Benefits from Web-Based Searching" (1998). *AMCIS 1998 Proceedings*. 342.
http://aisel.aisnet.org/amcis1998/342

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 1998 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Extracting Maximum Benefits from Web-Based Searching

M. Montebello
Computer Science Department
Cardiff University

Abstract

With huge amounts of information connected to the Internet, efficient and effective discovery of resource and knowledge using the Internet has become an imminent research issue. A vast array of networks services is growing up around the Internet and massive amounts of information is added everyday. Users can now access massive amounts of information in various forms, thereby creating an equally massive problem. This rapid growth in data volume, user base, and data diversity render Internet-accessible information increasingly difficult to be used effectively. Therefore, search for a specific information on this massive and exploding Internet information resource base becomes highly critical. In this paper we discuss the issues involved in the application of machine learning techniques to the problem of Internet-based information overload. We present a general architecture and describe how it has been instantiated in a functional system we developed. The system attempts to concurrently maximize and optimize the resource/knowledge discovery, and custimize the information to individual users. We discuss the design issues involved in the attempt to develop an evolvable architecture which can easily and inexpensively accommodate future generations of web-based systems and technologies.

Background and Motivations

The immense size of the distributed World-Wide Web (WWW) [Ber94] knowledge-base and the dramatic rapid increase in the volume of data on the Internet, requires techniques and tools that reduce users' information overload and improves the effectiveness of online information access. The size of the Internet is exploding and an exponential increase in the number of users and web servers is no surprise, resulting in an eventual feedback loop. As more users come on the net, they provide their own information which in turn encourages even more people to join in. With this massive increase of the Internet usage, a vast array of networks services is growing up around the Internet and huge amounts of information is added everyday. Users can now access enormous amounts of information in various forms, thereby creating an equally massive problem. This rapid growth in data volume, user base, and data diversity render Internet-accessible information increasingly difficult to be used effectively, thereby rendering the task of extracting maximum benefits from web-based searching highly critical. Users are faced with the problem of search engines being too generalised and not focused enough to their real and specific needs. This triggered further research to develop more sophisticated techniques and agent like systems that make use of the user profile to personalise the service they provide and add value to the information they presented [Paz96].

In Section 2 we briefly present the rationale underlying our system which extracts maximum benefits from the web by reusing information generated by search engines and previously developed retrieval systems. Conceptually, it is similar to a meta-search engine, but with the major difference that it employs user profiling to specifically target documents for individual users by making use of a number of machine learning techniques [Sal83]. Some results of pilot tests performed to evaluate their effectiveness are presented. We underline the major components of our system in Section 3, and point out how this web-based application has been designed and developed to evolve in Section 4. This will have considerable and significant implications for an organization that intends to develop a long lasting web-based application without having to redevelop the software every time the slightest change occurs on the WWW. Finally, in Section 5 we present our conclusions.

MetaSearching and Machine Learning

Our goal is to achieve a high recall and high precision performance score on the information presented to the user. Recall measures how efficient the system is at retrieving the relevant documents from the WWW, while precision measures the relevance of the retrieved set of documents to the user requirements. In order to obtain a high recall execution we make use of the metasearch approach, namely, hits returned by a number of traditional search engines together with previously developed retrieval systems are blended, aggregated and utilised by our system. On the other hand, to achieve a high precision execution we employ machine learning techniques to extract features from documents specific users find interesting, generate a profile and predict other documents that fit this profile. The search engines that are currently employed are AltaVista[Alt], Excite[Exc], WebCrawler[Web], Infoseek[Inf], Lycos[Lyc], and Yahoo[Yah], but the capability to incorporate others is already available as we shall see in Section 4. Machine learning techniques applied at this stage are:

1. Term Frequency,
2. Term Frequency/Inverse Document Frequency, and
3. Term Frequency/Discrimination Value.

For a full describtion of these techniques see [Kow97]. To compare the three machine learning techniques employed and assess the overall performance of our approach we ran a set of data tests on a number of different topics, once on documents returned from a single search engine and consecutively on our system. The pilot test involved the evaluation of the quality and accuracy our system predicted documents to the user. The number of documents selected from those retrieved in both cases started from 10 and went up to 100 in steps of 30. The accuracy of prediction was assessed on the relevance, correct domain, and user interest. The results (Figure 1) indicate that on average the machine learning techniques employed did not out-perform each other, but rather complemented one another. An additional binary measure was included in the testing just to give an idea of what accuracy a simple weighing technique (1 if the term is present, 0 otherwise) could achieve. Clearly, as the number of documents selected by the user increased then the profile generator could progressively improve the predictions. To note that some inconsistencies in the graphs on the left (the single search engine) could be due to the fact that even though the recall is high the precision starts to drop as the number of documents increased. This will automatically reflect in the quality of predictions made by our system due to the decline in the quality of documents retrieved by the single search engine.

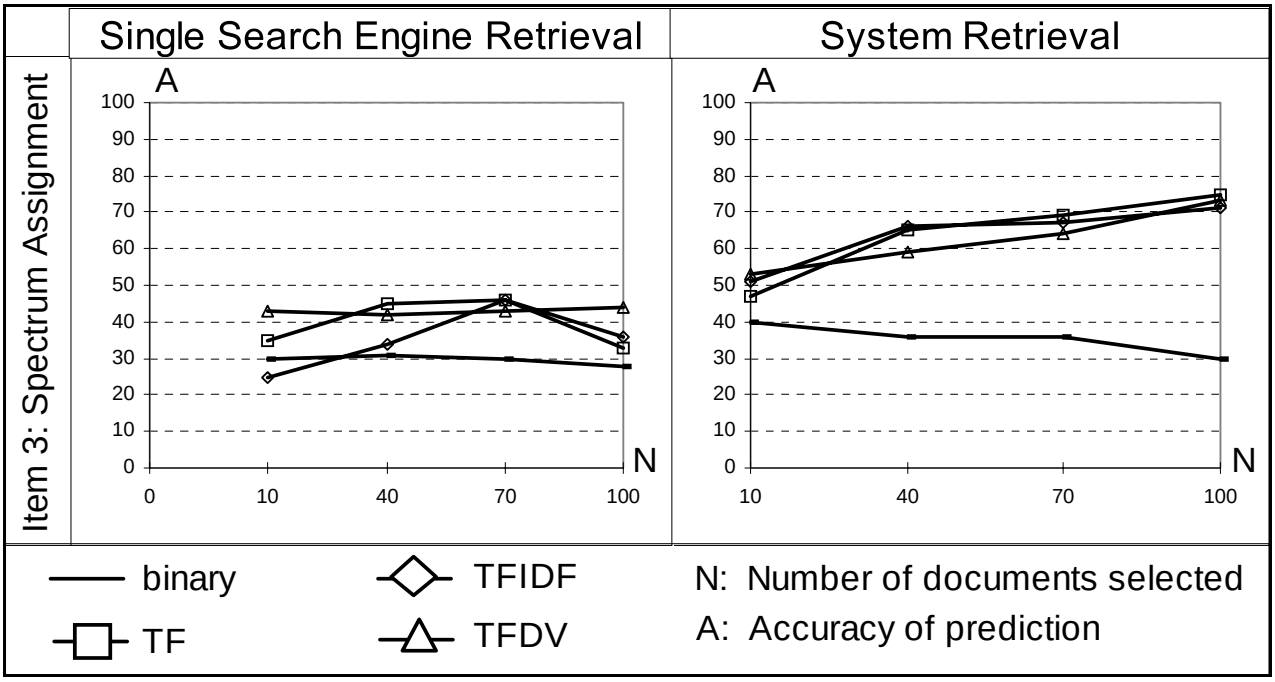

Figure 1. Results from Pilot Test

Architecture

The task performed by the system is decomposed into a number of simpler tasks. Figure 2 shows the major components of the system: the WWW and the external systems at the bottom level, the underlying application software on the next level up, and the GUI at the top. All the external systems are considered to be black boxes and action is taken upon the information they output. Wrappers are used to manage the appropriate and proper handshaking between the diverse search engines together with the other retrieving systems and the application layer.

The system requires an administrator to manage the general needs and demands of a specific interest group of users. The administrator can initialise search terms tailored to any type of interest group and furthermore users will be able to suggest any other terms to add to the main search list. Documents relevant to the specific area of interest are retrieved and stored by the underlying application within the main index, and when a user logs-in he/she is able to benefit from the systems' high recall fidelity. Having analysed the documents, individual users can bookmark and highlight specific items as interesting and appealing. These will be saved inside their personal database index. At this stage the underlying application plays an important role in attaining precise targeting of documents to individual users by generating a profile from the personal database index and predicting other documents from within the main index. Users can decide to add the suggested documents (Figure 3) to their personal database index or remove them completely. As new and suggested documents are entered in the personal database index the user profile becomes more focused and finely tuned, as a result of which higher precision results will be achieved.

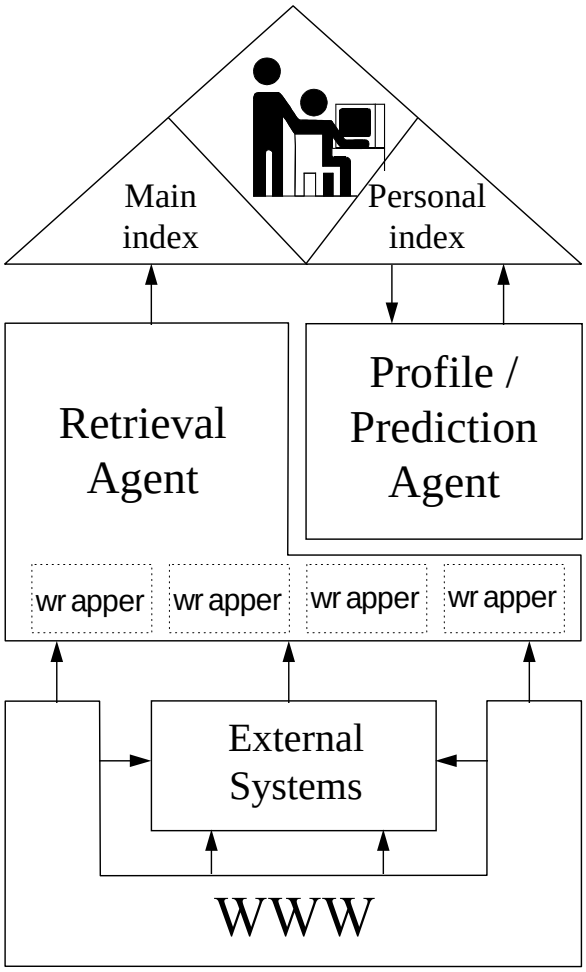

Figure 2. Architecture

Evolvability

The scramble for commercial success on the Internet has brought many technology vendors into the Web trade, resulting in the development of new tools and methods [Jac95]. This renders the issue of building evolvable systems into perspective [Moo97]. Our system is capable of evolving in two ways - additional external systems coupled with the appropriate wrappers within the retrieval agent, and additional machine learning techniques employed in the application layer. This is of utmost importance and has significant development, financial and time-consuming implications when new information sources become available and superior machine learning techniques are developed. The basic idea is the application of specific wrappers to cater for particular information sources and facilitate the translation of complex queries into their respective command language. As new sources and techniques develop, appropriate wrappers will be required. Similarly, other profile generation learning techniques can easily be inserted within the application layer together with the other machine learning techniques. Eventually only one of them will be utilised, but the user is capable to select which particular one (just like selecting a particular agent or actor) will be employed to extract features, generate user profile, and predict/suggest documents.

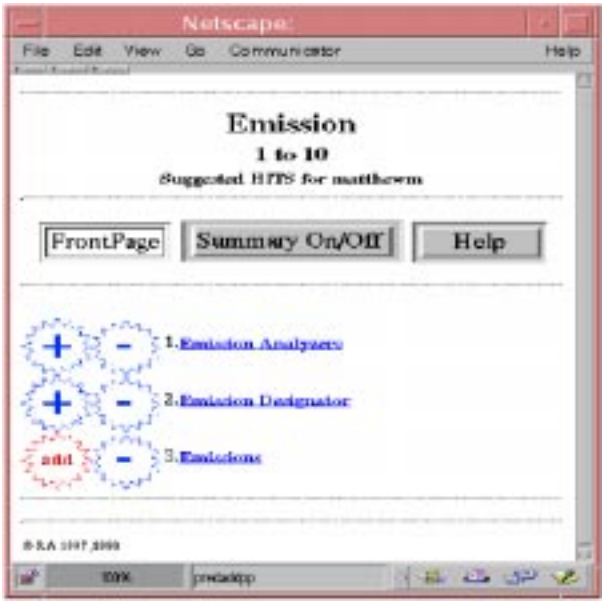

Figure 3. Suggested Documents

Concluding Comments

In this research article we have presented a system that adds value to the information traditional search engines and other metasearch engines generate from the WWW. We argue that by reusing the information output from several retrieving/indexing systems we ensure a high recall score, while generating a specific user profile to predict and target other documents to specific users, we also ensure a high precision score. Users are able to select their own profile generator/prediction agent from a number of alternatives, reflecting different machine learning techniques employed. The system administrator, who can also easily maintain the system's resources and update the search terms specific to a user group, can integrate new retrieving techniques into this evolvable system. With the use of wrappers we demonstrated that such techniques have considerable implications for organizations which are competing to design and develop high quality (software-wise) web-based applications. In the future we will be investigating the integration of other machine learning techniques that have been developed and employed by other systems. This will help us to evaluate which technique is best suited to cater for the needs of different users together with ensuring that the system easily and successfully accommodates these changes. Evaluation of the recall/precision scores is also required to ensure that value is added to the normal services provided by the search engines and the meta-search engines. This will be done by analysing the feedback that will be given by a group of users who are presently testing the system and who will eventually assess the extent to which the information presented is accurate and of high recall/precision quality.

References

Reference available upon request from author(m.montebello@cs.cf.ac.uk).