MetaSEEk: A Content-Based Meta-Search Engine for Images

Mandis Beigi, Ana B. Benitez, and Shih-Fu Chang
Department of Electrical Engineering & New Media Technology Center
Columbia University, New York, NY 10027

ABSTRACT

Search engines are the most powerful resources for finding information on the rapidly expanding World Wide Web (WWW). Finding the desired search engines and learning how to use them, however, can be very time consuming. The integration of such search tools enables the users to access information across the world in a transparent and efficient manner. These systems are called meta-search engines. The recent emergence of visual information retrieval (VIR) search engines on the web is leading to the same efficiency problem. This paper describes and evaluates MetaSEEk, a content-based meta-search engine used for finding images on the Web based on their visual information. MetaSEEk is designed to intelligently select and interface with multiple on-line image search engines by ranking their performance for different classes of user queries. User feedback is also integrated in the ranking refinement. We compare MetaSEEk with a base line version of meta-search engine, which does not use the past performance of the different search engines in recommending target search engines for future queries.

Keywords: MetaSEEk, meta-search engine, content-based visual query, color search, texture search, performance monitoring, World Wide Web

1. INTRODUCTION

The explosive growth of the World Wide Web has motivated the development of many search engines to assist the unmanageable task of navigating the Web. They try to satisfy the users' information needs for newspaper articles, software, movie reviews, books, music recording, images, video, etc. Two types of search engines can be found on the Web: large-scale robot-based and specialty search engines. Large-scale search engines try to index the contents of the entire World Wide Web, but usually fail to disseminate between desired data and unneeded information. On the other hand, specialty search engines are more focussed databases, which can not be applied to general topics.

Experienced users of the Internet would begin to query the appropriate specialty search engines to obtain desirable results, and continue querying general search engines when the specialized engines fail to yield helpful information. Nevertheless, the proliferation of search engines has replaced the problem of finding information on the Internet with the problem of knowing where search engines are, what they are designed to retrieve, and how to use them. Consequently, searching the Web for specific information has become a very time consuming and inefficient task for even the most expert users.

This situation has motivated the recent research and development in integrated search or meta-search engines [1]. Meta-search engines serve as common gateways, which automatically link users to multiple or competitive search engines. They accept requests from users, sometimes, along with user-specified query plans to select target search engines. The meta-search engines may also keep track of the past performance of each search engine and use it in selecting target search engines for future queries. Many approaches have been proposed for meta-searching. Section 2 presents an overview of these approaches, the majority of which have been designed for text databases.

Digital images and video are becoming an integral part of human communications [2]. The ease of creating and capturing digital imagery has trigged the recent development of visual information retrieval (VIR) systems on the web

Further author information –
M.B.: Email: mandis@ctr.columbia.edu
A.B.C.: Email: ana@ctr.columbia.edu
S.C.: Email: sfchang@ctr.columbia.edu

SPIE Vol. 3312 • 0277-786X/97/$10.00

[3,4,5]. These systems usually provide methods for retrieving digital images by using examples and/or visual sketches. In order to query the visual repositories, the visual features of the imagery, such as colors, textures, shapes, etc, are used in combination with text and other related information. Everyday, users are finding new VIR systems on-line what leads, once more, to the problem of efficiently and effectively retrieving the information of interest.

We have developed a prototype meta image search engine, MetaSEEk, to investigate the issues involved with efficiently querying large, distributed on-line visual information sources. Our meta-search engine, MetaSEEk, adopts the principle that Web resources should be used efficiently. For each query, MetaSEEk selects the target engines that may the desire results by weighing search tools' successes and failures in similar query conditions. The implementation of the meta-search engine is described in section 2.

Section 3 describes the experiments, the evaluation measures and the comparison results between the MetaSEEk prototype and a base line search-engine that randomly selects the search engines to send the queries to. An interesting issue examined in MetaSEEk is the reliability of the selection and ranking of the remote search engines for different type of queries. Another important technical aspect is the heterogeneity among the different remote search engines and possible technical approaches to enhance interoperability. Finally, section 4 closes with concluding remarks and open issues for future research.

2. RELATED RESEARCH

Meta-search engines serve as common gateways, linking users to multiple search engines in a transparent manner. Working meta-search engines include three basic components, as depicted in Figure 1[1]. The dispatching component selects target search engines for each query. The query interface component translates the user-specified query to compatible scripts to each target search engine. The display interface component merges the query results from each search engines, removes duplicates and displays them to the user in a uniform format.

At the present time the wealth of meta-search engines on the WWW is still growing. Many approaches have been proposed for meta-searching. We overview a few of these efforts.

The GlOSS (Glossary-of-Servers Server) project [6] uses a meta-index to estimate which databases are potentially most useful for a given query. This meta-index is constructed by integrating the indexes of each one of the target databases. For each database and each word, the number of documents containing that word is included in the meta-index. The two main drawbacks of this approach are: first, it requires each of the search engines to cooperate with the meta-searcher by supplying up-to-date indexing information, and second, as the number of databases increases, the administrative complexity may become prohibitive.

The Harvest system [7] is being designed and built by the Internet Research Task Force Research Group on Resource Discovery. Harvest consists of several subsystems: a Gatherer collects indexing information and a Broker provides a flexible interface to this information. It is intended to be a scalable form of infrastructure for building and distributing content, indexing information, as well as for accessing Web information

Wide Area Information Servers (WAIS) [8] divide its indices among the databases into multiple levels with the top-level index containing a "directory of servers". Given a query, the "directory of services" is searched and the query is then forwarded to selected databases.

MetaCrawler [9] is a meta-search service developed at the University of Washington that integrates a set of general Web search engines. When a query is submitted, MetaCrawler dispatches queries to each one of those search engines, retrieves the HTML source of all the returned documents, and applies further analysis to clean up unavailable links and irrelevant documents. MetaCrawler obtains high precision but at the cost of network utilization.

The SavvySearch meta-search tool [1] employs a meta-index approach for selecting relevant engines based on the terms in a user's query; previous experience about query successes and failures is tracked to enhance selection quality. SavvySearch selects resources for an individual user's query and balances resource consumption against expected result quality by querying the most relevant search engines first. Their experimental finding suggest that a meta-index approach

can be effective in making search engine selection decisions. However, the potentially large amount of knowledge required to make decisions raises some questions about the overall efficiency of the system.

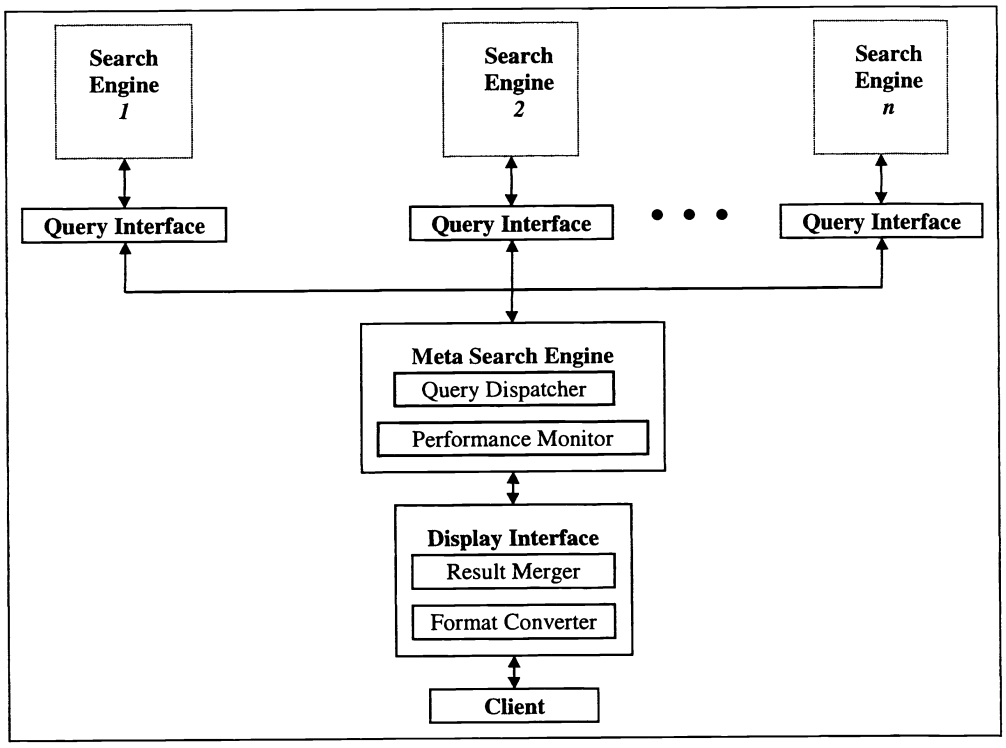

Figure 1: Basic components of a meta-search engine

Other automated Web meta-searchers are Dogpile, Metafind and Metasearch. These systems basically dispatch queries to each one of their search engines that they target and present the returned documents to the user in a uniform manner. Many manual query dispatch search engines are also available on the Web. Tools such as All-in-One, CUSI, search.com, Infi-Net's META search and InterNIC are essentially pages full of forms to sending queries to a number of different search engines. The selection process is entirely up to the user – they must type their query into a separate form for each query submission. Only one search engine is activated at a time, and the results appear in the native format of whichever search engine produced them.

The ProFusion system [10] is a Web meta-search engine that supports both manual and automatic query dispatch. In automatic query dispatch, ProFusion analyzes the incoming queries, categorizes them, and automatically picks the best search engines for the query based on a priori knowledge (confidence factors) which represents the suitability of each search engine for each category. It uses these confidence factors to merge the search results into a re-weight list of the returned documents, removes duplicates and, optionally, broken links and presents the final rank-ordered list to the user. ProFusion's performance has been compared to the individual search engines and other meta searchers, demonstrating its ability to retrieve more relevant information and present fewer duplicate pages.

3. METASEEK

MetaSEEk is an integrated search engine, which serves as a common gateway, linking users to multiple image search engines. It includes three main components as shown in Figure 1. The query interface component accepts search queries from the user and translates them to the specific query interfaces used by each target search engine. The dispatching component decides which search engines the query should be sent to. The display component merges the results and ranks them for displaying. MetaSEEk evaluates the performance of each query method on a search engine for future queries based on the user's feedback.

Queries can be submitted to MetaSEEk at http://www.ctr.columbia.edu/MetaSEEk. The underlying system is implemented in C and currently runs on a HP platform. MetaSEEk uses socket programming for opening ports to send the queries to the individual target search engines and to download their results. HTTP commands are sent to the remote search engines in a similar manner to web browsers such as Netscape and Mosaic.

3.1. Content based image query

There are several methods, which may be used to retrieve images based on their visual contents. Several systems use visual features such as texture, color, shape, and structure [3,4,5]. For example, texture can describe the coarseness, contrast, roughness, and presence/absence of directionality of each image. Another method may be based on the amounts of different colors in each image. The color amounts can either be used to describe the entire color content of each image or they can describe the color amounts in local regions of the image. These methods can be used separately or can be combined in calculating the similarity measures for the content-based image query. Different search engines use various methods and support alternate combinations. They also use different algorithms for calculating the similarity measures and their distances.

3.2. The query interface component

MetaSEEk currently supports the following target search engines: VisualSEEk, WebSEEk, QBIC and Virage. In the current version of MetaSEEk, the user interface allows for browsing of random images retrieved from the remote search engines. The user can select a method for querying such as color and texture. These two methods can be selected individually or they can be combined. Another popular search technique used in the image search engines is search based on keywords. This kind of search is used in search engines for querying documents as well as images. Image search based on visual content usually returns a ranked list of images which have the highest similarity to the query input, which could be an example image or a visual sketch. Keyword-based search may be used to match images with particular subjects (e.g., nature and people) and narrow down the search scope.

MetaSEEk allows a search based on example images, URLs or keyword text. No all the search engines support all these options. The query dispatching component of MetaSEEk makes the decision on which search engines the queries should be sent to. This component is explained in detail in the next subsection.

The user can specify a value for the maximum waiting time which is used to prevent the query system from stalling if a target search engine happens to be down or unreachable. Figure 2 shows the user interface for the MetaSEEk search engine.

3.3. The query dispatching component

MetaSEEk queries the search engines that first, support the method of the query selected by the user (i.e. color and/or texture), and second, have high past performance scores. The performance scores are calculated every time a query is made and are based on the user's feedback. The calculation of the performance scores is explained in section 3.5.

MetaSEEk keep track of the performance scores of the search engines with respect to each query on every image. These performance scores are indexed in the database and will be used to select the target search engines for each new query. MetaSEEk also stores the visual feature vector for each queried image in case the queried image is not already in the database. In this case, when the user issues a new query, a set of old queries with the most similar feature vectors with that of the new query will be used. The queried images' performance scores will be used to select the search engines for the new query. This approach basically finds the best query examples from the past and follows its route to selecting the remote search engines, which have done well for that past queries.

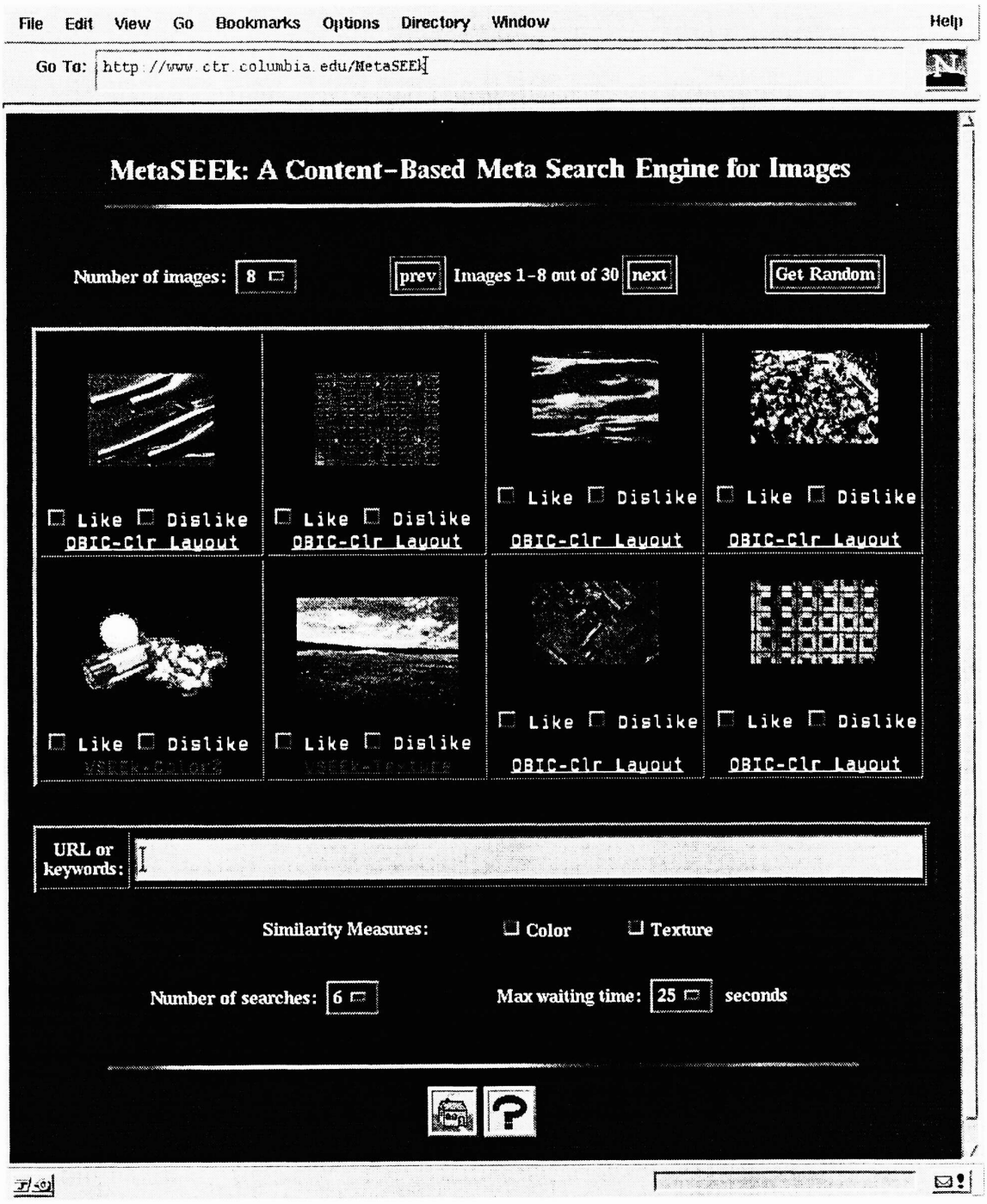

Figure 2: The query interface of MetaSEEk

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.