Search Platform (Codename: Nereus)

**Introduction**

The Search Platform is an application that helps researchers methodically search the web. It leverages existing generic and topical web search engines as well as internal search engines by providing the ability to search multiple search engines at once. By keeping track of web pages that have been seen and optionally hiding those from view it helps to expedite a deeper search of the web. Additionally, efforts spent doing web research by an individual are visible by the group in order to reduce duplication of research. Projects allow compartmentalizing web research and help with collaboration in order to obtain both broad and deep searches of the web. Ancillary to a researcher’s immediate and direct needs is that this platform is used to train computer algorithms and web crawlers to recognize relevant information.

This document explains the more in-depth features of the Nereus.

**Search Projects**

Projects allow keeping track of query and browsing history while working on a web research task. This information is used to help the user more quickly dig deeper into the web, find more relevant content, refine queries, filter results, jump start future web research tasks, train computer algorithms, create automated web crawler tasks, and identify and alert users when web content changes.

All queries performed in the context of the project are recorded for that project. In addition to query history many settings for individual URLs are also specific to the project. Users may view the query history and URL settings for a given project. Within a project’s context, queries can be altered and removed and URL settings may be changed.

Projects can be created from scratch or they may be created from one or more other projects. New projects can be created by typing in a query and selecting to run the query in a new project. When creating from another project, the new project is created with the same preferences and query list of the original project. When a new project is created from two or more projects the query list and settings are merged to create the new project.

Project settings include per web page settings, which search engines to use, the list of queries, group members, sorting (page rank, newest first, and relevance), page ranking algorithm settings, many generic search engine settings (Google’s search settings), and which categories to include in results.

Projects can be created for a group or for an individual. Each time a group project is created, a separate project is created for each member of the group. Members can be added to a group project at any time. All of the search and browsing activity of individuals in the group is tracked in the group project. At the same time, an individual’s activity is also tracked in their respective individual project. Individual projects that are part of a group project are visible to all members of the group. This allows members of the group to monitor how the web search is progressing and verify its completeness.

A user can create private projects that are only accessible to that user. Private projects can be made public at any time.

There is a one-to-many cardinal relationship between a major project (e.g. country study, company expansion options) and the search projects created in the platform while doing this research. Many small web search projects may be necessary to complete a single major research project. In the country study example a user may create a project for searching for information on just one of the country’s executive cabinet members. Multiple projects can be open at once in a different browser windows or tabs. Projects may be organized hierarchically in order to reduce clutter created by having a large number of projects.

**Federated Search**

A number of different search engines and sources internal and external to can be searched at once using a uniform search query syntax. External examples include: Google, Bing, lesser known search engines, specific website search engines, Google news, Yahoo! news, LexisNexis, patents, Twitter, Facebook, etc. Foreign language search engines such as Baidu and Daum can also be included in the search; the titles and snippets are translated to English. Internal examples include: Intelliwiki, Lotus Notes, Webi, SharePoint sites, Spark, daily brief, network drives, etc. Additional sources that are not often found in mainstream search engine results include: non-visible content such as HTML attributes (e.g. email addresses), RSS feeds, Google's competitors (Ask.com), internet archive, DMOZ (internet directory), Wikipedia talk and history pages, etc. In addition, all indexed and stored content from the web can be included in the search.

When creating a project, a user can select any number of search engines and sources to search when performing queries. Each query searches the selected sources and then combines and intermingles the results. Ranking is done using the combined search engine rankings, user preferences, and “learned” information. The ranking algorithm is not yet specified in detail.

**Custom Search Engine**

Users can create their own search engines by selecting a list of websites and directories to search. This goes beyond federated search by allowing the user to not only select the search engines or sources to use but to also constrain the search to a discrete number of sites. For instance, a user may have identified a number of good websites for information on international trade. They can create an “international trade” custom search engine that only returns results from those websites in order to reduce the amount of noise in the search results. Custom search engines are added in the list of search engines available when creating a new project.

**Query Interface**

The query interface is nearly identical to a standard search engine, consisting of a single text box for entering a query. In addition a list of previous queries is provided to quickly select or compare to when formulating new queries. The date and time each query was run by the user and the group is displayed with the query text. Additional options are available at query execution time to change which URLs are shown. Some of these include: show seen, show irrelevant, and show all.

An advanced query page may be made available.

**Query Results**

Each web page can have per query, per group project, per user project, and global settings associated with it. Per query settings include the ability to mark a URL returned by a query as relevant or irrelevant to the query. Marking a URL irrelevant hides it in that particular query’s results for that project.

Per group project settings include the ability to mark a URL as irrelevant to the project and effectively hiding it in all subsequent queries. URLs may also be marked as “seen” in order to hide them for the duration of the project.

Per user settings include the ability to mark a URL as “seen” by that user. The URL is then only hidden for that user. Additionally, the user can be alerted when the web page changes and the differences since the last time it was viewed can be shown.

Global settings persist across all projects. Global settings for a URL include the ability to place a URL in a category (e.g. news article, blog, profile, wiki page, etc.). This helps train the automated page classifier to classify future pages.

Search engine index information is displayed for each URL. The ranking for each search engine is shown. 0 means it was not indexed on account of the robots exclusion standard (robots.txt). An indication if has indexed the page is also displayed. Results may be re-ranked by the user, perhaps via drag-and-drop.

Historical information is displayed for each URL including: the date and time when the last person in the group visited it, when it last appeared in the search results for the group, when the user last viewed it, and when it last appeared in the search results for the user. In addition, the date it was used in the daily brief (if it was used) and its status in WebSense can also be displayed. The application can specify if it is blocked or flagged by WebSense and for what reason. An image of the web page can be displayed instead for WebSense blocked URLs.

Additional operations include notifying other members of the group about a web page. New queries can be created by using the "more like this" feature. An option to immediately add the page to the indexer for in-depth analysis and search is available.

**Browsing**

Web pages are viewed in a frame within a web browser. The per-page settings are always available in the outer frame. This allows tracking followed links in order to optimize future ranking.

All operations and capability of a standard web browser are leveraged to make the application as transparent as possible to the user. The browser’s default search engine can be set to use the search engine and a browser plug-in can be used to always open new URL’s in the frame. A browser plug-in can be used to select and show the project.

**Web Crawling**

Initially there are two crawler implementations available: Deep crawl and proximity crawl. The deep crawler is based on a custom search engine the user created. It attempts to index every page/file found in the websites specified. It will not index any pages/files that are not within the websites specified. Before a deep crawler is launched an estimate of the time required to complete the crawl is provided to the user.

The proximity crawl is based on a search project. It is seeded with the URLs that are marked relevant and it decides how to crawl based on the queries that were created for the project. The queries and relevance information are used to identify the relevancy of each page/file to the project. All crawled pages are indexed and can be searched the next time a user runs a query. The user may specify how deep to crawl and this will limit how long the crawl takes. There is no way to estimate how long the entire crawl will take. Queries performed against the index before the crawl is complete return results found up to that point in time.

**Future Features**

There are ways to mark content or add comments to content for sharing with a group (content curation). See: reframeit.com, icyte.com, memonic.com, scribble.com, diigo.com

Integrate tools from Stach & Liu

**See Also**

<http://www.stachliu.com/resources/tools/google-hacking-diggity-project/>

<http://www.stachliu.com/2012/08/quick-intro-to-notinmybackyard-diggity/>