DOCTORDIGEST: A SYSTEM FOR CLASSIFICATION OF MEDICAL INTERNET RESOURCES AND REVIEWS*

Tajana Lucic, Naphtali D. Rishe, Oksana Dyganova, Andriy Selivonenko, Ouri Wolfson+
Division of Medical Informatics
High Performance Database Research Center
Florida International University
Miami, FL 33199, USA
Tel: (305) 348-2025, (305) 348-1706, Fax: (305) 348-1707, E-mail: rishe@fiu.edu

+Department of Electrical Engineering and Computer Science
University of Illinois at Chicago
851 S. Morgan St., Chicago, IL 60612, USA.
Tel: (312) 996-6770, Fax: (312) 413-0024, E-mail: wolfson@eecs.uic.edu

Abstract: *A team of programmers and medical doctors at the High Performance Database Research Center at Florida International University and NOA Inc has developed the Medical Resource Reviews Database, which facilitates access to information resources by medical professional and patients.*
This database contains expert summaries, reviews, ratings, and classification of electronic medical information resources available by any of the following means: browsable WWW sites, or downloadable datasets and software, or on CD-ROMS.
We have developed an automated system to categorize, review, and index medical Web sites to facilitate the Internet user's access to particular topics of interest. This automation minimizes deployment time and expands the ability to easily implement Internet users' personal interface preferences.
For developers and site contributors, we have created a multi-mode, user-friendly interface to post or add new information to our database, and to help interactively walk the user through the available categories and directories to find the desired information quickly. This paper presents novel algorithms used by DoctorDigest Medical Resource Reviews Database.

KEYWORDS: Medical Resource Reviews Database, DoctorDigest, medical Internet

INTRODUCTION

DoctorDigest Medical Resource Reviews Database (http://www.doctordigest.com/) is well known to the Medical Internet Community. Our flagship Web sites:
http://www.doctordigest.com/ and http://www.drrecommend.com
have 1,000,000 hits per month as of Summer 2001.
Our staff medical doctors select and evaluate electronic educational medical resources, write reviews, and classify resources by audiences targeted, types of material, medical specialties, and other search criteria.
Our goal is to provide medical reference with reliable sources as well as the relevance and accuracy of data retrieved from Internet.

FEATURES of DrRecommend.com and DoctorDigest.com:

1. We maintain a database of reviews and classifications of medical sites by our Staff MDs.

2. We also maintain public information about 32,000 major medical sites.

3. A "Reference index" is periodically computed for each site and is provided as additional information to users. This reference index of a site is the number of links thereto from other sites. It represents the site's level of visibility on the Web.

4. Our directories allow the user to:
   (1) interactively walk through the available categories and subdirectories to find the desired information;
   (2) use search capabilities within a particular category or set of categories;
   (3) pose a user-friendly, multi-parameter query

5. The user can query our database by the following criteria:
   - purpose (learning, simulation or reference)
   - keywords (appearing in reviews, descriptions, or anywhere on the front page of the site)
   - any combination of the above

6. Smooth translation of all of our pages, external pages, and user queries, using our TranslationWrapper technology

7. Internet search by keywords in conjunction with knowledge categories using our SmartWebSearch technology

8. Automatic evaluation of site availability and speed using our SpeedometerIcons technology

9. External web site mirroring: if a Web site referenced in our database is temporarily down, users can be provided with the full textual content of the Web site.

10. We increase the website visibility using our MiniMirror technology

11. The Medical Doctors on our Consulting Staff can produce reviews and classifications remotely over the Internet using regular browsers and automatically submit for approval by our resident Editorial Board.
   - health specialties (e.g. "Dermatology or Cardiology")
   - audience targeted (e.g. "Nurses or Dentists")
   - price range (e.g. CD-ROMS for $100-$200; free websites)
   - reference index (e.g. sites referred to from at least 1000 other sites)
   - quality rating (e.g. 3-4 stars awarded by our Staff MDs)

12. MEDICAL SUPERSEARCH

We provide query results from various Internet medical databases which do not store their data in static HTML pages. Generic Internet search engines are unable to access, and thus index, this data. The user poses a keyword query and as result we bring many entries from various medical databases, in combination with the more conventional web page search via search engines.

1. GENERAL APPROACH

We review electronic medical information resources available in the following media: browsable WWW sites, or downloadable datasets and software, or on CD-ROMS. Our medical doctors evaluate the resources, write reviews for the medical Internet community, and classify resources by audiences targeted, types of material, medical specialties, and other search criteria.

DrRecommend provides some unique medical and Internet specific features:

The medical doctors who produce the medical review database work remotely over the Internet using regular browsers. In addition they can enter the information in their native language due to the translation capabilities available on our web site.

A comprehensive set of attributes, which describes each medical site is provided to make queries more precise.
Specific, comprehensive categories can be used to "drill-down" to the desired site or information via: health specialties; targeting audience; price range; reference index; quality rating; purpose (learning, simulation or reference); and multiple combinations of the above mentioned queries.

In addition, keywords can be combined with the above queries.

Through the use of our advanced Internet search technology, we provide query results from various Internet medical databases, which do not store their data in static HTML pages [3]. Generic Internet search engines are unable to access, and thus index, this data. Thus, this capability allows us to significantly enrich the results of user's queries.

For developers and site contributors, we have created a multi-mode, user-friendly interface to post or add new Web sites or information to our database. This interface allows Internet viewers to:
(1) interactively walk through the available categories and directories to find the desired information;
(2) use search capabilities within a particular category or set of categories;
(3) pose a user-friendly, multi-parameter query which has the same functionality as approach (1), but takes less time to reach to the desirable information.

Like most search engines, our search engine can employ the use of keywords to search for sites and information on the Internet. In addition, however, our site allows users to narrow their searches within a human selected knowledge category or combination of categories. This combination of the using a human selected category plus an automated keyword search gives significantly more precise results. More specifically, our engine searches:
- within the names and titles of Internet sites, as Yahoo does
- within the entire textual contents of Internet sites, as Altavista does
- using a combination of both of these approaches
- using these approaches along with category information
- using these features with the translation features mentioned above.

On our site all links are monitored in real time using our Live-Speed/Speedometer checking feature. Each time a page is opened or query results are displayed, this mechanism checks the status (i.e., is the link alive) and speed capability of each external link on the page. If the link is alive, a speedometer is displayed which indicates the current speed capability of the external site, otherwise, a “stop” sign is displayed. This feature is particularly useful in reducing user frustration and saving valuable time when external sites are temporarily down or very slow as users can choose to focus on other links. Other Web sites and search engines force users to check for themselves, if the web sites are working.

If a Web site referenced in our database is temporarily down, users will be provided with the full textual content of the Web site, if they desired that information. A "Reference index" is periodically computed for each site and provided as additional information to users. This index is

the number times the particular site is referenced by other sites. It is used to represent the site's level of visibility on the Web.

2. ARCHITECTURE ISSUES

There are 3 classes of data, which are available for each particular medical site:

- invariable data such as audience, purpose, health specialty, rating, base URL;
- periodically recomputed data: reference index and full-text keyword search data;
- real-time recomputed data: current connection speed to the medical site (speedometer feature).

Scenario of work of reference index automated calculation and full-text search support:

To enable 2 search features (reference index range query and full-text keyword search option) DoctorDigest carries automated procedures depicted on collaboration diagram (Fig.1). The sequence of actions is initiated by entering some Web site record by the reviewer. The reviewer puts various values into WebEntryForm. Some of these values are based on expert opinion and are not re-checked or recalculated by the system: targeted audience, site purpose, health specialty, quality rating. Other values are calculated automatically based on the base URL of the medical site. This reviewer's action disposes the record in the database to the SpiderPopScheduler. The former is a timer-driven routine, which checks whether database records are due for recalculating of Reference Index and Full-Text Keyword Search data. Every record, which was not updated during last week, is due for an update. Updates are scheduled in batches to run overnight, by preference that the oldest records will be scheduled for update first. A record, which was just entered by reviewer, is automatically placed into updateQueue to be processed with the highest priority.

Algorithm of the Reference Index calculator is:

- fetch a record from database;
- determine base URL from this record;
- send a query Web search engines (currently Google.com and AltaVista.com): how many external Web sites have links to a particular medical site.
- Find maximal value of these 2 results and store it to database

Spider Algorithm is:

- fetch a record from the database;
- determine baseURL from this record;
- call spidePage(baseURL);
  End of Main Routine
- function spidePage(URL URL2Spide)

- grab HTML content from URL2Spide;
- parse HTML;
- remove tags from HTML and store remaining text into database;
- find hyperlinks inside the HTML page;
- For each hyperlink:

    - If baseURL is substring of thisHyperlink Then
        - recursirve call of spidePage(thisHyperlink)
    - End If
- End For
- Return
- End of function

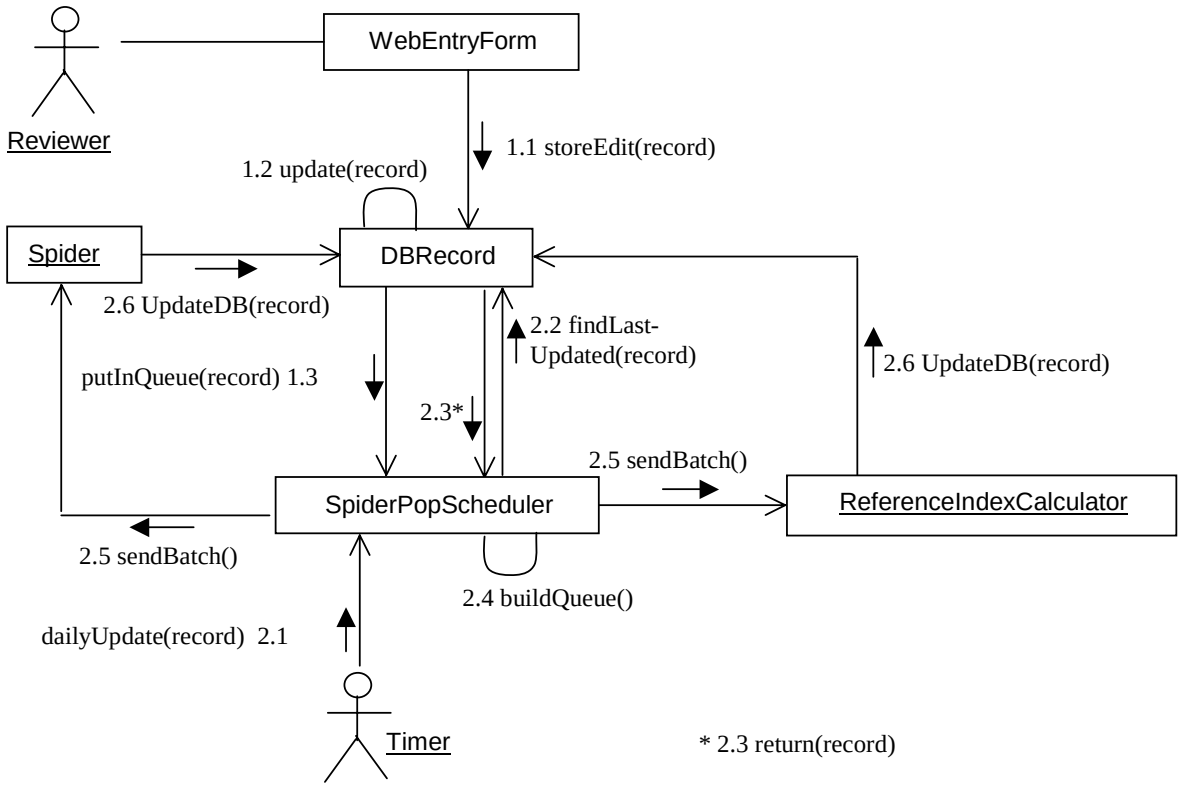

Figure1. Reference index calculator collaboration diagram.

For example - the record “Learn CPR” with base URL=
” http://depts.washington.edu/learncpr/ “ will be spidered as such:

- Page http://depts.washington.edu/learncpr/ will be stored in DB
- Link http://www.washington.edu/medical/som/ will be rejected for storing into DB and further spidering as far as http://depts.washington.edu/learncpr/ is not substring of http://www.washington.edu/medical/som/;
- Link http://depts.washington.edu/learncpr/askdoctor.html will be stored in DB and so on.

Real-time recomputed data: current connection speed to the medical site (speedometer feature).

Speedometer Module functions as follows (Fig. 2):

- user’s browser displays the HTML page;
- each paragraph on this page references the pictogram, which should display the actual connection speed to a particular Web site as a speedometer with indicated speed or stop-sign, if the site is down;

- the browser sends an http request to the Web server to fetch the image of speedometer pictogram;
- Web server forks “Speedometer CGI” process to generate required image of speedometer pictogram;
- “Speedometer CGI” process requests homepage of a particular Internet site;
- when HTML content of this Internet site homepage is received, “Speedometer CGI” process calculates connection speed in Kbytes/sec;
- if connection speed is more than 1 Kbyte/sec – an appropriate speedometer image is generate to depict actual speed, otherwise the site is considered down and “Stop sign” pictogram is generated;
- pictogram is returned back to the Web server, which sends it back to the client’s browser;
- pictogram, which represents connection speed to the site, is displayed on the screen.

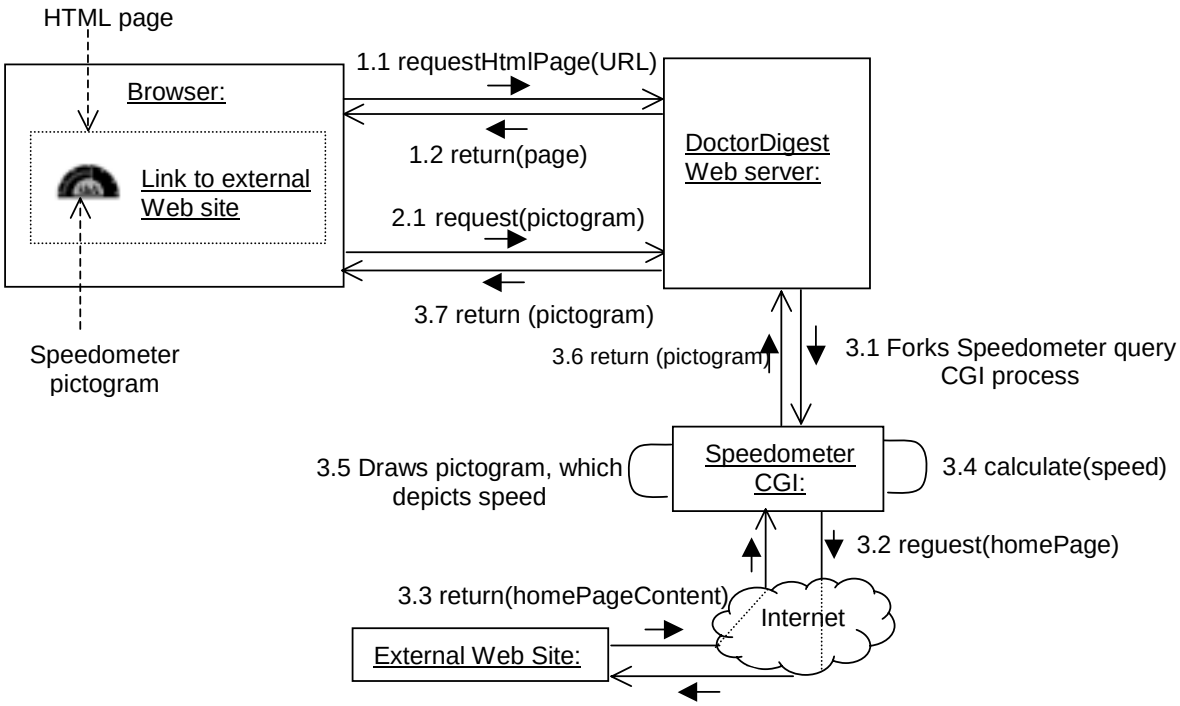

Figure 2. Speedometer module collaboration diagram.

3. DOCTORDIGEST FEATURES

Combined queries.

DoctorDigest allows browsing data according to various criteria, which are set by the user. There are such groups of categories:

- site purpose;
- targeted audience;
- health specialty.

Each particular site could have
- attributes of “site purpose” category:
Learning; Simulation; Reference
- attributes of “targeted audience” category:

Medical Doctor; Dentist; Nurse, Medical Student, Health Consumer/Patient.
- attributes of “health specialty” category:
Anatomy; Physiology; Internal Medicine; Pharmacology and so on.

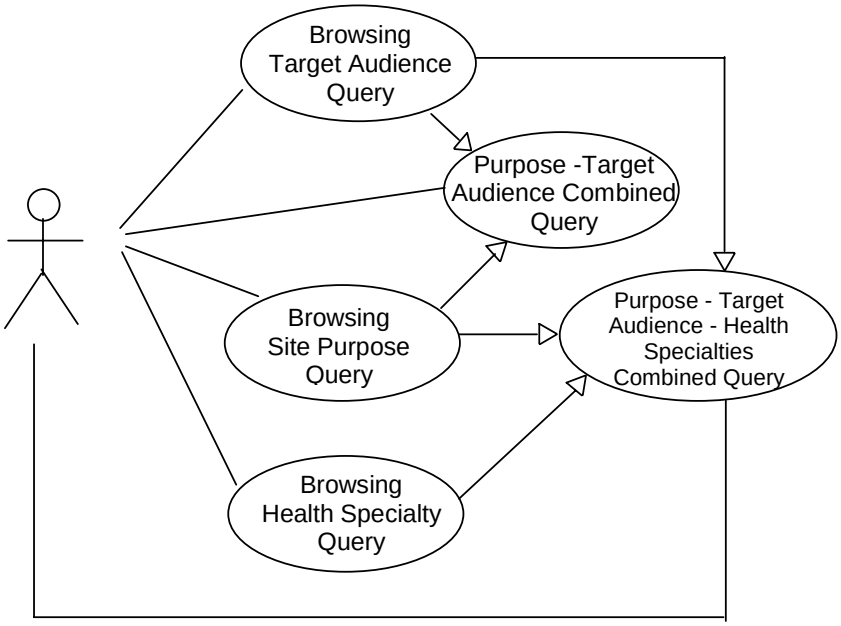

Figure 3. Combined query Use Case diagram

The user can assign desired values of “site purpose” category in “browsing site purpose” query and browse sites (Fig. 3). The user can combine a query by assigning a desired value to two or more categories – effectively performing a combined query.

Range queries.

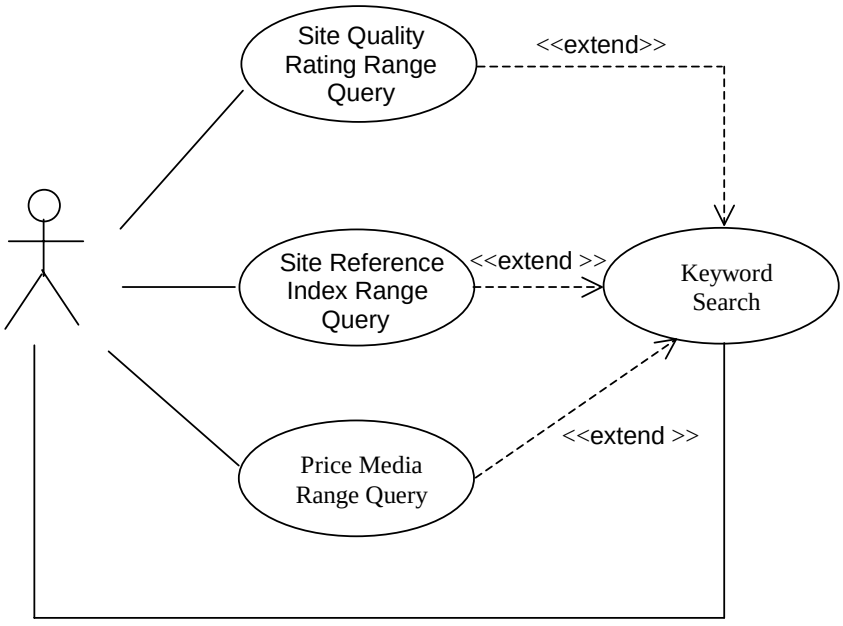

Figure 4. Keyword search with range queries Use case diagram

DoctorDigest employs a range query strategy for numeric data attributes such as: site quality rating; site reference index and price media (Fig. 3). For example, a user can search for medical resources that are within a $10 to $100 price range.

Keyword Search with category/range queries

Site selection criteria could be enriched by applying a keyword search to any of category queries or range queries (Fig. 4). DoctorDigest uses traditional keyword search as well as category-specific keyword search. This makes the results more precise.

4. IMPLEMENTATION AND RESULTS

DoctorDigest is implemented using the concepts of database storage of Web content. The DoctorDigest Web server is based on the Microsoft Internet Information Server technology. Real-time pages' generation is done via VBScript Active Server Pages, connected to a database via ADO/ODBC. Technological procedures -- full-text search indexing and reference index data generation -- are done with Sun Java applications, connected to a database via JDBC. DoctorDigest can work with any SQL/ODBC/JDBC compliant database. In particular, we utilize the Semantic ObjectDatabase Sem-ODB [1] as the database back-end for DoctorDigest, as it is ODBC/JDBC compliant and at the same time offers advanced semantic features.

REFERENCES

[1] N.Rishe Database Design: The Semantic Modeling Approach. McGraw-Hill, N.Y., N.Y, 1992.

[2] DoctorDigest Medical Resource Review Database: http://www.DoctorDigest.com

[3] N. Rishe, O. Dyganova, A. Selivonenko, M. Chekmasov, A. Mendoza. MedFerret: Client-Based Semantic Query Integrator.Proceedings of the 5th World Multiconference on Systemics, Cybernetics and Informatics. Orlando, FL, 2001