Homework 5
Indexing XML data
Wiki

Due date: Tuesday, November 20, in-class

Note

The deadline for **Homework 4, XQuery** is moved to **Monday, November 26, midnight**.

Wiki

Our course now has a wiki, which can be found at

> http://wiki.csc.calpoly.edu/csc560

(there is also a link from the main course page.)

The theme for the wiki is XML Indexing. Here, the term *XML indexing* is understood in the broad sense to mean

> methodology, techniques, data structures and associated algorithms for representing XML data in secondary (and, under some circumstances, primary) storage.

This definition encompasses *shredding* (storage of XML documents in relational databases), *native indexing* (paginated storage of XML documents), *indexing for query processing* (persistent and main-memory indexes for efficient query processing) and more.

XML Indexing

In class, each of you will be given 1–3 research papers in the are of XML indexing. Most of these papers will be assigned to individual students, although some papers may be assigned to two or more. (also, one or two survey papers will be given to each student and discussed in class).

The specific number of papers assigned to each student depends on the size of the papers and their purpose.

You have to do the following:

1. Establish the full citation for each paper given to you. This can be done by studying the Bibliography sections of the papers handed to you (the survey papers contain references to most of the papers distributed in class), as well as by using bibliographic resources on the world wide web. I recomend DBLP:

   http://www.informatik.uni-trier.de/∼ley/db/

   Some papers may contain their publication information in the footnotes on the first page (journal papers and papers in ACM conferences often would have this information).

2. Enter the full citations for all the papers assigned to you on the wiki. A special page, Reading Assignments,

   http://wiki.csc.calpoly.edu/csc560/wiki/Reading%20Assignments

   accessible from the root of the wiki is reserved for this activity. You need to add your name to the list (try doing it in alphabetical order), and under the name, list the full citations of the papers. If you were able to locate the soft copy of the paper, put a link to the soft copy.

   Electronic copies of papers can be found using the following resources:

   - Google Scholar (http://scholar.google.com)
   - Citeseer (http://citeseer.ist.psu.edu/)
   - authors' home pages (use DBLP, then google)
   - instructor (if you were unable to locate the electronic copy, but believe it must be somewhere, talk to me).

   Note, that some resources may be available to you if you are accessing them from a Cal Poly IP address (e.g., ACM Digital Library), while some other resources may appear to have the electronic copy, but would not be accessible directly (e.g., SpringerLink, for which Cal Poly does not seem to have a subscription).

3. Read the papers. Study the XML indexing techniques proposed in the papers. Not all papers are devoted exclusively to indexing. Some papers describe indexing techniques as part of an overall query processing

framework. Your goal is to concentrate on the XML indexing portions of the paper. While I encourage everyone to read full papers they were assigned, if a portion of the paper's content is out of the scope of this assignment, thorough reading of this portion is not required.

4. Post articles on the wiki on the indexing techniques you read about. Each student **must** become an originator of at least one wiki page on an indexing technique. If papers assigned to you describe more than one distinct indexing technique, create a separate wiki article for each technique. If some of the techniques you are describing have also been assigned to another student, make sure you and the other student use the same page to put information on. You can collaborate on such entries, or one student may edit/update the page created by the other student.

   The wiki articles **must contain text authored by you**. Any verbatim quotes from the source material must be explicitely identified (by putting the cited text in double quotation marks, changing font, using quote environment, etc.), and the source must be referenced. Using these rules, you may quote verbatim important text, e.g., formal definitions.

   The wiki article for each method must, at the minimum, have the following:

   - all sources used in the preparation of the article (both assigned papers, and any other sources used).
   - brief introduction to the method.
   - category of the method (shredding vs. native index; schema-aware vs. schema-independent, etc.)
   - a description of the method.
   - a small example.
   - a list of perceived advantages of the method.
   - a list of perceived disadvantages (drawbacks) of the method.

   Additionally, comparisons with other methods described in the wiki are encouraged, as well as any extra information about the method.

   You may also choose to include images to help you explain the workings of the specific methods. Any diagrams, graphs, etc. must be created by you, not scanned in from the paper, although you are allowed (with proper attribution) to use the examples from the papers.

5. Find more material or fact-check existing material. You have two options: (1) conduct a literature search (using references of papers assigned to you as a starting point, e.g.), select some more publications, find electronic copies, read, and describe them on the wiki or (2) select an already existing entry, read papers related to it, and edit the content of the entry to reflect your experiences/viewpoint.

Searches for additional literature may be conducted via studying the bibliography papers you have, or by searching bibliography servers/web search engines for specific papers, paper topics or authors.

6. Prepare an informal 5-10 minute presentation about the indexing techniques you are studying. You will be asked to present during the November 20 (Tuesday) class, or, if necessary, during November 27 (also Tuesday) class.

   Your presentation in class should be informal and informative. You are allowed to use computer in your presentation (e.g. to show slides, or to show the content of an XML file). However, you do not need to prepare a formal presentation based on PowerPoint slides. You can rely on the whiteboard and the markers as well.

Submission

Your wiki activity will be documented, and the log will allow me to determine who contributed what content to each page. When creating wiki pages, make sure that you create a meaningful hierarchy. Also note, that some pages may appear in more than one branch of the hierarchy - e.g., some research on tree index represenations of XML documents belong both to *shredding* and *native XML* categories.

Your oral presentation is due November 20, but you might be asked to actually give it on November 27. No deliverable is required for the oral presentation, but please make sure that any materials you want to use (slides, XML file examples, pictures, graphs, figures, software) are available on the laptop I use in class. The availability can be ensured by mailing me all the files by 8:00am on November 20. My email address is dekhtyar@csc.calpoly.edu.