533 Project Proposal

Matt Williams 46629960

Background and Motivation

XML documents provide structure (somewhat hierarchical) and description to the data that they contain. In a perfect world each content domain (e.g. music collection information such as artist, album, songs) would contain documents that all have exactly the same elements and structure; however, in actuality similar content is stored in heterogeneous and overlapping data collections across the world. Such heterogeneity can arise from differences in language, structure, data type, and breadth and depth of information.

When searching for data contained in XML documents, the additional information can be used to further direct a search but the underlying heterogeneity must be accounted for. A number of studies have focused on relaxing queries to find stored data that is similar to the query but not exactly the same. Various query languages have been built to allow for similarity in hierarchical structure or semantic similarity in both content and element names.

The added expressivity of the query languages offers additional power to the user but in many cases the user cannot efficiently apply the new expressivity without an understanding of the overall layout of the data collection. This layout is where visualization would seem to be helpful and is what I would like to focus my project on.

Solution:

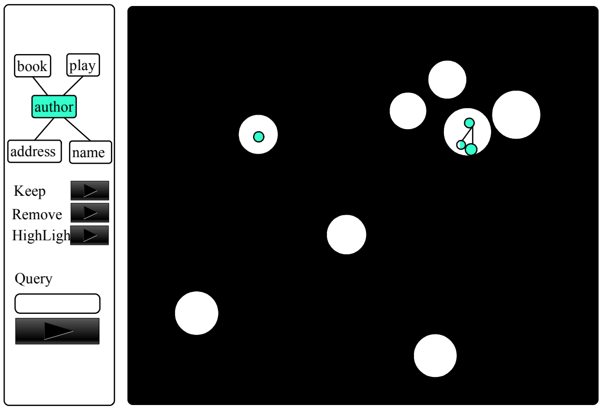
In general, I would like to provide the user a layout of the data using some of the measures above. Most likely the distance function will be based on word frequency using multidimensional scaling. The arrangement will be similar to Leuski's Lighthouse layout (without the ranked title list) with the addition of the XML structure information. In addition to the document display there will be an index of the element names on displayed on the left. This index will allow the user to get an overall understanding of the type of data stored in the collection as well as providing an addition interface for document navigation. Navigation could allow the colouring of or the removal of documents that match selections made from the index. Similarly, the user could colour or select documents based on a database style query (* if time permits). I would also like to add interactivity to allow the user to focus in on the data past the level of the document to show the underlying XML data structure (* if time permits).

Personal Expertise:

Unfortunately I am new to everything: new to XML, new to databases, and new to visualization. But, of course, very interested in learning ☺.

Scenario:

After behind the scenes indexing of an XML data collection, a user will be presented with a galaxy representing the documents and/or document subcomponents. A user can use this information to query the collection to further refine the presentation.



Implementation:

Possiblities

- C++ and Tulip for presentation,
- Java and Shrimp
- Java and yFiles
- Others??

Indexing Database

- Oracle??

Milestones:

1. Create a prototype of the visualization using dummy components and similarity measures
2. Develop Similarity Measure and decide on what will constitute an XML component to be visualized.
3. Prepare test data collection
4. Develop indexing system and index the test collection
5. Implement the visualization of the test collection
6. Add User interactivity to focus and query the display