Ontuition: Intuitive Data Exploration via Ontology Navigation (Demo Paper)[*]

Marco D. Adelfio Michael D. Lieberman Hanan Samet
Center for Automation Research, Institute for Advanced Studies
Department of Computer Science, University of Maryland
College Park, MD 20742 USA
{marco, codepoet, hjs}@cs.umd.edu

Kashif A. Firozvi
School of Medicine
Georgetown University
Washington, DC 20057 USA

ABSTRACT

We present Ontuition, a system for mapping ontologies. Transforming data to a usable format for Ontuition involves recognizing and resolving data values corresponding to concepts in multiple ontological domains. In particular, for datasets with a geographic component we try to identify and extract enough spatio-textual data that we can assign specific lat/long values to dataset entries. Next, a gazetteer is used to transform the textually-specified locations into lat/long values that can be displayed on a map. In addition, we discover non-spatial ontological concepts. This methodology is applied to the National Library of Medicine's very popular clinical trials website (`http://clinicaltrials.gov/`) whose users are generally interested in locating trials near where they live. The trials are specified using XML files. We show how to extract the location data and couple it with a disease ontology to enable general queries on the data with the result being of use to a very large group of people. The goal is to do this automatically for such ontology datasets with a locational component.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Performance

Keywords

Ontuition, ontology, spatio-textual, mapping

1. OVERVIEW

Tabular data is present in many forms on the Web, such as HTML tables, Excel spreadsheets, and CSV files. In addition, many XML files also effectively contain tabular data, where each row of a table can be represented by a subtree of the XML file. Websites and initiatives such as Data.gov[1] are making available a wide variety of datasets in these and other formats. However, making this data deluge easily browseable and searchable by laymen and expert users alike is an ever-present problem.

One such dataset, which serves as a motivating example for this study, is available at ClinicalTrials.gov[2]. This dataset contains information about ongoing and completed clinical trials for drugs and treatments in development, and includes information about each trial such as medical conditions under study, trial locations, and study timelines. Accessing this data is vital for patients seeking alternative treatments available through clinical trials, as well as physicians and other medical personnel, which is evidenced by the website's 90,000+ listed trials and 65,000+ daily visitors at the time of writing. However, the limited search functionality on the website allows only keyword searches, which prevents users from easily finding studies relevant to a certain disease or disease family, or in a certain geographic region (like their neighborhood). For example, a search for "Heart Attack AND Los Angeles" would only return results whose location matched "Los Angeles", even though spatial synonyms would be useful as studies located in nearby "Long Beach" or "Riverside" may also be acceptable.

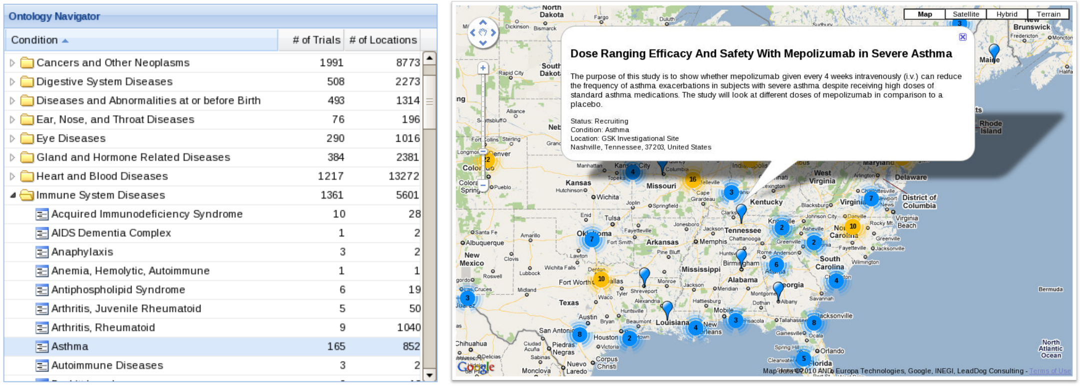
To answer this and related queries effectively, we must understand that "Long Beach" and "Los Angeles" are related data values that correspond to nearby locations. More generally, one of the key ideas here is deciding whether the data in a column of a table contains values that correspond to concepts in an *ontology*, a database containing concepts in some knowledge domain, as well as relationships among concepts. Furthermore, these relationships are often hierarchical in nature, and these hierarchies can be leveraged to effect more useful and intuitive querying. For example, in the clinical trial dataset, the location attribute contains data values which correspond to locations in a *gazetteer*, a database of geographic location names and their corresponding lat/long values, as well as hierarchical relationships among the locations in the sense of containment (i.e., street address, city, county, state, country, continent). Locations are specified using a combination of elements of the hierarchy and sometimes the specification can be ambiguous, in which case, at times, one resorts to using a *geotagger* [1] to resolve the ambiguity. The hierarchy is exploited by queries that range from a point location to a range query which can be in the form of a rectangular window or some other arbitrary shape. Likewise, the disease attribute corresponds to concepts in an ontological hierarchy of disease families, containing disease groupings by e.g. affected organ, as well as synonyms for disease and family names. Here queries can be posed in terms of a specific disease or a family of diseases.

[*] This work was supported in part by the National Science Foundation under Grants IIS-09-48548, IIS-08-12377, CCF-08-30618, IIS-07-13501, and IIS-10-18475, as well as the Office of Policy Development & Research of the Department of Housing and Urban Development, Microsoft Research, Google, and NVIDIA.

[1] `http://data.gov/`

[2] `http://clinicaltrials.gov/`

GIS '10, 03-NOV-2010, San Jose CA, USA



Ontology Navigator

Condition	# of Trials	# of Locations
Cancers and Other Neoplasms	1991	8773
Digestive System Diseases	508	2273
Diseases and Abnormalities at or before Birth	493	1314
Ear, Nose, and Throat Diseases	76	196
Eye Diseases	290	1016
Gland and Hormone Related Diseases	384	2381
Heart and Blood Diseases	1217	13272
Immune System Diseases	1361	5601
Acquired Immunodeficiency Syndrome	10	28
AIDS Dementia Complex	1	2
Anaphylaxis	3	2
Anemia, Hemolytic, Autoimmune	1	1
Antiphospholipid Syndrome	6	19
Arthritis, Juvenile Rheumatoid	5	50
Arthritis, Rheumatoid	9	1040
Asthma	165	852
Autoimmune Diseases	3	2

Dose Ranging Efficacy And Safety With Mepolizumab in Severe Asthma

The purpose of this study is to show whether mepolizumab given every 4 weeks intravenously (i.v.) can reduce the frequency of asthma exacerbations in subjects with severe asthma despite receiving high doses of standard asthma medications. The study will look at different doses of mepolizumab in comparison to a placebo.

Status: Recruiting
Condition: Asthma
Location: GSK Investigational Site
Nashville, Tennessee, 37203, United States

(a)

(b)

Figure 1: Screenshots of (a) the ontology navigator and (b) the map visualization for the clinical trial dataset. The "Asthma" concept has been selected in the ontology navigator, so only trials related to "Asthma" appear on the map. Geographically proximate trials are represented by a cluster icon, with colors and numbers indicating the number of trials in that cluster.

So, if the data in a column fits into an ontology (we try to discover this — see Section 2), our dynamic visualization tool, called *Ontuition*, allows users to use the ontology to query and visualize data values. Queries can be specified in the form of looking for particular concepts or groups of concepts in the ontology. For example, an interactive tree visualization of the disease hierarchy may be used to select studies by disease, while panning and zooming on a map will select studies by location. Query results can likewise be expressed and accessed via ontological concepts. In other words, users may use concepts from one hierarchy (e.g., diseases) to select data values, and use another (locations) to display results, or vice versa — essentially, using one hierarchy to traverse another. When the disease hierarchy is used to traverse the location hierarchy, the result is a set of maps which show the spatial variability of the various selected studies. On the other hand, when locations are used to navigate diseases, the result is analogous to a zoom-in/zoom-out operation. These visual, parallel traversal options allow for simple and powerful querying and exploration (see Section 3).

2. CONCEPT MATCHING

Transforming data to a usable format for Ontuition involves recognizing and resolving data values corresponding to concepts in multiple ontological domains. In particular, for datasets with a geographic component such as ours, we wish to identify and extract enough spatio-textual data that we can assign specific lat/long values to dataset entries. For the ClinicalTrials.org dataset, each trial location includes the city, state (where applicable), and country of the trial, and often the ZIP code and host hospital's name, all of which make geocoding these locations relatively simple. However, many datasets have locations that are not as fully specified, and require more sophisticated techniques (see Section 4). After this step, we ignore entries which have no associated locations.

In addition, we discover non-spatial ontological concepts. In the clinical trials dataset, ontology attributes include the list of conditions under study, medications, and dietary supplements being used. For this dataset, we manually extracted the ontological hierarchy by crawling the source website, since this hierarchy was created specifically for the dataset. However, it is easy to imagine datasets with other attributes where the ontology defining the attribute values is externally available, thus enabling some degree of automation. It may also be possible to perform approximate matching between data values and ontology concepts using textual similarity measures, although we did not pursue this approach, since our ontology was derived from the data and thus approximate matching was unnecessary.

3. VISUALIZATION

We have implemented a prototype of Ontuition, accessible at `http://ontuition.umiacs.umd.edu`, that allows querying, exploration, and visualization of a dataset using ontologies. A screenshot of our prototype, used to explore the clinical trials dataset, is presented in Figure 1. Figure 1a shows a tree-based *ontology navigator*, initialized with our ontological hierarchy of disease families. In the figure, the user has indicated interest in selecting studies involving "Asthma", which total 165 studies. Selecting one or more diseases or disease families is as simple as selecting the appropriate concepts in the hierarchy. Figure 1b shows a map-based visualization of the studies selected in the ontology navigator, by which the user can explore the spatial extent of the selected studies.

4. FUTURE WORK

The main goal of this system is to allow exploration of datasets using a combination of an ontology navigator and a map. To enable this exploration, geographic and ontological attributes of the data must be identified, and ontological concepts resolved appropriately. As mentioned in Section 2, we manually generated the disease ontology from our dataset, but such manual effort is not feasible for large collections of heterogeneous data. A successful automated technique requires more advanced methods. One possibility is to apply the concepts of *row coherence* and *column coherence* [2], both of which use spatial relationships between cells in the tabular structure to identify spatio-textual attributes in tabular data and resolve the textual specifications to precise lat/long values. We could apply similar principles for geocoding trial locations and for more general ontological concept identification using the tree structure of XML documents.

5. REFERENCES

[1] M. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proc. of ICDE*, pages 201–212, Long Beach, CA, Mar. 2010.

[2] M. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. Spatio-textual spreadsheets: Geotagging via spatial coherence. *Proc. of ACM SIGSPATIAL GIS*, pages 524–527, Seattle, WA, Nov. 2009.