Ontologies as Nested Facet Systems for Human-Data Interaction

Guo-Qiang Zhang a,*, Shiqiang Tao a, Ningzhou Zeng b, and Licong Cui a

a The University of Texas Health Science Center at Houston, Houston, Texas, USA
E-mails: guo-qiang.zhang@uth.tmc.edu, shiqiang.tao@uth.tmc.edu, licong.cui@uth.tmc.edu
b Department of Computer Science, University of Kentucky, Lexington, Kentucky, USA
E-mail: ningzhou.zeng@uky.edu

Abstract. Irrespective of data size and complexity, query and exploration tools for accessing data resources remain a central linkage for human-data interaction. A fundamental barrier in making query interfaces easier to use, ultimately as easy as online shopping, is the lack of faceted, interactive capabilities. We propose to repurpose existing ontologies by transforming them into nested facet systems (NFS) to support human-data interaction. Two basic issues need to be addressed for this to happen: one is that the structure and quality of ontologies need to be examined and elevated for the purpose of NFS; the second is that mappings from data-source specific metadata to a corresponding NFS need to be developed to support this new generation of NFS-enabled web-interfaces. The purpose of this paper is to introduce the concept of NFS and outline opportunities involved in using ontologies as NFS for querying and exploring data, especially in the biomedical domain.

Keywords: Web-interface, Ontology, Biomedical Big Data, Nested facet system, User experience

1. Introduction

When it comes to exploring and accessing biomedical data, often is the question asked: “Why can’t it be as easy as shopping on Amazon?”

To answer this question, we need to identify the core technologies that made online-shopping experience “pleasant,” and then hope to be able to apply a similar strategy for exploring and accessing biomedical data, big or small. Among many drivers of online-shopping [1], faceted search [2, 3] capability is perhaps one of the most ubiquitously applied information-retrieval techniques. Indeed, studies show that faceted search can help enhance user experience in a variety of settings [4–7].

Semantic labeling is the missing link between an entity (such as consumer goods for online shopping or study subjects in a clinical data warehouse) and ways to identify and accessing it through means such as a web-based user interface. This is well-articulated in a recent article by Balog [8].

Semantic labeling enables facets, such as size, color, make, price to be annotated for entities such as shoes in an online store. Faceted organization and presentation of metadata on products is the key mechanism that allowed consumers of web-sites to quickly narrow down from millions of products to items of interest using such simple facets. The entities for biomedical data, however, are highly complex and there does not exist a corresponding small set of semantic labels to support faceted search. For example, clinical data, captured as a part of patient care, are highly complex, and includes demographics, medical history, lab reports, diagnosis, medication, and discharge summary.

Biomedical ontologies are suitable as semantic labels for biomedical entities. However, these ontologies, intended to model and capture concepts and their relations in the biomedical domain, are broad and complex. For example, SNOMED CT [9, 10], the largest clinical terminology used worldwide, contains over 300,000 concepts and over 1.5 million relations. The
National Cancer Institute thesaurus (NCIt) [11, 12], on the other hand, is a biomedical terminology produced by NCI Enterprise Vocabulary Services, containing more than 140,000 concepts related to cancer. Such size and complexity raise basic questions related to their potential role as facets for web-based user interfaces: What, if any, structural transformations are needed for ontologies to play the role of facets for information retrieval? Is it feasible to have ontologies to play the role of facets? What kind of desirable properties are required for ontologies to support facet-oriented user interaction? How to measure and evaluate the performance of this approach?

In this paper we propose the concept of nested facet system (NFS), outline a strategy to transform existing ontologies into NFS to support human-data interaction, and identify exemplar research questions related to the use of NFS to enhance user experience in human-data interaction.

2. Nested Facet System

A facet is a semantic label of an entity along multiple possible axes or dimensions. Facets correspond to properties of the entity of interests. For example, online vendors use facets to label their product using readily available information about their type, brand, price, and support consumer shopping experience through faceted search [2].

A nested facet, or higher-order facet, is a facet that includes a (finite) collection of other facets as its components. In this context, traditional facets are primitive facets, those that are not made of other facets. A nested facet system is a set of nested facets (we call them facets from now on) with a taxonomy relation (i.e., subclass, subsumption, or hierarchical relation) among them.

Formally, given a set of entities \( E \), a primitive facet \( p \) with value space \( D(p) \) is a collection of parameterized semantic labels \( p(t) \), such that for each member \( e \in E \) and for each \( t \in D(p) \), \( e \) can be classified as having property \( p(t) \) or not. For each \( e \in E \), we write \( e \models p(t) \) if entity \( e \) has facet \( p \) with value \( t \). We write \( \[ p(t) \] \) for the set \( \{ e \in E \mid e \models p(t) \} \) for \( t \), and \( \[ p \] \) for the set \( \{ e \in E \mid e \models p(t), t \in D(E) \} \). In extreme cases, we allow \( D(p) \) to be empty, and \( p \) can be specified without a parameter.

A nested facet is either a primitive facet or a facet with other nested facet as components. For nested facet \( p \), we write \( p = \{ q_1, \ldots, q_n \} \), where \( q_i \)’s are the defining components of \( p \). In particular, \( p(\vec{t}) \) represents \( \{ q_1(t_1), \ldots, q_n(t_n) \} \), where \( \vec{t} = (t_1, \ldots, t_n) \). The meaning of a nested facet is defined as the aggregate of its components. So, for \( p(\vec{t}) = \{ q_1(t_1), \ldots, q_n(t_n) \} \), we define \( \| p(\vec{t}) \| \) to be the set union \( \bigcup_{1 \leq i \leq n} \| q_i(t_i) \| \) and \( \| p \| \) to be the set union \( \bigcup_{1 \leq i \leq n} \| q_i \| \).

In general, the value space \( I \) for facet \( p \) can be partially specified for only a subset of its component facets \( q_i \). In such cases, those components \( q_i \) without a specified parameter \( t_i \) are interpreted as the entire set \( \| q_i \| \). For example, if \( p = \{ q_1, q_2, q_3 \} \), then \( \| p(t) \| \) is interpreted as the set \( \| q_1(t_1) \| \cup \| q_2 \| \cup \| q_3 \| \). Such a convention corresponds to the intuition that when using facets to retrieve their corresponding entities, unspecified parameters, as well as un-selected facets, do not contribute to conditions for narrowing down the resulting entities.

When \( p = \{ q_1, \ldots, q_n \} \), we say that \( q_i \) is a subfacet of \( p \) for each \( 1 \leq i \leq n \), and write \( q_i \prec p \). A nested facet system \( F \) is a collection of primitive and nested facets closed under the sub-facet relation. We write \( \preceq \) for the reflexive, transitive closure of \( \prec \), which is a partial order on \( F \). By definition, if \( q \prec p \), then \( \| q \| \subseteq \| p \| \). Therefore, if \( q \preceq p \), then \( \| q \| \subseteq \| p \| \). We call this the soundness property for nested facet system.

3. Biomedical Ontologies and Nested Facet Systems

Biomedical ontologies serve as the semantic scaffolding for us to fully capitalize on the transformative opportunities of the increasingly large amounts of digital data produced by the biomedical research enterprise. For example, the BioPortal [13, 14], is the world’s most comprehensive repository containing over 600 ontologies and over 7 billion concepts that have been used to support a wide spectrum of scientific projects. Biomedical ontologies provide the basis for scientific rigor during the process of data collection, annotation, management, analysis, and sharing in biomedicine. They not only serve as metadata standards, but also play a vital role in down-stream systems as a declarative knowledge source [15]. For example, SNOMED CT [9, 10], the most comprehensive and precise clinical health terminology product in the world, facilitates the clear exchange of health information in Electronic Health Records (EHRs), leading to higher quality, consistency and safety in healthcare delivery [16, 17].
Ontological systems are not designed \textit{a priori} as nested facet systems. But what if we attempt to reuse them as facets to support user interfaces? An intuitive idea is to leverage the hierarchical or is-a relation, the structural backbone of most ontologies. For a given ontology such as SNOMED CT, we can treat each concept \( c \) as a facet \( p \), and build a nested facet system by defining \( p = \{ q_1, \ldots, q_n \} \) if the concepts corresponding to the \( q_i \)'s are the (immediate) lower neighbors of \( p \). In other words, if \( p \) is the facet corresponding to \( c \), and \( q_i \)'s are the facets corresponding to all the (immediate) lower neighbors of \( c \) with respect to the hierarchical relation, then make \( p \) a nested facet with \( q_i \)'s its components.

For this (very reasonable) intuition to work, the following questions must be answered:

1. Does this construction obey the soundness property mentioned at the end of the previous section, i.e., if \( q \sqsubseteq p \), then \( \| q \| \subseteq \| p \| \)?
2. Does this construction obey the completeness property, which states that for any \( e \in E \), if \( e \models p \), then for some \( i \), \( e \models q_i \) with \( 1 \leq i \leq n \)?
3. What are the primitive facets?

Intuitively, soundness means that all items below each facet are relevant to the facet. Completeness means that any items or facets relevant to a specific facet are already contained in and accessible through the facet. The soundness and completeness properties of NFS directly affects query performance in terms of precision and recall. Incomplete facets will reduce recall, while unsound facets will reduce precision.

In the following sections we discuss such questions in more depth using biomedical ontologies and clinical data resources as examples, and provide use cases to demonstrate the feasibility and work involved to implement this approach.

4. Clinical Data Resources and Related Ontologies

4.1. Clinical Data Warehouse

The entity \( E \) for clinical data consists of patients. Clinical data from electronic health records (EHR) are critical for analyses to improve health care delivery. Clinical data warehouses are EHR data made available for research. Examples include i2b2 data warehouses [18, 19], PCOR – the National Patient-Centered Clinical Research Network [20], and Observational Health Data Sciences and Informatics (OHDSI) research network [21] with an open, community data standard called the Observational Medical Outcomes Partnership (OMOP) Common Data Model. SNOMED CT is a common component of all these data sources.

4.2. Health Claims Data

Health claims data (also called administrative data) such as Cerner Health Facts, Truven Market Analytics, and Optum Health Data and Analytics, are those collected for the purpose of health insurance claims. They include information at the patient encounter level regarding diagnoses, treatments and billed and paid amounts. This is a valuable data source for research aimed at driving improvements in population health to address issues related to cost, quality and outcomes. The use of administrative data can complement EHR data by providing a regional or national scale view. Because of the health claims context, main vocabularies for health claims data involve diagnosis (ICD 9 and ICD 10), procedure code (CPT), and medication (RxNorm).

Clinical data and health claims data are domain-agnostic: they cover the entire spectrum of disorders and disease domains. Domain-specific data resources, however, are those cover a signal medical specialty, but with greater depth. We highlight several such resources next.

4.3. The National Sleep Research Resource

The gold standard for sleep diagnosis is polysomnography, which monitors physiological processes including electroencephalogram (EEG - brain waves), electromyogram (EMG - muscle tone), and electro-occulogram (EOG - eye movements). The recorded polysomnograms provide comprehensive data about biophysical changes that occur during sleep and characterize the association between sleep and other public health related problems. The National Sleep Research Resource [22, 23] is a retrospectively annotated repository of 30,000 overnight sleep recordings. The NSRR offers free and open web access to large collections of de-identified, well-annotated national repository of sleep data, including PSGs which are linked to risk factor and outcome data for participants in major NIH studies. Since its launching in 2014, 240TB of data have been downloaded by over 3,000 users around the world from the NSRR portal sleepdata.org.

NSRR uses the sleep domain ontology as the canonical vocabulary for across-study data mapping.
4.4. The Center for SUDEP Research

The Center for Sudden Unexpected Death in Epilepsy Research (CSR) [24] manages another domain-specific clinical research data resource. The CSR has prospectively collected high grade multimodal data including high-resolution electroencephalographic signal, research-grade brain MRI, biochemical and DNA samples together with detailed phenotypic data for more than 3,000 epilepsy patients. Similar to NSRR, a disease-specific ontology called Epilepsy and Seizure Ontology has been created as a part of the CSRinformatics infrastructure process.

4.5. Cancer Registries

For cancer research, the US National Cancer Institute’s Surveillance Epidemiology and End Results (SEER) program [25] coordinates a collection of state-based SEER registries. These state-centered cancer registry receiving data about new cancer cases from healthcare facilities and physicians within the state. Typically, five aspects of data are captured: patient data, case data, follow-up, therapy data and pathology reports. Patient data consists of variables including various patient-related information such as demographics, race, ethnicity, smoking, and clinical trial participation information. Case data captures variables for diagnosis, morphology, staging, biomarkers, and other categories. Follow up information contains variables including follow-up physician, date of last contact, survival status, and cancer status. Therapy data records variables with information on surgery, chemotherapy, radiation, and other treatment modalities.

In general, SEER data are considered to be among the most accurate and complete population-based cancer registries in the world that includes stage of cancer at the time of diagnosis and patient survival data. Cancer registries uses NAACCR data dictionary [26] for variable definition, and is only partially mapped to NCIt. This is where work on primitive facets is needed in order to use NCIt as NFS.

5. NFSs for Querying Clinical Data: Progress and Opportunities

For data resources based on SNOMED CT and NCIt, progress has been made [29] but gaps remain in the area of the basic questions mentioned in Section 3. For example, even though SNOMED CT and NCIt satisfy the soundness and completeness properties “for the most part” using the NFS constructs outlined in Section 3, enough facet instances exist where such properties are violated [30]. Such violations affect the soundness and completeness properties of facets, leading to reduced precision and recall for query interfaces using NFS. Interestingly, non-lattice auditing methods can precisely identify and potentially fix such issues [31–35].

6. Conclusion

We outlined a general approach for constructing nested facet systems from ontologies. We highlighted use cases for clinical data, and discussed progress and remaining challenges. Given the importance of faceted search, our proposed approach deserves further study. Efforts in developing experimental interfaces supporting NFS will be highly desirable and impactful for accessing biomedical data for research.

References


