Using Wikipedia to Enable Entity Retrieval and Visualization Concerning the Intellectual/Cultural Heritage

Athenikos, Sofia J.
sofia.j.athenikos@acm.org
Drexel University, Philadelphia, USA

At the 2009 Digital Humanities conference I presented my paper on the WikiPhiloSofia (http://research.cis.drexel.edu:8080/sofia/WPS/) project (Athenikos and Lin, 2009), which was concerned with extraction and visualization of facts, relations, and networks concerning philosophers using Wikipedia (http://www.wikipedia.org/) as the data source. In this proposal, I present a related, extended project in progress, entitled PanAnthropon, which incorporates the problems of retrieving entities in response to a query and retrieving entities related to a given entity and which extends the scope of application to domains other than philosophy.

1. Background

Traditional information retrieval is concerned with retrieving documents that are potentially relevant to a user's query. The relevance of a document to a given query is usually measured by lexico-syntactic matching between the terms in the query and those in the document (title). Familiar Web search engines, such as Google and Yahoo, for example, return a ranked list of Web pages that contain all or some of the keywords in the query entered by a user. The Semantic Web (Berners-Lee et al., 2001) initiative aims at transforming the Web of pages (documents) into the Web of entities (things in the broadest sense) (cf. OKKAM project (http://www.okkam.org/) (Bouquet et al., 2007)). Information retrieval on the Semantic Web is no longer a matter of retrieving documents via semantics-unaware keyword matching but a matter of retrieving entities that satisfy the semantic constraints imposed by the query, i.e. those that are of specific semantic type and that satisfy the given semantic conditions. Wikipedia has become an important semantic knowledge resource (cf. Zesch et al., 2007) thanks to its unique set of semi-structured semantic features and the huge amount of content covering a wide range of topics. What renders Wikipedia more interesting is the fact that it can be considered as a self-contained web of entities. Each Wikipedia article is concerned with one entity, and the given entity is connected to other entities via explicit semantic relations as in infoboxes and wikitables or via implicit semantic relations as in hyperlinks.

2. Motivation

Through the WikiPhiloSofia project I demonstrated extracting, retrieving, and visualizing specific facets of information, not documents, concerning entities of a selected type, namely, philosophers, by exploiting the hyperlinks, categories, infoboxes, and wikitables contained in Wikipedia articles. The interface that I created enables the users to select a focus of query in the form of an entity (philosopher) or a pair of entities (philosophers) and then to retrieve entities that satisfy specified conditions with respect to the given entity or pair of entities. However, the project did not consider the problems of retrieving entities as answers to queries, semantically typing entities, or retrieving related entities by type and condition.

The proposed PanAnthropon project takes up the aforementioned problems left out of the WikiPhiloSofia project. The dual objective is to enable retrieval of entities that directly answer a given semantics-based query and to enable retrieval of related entities by semantic type, subtype (role), and relation, by using information extracted/integrated from Wikipedia. The project applies the approach to the entities concerning the intellectual/cultural heritage – people, works, concepts, etc. The Web portal interface thereby constructed will allow the users to retrieve entities that directly answer their queries as well as to explore people, works, concepts, etc. *in relation to* other people, works, concepts, etc.

3. Conceptualization

In the proposed project, "entities" are conceived of as things of all kinds that have certain properties (or attributes). The “type” of an entity is considered as a generic kind (or class or category) into which the given entity is classified, e.g., person, work, etc. In general, the type of an entity is fixed and exclusive in the sense that an entity that belongs to one type does not or cannot belong to other types. The “subtype” of an entity refers to a subclass or subcategory into which the entity can be classified, under a given type. The subtype of an entity is fluid and non-exclusive in the sense that an entity may belong to more than one subtype (under a given type). This is especially so in the case of person-type entities, and thus a subtype may better be understood as a “role” in this case. In general, there are multiple subtypes under a given type, and the former can be further classified into still more specific subtypes. A type or subtype of an entity can be considered as a special kind of property. A “fact” concerning an entity refers to a tuple consisting of <entity, attribute, value, [context]>, which adds the optional “context” element to the <subject, predicate, object> triple model. An entity can have “relations” to other entities, given its properties. The kinds of properties and relations that are relevant or of interest concerning an entity, except certain basic facts, depend on the domain at issue. An entity may belong to multiple domains, but not every subtype, property, or relation is relevant or equally important in one domain as in another domain. The project therefore intends to build a portal consisting of sub-portals representing different domains.

4. Methodology

4.1. Data Extraction and Processing

The pre-processing stage of the project (for each domain) concerns: (1) compiling a seed list of entities of interest by extracting names from various lists and categories in Wikipedia, (2) downloading Wikipedia article pages for each entry on the list, and (3) inspecting typical attributes and (types/subtypes of) values contained in infoboxes, wikitables, etc. The main processing stage concerns extracting information on an entity and related entities from each Wikipedia page. The semantic type/subtypes of a given entity are extracted and/or assigned. Semi-structured templates and portions of the article are processed so as to extract attribute–value (or predicate–object or relation–entity) pairs. The related entities are matched to the entities on the seed entity list. Additional Wikipedia pages are downloaded for entities not matched on the list, and the information on those entities is extracted. The (optional) post-processing stage concerns converting the data stored in a MySQL database to XML files and RDF triples, thereby creating a semantic data repository that can be linked to other resources involved in the Linked Data (`http://linkeddata.org/`) initiative in the latter case.

4.2. Semantic Search Interface

The semantic search interface created will support three types of query and retrieval. The first type of query/retrieval concerns retrieval of entities that correspond to queries the expected answers of which are entities. The second type of query/retrieval concerns retrieval of entities related to an entity, according to type/subtypes and specified properties/values. The third type of query/retrieval concerns retrieval of facts concerning an entity. The interface will also incorporate some of the visualization features available in the WikiPhiloSofia portal interface.

5. Current Application

The film domain has been chosen as the initial proof-of-concept domain of application. In my presentation I will demonstrate the entity retrieval functionalities with 1.5+ million (and growing) facts about 11370 films, 69545 persons, 74545 film roles, 253 places, 6033 dates, etc.

6. Related Work

The task of retrieving entities in response to user queries using the information in Wikipedia has since 2007 been the focus of the INEX (Initiative for the Evaluation of XML Retrieval) XML Entity Ranking (XER) Track (de Vries et al., 2008; Demartini et al., 2009). Unlike in the INEX XER Track, the proposed project addresses the task by extracting information from the HTML Wikipedia files and building a knowledge base based on it. The task of constructing a knowledge base by extracting information from templates in Wikipedia such as infoboxes has been attempted in large scale by, e.g., Auer and Lehmann (2007) and Suchanek et al. (2007). There is also the DBpedia (`http://dbpedia.org/`) knowledge base, which contains the information extracted from Wikipedia. The proposed project, however, utilizes the information in the main content of Wikipedia articles, as well as templates, to enable and enhance entity retrieval. It will also provide a more flexible working search interface for both general and entity-specific queries.

7. Conclusion

The PanAnthropon project utilizes Wikipedia as a semantic knowledge source for entity retrieval and applies the approach to materials concerning the intellectual/cultural heritage. The semantic search interface created will allow the users to retrieve entities that directly answer their queries as well as to explore various semantic facets concerning those entities. As such, it will provide a useful resource for digital humanities.

References

Athenikos, S.J., Lin, X. (2009). 'WikiPhiloSofia: Extraction and Visualization of Facts, Relations, and Networks Concerning Philosophers Using Wikipedia'. *Conference Abstracts of Digital Humanities 2009.* Pp. 56-62.

Auer, S., Lehmann, J. (2007). 'What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content'. *Proceedings of 4th European Semantic Web Conference (ESWC 2007).* Innsbruck, Austria, June 2007.

Berners-Lee, T., Hendler, J., Lassila, O. (2001). 'The Semantic Web'. *Scientific American.* **5 (May 2001)**.

Bouquet, P., Stoermer, H., Giacomuzzi, D. (2007). 'OKKAM: Enabling a Web of Entities'. *Proceedings of the 16th International World Wide Web Conference (WWW 2007).* Banff, Alberta, Canada, 8-12 May 2007.

Demartini, G., de Vries, A.P., Iofciu, T., Zhu, J. (2009). *Overview of the INEX 2008 Entity Ranking Track, INEX 2008.* LNCS. Heidelberg: Springer-Verlag, Berlin. V. 5631, pp. 243-252.

de Vries, A.P., Vercoustre, A.-M., Thom, J.A., Craswell, N., Lalmas, M. (2008). *Overview of the INEX 2007 Entity Ranking Track, INEX 2007.* LNCS. Heidelberg: Springer-Verlag, Berlin. V. 4862, pp. 245-251.

Suchanek, F.M., Kasneci, G., Weikum, G. (2007). 'YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia'. *Proceedings of the 16th International World Wide Web Conference (WWW 2007).* Banff, Alberta, Canada, pp. 697-706.

Zesch, T., Gurevych, I., Mühlhäuser, M. (2007). 'Analyzing and Accessing Wikipedia as a Lexical Semantic Resource'. *Proceedings of the Biannual Conference of the Society for Computational Linguistics and Language Technology.* Tübingen, Germany, April 2007.