DISSERTATION ABSTRACT

Enabling Entity Retrieval by Exploiting Wikipedia as a Semantic Knowledge Source[1]

Sofia J. Athenikos
College of Info Science and Technology
Drexel University
Philadelphia, PA, USA
sofia.j.athenikos@acm.org

29 August 2011

Abstract

This dissertation research, PanAnthropon FilmWorld, aims to demonstrate direct retrieval of entities and related facts by exploiting Wikipedia as a semantic knowledge source, with the film domain as its proof-of-concept domain of application. To this end, a semantic knowledge base concerning the film domain has been constructed with the data extracted/derived from 10,640 Wikipedia pages on films and additional pages on film awards. The knowledge base currently contains 209,266 entities and 2,345,931 entity-centric facts. Both the knowledge base and the corresponding semantic search interface are based on the coherent classification of entities. Entity-centric facts are also consistently represented as <entity, attribute, value, note> tuples. The semantic search interface (http://dlib.ischool.drexel.edu:8080/sofia/PA/) supports multiple types of semantic search functions, which go beyond the traditional keyword-based search function, including the main General Entity Retrieval Query (GERQ) function, which is concerned with retrieving all entities that match the specified entity type, subtype, and semantic conditions and thus corresponds to the main research problem. Two types of evaluation have been performed in order to evaluate (1) the quality of information extraction and (2) the effectiveness of information retrieval using the semantic interface. The first type of evaluation has been performed by inspecting 11,495 film-centric facts concerning 100 films. The results have confirmed high data quality with 99.96% average precision and 99.84% average recall. The second type of evaluation has been performed by conducting an experiment with human subjects. The experiment involved having the subjects perform a retrieval task by using both the PanAnthropon interface and the Internet Movie Database (IMDb) interface and comparing their task performance between the two interfaces. The results have confirmed higher effectiveness of the PanAnthropon interface vs. the IMDb interface (83.11% vs. 40.78% average precision; 83.55% vs. 40.26% average recall). Moreover, the subjects' responses to the post-task questionnaire indicate that the subjects found the PanAnthropon interface to be highly usable and easily understandable as well as highly effective. The main contribution from this research therefore consists in achieving the set research goal, namely, demonstrating the utility and feasibility of semantics-based direct entity retrieval.

1 This research has been partially supported by a 2011 Eugene Garfield Doctoral Dissertation Fellowship awarded by Beta Phi Mu the International Library & Information Studies Honor Society.