Definition of User Profiles based on the YAGO Ontology

Silvia Calegari and Gabriella Pasi

DISCo, Universitá degli Studi di Milano-Bicocca,
vle. Sarca 336/14, 20126 Milano (Italy),
{calegari,pasi}@disco.unimib.it

Abstract. In this work, we consider the problem to personalize user's Web searches for improving the quality of results. To this aim, we propose a preliminary methodology that allows to define a conceptual user profile based on the YAGO ontology.

1 Introduction

To overcome the limitations of the "one size fits all" approach of search engines, personalized approaches to Information Retrieval have been proposed. Personalized search is based both on modeling the user's context by a user's profile that represents the user's preferences, and on the definition of processes that exploit the knowledge represented in the user profile to tailor the search outcome to users' needs. The accurate definition of a user profile plays then a central role to define effective approaches to personalization. Up to now, bags of words, and vectors or graph-based representations have been mainly used to define users' profiles. To improve the quality of the knowledge represented in user profiles, in some recent works, external knowledge sources (i.e., WordNet [2], or Web directories as the ODP [7] and the Yahoo! Web directory [4]) have been considered to represent in a more structured way the user context. The use of an ontology allows to give a more structured and expressive knowledge representation with respect to the above mentioned approaches [3].
A user profile is defined based on the analysis of the information characterizing the user's interests and preferences. Elicitation of user's interests and preferences is not the focus of the research reported in this paper. Numerous approaches have been proposed in the literature to this aim [6]. Our objective is the formal definition of an ontological user profile based on the use of YAGO as an external reference knowledge. YAGO [8] is a general purpose ontology containing several millions of entities and facts. Only the entities and facts which match the appropriate user's interests are used to derive the user profile. To this aim a preliminary methodology aimed at the extraction of the appropriate fragment of the YAGO ontology has been defined. Then the main objective of the research reported in this paper is (assuming to have the user's interests specified as a bag of words) both to extract the portion of YAGO useful for the definition of a user profile, and to organize it into a coherent ontological representation expressed by a language such as RDFS.

2 Building the YAGO-based profile

The novelty of the research reported in this paper is to employ the YAGO [8] ontology as external reference knowledge for building a conceptual user profile. YAGO is a general purpose ontology, and it consists of more than 1.7 million entities (like books, movies, ...), and over 14 million facts about them. The triple $< entity, relation, entity >$ is called a $fact$. All facts are grouped in 99 relations such as *FamilyNameOf,subClassOf,actedIn, etc.* To build the YAGO - based user profile, our methodology is articulated in four phases as sketched in Fig. 1. Our investigation addresses the methodology defined for extracting the sub-part of YAGO related to the user's interests. To produce a bag of words that represents the user's interest, we have decided to consider a set of documents residing on the user's PC related to his/her topical preferences. We have then analyzed them with standard IR techniques in order to extract meaningful terms, i.e. terms representative of the user's preferences (*interest-terms*). Thus, we have developed a strategy that allows to semantically extract the sub-YAGO ontology starting from the interest-terms. A similar approach has been reported in [1], where a set of documents are indexed, and the obtained index terms are semantically linked to a network of concepts, but to the different aim of the automatic construction of hypertexts. Moreover in [1], the external knowledge resource is a taxonomy, i.e. the ACM classification that defines a hierarchy of topics where each topic is a concept. Instead YAGO is an ontology with millions of entities (concepts plus individuals), and several relations with a different semantics; to this aim several rules have to be defined related to the possible relations for associating the index-terms with the right entities.

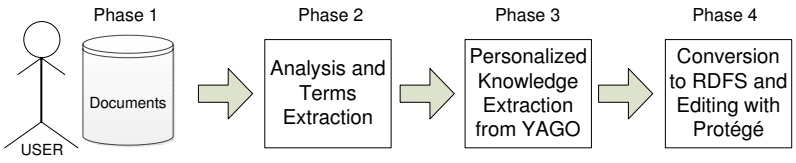

Fig. 1. Phases of profile building

Phase 1. This first phase consists in individuating the user's knowledge that has to be considered to extract the user's interests. In this specific case, we analyzed a set of documents collected by the user and stored in his/her personal computer.

Phase 2. Each document is analyzed in two steps: (1) document preprocessing and (2) term frequency analysis, respectively. In the first step, standard text processing techniques are applied such as stop-word removal, and stemming. In the second step the open source software Lucene is used for indexing the documents; a standard normalized Tf-Idf formula is adopted to compute the index terms weights, but other approaches will be taken into account for further investigations.

Phase 3. The outcome of the previous phase is a list of interest-terms with index terms weight over a given threshold α. To enrich the knowledge of the user's interests a process of knowledge extraction from the YAGO ontology is performed. This process is articulated in 3 sub-phases: (1) individuals and facts extraction, (2) direct concepts extraction and expansion to their child nodes, and (3) addition of new synonyms, respectively.
The fact extraction process is logically divided into non-taxonomic and taxonomic relations extraction. Non-taxonomic relations are defined in the YAGO ontology over entities which are referred to as *individuals*, while taxonomic relations can hold between an individual and its parent concept (class), or between two concepts. As previously stated a fact is a triple defined as $< entity, relation, entity >$, so the first step of the algorithm consists in locating the facts where an interest-term (obtained based on phase 2) matches with an entity. The outcome of this step is constituted by a set of facts and entities extracted from YAGO. From the analysis of the taxonomic relation different considerations have been made. In fact, it is possible that some facts based on the taxonomic relation $SubClassOf$ do not report useful information with respect to the considered term. For example, the fact $< relational\,database\,systems, SubClassOf, database\,systems >$ contains the knowledge that "relational database systems" are sub-class of "database systems", which is not very informative. For this reason, in case of a direct concept match, the algorithm takes all the first level children (individuals) of the matched concept. Referring to the previous example, for the term "relational database systems" the following instances - MySQL, Oracle, PostgreSQL etc., will be added in the user's profile.
A possibility is that the term is not found in YAGO. When this happens, our algorithm analyzes WordNet for checking the existence of synonyms. In case multiple synset exist, we adopt the methodology used by the authors of YAGO, where the most probable synset (i.e., the synset having higher probability of occurrence) is selected.

Phase 4. At this step, the resulting personal ontology is converted into the ontological language RDFS [1], and its graph portions are visualized by the ontology editor Protégé [5]. In the conversion process, every relation is exported into a single RDFS file, and afterwards all files are gathered into a single schema representing the personal ontology. A problem may arise related to the quality of the obtained profile. In fact, by the process of index analysis and facts extraction from YAGO, an unavoidable amount of noise is gathered into the final ontology. A first and preliminary solution was to manually improve its quality by using the Protégé editor.

Preliminary Experiment A preliminary analysis has been made for defining a conceptual user profile based on the YAGO ontology by considering 35 documents. This set of documents is related to several user interests such as art, literature, music, cinema and work. The second phase of the proposed methodology has been conducted by using Lucene, and the threshold for scoring was

[1] http://www.w3.org/TR/PR-rdf-schema

set to 0.5, thus obtaining 306 terms. At the end of phase 3, 578950 entities (i.e., individual plus concepts) have been counted in the user profile, where 11 new terms are added from WordNet. The last phase has consisted in converting the obtained profile in an ontological language (i.e., RDFS) in order to improve it by, for example, reducing noise or adding relations between terms. For example, if a term was related to the actor "Brad Pitt", all the corresponding information defined in YAGO are extracted such as categories it belongs to (i.e., *Action_film_actors, American_male_model, ...*), as well as its non-taxonomic relations (i.e., *hasWonPrize, produced, actedIn, ...*). By editing this ontological profile in Protégé, the user is allowed to delete non relevant information, for example the ones related to Brad Pitt as a model.

3 Conclusions and Future Works

The aim of this work is to create user profiles based on the YAGO general purpose ontology, to the aim of Web search personalization. We believe that ontologies are worth to be investigated as an interesting support for structuring knowledge in user profiles. To this aim, in a first preliminary application, the documents collected by a user are considered as the evidence of his/her interests. We plan to improve the methodology presented in this paper by following three main directions: the first is to automatically remove some noise from the profile (e.g., by deleting non relevant entities and the relations involving them), the second is to add new relations and facts between terms not defined in YAGO, and the last one is to consider other sources of information (i.e., past user's queries) to extract user's interests. Furthermore we will test the obtained YAGO-based user profile for expanding the user's queries to contextualize his/her Web searches.

References

[1] Agosti, M., Melucci, M., Crestani, F.: Tachir: A tool for automatic construction of hypertexts for information retrieval. In: Funck-Brentano, J.L., Seitz, F. (eds.) RIAO. pp. 338–358. CID (1994)

[2] Degemmis, M., Lops, P., Semeraro, G.: A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation. User Model. User-Adapt. Interact. 17(3), 217–255 (2007)

[3] Gauch, S., Chaffee, J., Pretschner, A.: Ontology-based personalized search and browsing. Web Intelligence and Agent Systems 1(3-4), 219–234 (2003)

[4] Labrou, Y., Finin, T.W.: Yahoo! as an ontology: Using yahoo! categories to describe documents. In: CIKM. pp. 180–187 (1999)

[5] Noy, N., Fergerson, R., Musen, M.: The knowledge model of protege-2000: Combining interoperability and flexibility. In: EKAW 2000. pp. 17–32 (2000)

[6] Pasi, G.: Issues in personalizing information retrieval. IEEE Intelligent Informatics Bulletin 11(1), 3–6 (December 2010)

[7] Sieg, A., Mobasher, B., Burke, R.D.: Ontological user profiles for representing context in web search. In: Web Intelligence/IAT Workshops. pp. 91–94 (2007)

[8] Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from wikipedia and wordnet. Journal of Web Semantic 6(3), 203–217 (2008)