Interactive Information Access in Semi-Structured Datasets with Rich Semantic Annotations

Eugen Popovici
Interactive Information Access, CWI Amsterdam, The Netherlands
eugen.popovici@cwi.nl

April 6, 2009

Abstract

This document is a very early draft of my future research work within the Interactive Information Access group at CWI.

1 Introduction

Digital documents represent a complex and heterogeneous mixture of text, structure and multimedia information. The management of semi-structured documents in digital libraries, product catalogs, scientific data repositories and across the Web requires the development of appropriate content description, indexing, filtering, searching and browsing methods and tools. To satisfy these requirements, several focused search paradigms such as XML element retrieval and entity ranking have been proposed.

Focused information retrieval allows users to have access not only to relevant documents, but also to relevant fragments/passages/elements within these documents. In addition, retrieving entities or a ranked list of entities instead of only documents or elements is becoming increasingly important for the current search engines. Entities are phrases with an associated semantic type or category (e.g. \CITY:Amsterdam", and \DATE:April 2009"). A plethora of new applications are currently under development by taking advantage of the available entity extraction and annotation techniques and services.

The automatic annotation of the mined entities by using appropriate semantic Web technologies could provide very useful information to aid navigation and search (for instance in faceted search and results clustering) especially in the cases where explicit meta data has a poor quality or is not available. The challenge is then to exploit the rich information from text, structure, the type/category of the retrieved entities as well as the links and the relations existing between them.

In this work we propose to study the interaction between focused/structured information retrieval and entity extraction and annotation approaches with semantic Web technology, faceted search and visualization and interaction techniques in order to support content representation and discovery, and information seeking and browsing in digital libraries/vast on line document repositories.

2 Data Annotation and Linking

A large number of the current available XML document collections contain poor structural annotations. Essentially these consist in presentation tags, like italics and bold, and of several logical tags, like sections and paragraphs. The semantic tags, like persons or locations are less frequent. Natural language processing techniques as entity extraction and annotation could help to improve the users access to the sought information by providing more meaningful annotations. Users profiles or domain ontologies could be used to guide the annotation process for a specific task or application.

The explicit underlying relations among the extracted entities could also be taken into account in the browsing and information retrieval process by using appropriate semantic web technologies and linked data resources. The challenges are how to rank and retrieve, and present and organize the potentially large number of results and how to differentiate the interesting from the large number of trivial relations, since these notions are subjective and context dependent.

3 The Need for Data Visualization and Interaction

The users have a vague, imprecise, erroneous or any knowledge at all (they simply ignore) the structure of the data (and usually they do not want to spend time to open a file and look inside for clues about the structural organization of the collection). The formulation of queries involving structural constraints seems to be a difficult task even for experts. In the same time, we start from the hypothesis that the users are able to recognize (and re/use) a useful structure/type/category in relation with their sought information.

Therefore, we propose to apply adapted visualization and interaction techniques that let the user explore and understand the data and the effects of its queries on that data. The purpose is to expose the structure and the semantic types/categories of the retrieved entities within their context. This could help to refine the faceted search and browsing of the extracted entities based on their type/category with their structural context of occurrence (where this context may be relevant for the given task/application). By highlighting the relations between the extracted entities and between the extracted entities and their structural context, we provide the user with an useful tool for assisting him in formulating content and structure queries and for discovering interesting patterns within the data.

4 Research Questions (Among the)

- How to enrich the documents with valid semantic tags (for the user)?
- How to support the users in exploring/understanding the (structure/content/organization of the) data ? and how their queries (both in efficiency and relevance) are affected by this structure ?
- How to support the users in finding the sought information ? and new information related to this (recommendations?, patterns?)
- How to evaluate the proposed system/interface?

5 Applications

Among the potential applications that could exploit the explicit structural contexts and rich semantic annotations within an exploratory/search task we can enumerate:

- Book content analysis, exploration and search tasks
- (Dynamic/Personalized) Semantic Rich Site Map Navigation
- Visual Document/Collection Summarization
- Visual exploration and retrieval of XML document collections
- Semantic and structural documents/results clustering
- (Visual) Schema matching with semantic clues/RSS feeds integration (applied on news?)
- Relevance feedback
- Recommendation systems

6 Evaluation

One of the research questions that must be address is how to validate and evaluate the proposed system/interface? To this end two approaches could be explored:

- The use of a standard IR collection with existing relevance assessments. Annotate both the topics and the relevant documents with semantic information about the entities (link the entities with domain ontologies). Answer IR tasks that can be evaluated on the original collection with and without the use of the extracted entities and with and without the relationships extracted from ontologies. Try to find statistical evidence that

the new features improved the quality of the retrieved results. If an improvement could not be observed in average for all the topics, analyze the topics in order to emphasize and characterize classes of topics for which the enhanced approach is effective.

- Users studies on a specific task:
  - (Semantic rich) Site map navigation
  - Content and structure query formulation
  - Summarization
  - Recommendation
  - Clustering
  - Schema matching / News feeds integration
  - ...