

CROEQS: Contemporaneous Role Ontology-based Expanded Query Search

Stijn Vandamme

Supervisors: Filip De Turck, Tim Wauters

To make full use of a multimedia database with annotated items, the possibility of searching the database is primordial. The traditional approach is to build a search engine based on textual metadata. However, the metadata of relevant items to a query may describe the content on a different conceptual level as the query formulation. To address this problem, we present CROEQS, a semantically enhanced search engine. It allows the user to query the annotated persons not only on their name, but also on their roles and affiliations at the time. We also present an ontology used to expand such queries: it allows us to semantically represent the domain knowledge on people fulfilling a role during a temporal interval in general, and politicians holding a political office specifically. The use of such a semantic clause in a query significantly filters the result set, allowing for targeted search, at the cost of a performance penalty.

I. INTRODUCTION

There is still a large gap between the information that can automatically be extracted from the multimedia data and the interpretation that the same data has for the users in different situations. This gap is called the *semantic gap*. To overcome this gap, and facilitate search and retrieval, items in multimedia databases are generally manually annotated with textual metadata.

Stijn Vandamme is with the INTEC Broadband Communication Networks (IBCN) research group of the Department of Information Technology (INTEC), Ghent University (UGent), Gent. E-mail: Stijn.Vandamme@intec.ugent.be.

But the textual metadata may describe the content on a different conceptual level as the query formulation. For example, when searching for political news items, users don't always know which politician has spoken out on a certain subject: they may want to express queries as "show me items about (previous) prime ministers co-occurring with a given political subject" or "show me items about a politician from a given party in connection with a given political subject".

CROEQS aims to answer such queries. Our approach is based on the combined use of annotations about the video fragments, and (separately stored) political domain knowledge.

Our work was performed on a database that contains about 5 000 hours of broadcasted television material from Flemish public and commercial broadcasters, for the most part news. This content spans 70 000 items, each individually and manually annotated with both keywords from a dictionary and a "free text" description. Additional metadata for each multimedia item includes the programme title, the date on which the item was broadcasted, etc. The textual metadata comprises 440 MiB in total.

II. SYSTEM OVERVIEW

We start with a traditional search engine, indexing the textual metadata. We implement one using Lucene.

The CROEQS search engine has a similar interface: it accepts the same queries as the text-based search engine, but in addition it also accepts queries with a semantic clause. Our system provides 3 kinds of semantic clauses: "a politician belonging to party x ", "a politician responsible for competence y ", and

“a politician with role z ”, where z can be Minister, Secretary of State, etc.

Queries with semantic clauses are translated to standard text queries, and forwarded to the traditional search engine. The results are sent back to the user unaltered (figure 1).

Query translation is done by querying the ontology for the politicians that match the clause, using SPARQL [1]. Then the semantic clause is then replaced with a long disjunctive (“or”) expression of all the politicians that match.

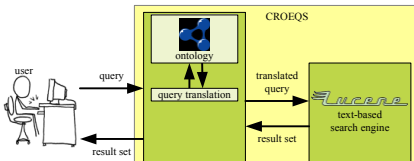


Figure 1. CROEQS schematic overview

III. CONTEMPORANEOUS ROLE ONTOLOGY

In computer science, an *ontology* (also referred to as a *domain model*) is a representation of a common conceptualization of a particular domain. It provides a shared understanding of a domain, and can be communicated across people and application systems.

An ontology may define a hierarchy of *classes*, and describe the *properties* of each, and *relations* between these classes. Classes, properties and relations constitute the ontology’s *model*. Once the model is defined, an ontology may contain *individuals* belonging to these classes, having specific values for the properties, and related (in ways defined by the relations) to other individuals.

In RDF [2], all these data are stored in *triples*, consisting of three resources: a subject, a predicate and an object.

We created a core ontology about people fulfilling roles during certain time intervals. It defines concepts as *Person*, *Role*, *Organization* and *Time_Interval*. Furthermore, a *Named_Organization* represents the holding of a name by an

organization during a certain time interval. Since an organization can change its name, an organization may be associated with more than one *Named_Organization* with different time intervals. Lastly, a *Role_Fulfillment* represents one specific person fulfilling one specific role during a specific time interval (figure 2).

Next, we extended this core ontology for the specific domain of (Belgian) politics. For example, we defined *Minister* as a subclass of *Role*, *Legislative_Body* as a subclass of *Organization*. We also introduced the class *Competence*, and the relation *responsible_for*. We populated this ontology with hundreds of Belgian politicians from the last decades, together with their roles, responsibilities and party affiliations.

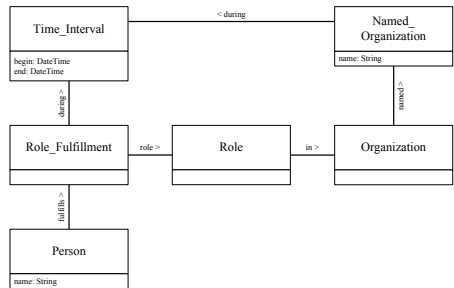


Figure 2. Core ontology about roles

IV. CONCLUSION

Using semantic clauses, users can express additional constraints to the queries that cannot be expressed in traditional text-based search engine. These clauses filter result set significantly and can be used for more targeted results. Since the system ultimately searches through text, irrelevant results can still show up because of polysemy.

REFERENCES

- [1] SPARQL Query Language for RDF, <http://www.w3.org/TR/rdf-sparql-query/>
- [2] RDF Primer, <http://www.w3.org/TR/rdf-primer/>