SQR: A Semantic Query Rating Scheme

Hany Azzam, and Thomas Roelleke
Dept. of Computer Science, Queen Mary
University of London
London, UK
hany@dcs.qmul.ac.uk
thor@dcs.qmul.ac.uk

ABSTRACT

We introduce a query rating scheme that identifies the possible interpretations which can be assigned to a semantic query. The interpretations range from the traditional bag-of-words interpretation to more context- and semantic-aware interpretations. The aims of this scheme are to communicate the extent of semantics that is being interpreted for a query and to assign suitable query processing methods for each level of interpretation accordingly.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Standardization, Measurement, Performance.

Keywords: Semantic search, query interpretation, logic.

1. INTRODUCTION & MOTIVATION

Modern retrieval systems have become more complex and semantic-aware by exploiting more than *just* the text, e.g. [1, 3]. Nowadays, large-scale knowledge bases can be automatically generated from high-quality text sources such as Wikipedia or rich semantic data repositories such as ontologies and taxonomies that contain entities (e.g. people, movies and companies) and relationships (e.g. bornIn, actedIn and isCEOof).

Such knowledge bases not only affect data/query representation and ranking strategies, they also affect query processing strategies. For example, the query "find actors who acted with Woody Allen and obtained an award for best-supporting actor" requires a search over instances of classes, such as "actor", a search over instances of relationships, such as "actedIn" and, then, aggregation (join) of the results from the different searches. The processing required to answer such a query differs from instances when no semantic knowledge is incorporated. The same query can be answered using a bag-of-words interpretation, but it will not necessarily achieve the same effectiveness as when semantic knowledge is interpreted [7].

This paper attempts to formalise a rating system for the different interpretations which can be given to a semantic query. The ratings serve as an indicator of how and which type of semantic knowledge has been considered when processing a query. Furthermore, they help to determine the best representation and processing strategy required to answer a query. The scheme in general helps to quickly communicate an agreeable interpretation of a given query. Such a feature can be desirable because it helps to associate several properties, such as semantic complexity, with a query.

The motivation behind this scheme is that a semantic-aware retrieval process (semantic retrieval in short) impacts the query processing strategy. On a physical level, a semantic query usually requires a search over different knowledge bases and indexes and then a join of the retrieved results. On a logical level, a semantic query resembles a natural language question that contains words/terms and classifications/relationships and which can then be further extended, either explicitly or implicitly, e.g. [4], with more terms, classifications and relationships. However, each retrieval system or retrieval process need not necessarily be aware of the semantic knowledge contained in each query. Therefore, we propose four levels of rating for a semantic query along with their definitions to help identify the semantic awareness assigned to a query.

ESAIR'10, October 30, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0372-9/10/10.

2. SEMANTIC QUERY RATING SCHEME

The Semantic Query Rating (SQR) scheme relies on two types of data: textual data embodied in terms (words) and semantic data embodied in classifications, relationships and attributes. We describe these two types of data using terminology similar to [5]. A classification, a class name with an object, e.g. "actor Woody Allen", a relationship, a subject with a relationship name and object, e.g. "Diane Keaton worked with Woody Allen" and an attribute, an object with an attribute name and an atomic value, e.g. "Woody Allen directed Husbands and Wives" are referred to as *propositions*; terms (a term and a context) are also considered to be propositions because one can probabilistically reason about the "importance" (power) of a term for representing the content of a certain context (e.g. a document).

2.1 How Does the Scheme Work?

The SQR scheme rates a query's level of semantic complexity. For example, what is the rating of a query such as "find titles of famous movies directed by Woody Allen where Woody Allen also plays an actor in the movie"?

Figure 1 illustrates keyword-based and logical formulations of the aforementioned pseudo query. Each formulation illustrates a certain interpretation assigned to the example query. A higher rating demonstrates greater semantic awareness.

The first row of Figure 1 gives the simplest interpretation of the pseudo query. The keyword representation and logical representation coincide. Such an interpretation resembles a topical query where the search is based solely on the term proposition. This query, which is common in traditional document retrieval, is rated as SQR-0 (Semantic Query Rating 0) because "zero" semantics is interpreted.

DEFINITION 1. *SQR-0: Query q is an* SQR-0 *query iff each proposition $\varphi_i \in q$ is a* term *proposition. Formally, this is denoted:*

$$q \text{ is SQR-0} : \Longleftrightarrow \forall \varphi_i \in q : \varphi_i \in \Phi_{term}$$

Rating	Textual Description	Propositions Used	Logical Representation
SQR-0 words	contexts containing the *term* "woody" and the *term* "allen"	Φ_{term}	retrieve(X) :- X[woody & allen];
SQR-1 structure	objects of type movie and display the *attribute* title	$\Phi_{\text{attribute}}$	retrieve(Y) :- X.type(movie) & X.title(Y);
SQR-2 semantics	contexts in which Woody Allen is *classified* as an actor	$\Phi_{\text{classification}}$	retrieve(X) :- X[actor(woody_allen)];
SQR-3 vagueness	*famous* movies	Φ_{vague}	retrieve(X) :- famousMovie(X);
SQR-123	titles of *famous* movies directed by Woody Allen and containing a story with Woody Allen *classified* as an actor	$\Phi_{\text{attribute}}$, $\Phi_{\text{classification}}$, $\Phi_{\text{relationship}}$, Φ_{vague}	retrieve(Y) :- famousMovie(X) & X.directedBy(woody_allen) & X.title(Y) & X[story[actor(woody_allen)]]];

Figure 1: The Semantic Query Rating Scheme (The '?' Denotes a Query and '&' Denotes the Boolean 'AND' Operator)

In the next rating, SQR-1, the unit of retrieval (answer type), which can be one of the document's elements (e.g. title, section), is interpreted. Such an interpretation of query words is similar to meta-data retrieval because it identifies the attributes (characteristics) of an object. This is demonstrated in the logical representation of SQR-1 in Figure 1 where "X" of type "movie" with a title "Y". The resulting rating, SQR-01, from combining SQR-0 and SQR-1 is similar to the content-and-structure classification of queries in [6].

DEFINITION 2. *SQR-1: Query q is an* SQR-1 *query iff each proposition $\varphi_i \in q$ is an* attribute *proposition.*

$$q \text{ is SQR-1} : \Longleftrightarrow \forall \varphi_i \in q : \varphi_i \in \Phi_{attribute}$$

For the third level of interpretation, the list of contexts is restricted to the ones in which actor "Woody Allen" occurs. In SQR-2 classifications and relationships are interpreted. In the logical representation, the context "X" (expressed by the square brackets) contains the semantic knowledge that "woody_allen is an actor". "woody_allen" is a unique object Id (e.g. URI) that identifies the object "Woody Allen".

DEFINITION 3. *SQR-2: Query q is an* SQR-2 *query iff each proposition $\varphi_i \in q$ is either a* classification *proposition or a* relationship *proposition.*

$$q \text{ is SQR-2} : \Longleftrightarrow \forall \varphi_i \in q : \varphi_i \in \Phi_{classification} \cup \Phi_{relationship}$$

The fourth rating takes vague propositions into account. Vague propositions, which can, for example, be expressed using logical rules [2], include notions such as "recent" and "famous". Interpreting such notions means utilising frequency- or probability-based estimates to define a particular vague proposition. In short, the last formulation consists of vague propositions (e.g. famous movies), which may or may not occur within a particular context.

DEFINITION 4. *SQR-3: Query q is an* SQR-3 *query iff each proposition $\varphi_i \in q$ is a* vague proposition.

$$q \text{ is SQR-3} : \Longleftrightarrow \forall \varphi_i \in q : \varphi_i \in \Phi_{vague}$$

The final row demonstrates one possible combination, SQR-123, of the "basic" ratings. The rating contains the components of SQR-1, SQR-2 and SQR-3 and, hence, describes a query that has attributes, classifications and relationships.

Maintaining the four basic query ratings as atomic ratings provides reasonable granularity. It helps to easily and precisely pinpoint the components of each query and to distinguish between a wide range of queries, including textual, textual and structural and textual, structural and semantic queries.

2.2 What are the Scheme's Benefits?

One of the benefits of the SQR scheme is the ability to communicate clearly and quickly the extent of semantics that is being interpreted for a query. This allows one to refer to a query as "this is an SQR-0 query" or "this is an SQR-2 query" and by doing so, to associate several properties, such as semantic complexity, with a query. In general, the proposed rating scheme can facilitate discussion about semantic search, and on a technical level, it can help to determine the right tools for effective processing.

Evaluating a retrieval model's performance while considering the level of semantic complexity of the queries can provide a more "targeted" evaluation. For example, if the evaluation is concerned with assessing the performance of a semantic-aware retrieval model, it could be set to be biased towards queries that incorporate SQR-2 (semantic) components. This, in return, would allow for a more accurate assessment of a model's effectiveness and for identifying the best model for each query or query type.

3. SUMMARY

We introduced and formally defined a scheme for rating queries according to their semantic complexity. The scheme aims to achieve a middle ground by being neither too complex to communicate and explain nor too straightforward and elementary to implement and utilise. Furthermore, the scheme helps to distinguish between a more semantic-aware processing of a query and a bag-of-words like interpretation. The latter requires only a one-step retrieval process based on an inverted file structure while the former involves several processing steps. Overall, this scheme highlights the gradation between these two extremes and helps to connect the processing and evaluation methods for semantic retrieval with the level of semantic expressiveness of a query.

4. REFERENCES

[1] H. Bast, A. Chitea, F. M. Suchanek, and I. Weber. Ester: efficient search on text, entities, and relations. In *SIGIR*, 2007.

[2] N. Fuhr, N. Gövert, and T. Rölleke. Dolores: A system for logic-based retrieval of multimedia objects. In *SIGIR*, pages 257–265. 1998.

[3] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. Naga: Searching and ranking knowledge. In *ICDE*, pages 953–962, 2008.

[4] J. Kim, X. Xue, and W. B. Croft. A probabilistic retrieval model for semistructured data. In *ECIR*, pages 228–239, 2009.

[5] C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of information retrieval based on a terminological logic. In *SIGIR*, pages 298–308, 1994.

[6] A. Trotman and B. Sigurbjörnsson. Narrowed extended xpath i (nexi). In *INEX*, pages 16–40, 2004.

[7] R. van Zwol and T. van Loosbroek. Effective use of semantic structure in xml retrieval. In *ECIR*, pages 621–628, 2007.