An Attribute-based Model for Semantic Retrieval

Hany Azzam, and Thomas Roelleke
School of Electronic Engineering & Computer Science
Queen Mary University of London
London, UK
{hany,thor}@dcs.qmul.ac.uk

Abstract

This paper introduces a knowledge-oriented approach for modelling semantic search. The modelling approach represents both semantic and textual data in one unifying framework, referred to as the probabilistic object-relational content modelling framework. The framework facilitates the transformation of "term-only" retrieval models into "semantic-aware" retrieval models that consist of semantic propositions, such as relationships and classification of objects. To illustrate this facility, an attribute-based retrieval model, referred to as TF-IEF-AF-IDF, is instantiated using the modelling framework. The effectiveness of the developed retrieval model is demonstrated using the Internet Movie Database test collection. Overall, the probabilistic object-relational content model can guide how semantic search and semantic data are modelled.

1 Introduction

Modern retrieval systems have become more complex and semantic-aware by exploiting more than *just* the text, e.g. [Bast *et al.*, 2007; Kasneci *et al.*, 2008]. Nowadays, large-scale knowledge bases can be automatically generated relatively easily from knowledge sources such as Wikipedia or other semantically explicit data repositories such as ontologies and taxonomies that explain entities (e.g. mark-up of persons, movies and locations) and record relationships (e.g. bornIn and actedIn).

Obstacles arise, however, when developing ranking functions and, in a broader sense, search strategies that combine query and document text with other types of evidence derived from semantic-rich knowledge bases. In particular, it is challenging when the ranking function is implemented directly on top of a standard physical document representation, as in the standard information retrieval (IR) engineering approach [Cornacchia and de Vries, 2007]. Consequently, an alternative approach, or more ambitiously, an alternative standard, is required to reduce the complexity of building and maintaining information systems and re-using their retrieval strategies [Fuhr, 1999; Hiemstra and Mihajlovic, 2010]

Design re-use is particularly important since IR already has a well-established family of retrieval models, namely TF-IDF, BM25 and language modelling (LM) that are used in many tasks but in slightly different ways. Ideally, these standard retrieval models should be re-used and adapted to solve complex and semantic retrieval tasks, and overall, to maximise the benefit gained from the underlying data.

How to link the the world of retrieval models with the world of semantic data, while avoiding an extensive engineering process, is not a straightforward process. However, the first step towards transferring the achievements of text retrieval models and so maximising the impact of semantic data, will be to simplify the process of tailoring search to a specific work task [Hawking, 2004].

This paper revisits a framework that may help to establish a standard for developing semantic retrieval models. The framework, referred to as the probabilistic object-relational content model (PORCM, [Roelleke, 1999]), is closely linked to engineering initiatives espoused in the development of business automation solutions using relational database management systems: design a conceptual schema; express the user application in terms of this schema; and design the user interface.

Such a framework, thus, gives creative freedom to designers to invent and refine semantic-aware retrieval models. It also allows for *more than one* semantic-based retrieval model and avoids the need to propose a single retrieval model for semantic search. As such, the framework provides a platform for developing effective semantic retrieval models applicable to textual and semantic data.

1.1 Contributions & Structure

The main contribution of this paper is to use a modelling framework to demonstrate a semantic variant of a standard retrieval model. The framework acts as a logical layer, which decouples the retrieval models from the physical representation of the data (document structure and content), bringing what the database field calls "data independence" to IR systems.

A BM25 motivated semantic retrieval model is instantiated using the probabilistic object-relational content model. This particular model exemplifies how, by taking a knowledge-oriented approach, retrieval models which are traditionally designed for *terms*, i.e. for keyword-based retrieval, can exploit terms and semantic evidence while ensuring data independence.

The feasibility of developing workable retrieval models for semantic search and the effectiveness of the developed retrieval model is demonstrated on the Internet Movies Database (IMDB) collection.

The remainder of this paper is organised as follows. Section 2 highlights some similarities and differences between structured document retrieval and semantic retrieval. The section also outlines related work in the literature. Section 3 details the probabilistic ORCM and its components. Section 4 showcases a retrieval model for semantic retrieval based on the proposed framework. Section 5 evaluates the retrieval model, and Section 6 concludes the discussion.

	Structured Document Retrieval (SDR)	Semantic Retrieval (SR)
Queries	Keyword-oriented, with structural components (e.g. XPATH with "contains" predicate)	Knowledge-based, with keyword-based components
Retrieval Unit	Documents, Sections, etc: structural objects	Documents, Actors, etc: any object
Evidence Spaces	Terms and Element Types (e.g. section, title)	Terms, Class names, Classifications, Relationship names, Relationships, Attribute names, attributes

Figure 1: Structured Document Retrieval versus Semantic Retrieval

2 Background and Related Work

2.1 Structured Document Retrieval & Semantic Retrieval

Figure 1 illustrates some characteristics of structured document retrieval (SDR) and semantic retrieval.

SDR is not limited to a particular structural markup language or a particular forum and so has general applicability. Our discussion, however, focuses on XML as the structural markup language because it is the most widely used standard. Additionally, we focus on INEX-related work and contributions as INEX is characterised by research about SDR, and it is the initiative for the evaluation of XML retrieval [Fuhr *et al.*, 2002].

The first point of comparison is the type/formulation of queries. In SDR the user information need can be formulated using either text-only or text-and-structure approaches. The text-only approach is keyword-based (e.g., the content-only approach in INEX [Amer-Yahia and Lalmas, 2006]). The text-and-structure approach combines textual and structural clues (e.g. XPath, NEXI [Trotman and Sigurbjörnsson, 2004]). Queries in semantic retrieval can be expressed using the aforementioned approaches in addition to the semantic structures found in the formulated query and/or the document representation (e.g. semantic content-and-structure approach [van Zwol and van Loosbroek, 2007]). Semantic queries can also be expressed using graph-patterns such as SPARQL [Prud'hommeaux and Seaborne, 2006]. Extending SPARQL for full-text and semantic search (e.g. [Bast *et al.*, 2007; Kasneci *et al.*, 2008; Elbassuoni *et al.*, 2009]) is analogous to the prior work on SDR which has enhanced XPath and XQuery by various forms of text-search and ranking capabilities. The graph-based approaches, in contrast to XML trees, are independent of the physical representation of the underlying data. If a query is expressed via XQuery, the application would have to know the particular data representation.

The second aspect illustrated in Figure 1 is the retrieval unit (answer type). In SDR the document structure is explicit and, therefore, the retrieval unit is based on the document's presentation and logical structures, such as chapter and section. Semantic retrieval focuses instead on retrieving *objects* which have a particular meaning. This is particularly evident in the discussion of query formulations when semantic text-and-structure topics are introduced to conduct *semantic* retrieval [van Zwol and van Loosbroek, 2007]. Therefore, in semantic retrieval, the answer type has a more general and semantic form, which includes objects such as a person, a product, or a project.

The third difference is the evidence (ranking criteria) used in retrieval. Unlike in traditional IR, in SDR the additional structural evidence not seen in unstructured (flat) text documents is exploited. For example, in XML retrieval the logical document structure is used to estimate the relevance of an element according to the evidence associated with this element only. When the goal is to rank documents, the probabilities of estimating the relevance of each element are combined to produce a single probability for the document. There are two strands of models for element-based ranking and evidence combination: variants of probabilistic models based on the probability ranking principle, such as [Robertson *et al.*, 2004; Lu *et al.*, 2005]; and variants of the statistical language modelling technique proposed by [Ponte and Croft, 1998], such as [Ogilvie and Callan, 2002].

In semantic retrieval, structural elements can be also utilised to estimate the relevance of an object. However, these elements bear a semantic meaning (e.g. actor, director Figure 2) and, thus, a distinctive term distribution. For example, [Kim *et al.*, 2009] extend [Ogilvie and Callan, 2002] to demonstrate how varying weights for different semantic elements across query terms can improve retrieval performance. Evidence associated with semantic annotations in the form of linguistic structures is also used in semantic retrieval, especially in question answering applications. [Zhao and Callan, 2008; Bilotti *et al.*, 2007] propose ranking answer-bearing sentences to questions by incorporating the semantic annotations in both the sentences and queries into the retrieval process.

Semantic annotations expressed in the form of Resource Description Framework (RDF) graphs, referred to as entity-relationship graphs, are also used as ranking criteria. These graphs are used as a source of evidence to construct graph-based ranking models and queries for retrieving semantic objects. For example, [Kasneci *et al.*, 2008; Elbassuoni *et al.*, 2009] propose LM variants for ranking the results of keyword-augmented graph-pattern queries over entity-relationship graphs.

```
<Movies>
  <movie id="329191" >
    <title> Gladiator </title>
    <yeart> 2000 </year>
    <actors><actor id="russell_crowe">Russell Crowe </actor> </actors>
    <team>  <director id="ridley_scott"> Ridley Scott </director> </team>
    ...
    <plot> Maximus is a powerful Roman general ... </plot>
  </movie>
</Movies>
```

Figure 2: XML-based Representation of a Movie

In this paper we utilise full-text and semantic evidence for semantic retrieval. These types of evidence are represented by a set of propositions: terms, classifications, relationships and attributes. For example, in Figure 2 the XML-based representation of a movie contains several types of evidence. If explicated according to the aforementioned set of propositions, then the term proposition would represent the full-text evidence, which is similar to the common keyword-based IR representation. The classi-

fication proposition would capture the "class of' relationship between objects and classes (e.g. director "Ridley Scott"). The relationship proposition would associate two objects, and the attribute proposition would contain the relationship between an object and an atomic value.

The aforementioned propositions stem from object-oriented and content modelling. They support the retrieval of structural elements, semantic elements and heterogeneous objects. The information about the structure and the specifics of objects is represented in the unifying framework of text (terms) and object-oriented modelling (Section 3 discusses the framework in more details).

2.2 Related Work

Related work can be found primarily in investigations of the XML retrieval task, which has been addressed from both IR and database perspectives. The proposed semantic retrieval model is akin to XML retrieval models such as BM25f [Robertson *et al.*, 2004] and hierarchical language modelling [Ogilvie and Callan, 2003] in that it is based on combining different evidence spaces and probabilities.

Our retrieval model is also similar to the models in [Kasneci *et al.*, 2008; Elbassuoni *et al.*, 2009] in the sense that these models focus on retrieving semantic data. However, these approaches mainly propose a variant of only one traditional retrieval model to answer semantic queries; moreover, they do not combine other sources of evidence such as full-text and/or structural elements. On systems-side, approaches such as ESTER [Bast *et al.*, 2007] support a similar class of semantic queries as the proposed model; however, efficiency is their primary design goal.

Our knowledge representation approach shares some aspects with logical approaches for modelling IR such as MIRTL (Multimedia Information Retrieval Terminological Logic, [Meghini *et al.*, 1993]) where "terms" are used to represent concepts and roles. In our framework, however, content is considered separate from the concepts of object-oriented modelling. To implement the retrieval model designed using the proposed knowledge representation, an integrated database and IR approach is used. This approach is similar to probabilistic database approaches found in [Dalvi and Suciu, 2004; Chaudhuri *et al.*, 2006; Roelleke *et al.*, 2008].

3 Knowledge Representation

This section proposes a schema component for semantic retrieval. The notion of a "schema" highlights the difference between keyword-based and semantic retrieval where the former requires a search over *only* an inverted file structure, while the latter requires several processing steps and different representations. The proposed schema is based on object-relational modelling principles. Traditionally, the object-relational model (e.g. [Stonebraker *et al.*, 1998]) uses relations such as "memberOf(...)", "relationship(...)" and "attribute(...)" to model concepts such as classification, relationships, and attributes. We extend this approach and introduce an object-relational content model. The model integrates object-relational modelling and content-oriented (term-based) modelling into one framework. Consequently, we extend the model to create its probabilistic variant, namely the probabilistic object-relational content model. This model includes relations which represent probabilistic parameters to model IR-like retrieval models.

3.1 Object Relational Content Model (ORCM)

Figure 3 uses the ORCM to represent the movie in Figure 2. Ellipses indicate that some data have been omitted to conserve space. The location where different elements occur are stored as paths, expressed in XPath. For readability we use a simplified syntax, e.g., "imdb/movie_1/title_1" points to the attribute describing a movie's title.

term	
Term	Context
gladiator	329191/title[1]
2000	329191/year[1]
russell	329191/.../actor[1]
crowe	329191/.../actor[1]
ridley	329191/.../director[1]
scott	329191/.../director[1]
maximus	329191/plot[1]
powerful	329191/plot[1]
roman	329191/plot[1]
general	329191/plot[1]
...	...

term_doc	
Term	Context
gladiator	329191
2000	329191
russell	329191
crowe	329191
ridley	329191
scott	329191
maximus	329191
powerful	329191
roman	329191
general	329191
...	...

(a) Term propositions in the element and root contexts

classification		
ClassName	Object	Context
movie	329191	movies[1]
title	329191/title[1]	329191
year	329191/year[1]	329191
actors	329191/actors[1]	329191
actor	329191/.../actor[1]	329191/actors[1]
team	329191/team[1]	329191
director	329191/.../director[1]	329191/team[1]
plot	329191/plot[1]	329191
...	...	...

(b) Classification propositions

attribute			
AttrName	Object	Vatlue	Context
id	329191	"329191"	movies[1]
id	329191/.../actor[1]	"russel..."	32.../actors[1]
id	329191/.../director[1]	"ridley..."	32.../team[1]
...	...	...	...

(c) Attribute propositions

Figure 3: An Object-Relational Content Model Representing a Movie

Traditionally, in order to model the task of document retrieval, a term-document representation based on a data structure such as "term(Term, DocId)" would suffice. For example, in Figure 2 the terms in the XML fragment can have a flat representation (the XML elements are not interpreted), such as "term(movie,329171)". Consequently, for document retrieval the retrieval models (e.g., TF-IDF, language modelling, BM25) for ranking documents are primarily based on a "term(Term,DocId)" relation.

In the case of SDR, a structural representation, such as "term(Term, SecId)", is necessary. This is because the contexts or the document structure (e.g. abstract, section, paragraph) are explicit. Additionally, the retrieval models are usually based on combining the scores obtained from scoring every context, or combining the term frequencies. What then, are the data structures which we use to model (implement) semantic retrieval?

We first review the design process of the ORCM [Roelleke, 1999] and then demonstrate how it can be utilised for semantic retrieval.

The ORCM combines object-oriented and content-based modelling concepts. The object-oriented concepts include classification, relationships and attributes, which are more generally referred to as propositions – a specification stemming from object-oriented modelling and terminological logics [Meghini *et al.*, 1993]. Content modelling is analogous to the traditional IR representation of text, which is usually a term-context-based representation. However, unlike the conventional modelling approaches for IR, such as terminological logic, [Meghini *et al.*, 1993]), "content" is viewed as separate from the concepts of object-oriented modelling. This separation helps content to be described in a more formal and knowledge-oriented way.

There are two design steps taken in order to achieve this separation. Each predicate within a proposition is associated with a context[1] and a term proposition is used as the keyword-based IR representation for text.

Other propositions (components) that can be taken into account include generalisation and aggregation, where generalisation is a relationship between classes, and aggregation is a particular relationship between objects (entities). However, it is the four aforementioned propositions with which we shall be mainly concerned here. The pillars of the probabilistic object-relational content model can be summaries as follows:

- classification of objects: monadic predicate of the form "ClassName(Object)", for example: "actor(russell_crowe)".
- relationship between objects: dyadic predicate of the form "RelationshipName(Subject,Object)", for example: "directedBy(329191,ridley_scott)".
- attribute of objects: dyadic predicate of the form "AttributeName(Object, Value)", for example: ridley_scott.name("Ridley Scott").

In order to implement the object-oriented modelling concepts, a relational approach is used, resulting in the proposed object-relational schema. Relations such as "classification" and "relationship" are devised. Moreover, in order to join object-oriented modelling with keyword-based and content-oriented modelling, an additional predicate, namely "term", is used, and an additional attribute column, namely "Context" is adjoined to term, classification, relationships and attributes. This yields the ORCM modelling paradigm.

Below we contrast the conventional object-relational model with the object-relational content model[2].

Object-relational modelling (ORM):

- classification(ClassName, Object)
- relship(RelshipName, Subject, Object)
- attr(AttrName, Object, Value)

Object-relational content modelling (ORCM):

- classification(ClassName, Object, *Context*)
- relship(RelshipName, Subject, Object, *Context*)
- attr(AttrName, Object, Value, *Context*)
- *term(Term, Context)*

[1]Context is a general concept that refers to documents, sections, databases or any other object with a content

[2]In the schema design process we often opt to use shorter names for relation and attribute names so that to achieve a slimmer form of the schema.

Terms are complementary to classification and relationships. Most importantly, and one of the main emphases of this modelling paradigm, is that content is not modelled, for example, as a relationship "contains(DocumentId, Term)", but rather content is modelled by maintaining an attribute column "Context" in the schema for each proposition. In other words, content is modelled separately from existing concepts, such as classification and relationships.

There are several fundamental benefits of utilising the object-relational content model. One benefit is that knowledge modelling, in general, aids "knowledge architects" to build information and knowledge management systems that are both flexible and scalable. Another benefit is that it facilitates the transformation of term-document-based IR retrieval models into retrieval models founded on the probabilistic object-relational content model, thus resulting in a strand of retrieval models suited for semantic retrieval. Lastly, the model enables the representation of textual, structural and semantic data in one unifying framework. The uniform representation of the data, the semantic retrieval models and the decoupling between the two using the object-relational content model results in data independence. This is a desirable feature when designing complex retrieval systems.

In summary, the ORCM is to be understood as a conceptual model with a set of relations – a relation for each basic concept of object-oriented and content-based modelling.

3.2 Probabilistic Spaces for Semantic Retrieval

The evidence space (ranking criteria) construction is facilitated by the probabilistic object relational content model. The probabilistic ORCM comprises the relations of the ORCM, as well as relations representing probabilistic parameters. For example, for the basic relation "term_doc(Term,Doc)", there can be term-based and document-based probabilities.

Some of the probabilistic relations for "term_doc" are:

p_DF_t_term_doc(T): Document frequency-based probability of term t derived from relation "term_doc(Term,Doc)"

p_TF_t_term_doc(T): Tuple frequency-based probability of term t derived from relation "term_doc(Term,Doc)"

pidf_term_doc(T): Inverse document frequency (IDF)-based probability of term t derived from relation "term_doc(Term,Doc)"

In probabilistic ORCM the techniques and models of IR devised for term-based retrieval models become available for class-based, relationship-based and attribute-based retrieval. Concepts such as the tuple frequency-based probability of a class (class-frequency, *CF*) and the IDF of a class name (similar to the IDF of a term) make immediate sense. Similarly, the tuple frequency-based probability of an attribute name (attribute name-frequency, *AF*) and the IDF of an attribute name become possible.

The ability to transfer the achievements of term-based retrieval directly to semantic retrieval models makes the probabilistic ORCM a potential base for semantic retrieval. Moreover, the way probabilistic spaces can be combined in probabilistic ORCM can lead to new and effective retrieval models. For example, the frequencies of attributes names are exploited to define an attribute-based retrieval status value (RSV), and this RSV can be combined with other RSV's, such as the term-based one.

The next section provides an example of a retrieval model constructed on top of the discussed representation.

4 An Attribute-based Retrieval Model

The proposed model, TF-IEF-AF-IDF, for semantic retrieval focuses on combining evidences from attribute name and term predicate spaces. The model is implemented in three phases. Figure 4 (Page 6) is a snapshot of the proposed model's components when answering query number 28, "gladiator action maximus scott", from the IMDB test collection. We detail the three phases below.

Phase 1 retrieves "Term-Document's Element-Query" triplets for each query term. Such retrieval can be performed using any term-based retrieval model (e.g. TF-IEF, TF-IDF, BM25, LM). We choose here TF-IEF, where TF is the within-element term frequency, and IEF is the inverse element frequency of a term. TF-IEF is defined as follows:

Definition 1 *TF-IEF:*

$$\mathrm{RSV}_{\textit{TF-IEF}}(e,q) := \sum_{t \in e \cap q} \mathrm{TF}(t,e) \cdot \mathrm{IEF}(t) \qquad (1)$$

"t" stands for term, "e" for element, such as "title", and IEF is inverse element frequency.

Phase 2 consists of two parts. The first part infers the attribute name and root context (root nodes) from the document's elements in the "Term-Document's Element-Query" triplet. This yields an intermediate and query-dependent attribute-based index (this index corresponds to the tf_ief_match_augmented in Figure 4).

The second part infers for each query term its top-k corresponding "context type". For example, for a query such as "fight brad pitt" the inferred top–1 context type would be "title" for query term "fight" and "actor" for query terms "brad" and "pitt". This is because "fight" occurs in the context of type "title", and "brad" and "pitt" occur in the context of type"actor". The attribute-based index constructed in the first part of Phase 2 is used to infer the mapping between each query term and its type. The result of this inference is represented in Figure 4 ("AttrName" corresponds to the context type in which the query term occurs).

qTermAttr(Term, AttrName, Query)

The probability of the mapping between a query term and an attribute type is estimated using the number of mappings between a term and an attribute name divided by the total number of mappings in the intermediate index. The intuition behind the mapping is that if a term occurs frequently within a certain context type then the term is more likely to be "characterised" by that particular context [Kim *et al.*, 2009].

Phase 3 combines an attribute-based retrieval score with a traditional topical (term-based) score resulting in the TF-IEF-AF-IDF model. The motivation to do so is that in some cases an attribute-only retrieval score is deemed unsuitable for estimating the relevance of a particular semantic object with respect to a query. In other words, not all queries are issued with a particular semantic predicate or relationship in mind which the query terms can be mapped to. Therefore, a document-based retrieval score would provide a more realistic setting whereby both the attribute-based and document-based retrieval scores are considered. The intuition behind this combination is comparable to the mixture of document-based and element-based language model scores [Zhao and Callan, 2008].

We formally define the model and its components below.

Definition 2 *The TF-IEF-AF-IDF Model: TF-IEF-AF-IDF model is a multi-stage retrieval model. In the first stage, TF-IEF is used to associate for each query term the document elements and query. Also, each query term is associated with an attribute name. Then, the RSV's of attribute-based retrieval and term-based retrieval are combined into an overall score.*

Let d' be the document inferred from d, where the inference assigns several attribute names (context types) to each "Term-Document's Element-Query" generated in phase 1.

Let q' be the query inferred from q, where the inference assigns several attribute names (context types) to each query term.

$$\begin{aligned}\mathrm{RSV}_{\textit{TF-IEF-AF-IDF}}(d,q) := \qquad (2)\\ \mathrm{RSV}_{\textit{TF-IDF}}(d,q,\textit{term-based-index-all-docs}) +\\ \mathrm{RSV}_{\textit{AF-IDF}}(d',q',\textit{attribute-based-index-retrieved-docs})\end{aligned}$$

The term-based score is defined as follows:

Definition 3 *TF-IDF:*

$$\mathrm{RSV}_{\textit{TF-IDF}}(d,q) := \sum_{t \in d \cap q} \mathrm{TF}(t,d) \cdot \mathrm{IDF}(t) \qquad (3)$$

The attribute-based score is defined as follows:

Definition 4 *AF–IDF:*

$$\mathrm{RSV}_{\textit{AF-IDF}}(d,q) := \sum_{a \in d \cap q} \mathrm{AF}(a,d) \cdot \mathrm{IDF}(a) \qquad (4)$$

$\mathrm{AF}(a,d)$ and $\mathrm{TF}(t,d)$ correspond to the within-document *attribute name* frequency and the within-document *term* frequency components, respectively. "aName" is an attribute name, "d" is a document and "q" is the query. The frequencies are estimated using BM25's TF_K(t,d) ($\mathrm{tf}_d/(\mathrm{tf}_d + K_d)$) quantification [Robertson, 2004]. tf_d is total frequency and K is a normalisation factor reflecting the document length. In the attribute-based aggregation it is af_d instead of tf_d and K is the number of attributes in the intermediate index. The IDF in the AF-IDF is calculated over the set of *retrieved* documents only.

Figure 4 demonstrates for the IMDB's query number 28 the processing steps and relational instantiations of the TF-IEF-AF-IDF model (see Section 5 for details on the IMDB collection). "tf_ief_match" represents the retrieved Term-DocumentElement-Query triplets for each query term. "tf_ief_match_augmented" infers attribute names and root contexts for each Term-DocumentElement-Query triplet, and "attr_index" is the representation of the attribute names and root contexts. "qTermAttr" represents the query term and its inferred attribute name. Lastly, "AF-IDF" represents the predicate-based attribute retrieval scores and "TF-IEF-AF-IDF" represents the combined predicate-based attribute and term retrieval scores.

4.1 Discussion

The model described above integrates several models into a stepwise retrieval process. Such a retrieval process is similar to "database matching" which is achieved in several steps, using different representations at each database level. The stepwise approach is more complex than simple flat text matching and creates opportunities for more powerful matching specifications for semantic search.

The framework ensures that traditional models such as TF-IDF and aggregation techniques, such as BM25's TF_K quantification are transferrable to models where semantic knowledge is explicated, such as AF-IDF.

tf_ief_match			
$P(t,e\|q)$	Term	Retrieved_Element	Query
0.774433	gladiator	imdb/movie_128903/title[1]	q028
...	...	...	q028
0.392531	action	imdb/movie_113798/genre[1]	q028
...	...	...	q028
0.426180	maximus	imdb/movie_2112/plot[1]	q028
...	...	...	q028
0.187669	scott	imdb/movie_284995/team[2]	q028

(a) Element-based Retrieval Results

tf_ief_match_augmented: inferred attributes and contexts					
Prob	Term	Retrieved_Element	Attribute_Name	Root_Context	Query
0.77	gladiator	imdb/movie_128902/title[1]	title	imdb/movie_128902	q028
...	...	...	...	...	q028
0.39	action	imdb/movie_113798/genre[1]	genre	imdb/movie_113798	q028
...	...	...	...	...	q028
0.43	maximus	imdb/movie_2112/plot[1]	plot	imdb/movie_2112	q028
...	...	...	...	...	q028
0.19	scott	imdb/movie_284995/team[2]	team	imdb/movie_284995	q028

(b) Augmented Retrieval Result: Inferred Attributes Names and Root Contexts

attr_index(AttrName, Context)			
Prob	Attribute_Name	Context	Query
0.77	title	imdb/movie_128902	q028
...	...	...	q028
0.39	genre	imdb/movie_113798	q028
...	...	...	q028
0.43	plot	imdb/movie_2112	q028
...	...	...	q028
0.19	team	imdb/movie_284995	q028

(c) Attribute-based Index

idf: over attr_index		
P(aName)	Attribute_Name	Query
...	...	q028
0.48	title	q028
0.31	plot	q028
0.14	team	q028
0.09	genre	q028
...	...	q028

(d) IDF of attribute names over attr_index

qTermAttr (IMDB query 28)			
Prob	Term	Attribute_Name	Query
0.000426942	gladiator	title	q028
...	...	...	q028
0.616951	action	genre	q028
...	...	...	q028
0.00000521851	maximus	plot	q028
...	...	...	q028
0.14484	scott	team	q028

(e) Query Representation: Terms and Inferred Attributes

AF_K–IDF	
$Prob$	Context
...	...
0.17	imdb/movie_128902
...	...
0.27	imdb/movie_128908

(f) Attribute-Frequency-based Score

TF-IEF-AF-IDF	
RSV	Context
0.33	imdb/movie_128902
0.32	imdb/movie_128908
...	...
...	...

(g) TF-IEF-AF-IDF Score

Figure 4: TF-IEF-AF-IDF Retrieval Phases (IMDB Query 28)

In particular, the attribute-based aggregation (the AF_K component) is instrumental to the performance of the TF-IEF-AF-IDF retrieval model as will be illustrated in the evaluation section. The query terms and the retrieved elements (Phase 2) are mapped to their corresponding semantic predicates, which, in this case, are semantic attributes names. This mapping, then, results in an aggregation over the attribute names instead of the terms.

Our analysis suggests that this shift, hence, leads to an event space that contains less number of *distinct* events (attribute names) but that occur frequently. Such a feature is well-suited to the BM25-like aggregation of frequencies because if an event occurs in a context then the probability it occurs again is greater than the initial probability, i.e. the occurrence of an event depends on previous occurrences – [Wu and Roelleke, 2009] have proposed a probabilistic semantics for this feature, referred to as "semi-subsumed". The non-linear nature of the aggregation is key to the good retrieval quality of TF-IEF-AF-IDF.

To illustrate the proposed retrieval model we used an XML-based collection. Particular to this collection is that each element type has specific semantics and, thus, a distinctive term distribution. This is analogous to, for example, entity relationship graphs where the semantics of the data is is represented rather than the structural layout.

However, unlike its full-fledged semantic counterpart, the XML data "as it comes" does not explicate relationships between entities. This is reflected in the ORCM representation as there are mainly terms, classification and attribute relations. Furthermore, the "basic" representation of the XML data still uses XPath expressions to denote object Id's and contexts. This can be viewed as problematic since the main aim here is to achieve a semantic as opposed to structural representation and eventually semantic as opposed to structural retrieval models.

There are two main solutions to this problem. The first is that the XML data in itself consists of element types that have specific semantics and therefore differ from logical or layout element types that are concerned with a document's or a page's presentation. Secondly, the structural representation can be lifted to become a semantic representation. The following Datalog rule exemplifies how this can be done. The rule for "actor" underlines that a "semantic" object can be extracted by combining structural information about elements of type actor and their attributes (e.g. "russell_crowe" in Figure 2).

```
actorElement(XPath, Context) :−
    classification (actor, XPath, Context);

actorEntity (ObjectId, Context) :−
        actorElement(XPath, Context) &
         attribute (id, XPath, ObjectId, Context);
```

The above rules derive a semantic Id for an actor; moreover, it lifts the "structural" into a "semantic" classification. In a similar fashion, attributes can be lifted to become relations in "higher-level" layers of the ORCM. These layers can be derived from the "basic" ORCM and form an abstraction hierarchy from basic to structural to semantic schema. This helps to achieve data independence, as any data (XML, RDF, RSS) can be represented in the basic ORCM, and then, application-specific relations are derived.

Overall, this discussion emphasises that the ORCM schema supports reasoning over structural elements, semantic structures and, eventually, semantic information which is "naturally" present in entity relationship graphs.

5 Evaluation

The purpose of the evaluation is two-fold. Firstly, it proves the feasibility and applicability of the proposed knowledge representation, namely the probabilistic object relational model, for both term-based and semantic retrieval. Secondly, it investigates the quality of the proposed retrieval model, which is *one instance* of the proposed knowledge representation.

	MAP	RecipRank
TF_K–IDF	35.07	36.80
TF-IEF-AF-IDF-top-1	*52.11*	*53.96*
TF-IEF-AF-IDF-top-5	**60.32**	**62.04**
Improvement	+71.19	+68.59

Figure 5: Retrieval Performance per Query Mapping (boldface indicate best performing model and results are in percentages)

The experiment was performed on the IMDB collection[3], which consists of $437,281$ documents or XML records. Each document corresponds to a movie and was constructed from text data. The element types were "title", "year", "releasedata", "language", "genre", "country", "location", "colorinfo", "cast", "team" and "plot". Document content consists mostly of keywords, with the exception of the plot element.

We utilised the 40 queries and relevance criteria in [Kim *et al.*, 2009]. For each query, a query term is mapped to its corresponding semantic structure. This leads to a set of queries that contain keywords and semantic predicates (attributes). The unit of retrieval for all queries is the movie object. Below is a logical representation of query number 28 with top–1 mapping.

```
retrieve (X) :−
 X. title ( gladiator ) & X.genre(action ) &
 X.actor(maximus) & X.plot(maximus) & X.director( scott );
```

For the experiments, we used HySpirit [Roelleke *et al.*, 2001], a probabilistic reasoning system which supports the retrieval of text and (semi-)structured data. We chose HySpirit because it provides a framework with high-level and customisable concepts for modelling retrieval models. The framework provides an open-box approach for describing ranking models for any object.

Figure 5 shows the retrieval effectiveness for the test queries on the IMDB collection. To conserve space, only the performance of TF-IEF-AF-IDF with "top–1" and "top–5" mapping has been reported. The main observation is that the proposed method, an attribute-based model for semantic retrieval, significantly outperforms (p-value < 0.01 with two-tailed t-test) the TF_K–IDF baseline. The baseline is a document-oriented retrieval model where the XML elements are discarded (similar to the method reported in [Theobald *et al.*, 2005]).

The improvement performance can be accredited to the combination of term-based and attribute-based evidence spaces. Furthermore, the TF and AF parameters of the model are set to the BM25-like quantification that delivers the best performance since it mitigates the sub-optimal independence assumption of the total count.

The AF component, in particular, reflects that if a term occurs in a document then the probability that it occurs

[3] http://www.imdb.com/interfaces#plain

again is greater than the initial probability, i.e. the occurrence of an event depends on previous occurrences. Switching from term to attribute space, which groups terms under a particular context type, is conducive to retrieval performance.

Overall, the evaluation demonstrates that the expressiveness of the probabilistic ORCM model can lead to an effective model for semantic retrieval.

6 Summary & Conclusion

This paper demonstrates an approach for representing knowledge, how to merge object-relational and term-oriented modelling and how to steer term-based modelling towards semantic modelling (the semantic knowledge is explicit); therefore, it contributes a discussion of how object-relational modelling meets content modelling and its effect on probabilistic retrieval models for semantic retrieval.

Semantic retrieval requires models that, in sound and transparent ways, mix various frequencies and probabilities. Whereas in text retrieval, probabilities of terms are the dominating players, in semantic retrieval, probabilities of terms, classes, relationships, attributes and objects are the parameters involved in the design of a retrieval model. This paper introduces a particular retrieval model, namely TF-IEF-AF-IDF, in which the attribute frequency of retrieved elements (attributes) is a crucial component of ranking retrieved objects (contexts). The AF component gathers evidence from different query terms into one attribute. This aggregation has a positive outcome on retrieval quality, especially when combined with traditional term-based retrieval, as shown in this paper for TF_K–IDF.

The proposed model is one instance of a large family of models that can be developed using the probabilistic object relational content framework. The framework helps to transform retrieval models that are traditionally designed for *terms*, i.e. for keyword-based retrieval to models that are based on *propositions* and, hence, tailored towards more complex and semantic retrieval tasks. Furthermore, the flexibility and openness of the framework encourages engineers to create a variety of retrieval models that combine textual, structural and semantic sources of evidence.

We have contributed to two related facets of semantic retrieval: knowledge representation and retrieval strategy modelling. Future work will investigate other retrieval models for semantic retrieval based on the probabilistic object-relational content model.

7 Acknowledgments

We would like to thank Jinyoung Kim of the University of Massachusetts Amherst for providing us with the collection and the queries. We would also like to thank the reviewers for their excellent suggestions.

References

[Amer-Yahia and Lalmas, 2006] Sihem Amer-Yahia and Mounia Lalmas. Xml search: languages, inex and scoring. *SIGMOD Rec.*, 35(4):16–23, 2006.

[Bast *et al.*, 2007] Holger Bast, Alexandru Chitea, Fabian M. Suchanek, and Ingmar Weber. Ester: efficient search on text, entities, and relations. In *SIGIR*, 2007.

[Bilotti *et al.*, 2007] Matthew W. Bilotti, Paul Ogilvie, Jamie Callan, and Eric Nyberg. Structured retrieval for question answering. In *SIGIR*, pages 351–358, 2007.

[Chaudhuri *et al.*, 2006] Surajit Chaudhuri, Gautam Das, Vagelis Hristidis, and Gerhard Weikum. Probabilistic information retrieval approach for ranking of database query results. *ACM Trans. Database Syst.*, 31(3):1134–1168, 2006.

[Cornacchia and de Vries, 2007] R. Cornacchia and A. P. de Vries. A Parameterised Search System. In *Proceedings of the European Conference on IR Research (ECIR)*, 2007. Best student paper award.

[Dalvi and Suciu, 2004] Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. In *VLDB*, pages 864–875, 2004.

[Elbassuoni *et al.*, 2009] Shady Elbassuoni, Maya Ramanath, Ralf Schenkel, Marcin Sydow, and Gerhard Weikum. Language-model-based ranking for queries on rdf-graphs. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 977–986, New York, NY, USA, 2009. ACM.

[Fuhr *et al.*, 2002] Norbert Fuhr, Norbert Goevert, Gabriella Kazai, and Mounia Lalmas. Inex: Intitiative for the evaluation of xml retrieval. In *ACM SIGIR Workshop on XML and Information Retrieval*, 2002.

[Fuhr, 1999] N. Fuhr. Towards data abstraction in networked information retrieval systems. *Information Processing and Management*, 35(2):101–119, 1999.

[Hawking, 2004] David Hawking. Challenges in enterprise search. In *ADC*, pages 15–24, 2004.

[Hiemstra and Mihajlovic, 2010] Djoerd Hiemstra and Vojkan Mihajlovic. A database approach to information retrieval: The remarkable relationship between language models and region models. Technical Report arXiv:1005.4752, May 2010. Comments: Published as CTIT Technical Report 05-35.

[Kasneci *et al.*, 2008] Gjergji Kasneci, Fabian M. Suchanek, Georgiana Ifrim, Maya Ramanath, and Gerhard Weikum. Naga: Searching and ranking knowledge. In *ICDE*, pages 953–962, 2008.

[Kim *et al.*, 2009] Jinyoung Kim, Xiaobing Xue, and W. Bruce Croft. A probabilistic retrieval model for semistructured data. In *ECIR*, pages 228–239, 2009.

[Lu *et al.*, 2005] Wei Lu, Stephen E. Robertson, and Andrew MacFarlane. Field-weighted xml retrieval based on bm25. In *INEX*, pages 161–171, 2005.

[Meghini *et al.*, 1993] C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of information retrieval based on a terminological logic. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298–308, New York, 1993. ACM.

[Ogilvie and Callan, 2002] Paul Ogilvie and Jamie Callan. Language models and structured document retrieval. In *INEX Workshop*, pages 33–40, 2002.

[Ogilvie and Callan, 2003] P. Ogilvie and J. Callan. Language models and structured document retrieval, 2003.

[Ponte and Croft, 1998] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, New York, 1998. ACM.

[Prud'hommeaux and Seaborne, 2006] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. Technical report, W3C, 2006.

[Robertson *et al.*, 2004] Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. Simple BM25 extension to multiple weighted fields. In *CIKM*, pages 42–49, 2004.

[Robertson, 2004] S.E. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:503–520, 2004.

[Roelleke *et al.*, 2001] Thomas Roelleke, Ralf Luebeck, and Gabriella Kazai. The HySpirit retrieval platform, demonstration. In Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, USA*, New York, August 2001. ACM.

[Roelleke *et al.*, 2008] Thomas Roelleke, Hengzhi Wu, Jun Wang, and Hany Azzam. Modelling retrieval models in a probabilistic relational algebra with a new operator: the relational Bayes. *VLDB J.*, 17(1):5–37, 2008.

[Roelleke, 1999] Thomas Roelleke. *POOL: Probabilistic Object-Oriented Logical Representation and Retrieval of Complex Objects*. Shaker Verlag, Aachen, 1999. Dissertation.

[Stonebraker *et al.*, 1998] Michael Stonebraker, Dorothy Moore, and Paul Brown. *Object-Relational DBMSs: Tracking the Next Great Wave*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[Theobald *et al.*, 2005] Martin Theobald, Ralf Schenkel, and Gerhard Weikum. Topx and xxl at inex 2005. In *INEX*, pages 282–295, 2005.

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

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

[Wu and Roelleke, 2009] Hengzhi Wu and Thomas Roelleke. Semi-subsumed events: A probabilistic semantics for the BM25 term frequency quantification. In *ICTIR (International Conference on Theory in Information Retrieval)*. Springer, 2009.

[Zhao and Callan, 2008] Le Zhao and Jamie Callan. A generative retrieval model for structured documents. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1163–1172, New York, NY, USA, 2008. ACM.