A Context – based Semantically Enhanced Information Retrieval Model

Tudor Cioara, Ionut Anghel, Ioan Salomie, Mihaela Dinsoreanu
Technical University of Cluj-Napoca, Computer Science Department
{Tudor.Cioara, Ionut.Anghel, Ioan.Salomie, Mihaela.Dinsoreanu}@cs.utcluj.ro

Abstract

This paper approaches the use of both context and semantic information in the information retrieval process with the goal of developing context-based semantically enhanced information retrieval systems. To achieve our objective we have identified, defined and formalized three distinct types of context information relevant for an information retrieval system: knowledge context information, user context information and constraint context information. The context information is represented in an information system interpretable way by mapping it onto our RAP context model elements. The proposed information retrieval model is tested using the arhiNet system, our integrated information retrieval system for archive content, based on semantic enhancements.

1. Introduction and Related Work

The information retrieval systems deal with searching for documents or information within documents on user's request.

One of the most common information retrieval information searching techniques is looking up the documents for keywords [1], [2]. The extracted text data is unstructured, meaningless and difficult to process by computer programs. Using this type of techniques the software agents face difficulties in understanding the semantic meaning of user queries and therefore poor values are obtained for system's *precision* and *recall indicators* [3].

Another approach for the information retrieval research domain is to address the information retrieval challenges by adopting and using the Semantic Web techniques [4]. In these approaches two distinct steps can be identified: (i) semantic text refining, which transforms free text into an intermediate machine-processable representation and (ii) semantic knowledge inferring using reasoning and learning algorithms. The machine-processable semantics, captured in a domain knowledge base, is further used as a support for intelligent searches that provide the most relevant results to user queries. These relevant results may represent documents, information and knowledge related to the semantics of the keywords specified in the queries.

In the semantic text refining direction the research is concentrated on developing models and tools for capturing the semantic information in a domain knowledge base and semantically refining the documents using annotations. The OntoPop methodology [5] provides a single-step solution to (i) semantically annotate the content of documents and (ii) populate the ontology with the new concepts and instances found in the documents. The solution uses domain-specific knowledge acquisition rules which link the results obtained from the information extraction tools to the ontology elements, thus creating a more formal representation of the document content (RDF or OWL). The OntoPop methodology has certain limitations regarding solving synonyms on one hand and multiple instances with the same lexical representation on the other hand. SOBA is a system designed to create a soccer specific knowledge base from heterogeneous sources [6]. The system performs (i) automatic document retrieval from the Web, (ii) linguistic annotation and information extraction using the Heart-of-Gold approach [7] and (iii) mapping the annotated document parts on ontology elements. Ontea performs semi-automatic annotation using regular expressions combined with lemmatization and indexing mechanisms [8]. The methodology was implemented and tested on English and Slovak content.

In the semantic knowledge inferring research direction, models and techniques based on reasoning and learning algorithms are proposed. The Ginseng approach [9] provides a natural language (NL) querying access to any knowledge base developed in the OWL language. The RACER reasoner (Renamed ABox and Concept Expression Reasoner) provides a query language nRQL (new Racer Query Language) that permits conjunctive queries with head projection

978-1-4244-5007-7/09/$26.00 ©2009 IEEE

operators, negation as failure and aggregation operators [10]. Pellet is an open source OWL DL reasoner which uses the SWRL language to describe first order query rules written in the form of an implication between an antecedent (body) and consequent (head) [11]. Both the antecedent and consequent consist of multiple atoms conjunctions. The use of reinforcement learning techniques to deduce new semantic knowledge information is proposed in [12], [13].

Overall, the discussed approaches fail to consider the context in which the queries are made as relevant information for the information retrieval process. In this paper we overcome this problem and take the information retrieval researches one step further by proposing an information retrieval model that considers both the context and the semantic information in the query process aiming at the development of context based semantically enhanced information retrieval systems. To achieve our objective we have identified, defined and formalized three distinct types of context information relevant for an information retrieval system: knowledge context information, user context information and constraint context information. The context information is represented in an information system interpretable way by mapping it on our RAP context model [14]. The proposed information retrieval model is tested using the arhiNet system, our integrated information retrieval system for archive content, based on semantic enhancements [15].

The rest of the paper is organized as follows: Section 2 presents the arhiNet information retrieval system; Section 3 gives a short overview of our RAP context model; Section 4 shows how the archival context information can be represented and formalized in a programmatic manner using the RAP context model while Section 5 concludes the paper and describes the future work.

2. The arhiNet Information Retrieval System

arhiNet is an integrated information retrieval system for developing, managing and retrieving archive content based on semantic enhancements. The semantic enhanced content allows for applying data mining, reasoning and learning algorithms to identify information correlation and retrieve the semantic relevant data.

The arhiNet system architecture defines three main layers (Figure 1): (i) the primary documents acquisition layer, (ii) the documents processing layer and (iii) the knowledge processing and retrieval layer. The primary documents acquisition layer includes both the data acquisition from the archive primary sources (raw archival documents) and also the structuring of the primary data for semantic enhancements. The document processing layer uses pattern-matching to extract relevant data from the raw documents. Based on the domain ontology and on a set of semantic rules, the documents are then semantically annotated. New concepts and instances are identified and added to the domain ontology as a result of this process. The knowledge processing and retrieval layer defines reasoning/learning and mining algorithms for executing intelligent queries that enable searching for the most relevant information available in archival documents.

The user input query triggers a complex reasoning

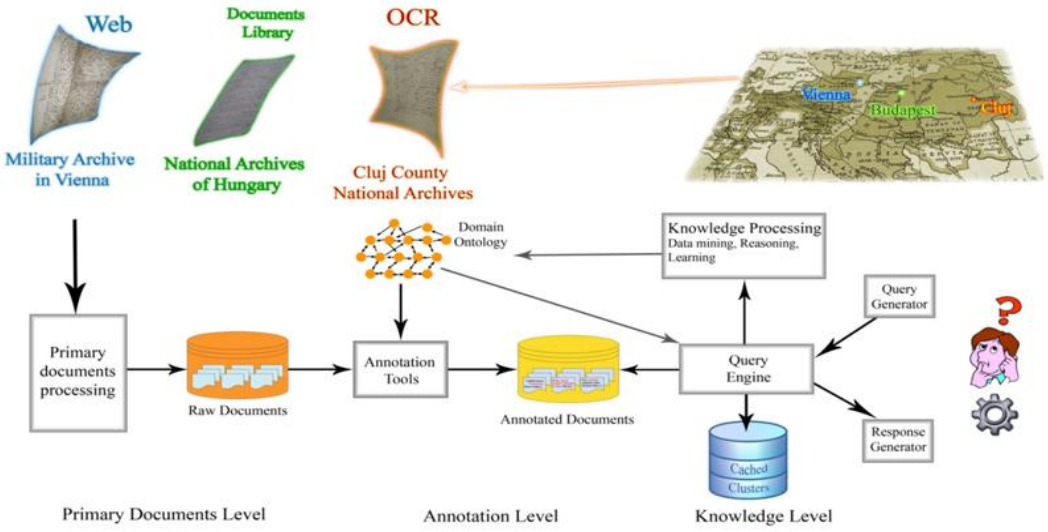

Figure 1. The arhiNet information retrieval system architecture

process that includes synonym search, logical inferences and subclass/super-class searches. As a result, the set of query relevant documents is identified and new query relevant knowledge may be generated.

3. The Context Information Model

To represent the context information in a programmatic manner we develop the RAP context model. This model defines to two ways for context information representation: a set based and ontology based.

In the set based approach the context information is modeled as a triple: **C = <R, A, P>** where **R** is the set of context resources, **A** is the set of context actors and **P** is the set of context related policies.

A **context resource** is a physical or virtual entity which generates and /or processes context information. A context resource has a unique identity, can be annotated with semantic information. A resource is characterized by its *properties*, *services* and *influence zone*. The *resource properties,* **K(r)**, describes the set of relevant context information provided by the resource. The *resource services,* **S(r)**, specifies the resource functionality as a set of services (for example a service that locates / updates an object). The actors interact with a context resource through its attached services. The *resource influence zone*, **Z(r)**, represents the physical or logical area in which the resource presence can be sensed (in other words, it becomes visible for an actor or for another resource).

A **context actor** represents a physical or virtual entity that interacts directly with the context or uses the context resources to fulfill its needs. The actor is a context information generator, has a unique identity and can be annotated with semantic information. An actor is characterized by: (i) its specific resources $\mathbf{R_{ai}}$, (ii) context related request $\mathbf{Req(a_i)}$, (iii) its preference $\mathbf{Pref(a_i)}$ and (iv) the actor-context contract $\mathbf{Ctr(a_i, CTX)}$.

A **context policy** (**p**), represents a set of rules that must be followed by the actors or resources located in the context influence zone.

In the ontology based representation the relationships between the context model sets are modeled in a general purpose context ontology core (see Figure 2). The domain specific concepts are represented as sub trees of the core ontology by using is-a type relations. A system context situation is represented by the core ontology together with the domain specific concepts sub trees and their instances in a specific moment of time.

The two ways of representing the context (set based and ontology based) are equivalent and need to be kept synchronized. The set based context model is used to evaluate the conditions under which the context management agents should execute self-* processes in order to enforce the autonomic properties at the middleware level (self-configuring, self-healing, self-optimizing and self-protection). The ontology based representation is used by the context aware applications for reasoning and learning purposes.

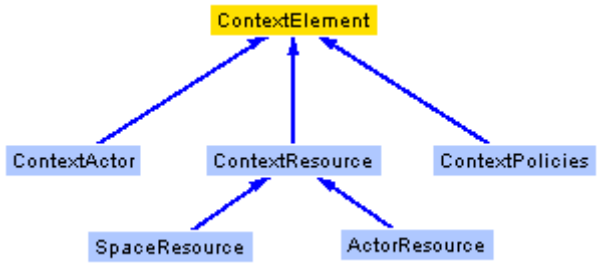

Figure 2. The context model core ontology

In order to provide an accurate representation of the real world context, the following context artifacts are defined (see Figure 3): *specific context model*, *specific context model instance* and *context – actor instance*.

The specific context model $\mathbf{C_S} = \mathbf{<R_S, A_S, P_S>}$ is obtained by mapping the context model onto different real contexts and populating the sets with real context specific elements.

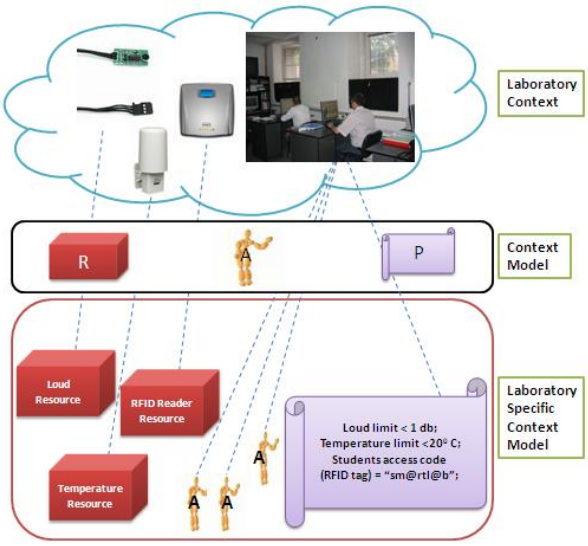

Figure 3. The RAP context model

A specific context model instance $\mathbf{C_{SI}} = \langle \mathbf{R_{SI}}, \mathbf{A_{SI}}, \mathbf{P_{SI}} \rangle$ contains the set of context resources with which the middleware interacts, together with their values in a specific moment of time **t**. The specific context model represents the context situation to which a pervasive application build onto the middleware must adapt.

The context – actor instance $\mathbf{CI_a^t} = \langle \mathbf{R_a^t}, \mathbf{a}, \mathbf{P^t} \rangle$ contains the set of context resources with which the actor can interact, together with their values in a specific moment of time **t**. A context – actor instance represents the projection of the specific context model instance onto a certain actor.

4. Representing the Retrieval process Relevant Context Information

In this section we identify, define and formalize the system execution context as relevant information for the retrieval process.

For an information retrieval system we define the context as the set of information, relevant for the document querying and retrieving processes, provided by its resources. We have identified three distinct types of context information that can be used for an information retrieval system: *knowledge context information*, *user context information* and *constraint context information*.

In order to represent the context information in a system interpretable way we use the RAP context model. The three types of context information are mapped onto the RAP context model specific concepts: *context resources*, *context actors* and *context polices*. The mapping is presented using our semantically enhanced information retrieval system for archive content, arhiNet, as an example. The proposed model can be easily extended for any information retrieval system.

The **knowledge context information** is provided by the set of archival documents used by the arhiNet information retrieval system. We represent the knowledge context information (see Figure 4 for details) by using a one to one mapping between a RAP model context resource and an archival document (notated as $\mathbf{Doc_{arc}}$).

In the arhiNet system an archival document becomes a context resource that provides relevant information for the retrieval system through its semantic annotations (notated as $\mathbf{SA(Doc_{arc})}$). The relation between the archival document and its semantic annotations corresponds to the *resource – property* relation in the RAP context model. The set of document specific annotations, represented in an ontological manner, are stored together with the annotated document. The RAP *context resource service* corresponds to a *document attached web service* through which the document can be accessed.

The *archival document influence zone* represents the set o archival documents that are correlated with the current document. We determine the correlation using the semantic similarity measure between the documents annotations ontology representation (see function **sim** in Figure 4).

$$R = < Doc^i_{arc} > \ i: 1..n$$
$$K(r) => K(Doc_{arc}) = < SA(Doc_{arc}) >$$
$$S(r) => S(Doc_{arc})$$
$$Z(r) => Z(Doc_{arc}) = \{ Doc^i_{arc} \mid \text{where } sim(Doc^i_{arc}, Doc_{arc}) = 1\}$$

Figure 4. The knowledge context information representation

The **user context information** is determined by the set of context actors that interact through queries with the information retrieval system. An information system user can be described by his cognitive capabilities and cultural background (**CCCB**), his computational capabilities (**ENV**) and motivation for using the information retrieval system (**M**) [16]. For query processing and information retrieval, the system must take into consideration the user's **CCCB**. The returned relevant data set or documents must be different if the same query is issued by a ten year child or by a historical expert. **ENV** describes the actor available computational resource used for querying the information retrieval system. The set of relevant data or documents returned by the information retrieval system is conditioned, for example, by the display type, memory allocation or the available software. In the domain literature the user motivation **M** is usually described using certain types of behavior [17]. Consider for example two computer science students that seek information about computer memory types. The first student seeks information for writing a paper and will explore the topic in depth. The second student needs the information only for buying a computer and wants immediate clarification without further details. In this case the set of relevant documents must be different for each student.

The cognitive capabilities and cultural background, together with the motivation, are mapped onto the RAP context actor preferences while the user computational capabilities are mapped onto the context actor resources (see Figure 5 where **a** represents an actor as an information retrieval user).

R_a = < ENV >
Req(a) = < Query >
Pref(a) = < CCCB, M >

Figure 5. The user context information representation

The **constraint context information** is determined by the set of context policies that drive the actor-context interaction. As an example, a user may have restricted access to documents or banned to use the system. In the process of query processing and information retrieval, the system must take into consideration this kind of limitations when determining the set of data or documents returned to the user. In our approach (Figure 6), the RAP actor – context contract **Ctr(a_i, CTX)** describes the user – system interaction limitations together with the user capabilities.

Ctr(a_i,CTX) => Ctr(a_i, IR)
Ctr(a_i, IR) = < R_{ai} , Pref(a_i),P >

Figure 6. The constraint context information representation

Figure 7 shows the arhiNet information retrieval system enhanced with the RAP context model mapping for taking into consideration in the querying process the context information.

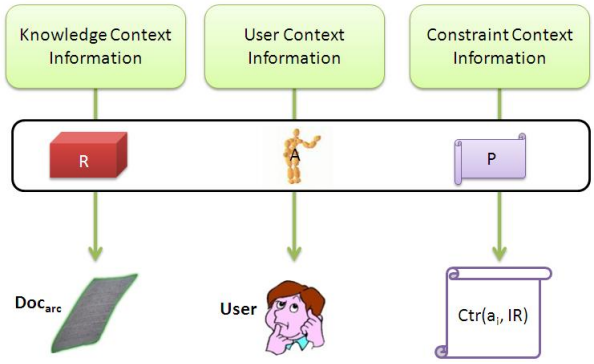

Figure 7. The RAP-arhiNet context modeling

5. Conclusions and Future Work

In this paper we have proposed an information retrieval model that considers the context information together with semantic information for the retrieval process. To achieve our objectives, we have defined and formalized three distinct types of context information using our RAP context model: the knowledge context information, the user context information and the constraint context information. The proposed information retrieval model is presented using arhiNet, an integrated information retrieval system for archive content based on semantic enhancements, as a test case system.

For future work we intend to formalize and develop a generic model for developing context based semantically enhanced information retrieval systems. By using Precision, Recall, Fall-Out and F-measure as performance metrics we intend to prove that our approach of considering the context information for query processing gives better results than other existing approaches.

6. References

[1] Guoliang Li, Jianhua Feng, Jianyong Wang and Lizhu Zhou, "Effective keyword search for valuable lcas over xml documents", *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 31-40, ISBN:978-1-59593-803-9, 2007.

[2] Paul Ogilvie and Jamie Callan, "Combining document representations for known-item search", *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 143 - 150, ISBN:1-58113-646-3, 2003.

[3] Guizhen Yang, Saikat Mukherjee, I. V. Ramakrishnan, "On Precision and Recall of Multi-Attribute Data Extraction from Semistructured Sources," *Third IEEE International Conference on Data Mining (ICDM'03)*, pp.395, 2003.

[4] Tao Jiang, Ah-Hwee Tan and Ke Wang, "Mining Generalized Associations of Semantic Relations from Textual Web Content", *IEEE transactions on knowledge and data engineering*, vol. 19, no. 2, 2007.

[5] Amardeilh F., "OntoPop or how to annotate documents and populate ontologies from texts", *Proceedings of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, June 12, 2006.

[6] Buitelaar P., Cimiano P., Racioppa S. and Siegel M., "Ontology-based Information Extraction with SOBA", *In Proceedings of the International Conference on Language Resources and Evaluation*, pp. 2321-2324, 2006.

[7] Schäfer U., "Integrating Deep and Shallow Natural Language Processing Components – Representations and Hybrid Architectures", *Saarbrücken Dissertations in Computational Linguistics and Language Te, DFKI GmbH and Computational Linguistics Department*, Saarland University, Saarbrücken, Germany, 2007.

[8] Laclavik M., Ciglan M, Seleng M and Krajei S., "Ontea: Semi-automatic Pattern based Text Annotation empowered with Information Retrieval Methods", *Tools for acquisition,*

organisation and presenting of information and knowledge: proceedings in Informatics and Information Technologies, ISBN 978-80-227-2716-7, pp. 119-129, 2007.

[9] Abraham Bernstein, Esther Kaufmann and Christian Kaiser, "Querying the Semantic Web with Ginseng:A Guided Input Natural Language Search Engine", *Proceedings of 15th Workshop on Information Technology and Systems*, 2005.

[10] Volker Haarslev and Ralf Möller, "Racer: An OWL Reasoning Agent for the Semantic Web", *Proc. Int'l Wkshp on Applications, Products and Services of Web-based Support Systems*, 2003..

[11] Evren Sirin, Bijan Parsia, Bernardo C Grau, Aditya Kalyanpur and Yarden Katz, "Pellet: A practical OWL-DL reasoner", *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 5, No. 2., pp. 51-53, 2007.

[12] Mehdi Amoui, Mazeiar Salehie, Siavash Mirarab and Ladan Tahvildari, "Adaptive Action Selection in Autonomic Software Using Reinforcement Learning", *Proc. of the Fourth International Conference on Autonomic and Autonomous Systems (ICAS'08)*, 2008, pp. 175-181.

[13] Neil O'Connor, Raymond Cunningham, "Self-Adapting Context Definition", *First International Conference on Self-Adaptive and Self-Organizing Systems*, 2007.

[14] Ioan Salomie, Tudor Cioara, Ionut Anghel and Mihaela Dinsoreanu, "RAP - A Basic Context Awareness Model", *Proceedings of 4th IEEE International Conference on Intelligent Computer Communication and Processing*, Cluj-Napoca, Romania, 2008, pp. 315-318.

[15] Ioan Salomie, Mihaela Dinsoreanu, Cristina Pop, Sorin Suciu, Tudor Vlad and Ioana Iacob, "ARHINET A System for Generating and Processing Semantically-Enhanced Archival eContent", *Proceedings of WEBIST 2009 - 5th International Conference on Web Information Systems and Technologies*, 2009.

[16] Krzysztof Janowicz, "Kinds of Contexts and their Impact on Semantic Similarity Measurement", Sixth Annual IEEE International Conference on Pervasive Computing and Communications, 2008.

[17] Jerald Hughes, "The Ability-Motivation-Opportunity Framework for Behavior Research in IS," *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, p.250a, 2007