An approach for ontology-enhanced query refinement in information portals

Nenad Stojanovic
Institute AIFB, University of Karlsruhe, Germany
nst@aifb.uni-karlsruhe.de

Abstract

In this paper we present an approach that uses domain knowledge in order to support of queries posted to an information portal. This approach enables a user to navigate through the information content incrementally and interactively. In each refinement step a user is provided with a complete but minimal set of refinements, which enables him to develop/express his information need in a step-by-step fashion. In a case study regarding searching a bibliographic database we demonstrate the benefits of using our approach in the traditional information retrieval tasks, especially the combination of the free-text based querying and the ontology-based query refinement.

1. Introduction

The growing nature of the (public available) information content implies a users behavior's pattern that should be treated in a more collaborative way in the modern retrieval systems: users tend to make short queries which they refine (expand) subsequently. Indeed, in order to be sure to get any answer to a query, a user forms as short as possible query and depending on the list of answers, he tries to narrow his query in several refinement steps. The main problem in modeling an efficient retrieval process is that a user cannot express his information need straightforwardly in a query posted to an information repository, i.e. a user's query represents just an approximation of his information need [1]. Consequently such a query should be refined in order to ensure the retrieval of as much as relevant products. Unfortunately, most of the retrieval systems do not provide a cooperative support in the query refinement process, so that a user is "forced" to change his query on his own in order to find the most suitable results. Indeed, although in an interactive query refinement process [2] a user is provided with a list of terms that appear frequently in retrieved documents, the explanation of their impact on the retrieval process is completely missing. Consequently, some redundant and/or failing refinements can be suggested to a user, what decreases the efficiency of the refinement process drastically.

Recently, in order to enable more precise searching, traditional web information portals employ more semantics for the description of the information content. Firstly, instead of a free-text query, some structuring of the content of a query is possible, e.g. according to the creation date and author of an information resource. Secondly, the content of the information resources can be annotated using terms from a predefined taxonomy, which enables using a controlled vocabulary for generating more precise queries. However, although these modelling primitives can help in increasing the precision of the retrieval process, none of them is used for the refinement of users' queries.

In our previous work we developed a logic-based approach for refining queries that uses an ontology for modeling an information repository [3]. The approach is based on the model-theoretic interpretation of the refinement problem, so that the query refinement process can be considered as the process of inferring all queries which are subsumed by a given query. Moreover, query refinements are ranked according to their relevance to user's needs, whereas these needs are on-line discovered by analysing a user's behaviour.

In this paper we extend that approach for a traditional information repository, i.e. for the case that the resources are not well structured. However, we assume that some semantic descriptions of the content of resources exist. For example, the hierarchical organisation of a taxonomy used for describing content of resources in an information repository can be treated as a light-weight ontology.

This approach enables a user to navigate through the information content incrementally and interactively. In each refinement step a user is provided with a complete but minimal set of refinements, which enables him to develop/express his information need in a step-by-step fashion. In this paper we illustrate the approach on a bibliographic database. Our evaluation study shows two main advantages of such a refinement: (i) a user can find relevant documents faster and (ii) he is more satisfied with the relevance of the documents for his information need.

The paper is structured in the following manner: In Section 2 we present the main details of our query refinement approach, whereas in Section 3 a bibliographic case study is presented. In Section 4 we give concluding remarks.

2. The approach

As a response to a user's query a search engine retrieves a set of resources (documents) that are in some

Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004)
1082-3409/04 $20.00 © 2004 **IEEE**

way relevant for that query. Usualy the documents are described using the terms that appear frequently in them [4], such that the documents retrieved for a query contain the terms from the user's query. However, such a syntactical retrieval model often leads to a semantic mismatch between a query and the documents, which results in a low precision of the retrieval system. Moreover, the query refinement that is based on such a model suffers from the same problem, such that the refinements provided to a user do not reflect his need in an approproate manner.

In this section we present an approach that uses more semantic in order to support the query refinement process. Domain ontology is used as the conceptual backbone of the approach. The approach consists of three phases:

(1) Filtering, in which the initial set of relevant documents is retrieved by a search engine,

(2) Disambiguation (Contextualization), in which the domain ontology is used for clarifying the contexts in which the retieved documents appear in order to define the model for refinement (based on the language model [5]) and

(3) Clustering, in which the set of most appropriate refinements is derived from the refinement model.

Figure 1 illustrates the whole process.

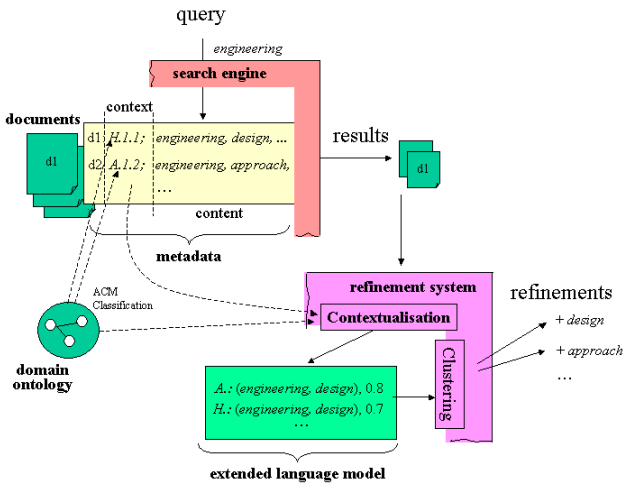

Figure 1. The query refinement "workflow"

Since the Filtering is done by a search engine, in the next two subsections we describe other two phases in the refinement process.

2.1 Disambiguation

The main problem in defining an efficient refinement system is that the vector space model, usually used for indexing documents, is not adequate for the query refinement task. Namely, the vector space model treats document terms in the isolation (i.e. it represent the content of a document as a set of terms). For example, regarding Figure 1, the document d1 is indexed with terms *engineering* and *design* whereas it can be possible that these two terms do not appear in the same context in that document. On the other hand, the task of the query refinement is to find some terms that are correlated with the terms from the query, which means that these terms can clarify the meaning of a query in a very efficient way. Regarding Figure 1, it is important to find terms that appear in the same context with the term *engineering*. For example, the terms *design* and *approach* can be treated as relevant for the refinement only if they constraint the meaning of the term *engineering* directly. Note that regarding the vector space model such a discussion is not possible.

Therefore, we introduce a novel model for representing documents in order to support query refinement process. It is based on the language model for information retrieval [5] and consists of sets of bigrams in the form (term1, term2). Moreover, these bigrams are extended by a relevance factor that describes the strength of the correlation, as well as by a information about the context in which this bigram exists (c.f. Figure 1, lefthand part).

Such a format enables us to define two types of metadata for a document (c.f. Figure 1): (i) content-based metadata that represent traditional indexes, which can be created by a search engine and (ii) context-based metadata that represent the preclasiffication of a document regarding ACM classification and usually is treated as background knowledge about a document.

Therefore, a document is described in the form:

docx: (context*; (bigrams, relevance)*)* ,

where * depicts the arbitrary number of repetitions

As already mentioned, this context information represents background knowledge about a document and seems to be very importanmt for the refinement. For example, this backgrpound information can be ACM classification very often used for describing bibliographic data. In that way it is possible to differ between two correlations which belong to two different contexts, e.g. (*engineering*, *design*) in the context of *databases* or in the context of *image processing*. Note that by condsidering inferencing a topic belongs to all topics that are on each of its paths to the root.

2.2 Clustering

The extended language model gives an overview which terms are suitable for the refinement including their relevance for it. Moreover, it ensures that some refinements that appear often, but in different context will not be treated as highly relevant. However, we assume that context-related metadata, usually provided by manual indexing, can contain some mistakes in terms of the wrong classification of documents. Inorder to bias such problems we do not constraint refinements derived from

extended language model to the context information, i.e. the queryies are extended only by content related terms.

The second task of the Clustering process is to ensure the minimality of the proposed refinements. It is possible that a potential refinement is subsumed by another refinement regarding list of results (e.g. all the documents retrieved for the query *engineering and design* are retrieved for the query *engineering and approach*). In such a case the more general results should be only presented. Subsumed refinement will be presented if the user requires refinement of the subsuming refinements.

3. Case study

In this section we present a case study regarding the bibliographic search we have done in the scope of the project SemIPort[1].

CompuScience[2] is a bibliographic database covering literature in the field of computer science, information- and communication technology, information management and science with about 160.000 citations. Citations are in English and contain bibliographic information and indexing terms. Many records also include an abstract. The citations are classified according to the Computing Reviews Classification Scheme of ACM. Therefore, for a publication not only the "traditional" bibliographic data (i.e. administrative, like author, publication year) but also the metadata w.r.t. ACM classification are given.

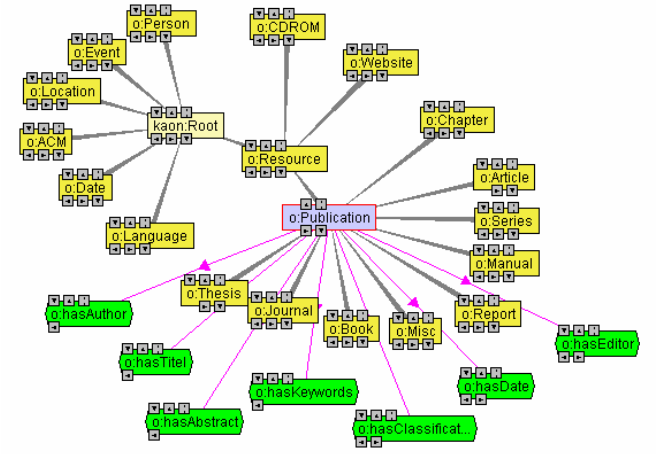

Figure 2. A part of the SemIPort ontology

Since the data are structured according to a schema defined by a database provider, we migrated this schema into an ontology using the approach described in [6]. By using this ontology the content of the database is translated into a knowledge base.

The ontology is which is partially presented in Figure 1. As the ontology modelling language we use KAON (kaon.semanticweb.org).

Figure 3 presents the simplified integration architecture. A user's query is executed against a full-text search engine (in this case Lucene - http://jakarta.apache.org/ lucene/docs/index.html). In the case that a user requires refinement of his query, the query string is transformed into an ontology based query (the task of the "conceptualisation" module in Figure 4) and processed using the approach presented in this paper (the task of the "query refinement" module in Figure 4). The generated refinements are translated into a set of query strings and retrieved to the user.

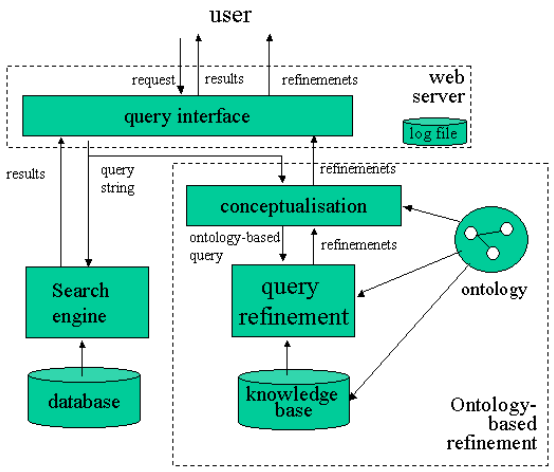

Figure 3. The integration of the logic-based refinement in a traditional information portal (a simplified model)

Our approach exploits the ACM hierarchy in order to structure the refinements in a more abstract way. In that process the ACM categories are decomposed in a step-by-step manner. The results are clustered firstly according to the top-level categories. After a user selected a category, it is decomposed on the lower levels. This process is repeated subsequently. Consequently, a user can define a query that corresponds to his information need more easily. Figure 4 illustrates this process.

Finally, Figure 5 represents the result of applying the approach presented in Section 2 on the *CompuScience* dataset. A user is provided with a complete and minimal list of refinements that can help him to refine his query according to his information need.

[1] SemIPort (http://km.aifb.uni-karlsruhe.de/semiport/) is a Semantic Web related project, funded by the BMBF, whose task is the development of semantic methods for the traditional information portals.

[2] http://www.fiz-informationsdienste.de/en/DB/compusci/index. html

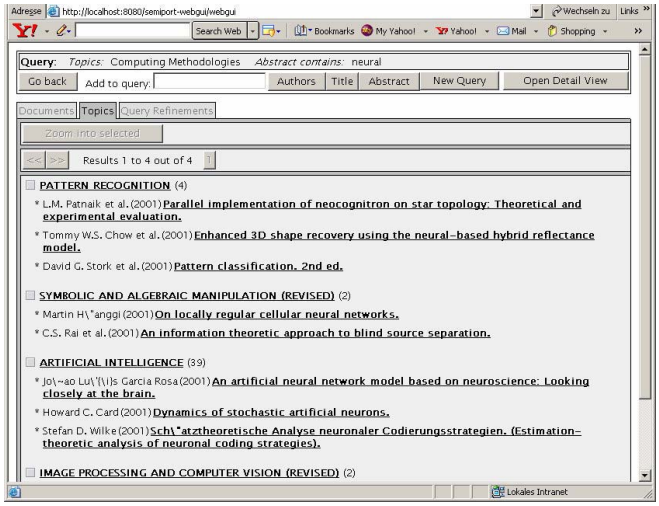

Figure 4. A screenshot from the test portal. The usage of the ACM Classification for the query refinement: the second level of the decomposition. In the first level the top-level category "Computing Methodologies" was selected

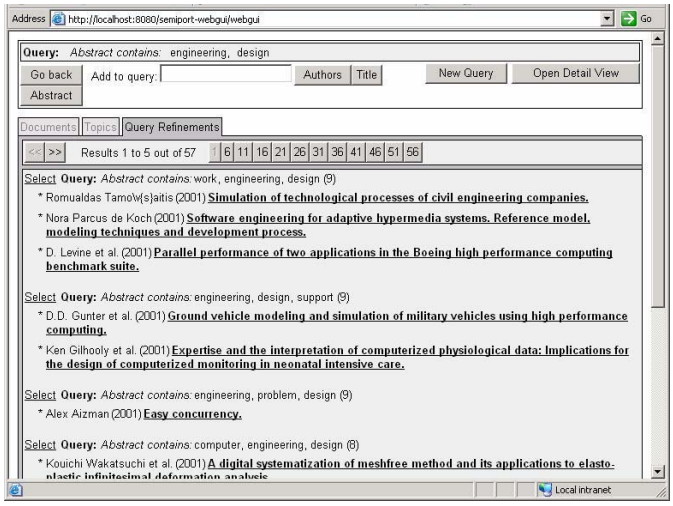

Figure 5. A screenshot from the test portal. The list of the refinements generated for the query "engineering and design"

4. Conclusion

In our previous work we developed a logic-based approach for refining queries that uses an ontology for modeling an information repository. The approach is based on the model-theoretic interpretation of the refinement problem, so that the query refinement process can be considered as the process of inferring all queries which are subsumed by a given query. In this paper we extend that approach for a traditional information repository, i.e. for the case that the resources are not well structured. However, the approach requires some background information about the domain at hand. For example, the hierarchical organisation of a taxonomy used for describing content of resources in an information repository can be treated as a light-weight ontology. This information is used for defining context of the refinement information which is extracted from the content of the repository. In that way the refinements that are more appropriate for a user's query are produced. Moreover, the approach enables a user to navigate through the information content incrementally and interactively. In each refinement step a user is provided with a complete but minimal set of refinements, which enables him to develop/express his information need in a step-by-step fashion. More precise tailoring to a user's need can be obtained by introducing user's relevance feedback in the query refinement process.

Acknowledgement. Research for this paper was partially financed by BMBF in the project "SemIPort" (08C5939) and EU in project "KnowledgeWeb" (507482). Special thanks to Eric Schwarzkopf from DFKI Saarbruecken, Germany for implementing GUI and interfaces to Lucene search engine.

5. References

[1] T. Saracevic, Relevance, "A Review of and a framework for the thinking on the notion in information science", *Journal of the American Society for Information Science*, 26, (6), 321-343, 1975.

[2] E.N. Efthimiadis, "User choices: A new yardstick for the evaluation of ranking algorithms for interactive query expansion." *Information Processing and Management*, 31(4), 605-620, 1995.

[3] N. Stojanovic, "A Logic-based Approach for Query Refinement", WI 2004, IEEE, in press

[4] R. Baeza-Yates, Modern Information Retrieval, Addison Wesley, 1999

[5] F. Song and W.B. Croft, "A general language model for information retrieval", In Proceedings of Eighth International Conference on Information and Knowledge Management (CIKM'99), 1999.

[6] Lj. Stojanovic, N. Stojanovic and R. Volz, "Migrating data-intensive Web Sites into the Semantic Web", ACM SAC 2002, 2002