

# Context-based Search in Topic-centered Digital Repositories

Christo Dichev and Darina Dicheva

Winston-Salem State University, 3206 E J Jones Computer Science Building,  
Winston Salem, NC 27110, USA  
{dichevc, dichevad}@wssu.edu

**Abstract** In this paper we address the issue of expressing and using contextual factors in information seeking tasks. Following our perception of context for the IR domain, we discuss the contextual support provided in TM4L - an environment for building and using topic-centered learning repositories. We propose an extension of the Topic Maps model with contexts and an approach for contextual retrieval that improves the search of relevant resources. The implemented support based on this extension enables users to retrieve resources by specifying contextual factors derived from their current task or goal. An evaluation of the proposed approach based on the precision and recall is presented.

## 1 Introduction

Topic-centered information organization provides a natural and meaningful way of structuring information. This fact coupled with the growing ontological support, stimulates the development of different types of concept-based repositories on the semantic web. Topic-centered organization is one of the most promising forms of digital repositories organization, as it allows users to explore unknown collections. Using this kind of exploratory search, users generally combine querying and browsing strategies to foster the investigation. Frequently, the needs for exploratory search are driven by contextual factors besides topicality. Examples of such factors are relevancy to the task at hand, level of difficulty, rigor, depth of coverage, recency, etc. Therefore an obvious way to improve the exploratory search is to allow users to submit queries augmented by contextual descriptions that can filter out non-relevant results and thus provide a focus to the exploration.

One weakness of the conventional search systems is the assumption that users search to find one or more individual resources and that each should satisfy the user's information need. In reality, however, non-trivial information needs normally can not be satisfied by a single document. When exploring a collection for information to complete a given task, one might need a set of resources, each covering part of the task requirements. Therefore exploratory search loosens this restrictive assumption built into the traditional search systems.

An obvious help to users involved in task-driven exploratory search is to provide them with a cue where the *most promising area for exploration* lies. One way to realize this idea is to select from the entire collection a set of documents satisfying par-

tially user's criteria for relevancy. The user can use this filtered set, in search for resources satisfying the remaining criteria. The driving idea in our approach is that the initial partial filtering of resources can be done by specifying user's contextual constraints. The basic intuition supporting this idea is that the key contextual aspects in information retrieval can be expressed in terms of topics and relations coupled with a chaining mechanism that identifies what is in and what is out of the context. We explore this idea in two directions: (i) defining a language that enables users to specify topic queries combined with contextual constraints in a simple way; and (ii) using the result of the contextual queries to improve query-browsing interaction.

In this paper we propose a context-based information retrieval framework aimed at topic-centered digital collections. We exemplify our approach in TM4L – an environment for building and maintaining Topic Map-based e-learning repositories [7], where this framework was incorporated and tested. However, the central idea underlying this work is not bound to any particular application domain; the ontological assumptions reflect the concept-based architecture of the repositories and the Topic Maps model.

We address the problem of searching relevant information in digital collections by focusing on two areas: how user's information needs can be described with the aid of contextual descriptors and how to combine querying and browsing. For this purpose, we extend the Topic Map model [3] with a notion of context beyond scopes and attempt to improve information retrieval by: (i) enabling users to define and explore topics and resources in specified regions, (ii) making topic map author's vocabulary available to the users for expressing their information needs, (iii) using *samples* identified by the users as pointers to locations of relevant resources within the collection.

By introducing semantically enhanced context-based search, we aim to improve both the precision and recall of the resource selection process. Precision is improved as the context specification allows more accurate description of the needed resources. Recall is improved by exploiting the collocation goal of Topic Maps - to make everything known about a subject accessible from one place – the topic reifying the subject.

The paper is organized as follows. In Section 2 we discuss briefly the Topic Map model and TM4L. In Section 3 we define contexts incorporated in topic maps. In Section 4 we propose our approach to searching in context, while in Section 5 we discuss the extension of TM4L to support contextual retrieval in learning repositories. Evaluation of the proposed approach is discussed in Section 6.

## 2 Topic Map-based Repositories

Topic Maps (TM) [3] are a semantic web technology that provides a flexible and intuitive modeling paradigm for defining a conceptual navigation layer that supports finding of web resources of various kinds, such as documents, images, database records, audio/video clips, etc. The advantage of using the TM technology for developing digital collections is twofold: from one side it provides a convenient and intuitive presentation of interrelated concepts embedded in information resources, and from another, the digital content is in a standards-based format, which makes it inter-

changeable and interoperable. Basically, topic maps are collections of topics, associations, resources, and scopes. In the TM model the concepts are reified in topics, and they can be categorized using types. Topic maps describe by means of topics what a resource is about. Associations express semantic relationships between topics, and the extent of validity of topics, associations, and resources is called scope. Topic Maps can be viewed as a method for structuring and organizing information on the semantic and metadata level.

TM associations typically interconnect topics in some kind of relevancy relations. It is much easier to discover information about a particular subject if you see it in the context of related information. Topics associations create ‘lateral’ relationships between subjects, allowing the user to see what other concepts, covered by the repository, are related to the subject of current interest and to easily browse to them.

Topic Maps are appropriate for modeling learning content as they serve as a light-weight ontology model providing learners with a browsable structure of the specific domain. An assumed purpose of the conceptual exploration of an educational topic map is that some form of learning will occur. By browsing the map, the learner will gain insight into the domain and access to information provided by the links to the learning resources.

TM4L [7] is an environment for building and using ontology-aware learning repositories represented as topic maps. The TM4L Editor allows creating hierarchies of topics, topic types and instances, as well as relations between topics and between topics and learning resources. It supports TM-related operations such as merging, browsing, searching, and scoping. The learning content created by the Editor is compliant with the ISO XML Topic Maps (XTM) standard [3]. TM4L is available for download at <http://compsci.wssu.edu/iis/nsdl/download.html>.

### 3 Context-Driven Topic Maps

The type of interaction, where an information seeking application would respond to its users presenting always the precise set of resources they might need to complete their current task, implies that the application is aware of the user’s personal *context*. There are many ways to model or represent context, and *scope* [3] is one means explicitly provided by the TM model. The scope mechanism enables any topic characteristic to be qualified by defining a range within which the information is valid. Scopes may be used to define different perspectives on the same set of information. For example, they may be used to separate “beginner” from “advanced” resources, thus enabling different sets of information to be presented to users with different knowledge levels. The scope concept is intended as syntactic shorthand in cases where more elaborate modeling of context by other means is beyond the intended expressiveness of Topic Maps.

Intuitively, we define a context as a collection of entities grouped on the basis of their relations to a common set of features determined by the current user’s goal. Entities can be any objects, facts, statements, resources, etc. The common set of features defines the *grouping criterion* for entities. The goal specifies the boundaries of the context, by limiting it to entities directly or indirectly related to the goal. The set

of assumptions that provides a background of a particular task, such as finding a book in a library, is an example of such a context. Here, the context includes the assumptions about the library organization, classification system, location of specific shelves, etc. Often certain assumptions are left implicit. That is why the notion of context is so elusive, despite of the various models proposed and developed [4, 5, 6, 12, 14]. In this work we don't intend to propose a general framework for modeling contexts, we rather take a pragmatic approach.

In the Topic Maps world all objects of interest are mapped to topics, associations, and occurrences. Our intuitive interpretation of context in TM terms is as the minimal environment that surrounds and gives meaning of the topics of interest. In this sense, context can be interpreted as the minimal set of all topics and occurrences related to a given task. Thus, we define a context as *a set of facts (topics or occurrences) grouped on the basis of their relations to a common set of topics, associations, or scopes describing a given goal*. The set of all topics related to a given topic is an example of such a context, where all objects that belong to the context share a common topic. Another example is the set of all pairs of topics related by the *whole-part* association. In this context all objects share a common association type. It can be used to explore digital content from the viewpoint of its natural parts, such as chapters, subchapters, etc. Learning materials for advanced students is yet another example, with context - the set of topics and occurrences sharing a common scope ("Advanced").

## 4 Searching in Context

The intended use of (Topic Map-based) digital repositories is to support users in their task of finding relevant resources. Therefore contextual support in the following discussions boils down to contextual support for information retrieval.

### 4.1 Contextual Aspects of Relevancy

Any combination of common factors defines a particular grouping of resources. Pre-determining all possible groupings is not a realistic approach, while grouping based on a single property is a feasible task. In a Topic Map-based digital repository, topic and resource grouping may be done on several ontological levels that reflect various contexts of digital content utilization. This is especially important for educational (learning) repositories. When a user searches for (learning) resources, key issues are:

- a) What is the resource about (i.e. topicality)?
- b) In what a form and level is the content presented?
- c) How is it related to the current user's task?

The traditional approach to relevancy is focused on topicality with less attention to other aspects, such as task relevancy, level of difficulty, technicality, depth of coverage, recency, etc. For example, if users are searching for detailed explanation of a topic, they might like to get material with more examples. To describe such contextual aspects of relevancy we can use the TM notion of *theme (scope)*. In its simplest

meaning a theme is a special topic that defines a semantic area in which selected topics and resources are valid (accessible).

Since most of the contextual aspects are subjective, different factors can make a user judge a document as relevant. Additionally, when the number of factors is large, it obscures the key factors. One factor that is ubiquitous is *topicality*.

In our context-based information retrieval framework, topicality is addressed by the Topic Map subject domain. It refers to the “aboutness” of the search as perceived by the users in relation to their information need. In addition to topicality, we introduce three predefined *themes* for modeling relevance: *task relevance*, *level of difficulty*, and *coverage*. Besides, the user can define any additional themes.

#### 4.2 Similar vs. Relevant Resources

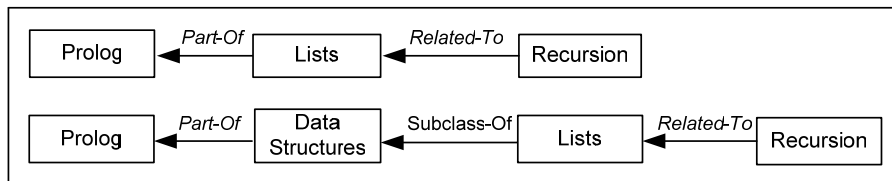
In information retrieval, the terms “similar” and “relevant” are often used interchangeably, usually without supporting justification. When we find a relevant page, i.e. a page which contains desired information, and want to know more, the best we can do is to try to find *similar pages*. There is a difference between relevance and similarity, though the search engines typically interpret both terms as synonyms and don’t support them with different functionality. “Similar pages” is probably a wrong phrase to use, as Google uses it in its algorithm to mean pages that are related to each other but based on what they are linked to (pages that are in the same neighborhood from the viewpoint of linkage).

We address the problem of seeking similar information from a contextual perspective. Concepts are related, and their relationships imply semantics. Using this property, our approach implies that the conceptually related resources are presented to the users and they are let to decide which resources are similar. Similarity, derived from conceptually related resources taken in context, can tell users more than the “same keywords” relation or uncontrolled outgoing and incoming links. Our approach to information seeking includes extending the contextual framework to support users in searching for similar resources.

#### 4.3 More Contextual Aspects: A Motivating Example

Assume a user is browsing a topic map on Programming Languages and wants to explore resources related to list processing in terms of Prolog. For a Prolog beginner even such a simple problem might not be a trivial task. In order to provide guidance to the user, the system should be aware where the relevant resources are. However, providing such kind of guidance for any kind of users and any kind of TM structuring might not be a feasible task. For example, resources linked to “Lists” (see Fig. 1) may cover only a portion of all resources related to list processing. On the other hand, resources associated with recursion are generally relevant to list processing, since recursion is the standard way of manipulating lists in Prolog. Assuming that according to our TM structure, the topic “Lists” is *Part-of* “Prolog” and the topic “Recursion” is *Related-to* “Lists”, we can show all resources that are reachable from “Prolog” following *Part-of* or *Related-to* relations. However, such type of guidance

will work only for this particular structure. An alternative organization of Prolog topics is shown on the bottom part of Fig. 1. In this case, to display all relevant resources that are reachable from “Prolog”, we have to follow *Part-of*, *Subclass-of* and *Related-to* relations. If the user’s interest includes more advanced topics, playing different roles in terms of Lists and List processing (e.g. “Lists” *Based-on* “S-expressions” or “Difference lists” *Originate-from* “Prolog Lists”), then to make the corresponding resources reachable from “Prolog” we have to change again the relation set by adding *Based-on* and *Originated-from* relations.



**Fig. 1.** Two alternative ways of relating *Recursion* to *Prolog*.

These examples illustrate some contextual aspects in information organization. In practice, we group contextually topics and resources based on a certain set of relation types. This allows the user to select from all related topics the ones that are related in a *certain way*, i.e., related on the basis of certain relation types. The drawback is that this makes information organization context-dependent, that is, dependent on the utilized set of relation types.

A possible solution to this problem is to adopt a *mixed initiative* approach, where users specify the relation types relevancy they are interested in, while the system draws the relevancy regions in the TM, restricted to the set of specified relations.

### 4.3 Context Expressions and Their Meaning

**Context expressions.** As we mentioned, our interpretation of context in Topic Map terms is as a *set of facts (topics or occurrences) grouped on the basis of their relations to a common set of topics, associations, or scopes describing a given goal*. In order to use effectively context in information seeking, we need to specify a notation for expressing it, that is, to provide operational means for defining context.

In our framework contexts are defined using a simple language for creating *contextual expressions*. Contextual expressions are built out of the following primitives: *topics*, *association types*, *themes*, and *resource types*. A contextual expression is interpreted as a query for information resources that represents a user’s current information needs. The evaluation of the query results in a set of resources satisfying the context expression. Each context expression is defined as follows:

```

<cont_expression> -> <topics> $ <relations>
    & <themes> & <resource_types> & <range>
<topics> -> [topic_name]
<relations> -> {relation_type_name}
<themes> -> {theme_name}
<resource_types> -> {resource_type_value}
<range> -> full | terminals
  
```

Pragmatically, context expressions can be interpreted as syntactic means for expressing a user's notion of "relevant resources".

For simplicity we restrict our consideration to binary relations. The intended use of relations within the contextual expressions is for generating "descendant topics". Generating "descendants" assumes some kind of directionality. For example, when using a *whole-part* relation we typically generate the *parts* from the *whole*. In a similar fashion, for other predefined relations in the topic map (if any) we assume the application of a default directionality, which defines the order of the topic traversal. Analogously, for relations defined by the TM authors, this directionality is determined by the order of the roles in the definition of the corresponding relation type. For example, *simpler(less-simple, more-simple)* specifies that, starting from the topic playing the role of *less-simple* we generate its children playing the role of *more-simple* topics.

Technically, the resource types and even themes can be expressed in terms of relations, e.g.

```
resource-type(resource: http://www.clawbox.com/lists/, type: Online notes).
```

The syntactic distinction is introduced here for continence and simplicity; these types of relations carry more special meaning in terms of (e-learning) repositories.

The term *range* refers to the level of resource inclusion/filtering given the set of descendant topics, generated from the set of topics, specified in the context expression. The term *full* denotes the set of resources linked to the full set of descendant topics, while the term *terminals* refers to the resources corresponding to the terminal topics generated in each branch of the topic generation procedure.

If the context expression includes no relations, themes or resource types, then the interpretation of such an expression is respectively *all relation types*, *all resource types*, *no themes*, and *full range*. The list of resources is obtained by adding to an initially empty list, all resources linked to the set of topics specified in the context expression, followed by generating all children topics of the latter set of topics, by applying the relations specified in the context expression. This procedure is recursively applied to the list of the newly generated topics until no more new children can be generated. An additional filtering is performed at each step, using the themes and resource properties specified by the remaining part of the context expression.

**Meaning.** In creating a context expression the user may wish to constrain the resources by using some of the predefined contexts/viewpoints such as "beginner" or "intermediate"; for example, one could restrict a query to resources considered as introductory material with easy explanation of the key concepts, by selecting "beginners" as a context of the query. In addition, users may want to constrain their queries by using attributes such as resource types; for example, by choosing the appropriate value for the resource type, one could restrict the query to online notes or examples. The user may wish also to constrain the query based on the resource coverage of the topical collection generated by the context expression. For example, the topic "List Processing" may include subtopics, "List Representation" and "Operations on Lists". The concept of *range* enables the user to restrict the query including the topic "List Processing" to return only resources linked to its terminal topics, "List Representation" and "Operations on list".

- A precise interpretation of the following context expressions is given in [8]:
- (a)  $t : R$ . The set of resources accessible from a topic  $t$  following the relation  $R$ .
  - (b)  $(t_1, t_2) : R \equiv (t_1 : R), (t_2 : R)$ . Note that “,” is interpreted as “or”.
  - (c)  $t : Q, R$ . The set of resources accessible from a topic  $t$  following the relations  $Q$  or  $R$ .
  - (d)  $E \& v$ ,  $v$  is a value of the property resource type.
  - (e)  $E \& T$ ,  $T$  is a predefined theme.

## 5 Implementing Context-based Search in TM4L

### 5.1 Defining Contexts in TM4L

A personalized, context-based search implies that the users are able to define their own context. TM4L supports “user-defined” contexts. Our approach to contextual retrieval involves two distinct stages. The first stage includes constructing a *contextual query* which is translated to a contextual expression (as described in Section 4). Contextual queries can be performed independently of the second stage; they have independent value in terms of locating resources. In the second stage, the set of resources found in the first stage are used as *seeds* for locating some other relevant resources. The idea is to use the resources corresponding to the user’s context in order to provide an orientation with regard to the organization of the collection.

Figure 2 shows TM4L Context Query interface. The user creates a query by selecting context specifiers from the Specify Context pane, and executes it by clicking the *Run* button. In the example below the context is defined by the selected context specifiers:  $\langle$ Topic: *Logic* and *Control*; Association Types: *Whole-part*; Resource Types: *Online notes*; Themes: *Advanced* $\rangle$ . The specified range of coverage is *All Nodes (Full)*.

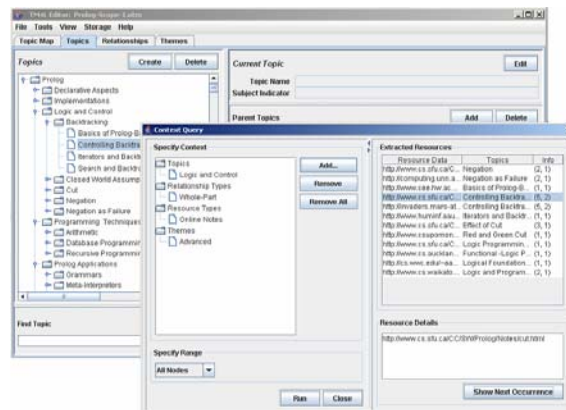


Fig. 2. Defining a context in TM4L.

The selected resources corresponding to this context are shown on the right pane where the numbers  $(m, n)$  indicate that  $n$  out of  $m$  resources linked to the topic  $t$  (se-

lected on the left pane) are included in the context. Resource details can be examined in the bottom right window. For example, if the resource is linked to several topics, the topics with co-occurring resources can be seen by clicking *Show Next Occurrence* button. For example, the resource <http://www.cs.sfu.ca/CC/SW/Prolog/Notes/cut.html> is displayed initially linked to topic “Backtracking and Cut”. After clicking *Show Next Occurrence*, the next co-location is displayed - the topic “Effect of Cut” in our example. If the same resource is linked to different topics, it is likely that they have something in common. It might be hard to articulate the commonality that caused sharing a specific resource but it makes this particular topic more likely to contain some other resources in the proximity to the current point of interest than any other arbitrary topic. The numbers  $(m,n)$  provide additional indication to the user when deciding where to look for other relevant resources. For example, when two out of five resources are selected in the defined context, this makes the remaining resources potentially interesting for exploration. Such type of common sense strategy can be applied to topic “Effect of Cut” when looking for resources explaining the Prolog pragmatics of narrowing the search space.

Semantic searches require understanding of the context in which a particular search occurs. Since this is even more important for exploratory search and especially in e-learning repositories, we enhanced the search support to help users in that regard. When results are organized in smaller groups, users are more inclined to explore [10]. By showing the resources, found during the first stage, in terms of their location within the source topic structure, we can provide an aid to schematic cognition of the structure of the collection. This process can be viewed as a partial mapping between the user’s view of the collection and the TM author’s view.

To clarify our idea, let us consider an analogy with the library search. Assume you want to locate some books on Programming Languages such as Java, Perl, Python, Prolog, and Lisp. You tell the librarian that you are looking for “Introduction to Java” showing the Dewey decimal classification, e.g. QA76.73.J38 L52 2005. The librarian knows where to look for the book and takes you to the right shelf. Now you are left in the right place for your task; you can locate other relevant titles, without any further assistance from the librarian. Following the analogy, assume that from the collection of resources found in the first stage the user selects certain relevant resources and asks the system to display them within the collection. If the system is able to show them linked to the minimal set of topics containing them, this display will be a guide to the right topics (shelves) for the user to explore further.

## 5.2 Context Mapping and Structure Cognition

Context mapping is a relation between a context called *source* and another context called *target*. The mapping can be formally represented by a triple  $\langle c_s; M; c_t \rangle$ , where  $c_s$  and  $c_t$  are the source and target contexts, respectively and  $M$  is the mapping, i.e. the relation between the explicit representations of  $c_s$  and  $c_t$ . For a TM representation we define the mapping (a special case of [4]) as a triple  $\langle c_s; f; c_t \rangle$ , where  $f$  is a function

$$r_1 \xrightarrow{f} o(t, r_2),$$

mapping a resource  $r_1$  from source context  $c_s$  to resource  $r_2$  from target context  $c_t$ , such that  $r_2$  is linked to a topic  $t$  through an occurrence relation  $o(t, r_2)$ . It is possible that  $f$  maps different resources from the source context into resources in the target paired with different topics. This would mean, for example, that all resources of the source collection are scattered in a disorganized fashion over the target structure. However, our intuition suggests that a disordered mapping is unlikely in related contexts; in most cases the mapping will be distributed between logically related topics.

We denote by  $A_s = \{r_i \mid r_i \in c_s\}$  the set of the source resources and by  $A_t = \{r_i \mid o(t, r_i) \in c_t\}$  their images in  $c_t$ , where  $f(A_s) \subseteq A_t$ .

Assume that there is a device such that for each  $r_j \in c_s$  it is able to display  $r_j$ 's image in  $c_t$ . Function  $f$  supplemented by such a device is able to reveal the structure of the target context  $c_t$  and make it cognizable for an external agent observing the distribution of  $f$ 's images. This task is problematic however as it presupposes that we know entirely the source context  $c_s$ .

Assume now that the agent carrying the context  $c_s$  is able to express partially fragments of its context using the vocabulary  $V_t$  of the target context. Let  $c_t$  denote a fragment (portion) of the context  $c_s$  expressed using the vocabulary  $V_t$ . Let  $A_p = \{r_i \mid r_i \in c_t\}$  be the set of all resources of  $c_t$ . This fact implies that  $f(A_p) \subseteq A_t$ . Denote next by  $T_p$  the set of all tuples  $o(t, r_i)$  with a fixed topic  $t$  (e.g. the set of all resources  $r_i$  linked to the topic  $t$ ):  $T_p = \{o(t, r_i) \mid r_i \in c_s, i = 1, 2, \dots, m\}$ . Assume now that there is a device such that for each  $r_j \in c_s$  it is able to display  $r_j$ 's image in  $c_t$  within  $T_p$  (e.g.  $r_i$  within the group of corresponding resources linked to  $t$  along with  $t$  and its ancestors).

The key insight here is that the knowledge acquired by observing the traces of  $f(A_s)$  (e.g. the *locations* and the *distribution* of resources from  $A_s$  inside the target context structure) will help the agent in exploring the target context  $c_t$  in a more systematic way and make the search for *relevant resources*  $r_j \in T_p$  more logical. The actual goal of this approach is to provide a framework for information exchange based on topic maps not by eliminating differences between TM authors and users or between users but by offering a system that enables *semantic interoperability* between different types of users.

### 5.3 Use of Contexts

TM4L allows formulation of *context queries*, which enable on one hand, to specify the boundaries of the contextual space, and on the other, to filter the result on the basis of certain semantic properties. The advantage of such a customized context is that it brings in results with double utilization. First, they can be used as resources satisfying the specified conditions; second, they provide a starting point for further exploration for relevant resources. Our approach to relevancy exploits the fact that resources on the same subject are typically *shelved* together. By submitting samples from the contextual query, the user is pointed to topics where he can find other relevant material.

Suppose that a user has defined a context as {"Prolog", "Advanced material", "Examples"} and has all corresponding resources displayed. If the user selects a resource that happened to be related to "Recursion" (e.g. <http://ktiml.mff.cuni.cz/~bartak/prolog/learning.html>), chances

are for this resource to be shown under the topic “List Processing” in the *Partonomy* (*chapter-subchapter*) view of the collection. Using then *neighborhood navigation*, the user can find some relevant examples in the topic map author’s original (default) context. Such an approach allows a user looking for resources from a “Programming Language” perspective to observe their duplicates, for example, in “Models of Computation”.

## 6 Evaluation

Performance evaluation of Topic Map-based contextual information retrieval methods is an important but challenging problem. Analytical performance evaluation is difficult, because many characteristics such as relevance, distribution of resources, etc., are difficult to describe with mathematical precision. Alternatively, performance can be measured by *benchmarking*. That is, the retrieval effectiveness of a system can be evaluated on a given set of resources, queries, and relevance judgments. However, no standard collections are available in this particular area, no gold standards are provided and still there is no good baseline for comparison.

Since the proposed contextual search is intended to be used as an enhancement of the available topic/keyword search in TM4L, one of the goals of the evaluation was to measure the performance of the TM4L topic search against the context-based search.

The first step is to specify what and how to measure. For topic/keyword search we consider as retrieved only those resources that are linked to the topic found by the keyword search. For example, the term *backtracking* will select the topics “Basics of Prolog Backtracking”, “Controlling Backtracking”, “Iterators and Backtracking”, “Search and Backtracking”. As a result, all resources that are linked to those three topics are considered retrieved (See Fig. 3).

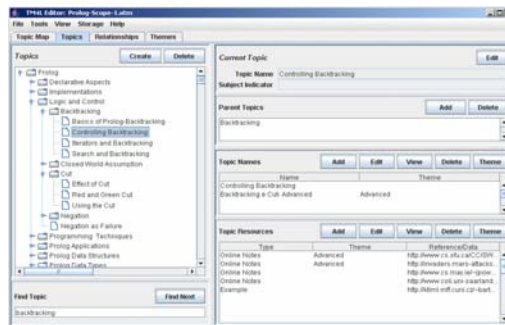


Fig. 3. Keyword search for topics containing the term “backtracking”.

Our performance evaluation is based largely on the conventional notions of *precision* and *recall*. Recall  $R$  is defined as the ratio of the number of relevant resources selected to total number of relevant resources available, while Precision  $P$  is defined as the ratio of number  $N_{RS}$  of relevant resources selected to number  $N_S$  of resources selected:

$$R = \frac{N_{RS}}{N_R} \qquad P = \frac{N_{RS}}{N_S}$$

For the evaluation of the proposed context-based search framework we selected three Topic Maps. The first one, Prolog TM (<http://gorams.wssu.edu/faculty/dichevc/Research/Prolog.xtm>), developed by the authors, was intended to represent a possibly biased collection. The second one, a topic map on Operating Systems (<http://gorams.wssu.edu/faculty/dichevc/Research/OperatingSystems-TM.xtm>), developed by graduate students, can be viewed as representing average TM authors. The last topic map is based on the ACM Computing Classification (<http://gorams.wssu.edu/faculty/dichevc/Research/ACM.xtm>) and was intended to represent a standard topic map. For each of the TMs we ran seven different pairs of queries: a keyword-based query and the corresponding context-based query. The assumed scenario was that users are searching for resources that are *examples* of particular concepts on a *beginner* level. For example, the keyword query *backtracking* can be paired with the context-based query <Topic: *backtracking*; Relations: *Whole-Part, Instance-of, Related*; Resource Type: *Example*; Theme: *Beginner*>.

Precision and recall depend on the accuracy of the separation of relevant and non-relevant resources. The definition of “relevance” and the proper way to compute it has been a significant source of argument within the field of information retrieval [13]. Most information retrieval evaluations have been focused on an objective version of relevance, where relevance is defined with respect to a query and is independent of the user. However, objective relevance makes a little sense in a learner task information support environment. Resources are relevant only if they meet a specific learner’s need or interest and he is the only person who can determine the relevance. Therefore, relevance is more subjective in learner task information support systems than in traditional document retrieval. Thus the most challenging part of this evaluation was the judgment of relevance. The thematic relevance was formulated as follows: A relevant resource is one, which contains information related to the queried concept and meets the user’s knowledge level along with other specified information needs.

For the actual performance evaluation we used 12 students divided into two groups: *good students* and *average students*. Each student was asked to run 7 queries. Six students – three from each group, ran keyword queries while the remaining six students ran context-based queries. For each of the seven queries, students were asked to provide the number of selected resources, the number of relevant selected resources and the number of all relevant resources. Based on these numbers we computed the corresponding precision and recall values for the topic maps (see Tables 1, 2, and 3).

**Table 1.**

Context Query	Precision User-1	Recall User-1	Precision User-2	Recall User-2	Query	Precision User-1	Recall User-1	Precision User-2	Recall User-2
Q1	(17,17) 100	(16,17) 94	(17,15) 88	(16,17) 94	Q1	(9,14) 64	(9,12) 75	(10,13) 77	(9,13) 69
Q2	(22,23) 96	(22,22) 100	(21,23) 91	(21,21) 100	Q2	(18,32) 56	(18,27) 67	(15,32) 47	(15,24) 63
Q3	(4,4) 100	(4,5) 80	(3,4) 75	(3,4) 75	Q3	(1,3) 33	(1,4) 25	(1,3) 33	(1,2) 50
Q4	(6,6) 100	(6,7) 86	(5,6) 83	(5,5) 100	Q4	(2,4) 50	(2,5) 40	(1,4) 25	(1,3) 33
Q5	(2,2) 100	(2,2) 100	(2,2) 100	(2,3) 67	Q5	(2,4) 50	(2,4) 50	(1,4) 25	(1,4) 25
Q6	(3,3) 100	(3,4) 75	(3,3) 100	(3,3) 100	Q6	(1,4) 25	(1,4) 25	(1,4) 25	(1,4) 25
Q7	(7,8) 88	(7,9) 78	(6,8) 75	(6,8) 75	Q7	(4,6) 67	(4,7) 57	(2,6) 33	(2,5) 40
Total	97.7	87.6	87.4	87.2	Q7	49.3	48.4	37.9	43.6

**Table 2.**

Context Query	Precision User-1	Recall User-1	Precision User-2	Recall User-2	Query	Precision User-1	Recall User-1	Precision User-2	Recall User-2
Total	89.3	90.4	89.6	87.3		58.6	55.6	49.3	51.1

**Table 3.**

Context Query	Precision Author	Recall Author	Precision User	Recall User	Query	Precision Author	Recall Author	Precision User	Recall User
Total	93.6	90.7	94	91.1		50	45.1	47.6	43.4

The evaluation indicates a notable improvement of precision and recall. The low precision and recall for keyword queries is due to the fact that they are not selective in terms of different levels of difficulty. Precision is improved as the context specification allows more accurate description of the needed resources. Recall is improved by exploiting the collocation goal of Topic Maps - to make everything that is known about a subject accessible from one place coupled with a relation traversal mechanism.

## 6 Related Work

In the first TM applications the most common way of searching topic maps was by walking through topics, occurrences, and associations. While such approach is suitable for small and not complicated maps, it does not turn out to be useful for large sets of topics. It became obvious that there was a need of creating a query language specialized for topic maps. There are several different proposals for TM query languages:

- Topic Maps Query Language – TMQL [17] - an XML-based extension of SQL, meeting the specialized data access requirements of Topic Maps.
- AsTMA? [1] – a language with queries in the form ‘Let – In – Where – Return’.
- Tolog [11] – a query language based on Prolog.

These languages are designed for selecting specified TM elements within the whole topic map. In our approach, we can constrain the query to certain localities defined by a set of topics and a set of relations.

SPARQL [9] is another similar to SQL language, which is based on RDF. Though it allows queries with second order properties, SPARQL is not designed with context queries in mind. It has to be tweaked beyond the current syntax to handle contextual queries. With our approach one can formulate properties that are valid for all relations from a given set. Thus its expressivity is above that of languages such as Tolog or versions of SPARQL (<http://kaon2.semanticweb.org/>), not supporting queries with variables at predicate positions. Despite the second order features, the semantics of our query language remains tractable; it is designed with efficiency in mind.

In terms of Web, a great part of the contextual search work is centered around automatic building of user profiles on the basis of user’s previous searches, search results, and Web navigation patterns [2, 16]. The information system uses the profiles, for refinement of future searches. In the Context Learning approach [15], the focus of learning is on judged relevant documents, query terms, or document vectors.

In the proposed approach we provide a vocabulary and language to enable users to express the task- and context-dependent features of the information of interest.

## 7. Conclusion

Efficient information retrieval requires information filtering and search adaptation to the user's current needs, interests, knowledge level, etc. The notion of context is intrinsically related to this subject. In this paper we propose an approach to context modeling in Topic Map-based digital library applications. It is based on the standard TM support for associations and scopes and defines the context as an abstraction of grouping of related information based on a specified task. The proposed model of context is utilized in an extension of TM4L, an e-learning environment aimed at supporting the development of efficiently searchable and reusable learning repositories, focused on aiding the retrieval of relevant information.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. DUE-0333069 "NSDL: Towards Reusable and Shareable Courseware: Topic Maps-Based Digital Libraries" and Grant No. DUE-0442702 "CCLI-EMD: Topic Maps-based courseware to Support Undergraduate Computer Science Courses."

## References

1. Barta, R. AsTMA? Language Definition. Bond University Technical Report, Online: <http://astma.it.bond.edu.au/astma%3Fec.dbk?style=printable>, 2003.
2. Bharat K.. Searchpad: Explicit Capture of Search Context to Support Web Search. *Int. Journal of Computer and Telecommunications Networking*, 33(1-6), 493-501, 2000.
3. Biezunski, M., Bryan, M., & Newcomb, S., ISO/IEC 13250:2000 Topic Maps: Information Technology, [www.y12.doe.gov/sgml/sc34/document/0129.pdf](http://www.y12.doe.gov/sgml/sc34/document/0129.pdf)
4. Bouquet, P., Serafini, L., Zanobini, S.: Peer-to-Peer Semantic Coordination. *J. of Web Semantics*, 2(1) (2005)
5. Dey, A.K.: Providing Architectural Support for Building Context-Aware Applications. Ph.D. Thesis, Georgia Tech (2000), <http://www.cc.gatech.edu/fce/ctk/pubs/dey-thesis.pdf>.
6. Dichev C., Dicheva D.: Contexts as Abstraction of Grouping, Workshop on Contexts and Ontologies, 12th Nat. Conf. on Artificial Intelligence, AAAI 2005, Pittsburgh (2005) 49-56.
7. Dicheva, D. & Dichev, C.: TM4L: Creating and Browsing Educational Topic Maps, *British Journal of Educational Technology - BJET*, 37(3) (2006) 391-404
8. Dichev C., Dicheva D. Contextual Retrieval of Digital Context in Topic Maps, to appear in the *Proc. of the Workshop on Modeling and Retrieval Context*, 13th AAAI Conf., 2006
9. Dodds, L. Introducing SPARQL: Querying the Semantic Web, *O'Reilly XML*, Online: 2005, <http://www.xml.com/pub/a/2005/11/16/introducing-sparql-querying-semantic-web-utorial.html>.
10. Fox, E. A. et al. Exploring the computing literature with visualization and stepping stones & pathways. *Communications of the ACM (CACM)*, 49(4): 52-58, April 2006

11. Garshol, L.M.. Tolog: Topic maps query language. *Proceedings of the XML Europe 2001 Conference*, IDEAlliance, <http://www.ontopia.net/topicmaps/materials/tolog.html>, 2001
12. Giunchiglia F.: Contextual reasoning, *Epistemologia*, Special issue on *I Linguaggi e le Macchine*, XVI (1993) 345–364
13. Harter, s. p. 1996. Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness. *Journal of the American Society for Information Science* 47, 37-49
14. McCarthy J.: Generality in Artificial Intelligence, *Comm. of ACM*, 30(12)(1987)1030–1035
15. Goker A, et al. User context learning for intelligent information retrieval. In Proc. EUSAI'04, ACM Press, 19–24, 2004.
16. Kraft R., Maghoul F., Chang C. C. . Y!Q: Contextual Search at the Point of Inspiration In the Proc. CIKM 2005, ACM, Germany, 2005.
17. Wrightson A. 2000. TMQL Draft (Topic Map Query Language). Ontopia, BSI. Online: <http://www.y12.doe.gov/sgml/sc34/document/0186.doc>.