

# Community Support Based on Thematic Objects and Similarity Search

Niels Pinkwart, Nils Malzahn, Daniel Westheide and H. Ulrich Hoppe  
*University of Duisburg-Essen, Germany*

**Abstract.** This paper describes an approach for community support based on similarity of learning objects: the current document a user is working on is used as a search template, which is matched against a learning object repository. The paper presents a simple similarity measurement, discusses potential enhancements, and shortly describes the results of a first usage study.

## 1. Introduction

Current educational practice shows a wide variety of computational tools being used by learners and learning groups in highly heterogeneous settings. The particular role of digital tools in these scenarios differs considerably. A common point for most of the networked applications is the use of digital media as a means for sharing and exchanging resources. This can be a very fruitful support for educational communities, since jointly used resources can play a key role for knowledge sharing and discovery. In addition, sharing resources offers potential for building and supporting communities of interest - groups of learners that have a joint interest in certain topics.

The construction of applications for these purposes of community support is a challenging task. Quite a number of applications and developments already exist in this field. Some approaches rely on user models to suggest communities and propose documents. These *recommender systems* typically have a weak point in that at least initial user models and/or document ratings have to be provided manually. Some techniques [4] try to address this problem with underlying *ontologies* - yet, still a manual rating of documents is necessary here. Finally, techniques like [1], which are able to dynamically recommend peers as interaction partners, usually need a detailed domain model for their calculations.

The approach presented in this paper relies on an alternative and simple conceptual model: it uses the learning objects created by the users as primary source of information. A repository service is able to propose *similar* learning objects - recommendations for artifacts which match the current context of the learner. These recommended documents can then either be accessed directly (anonymous object centered exchange), or can serve as a base for stimulating interaction among the users that created the “similar” objects.

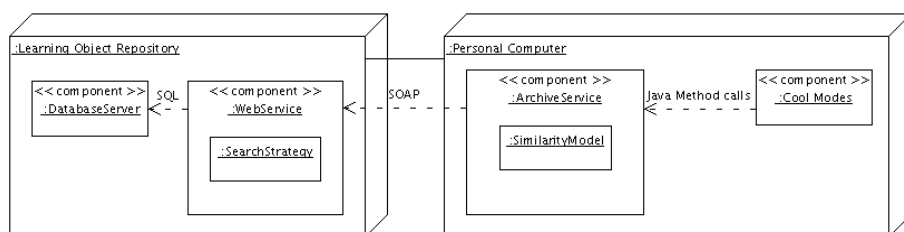
## 2. Approach for a Similarity Search on Learning Objects

A critical point of the approach outlined in the introduction is that the similarity calculation needs semantically rich data in order to produce meaningful results. Standards

like LOM, Dublin Core or IMS-LD are important contributions to syntactic and semantic interoperability, but they do not address three problems: First, the time-consuming *creation* of metadata is a necessity which most users try to avoid. Second, a *navigation* through document databases using traditional *retrieval* mechanisms and user interfaces is often based on complex electronic forms. In addition, free text input fields for specific metadata slots are of little help for retrieval of semantically similar documents. Third, a restriction to "standard" metadata is not likely to lead to fruitful retrieval, since the standards (have to) stay on a rather generic level.

To address these problems, the proposed approach relies on a partially automatic generation of metadata that exceeds current standards [5]. Using this generated data, the archive is queried for similar documents. This associative lookup enables users to find "interesting" documents without specifying exactly what they look for. Furthermore by applying ontologies potential collaborators sharing the same or topic-related interests can be pointed out to the user [2].

The results of these searches can then serve as a base for further navigation in the archive, which practically eliminates the need for manual input of search terms.



**Figure 1.** A deployment diagram of the system architecture

Figure 1 shows the system architecture. The flexible four-tier design, which allows each of the components to be exchanged provided that the technical interfaces are retained, includes two server-side components: the Learning Object Repository (LOR) [7], a central database where learning objects can be stored together with semantically rich metadata, and a web service which serves as an interface to transparently communicate with the repository. Two components are located on the client side: the concrete application used by the learner, and an archive service whose primary function is to access the web service. Details about the employed tools and the XML-based communication between them are described in [5].

Apart from the flexibility resulting from the multi-tier architecture design, also the core function of similarity search is customizable in two ways: on the client side, an exchangeable *similarity model* defines a measure for similarity of documents based on their metadata. This makes it possible to not only define which metadata of the source document are important, but more importantly to use any kind of analysis mechanism, from simple exact text matches connected with boolean logics to more sophisticated mechanisms. Metadata is preferred to full-text search because it abstracts over the concrete data format. So different sources of data can be compared. Similarly, an exchangeable *search strategy* component on the server side can be used to implement different retrieval methods reducing network traffic.

Our first implementation includes simple prototypes of similarity models and search strategies: the latter makes use of boolean retrieval in the sense that for each metadata slot that is considered important, a query with only this metadata slot is sent to the LOR. For each document, the number of (exact) matches is counted. Only documents trespassing a certain threshold are considered relevant. There, the prototype similarity model consists of a simple ranking by the number of “hits” (i.e., metadata slots that match).

### 3. Illustrative Example

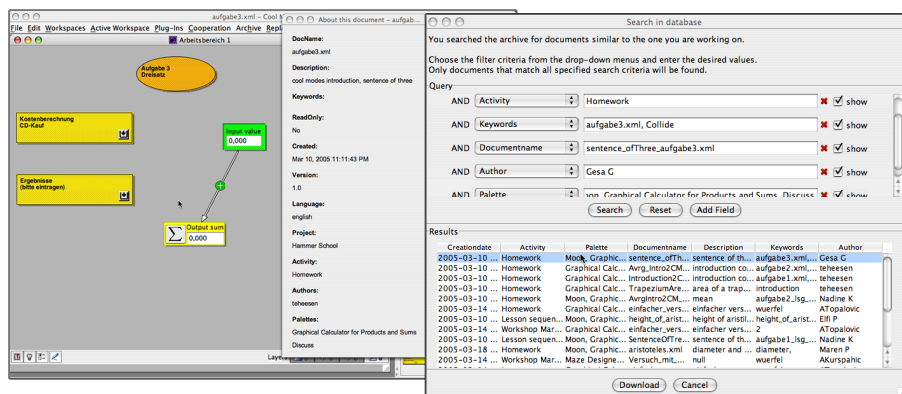


Figure 2. Similarity search example: a source document, its metadata and the search results.

To evaluate the presented architecture, the system was used in the maths lessons in a class of 20 students from a nearby higher education school. One of the students’ tasks was to solve a word problem which involved applying the rule of three (cf. fig. 2). If students had problems they were allowed to use the document repository. Students looking for help made use of the “similarity search”. Based on the semi-automatically generated metadata of the source document containing the task, the retrieval mechanism presented other students’ suggestions for a solution to the task (cf. fig. 2). Thus, the search for similar documents provides valuable results to students because they can consult others’ solutions to mathematical tasks in order to get a better understanding of the matter. In-depth evaluation studies are subject of subsequent research.

The similarity search also proved to be successful in more complex situations. For instance, when working on a document for calculating a diet based on human weight and energy needs, a similarity search finds documents from related domains of health, e.g. a system dynamics model for calculating people’s blood sugar.

### 4. Conclusion and Future Work

This paper presented a flexible architecture enabling users of the Cool Modes system to search for similar documents in a given repository. Speaking in abstract terms, we allow the users to define queries to a repository containing learning objects by defining

it in graphic notation using exactly the same elements they expect to find in the resulting documents.

Despite its simplicity, the results of the boolean retrieval mechanism currently used are promising. The next steps will be to implement more elaborated retrieval mechanisms.

The first step to a better retrieval method will be to weight the different attributes of the meta data currently used for retrieval purposes. For example, the author entry may be more decisive than the creation date. To get proper weightings, three approaches will be followed: (1) user defined weights, (2) TF-IDF-values[6] to improve the influence of characteristic entries, and (3) categorizing the documents into an ontology giving higher rankings to documents which belong to the same ontology node.

To prepare enhanced retrieval mechanisms, a measure must be defined. One approach is to define the measure per plug-in. The idea is that each particular plug-in has got certain semantics influencing the definition of similarity. For example, the exact position of the places and transitions does not matter when comparing two petri nets. In contrast to that, for a concept map the exact positions may be very important to decide if two different documents are similar.

While the proposed approach of defining distance measures on the basis of plug-ins is easily applied if only one kind of plug-in is used, some questions arise when using more than one plug-in. Since the use of multiple plug-ins is intended by our applications it must not be restricted. So we will establish a second level of weightings. The results of each plugin will be calculated and afterwards combined to get an overall result. This kind of approach has produced decent results on web documents [3] and seems promising in our case. The weights in later formula may then be adjusted based on implicit or explicit user feedback strategies.

## References

- [1] Mitsuru Ikeda, Shogo Go, and Riichiro Mizoguchi. Opportunistic group formation. In *Proceedings of the Conference on Artificial Intelligence in Education (AI-ED)*, pages 167–174, Amsterdam, 1997. IOS Press.
- [2] Nils Malzahn, Sam Zeini, and Andreas Harrer. Ontology facilitated community navigation – who is interesting for what i am interested in? In A. Dey et al., editor, *CONTEXT 2005*, Lecture Notes in Artificial Intelligence 3554, pages 292–303, Heidelberg, 2005. Springer-Verlag.
- [3] André Masloch. Ein intelligenter URL-Checker. Master’s thesis, Fachbereich Informatik, Universität Dortmund, 2003.
- [4] Stuart E. Middleton, Harith Alani, and David C. de Roure. Exploiting synergy between ontologies and recommender systems. In *Proceedings of the Semantic Web Workshop at the WWW Conference*, Honolulu, HI, 2002.
- [5] Niels Pinkwart, Marc Jansen, Maria Oelinger, Lena Korchounova, and H. Ulrich Hoppe. Partial generation of contextualized metadata in a collaborative modeling environment. In *Workshop proceedings of the Conference on Adaptive Hypermedia (AH)*, pages 372–376. Technische Universiteit Eindhoven, 2004.
- [6] Gerard Salton. *Automatic text processing — the transformation, analysis, and retrieval of information by computer*. Addison-Wesley series in computer science. Addison-Wesley, 1989.
- [7] M. Felisa Verdejo, Beatriz Barros, Jose I. Mayorga, and Tim Read. Designing a semantic portal for collaborative learning communities. In R. Conejo, editor, *Lecture Notes in Artificial Intelligence 3040*, pages 251–259, Berlin, 2004. Springer.