Topic-Centered Aggregation of Presentations for Learning Object Repurposing

Bilal Zaka[1] , Narayanan Kulathuramaiyer[1], Wolf-Tilo Balke[2], Hermann Maurer[1]

[1]Institute for Information Systems and Computer Media,
Graz University of Technology, Austria
{bzaka, nara, hmaurer}@iicm.edu

[2]L3S Research Center
Leibniz University of Hannover, Germany
balke@l3s.de

Abstract: Large collections of learning content, developed using applications such as Microsoft PowerPoint can be found in numerous institutions and organizations. The reusability of such custom-made learning content are however not supported in traditional applications. There is a need to annotate the available learning objects with suitable meta-data at various level of granularity. This paper investigates the provision of a tool that allows users to compare and efficiently repurpose presentations collected in some institution's content repository. Our preliminary experiments have shown that it is indeed possible to extend the reusability of learning objects beyond its originally intended usage. By adopting a layered approach of similarity checking, topic-centered content aggregation and repurposing has been achieved. The prototype system has also shed insights on the much-needed support mechanisms for the on-the-job, incremental elicitation of metadata.

Introduction

Content generation occupies a large chunk of the working time in both the academic and the business area. The content generated is usually either collected in reports or (more often) put together in some kind of presentation for immediate communication of salient topics to students, colleagues, or managers. Especially the last kind of communication is already since long supported by a variety of applications like Microsoft PowerPoint or Open Office Impress that allow for an easy creation of presentations. The content generated generally addresses a specific audience and thus is custom-made: even if the overall content is similar, different audiences may need different information blocks to create or refresh the understanding of a topic in the desired way. The reusability of such information blocks is however not supported in traditional applications.

This is particularly true in the area of learning or training, where a topic can be understood in a number of ways or seen from different angles and the usefulness of reusing blocks for creating new or adjusted slide sets is obvious. Therefore capturing these basic information blocks in so-called 'learning objects' (LOs) has spawned a tremendous body of research work discussing how to correctly model courses, learning units, etc. in an abstract way in a variety of levels of granularity ranging from topic level to media level and for different fields and institutions. An especially interesting part of this work deals with the question of reusability (or repurposing) of learning objects. The basic question here is how to annotate learning objects with suitable meta-data for later sensible reuse.

To standardize the meta-data several standards have been presented like the IEEE Learning Object Model (LOM) (IEEE, 2002), the Sharable Content Object Reference Model (SCORM), or National Education Training Group Learning Object Model (NETg). And already large repositories for learning objects ready for reuse have been built. For instance in the ARIADNE knowledge portal (Duval et al. 2001) tools for annotating and indexing learning objects using IEEE LOM, as well as a federated search over several repositories is offered. But still, all these repositories have to rely on a (mostly manual) annotation of learning objects. However, it has also been described (Cardinaels, et al. 2005) that most creators of presentations in fact do not annotate their content properly and already some approaches striving for automatic annotation have been presented like for instance (Cardinaels, et al. 2005), where the meta-data of Microsoft Office files is extracted as automatic annotations. Nevertheless, especially for smaller granularities of content units the problem is hard to solve.

In this paper we do not address the problem of actual meta-data generation, but investigate how to provide a tool that allows users to compare and efficiently repurpose presentations collected in some institution's or company's content repository. The basic idea is to integrate similar presentation in such a way that topically similar parts are interleaved in the target presentation and then can be easily edited by the author. The repurposing of content is thus basically broken down to a few simple steps:

1. choosing one or more topics for a presentation (in our system this is done by providing a sample presentation or providing a list of topics containing the relevant topics)
2. automatically integrating all available similar presentations from some repository and
3. finally selecting or deselecting the content parts relevant for the intended audience and filling in suitable transitions and additional topics

As a practical use case let us present two typical scenarios for which technical support for such an aggregation of presentations is necessary. As a first scenario we consider a university department which has a collection of courses on similar topics in a suitable learning repository. There can be similar lectures in different application fields, several versions of a course from previous years, etc. The main concern here is on how overlapping content across lectures can be assessed and how the lectures or updated versions are efficiently created reusing available content.

As a second scenario, we present a business organization that has a large repository of slide sets created for different target groups (e.g., product presentations, sales figures, etc.). The main issue in this case is on how the time-effective derivation of slide-sets can be aimed at new target audiences.

In the following we showcase our approach based on both document- and topic-similarity for the case of Microsoft PowerPoint slide sets and discuss the effectiveness and usability issues of our approach. Our preliminary experiments show that it is indeed possible to extend the reusability of LOs beyond its originally intended usage. We propose a context specific repurposing of content found in institutional archives. The idea is to enhance the reusability of these legacy contents by providing support for the guided discovery of matching material from a content repository. As opposed to (Najjar, et al. 2005), which explores the repurposing of LOs based on a domain specific ontology, our approach dynamically extracts similar LOs based on the implicit similarity measure in usage and structure of the proposed content. We employ a two-phased similarity checking scheme to suggest a set of similar documents and topics, created in the past.

Aggregation of Presentations based on Document and Topic Similarity

In this section we describe the layered approach employed in the extraction of similar content. The user submits a PowerPoint presentation to seek system input on related content that could be repurposed for a particular task. The system then performs document-level similarity checking to present to the user those documents closely matching some presentation or a planned course plan.

This process will result in a set of presentations that best match a source PowerPoint document in a query-by-example fashion. If as result of the similarity check there are no documents currently available with the needed grade of match. The user may either decrease the similarity threshold or make changes in the description of the LOs in the query document.

The system subsequently identifies groups of slides with the highest, content fragment-level similarity within the chosen documents at document level matching. Here we consider each slide as a minimum size content fragment. The order of slides is also taken into consideration in determining structure-level similarity within the documents discovered. As a result we present the user similar content at three levels, document level similarity, similarity at a topic level comprised of overlapping group of slides and sub-topic level similarity based on individual slide similarity. (Fig. 1) illustrates the layered similarity checking approach.

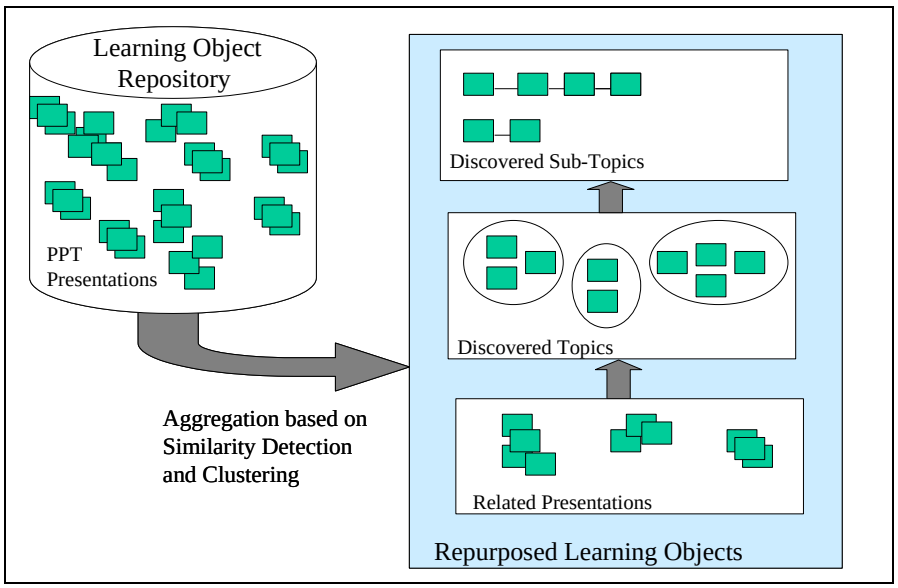

Figure 1: Layered Approach for Repurposing of Learning Objects

Slide Content X_i:
Runway safety in tough atmospheric conditions is poor, federal report says:
Providing pilots with more accurate information about icy or snowy runways is vital to reducing accidents, said a congressional report Wednesday that blamed safety problems at U.S. airports on sluggish government action.

Response from POS tagger:
Array ([0] => Runway() [1] => safety(NN) [2] => in(IN) [3] => tough(JJ) [4] => atmospheric(JJ) [5] => conditions(NNS) [6] => is(VBZ) [7] => poor(JJ) [8] => federal(JJ) [9] => report(NN) [10] => says Providing() [11] => pilots(NNS) [12] => with(IN) [13] => more(JJR) [14] => accurate(JJ) [15] => information(NN) [16] => about(IN) [17] => icy(JJ) [18] => on(CC) [19] => snowy(JJ) [20] => runways(NNS) [21] => is(VBZ) [22] => vital(JJ) [23] => to(TO) [24] => reducing(VBG) [25] => accidents(NNS) [26] => said(VBD) [27] => a(DT) [28] => congressional(JJ) [29] => report(NN) [30] => Wednesday(NNP) [31] => that(IN) [32] => blamed(VBD) [33] => safety(NN) [34] => problems(NNS) [35] => at(IN) [36] => U.S(NNP) [37] => airports(NNS) [38] => on(IN) [39] => sluggish(JJ) [40] => government(NN) [41] => action(NN))

Response from index service analyzer (tokenization, stop word removal):
Array ([0] => risky atmospheric conditions [1] => runway safety [2] => congressional report [3] => safety problems [4] => government action [5] => runways [6] => accidents [7] => airports [8] => federal report)

Response from normalization process:
•tough atmospheric conditions -> bad (bad, badness, tough, risky) weather (weather, weather_condition, atmospheric_condition)
•runway safety -> runway (runway, track) guard (guard, safety)
•congressional report -> congress (congress, United States Congress, U.S. Congress, US Congress) account (account, study, written report, news report, story, paper, write up)
•safety problems -> guard (guard, safety) problem (problem, trouble)
•government action -> government (government, authorities, regime, politics, governing, governance, government_activity) action (action, activity)
•runways -> runway (runway, track)
•accidents -> accident (accident, stroke, fortuity, chance event)
•airports -> airport (airport, airdrome, aerodrome)
•Federal report -> federal (federal) account (account, study, written report, news report, story, paper, write up)

Inverted Index data (concept vector) of Learning Object X_i:
(bad[1] weather[1] runway[2] guard[2] congress[1] account[2] problem[1] government[1] action[1] accident[1] federal[1])

Figure 2: Process of Generating a Normalized Representation for a Collection of PowerPoint Documents

The Similarity Checking Prototype

We developed a prototype for performing the actual integration of topically similar presentations. We will first describe the process of transforming the PowerPoint presentations in a repository of documents into an internal normalized form. The text from a collection of presentations is first extracted and stored as text files. The contents of these files include all text presented on slides together with user notes attached to slides and text from hidden objects maintained by PowerPoint system.

Text is then parsed and augmented with Part-of-Speech tags which are used to annotate each word with its syntactical form. The extracted word forms are then normalized into an internal form and represented as a synonym derivative (root form with standard word sense). We employed WordNet's synonym dictionary (Miller, 1995) to capture the most likely sense of each term. An inverted index is then built based on term frequencies of normalized root forms. The vector space of the normalized root senses both reduces dimensionality (as document fingerprints) and facilitates the retrieval of concept-level (synonymous) terms. (Fig. 2) illustrates this transformation process.

For performing the similarity check, the resolution of the fingerprints used for matching of target documents can be varied to either perform coarse or fine-grained similarity checking. For a document level similarity checking a document is used as fingerprint, whereas a slide is used as fingerprint at the topic level.

Experiments on the Aggregation of Learning Objects

We have conducted preliminary experiments to illustrate the workings of our prototype system for the repurposing of LOs. A collection of about 350 PowerPoint presentations with an average slide-count of 25 was used throughout our experiments. In our experimental scenario we explored the ability of the system to support the generation of content from repurposed objects. A sample presentation with some topic was presented to the system. For evaluating the capability of the system in repurposing of learning objects, we also carried out the same experiment by using text queries. In the showcase below we used the query terms 'future of computing', 'ubiquitous computing' and 'reading and writing'. We will first highlight the results of the document match experiment and will discuss the comparison with text-based queries in the evaluation section. We have also explored a comparison in performance when carrying out a presentation level matching of documents as opposed to a snippet-based matching at the slide level. In our system, the difference is given by the choice of fingerprint resolution.

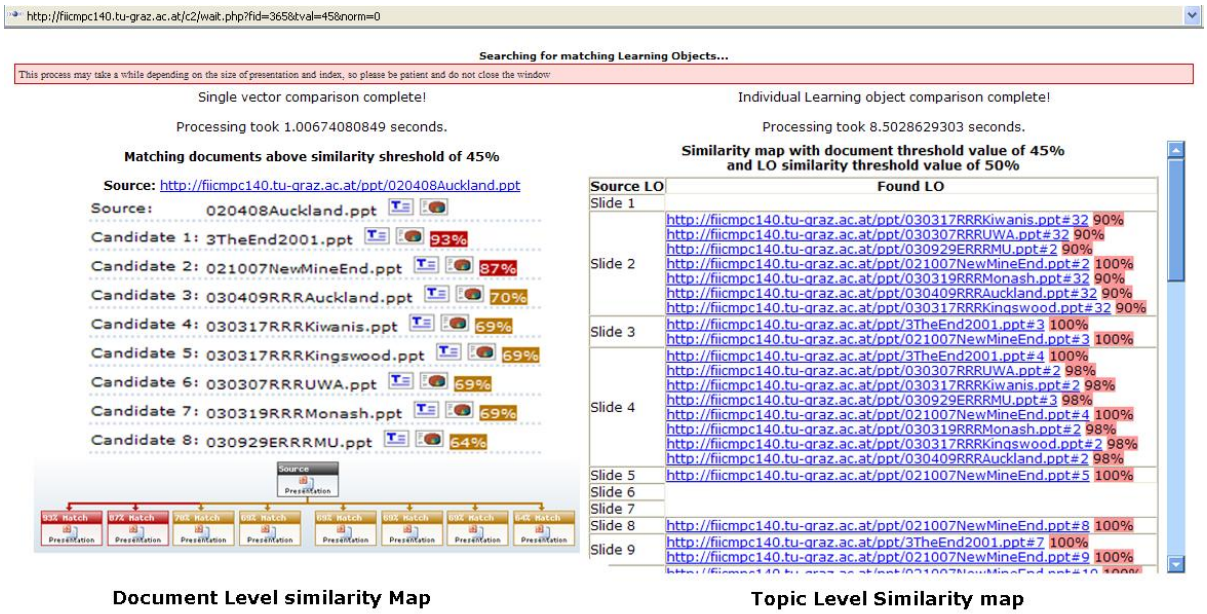

Figure 3: Results of Similarity Checking

(Fig. 3) shows both the document and slide level similarities between the source and the retrieved target documents found in the repository. The results show that the system has been able to perform an aggregation of contents at both the document and topic-specific levels. The threshold values shown provide an indication of the degree of similarity to facilitate further exploration.

(Fig. 4) then demonstrates the mapping between the presentations and the learning units discovered. The tag cloud style representation was used to depict the weights of terms. In our current implementation, the weights are merely based on term frequency. As demonstrated here, the environment is seen to enable a deeper analysis of available contents.

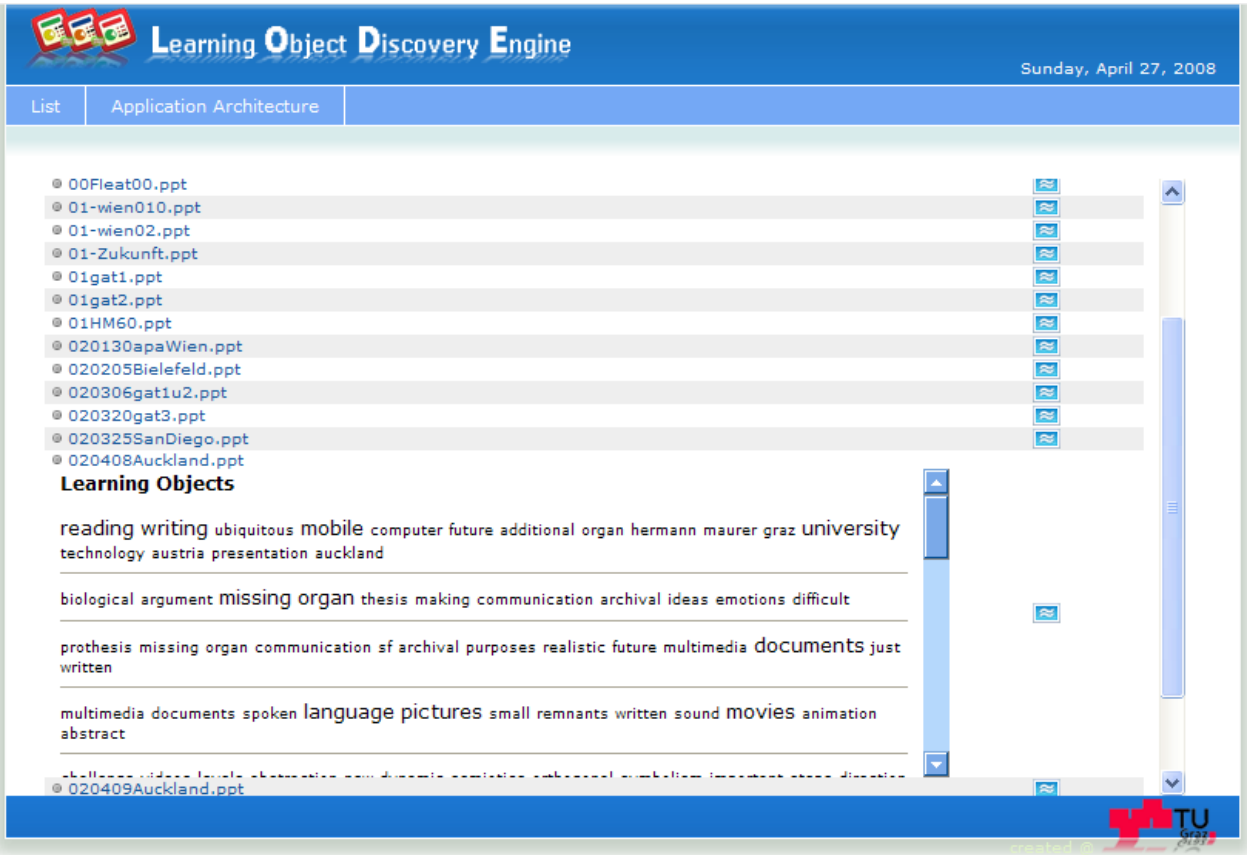

Figure 4: Environment for the Discovery of Learning Objects

Evaluation and Usability of the System

Apart from providing a presentation as query document, we also experimented with text string inputs as a query. In this experiment, we also explored various combinations of term sets. As expected, the use of just these terms produced a larger number of text matches as compared to the matching of whole documents. Exact term matching produced no results or few results in most cases.

A non-phrase specific matching was able to produce a large number of results from slides talking about the same area. The concept-level indexing was found to be useful, as the system was able to identity related slides despite the differences in words used in a query (e.g. 'mobile' as being synonymous to 'ubiquitous access'.) In this situation however, the user will still have to manually go through each relevant slide and decide individually how and where it can be applied. The proposed approach of allowing the use of PowerPoint documents as query object has thus proven to have immense value in that it provides a great deal of information about the context of work. (Fig. 5) demonstrates the three levels of discovery enabled by the proposed system.

There is great benefit that can be gained from a systems-enabled repackaging of content to serve the needs of a user's context of work. In helping an instructor in preparing a lesson plan, as described for scenario 1, it is important to present related learning content initially at a coarse-grained level (based on document-level similarity) with a gradual refinement of similarity results fine-tuned to take into consideration user feedback particular to the current task. The layered approach in presenting relevant content at multiple level of granularity is seen to be promising in assisting the user producing learning material.

Our experiments also revealed that a document-level similarity check (based on a larger vector space) required much less time. Our results in (Fig. 3) shows that document-level similarity took only 1 second, while the slide-level checking of the entire collection took 8.6 seconds. This was found to be particularly reflective of the sparse vector space of typical PowerPoint presentations. The abstract level could thus be use (even in a real-time environment) to serve as a first coarse filtering step or to provide a quick overview.

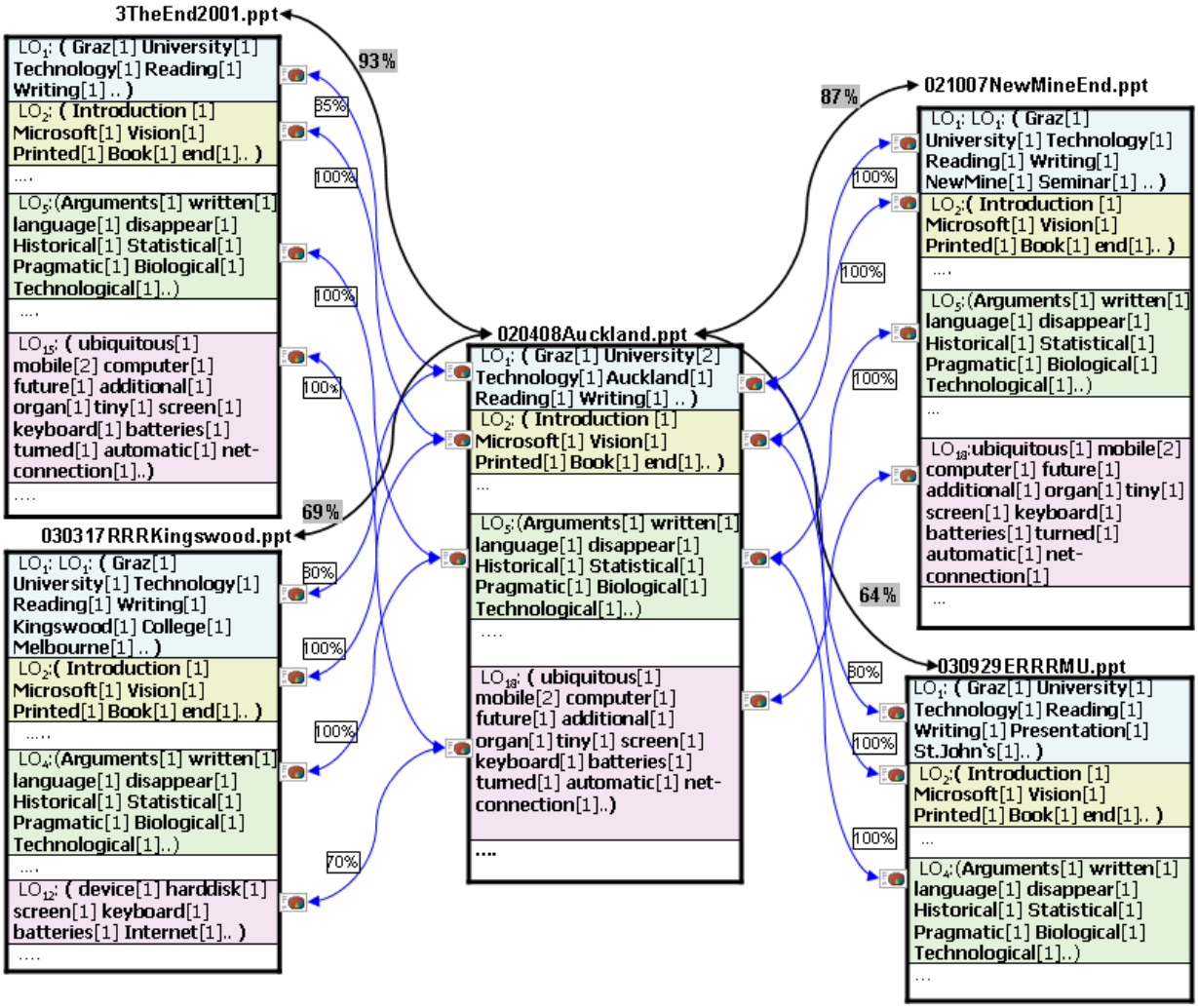

Figure 5: Illustration of the Three Levels of Contents Aggregation

Another interesting finding was that there were a number of hidden words associated with presentations stored as internal object, together with slides by PowerPoint, which resulted in surprising or unexpected results. There seem to be an internal representation of presentation objects that was still maintained internally, even after an object in a presentation was deleted. The presence of hidden data in documents has been discussed at length in works such as: (Forrester & Irwin 2005) (Byers, 2004).

Conclusion

This paper has presented a layered approach for the aggregation of content and the effective repurposing of learning objects. Our preliminary experiments have revealed the benefits and significance of the proposed approach. Dividing the similarity check into 3 layers of similarity, namely document-layer, topic-layer and slide-layer, nicely reflects the granularity of learning objects. By first doing a similarity check on presentation examples, already a large number of unrelated presentations can be ruled out. For the rest we perform a topic similarity check and interleave slides sets from different presentations based on the outcome. Finally, our slide similarity allows finding specific slides that are needed for customizing a presentation with respect to a certain target group.

The research presented here also leads to the provision of support mechanisms for the on-the-job elicitation of metadata. By first acquiring an overview of similar presentations, the user is able to gain quick insights into the relevance of the contents and also regarding the extent of re-usability. This also allows metadata annotation to be performed at both the presentation level as well as the slide level. In other words both context-specific information as well as task-specific information can be acquired from the user and applied as annotations to previously created content.

The discovery of related contents at multiple level of granularity holds the key to the discovery of deeper insights into large document archives (which typically contains PowerPoint presentations). The incorporation of implicit information of content abstraction and structure will further enhance the value of content analysis.

The application of content reuse and re-purposing however poses concerns of possible infringements of intellectual property rights or a possible negligence in the proper attribution to source documents being re-packaged. Even within a single organization, there is a need to ensure that due recognition of contributions has to be recorded and maintained as metadata to all re-packaged contents. Incentive schemes may then be drawn up to reward contributions to knowledge creation activities.

References

SCORM, Advanced Distributed Learning, http://www.adlnet.org.

Ariadne. http://www.ariadne-eu.org

Byers (2004) Byers, S., Information Leakage caused by hidden data in published documents, IEEE Security and Privacy, 2 (2) 23- 27 http://ieeexplore.ieee.org/iel5/8013/28622/01281241.pdf?tp=&isnumber=&arnumber=1281241

Cardinaels, et al. (2005) Cardinaels, K., & Meire, M., & Duval, E. Automating Metadata Generation: the Simple Indexing Interface, *International World Wide Web Conference*, 2005, Chiba, Japan 548 - 556

Duval, et al. (2001) Duval, E., & Forte, E., & Cardinaels, K., & Verhoeven, B., & Van Durm, R., & Hendrikx, K., & Wentland-Forte, M., & Ebel, N., & Macowicz, M., & Warkentyne, K., & Haenni, F., The ARIADNE Knowledge Pool System, *Communications of the ACM* 44 (5), 73-78.

IEEE, (2002) IEEE Standard for learning object metadata, Learning Technology Standards Committee of the IEEE, http://ltsc.ieee.org

Forrester & Irwin (2005) Forrester, J., Irwin, B., An Investigation into Unintentional Information Leakage through Electronic Publication, ISSA New Knowledge Today Conference, 2005, South Africa http://icsa.cs.up.ac.za/issa/2005/Proceedings/Poster/012_Article.pdf

Miller, (1995) Miller, G A., WordNet: a lexical database for English., *Communications of the ACM* 38 (11), 39 - 41. http://www.acm.org/pubs/articles/journals/cacm/1995-38-11/p39-miller/p39-miller.pdf

Najjar et al. (2005) Najjar, N., & Klerkx, J., & Vuorikari, R., & Duval, E., Finding Appropriate Learning Objects: An Empirical Evaluation, Research and Advanced Technology for Digital Libraries, Springer Berlin / Heidelberg, 3652, 323-335.

Verbert, et al. (2005) Verbert, K., & Gasevic, D., & Jovanovic J.,& Duval, E., Ontology-based learning content repurposing, *International World Wide Web Conference*, 2005, Chiba, Japan 1140 - 1141

Zaka, Maurer (2007) Zaka, B., & Maurer, H., Service Oriented Information Supply Model for Knowledge Workers, *I-Know*, 2007, 432-439.