Augmented Hyperbooks through Conceptual Integration

Gilles Falquet
CUI - University of Geneva
24, rue du Général-Dufour
1211 Geneva 4, Switzerland
falquet@cui.unige.ch

Luka Nerima
CUI - University of Geneva
24, rue du Général-Dufour
1211 Geneva 4, Switzerland
nerima@cui.unige.ch

Jean-Claude Ziswiler
CUI - University of Geneva
24, rue du Général-Dufour
1211 Geneva 4, Switzerland
ziswiler@cui.unige.ch

ABSTRACT
We describe the automatic transformation of a traditional electronic document into a augmented, virtual document. After converting the content into a "small-scale" hyperbook structure with an ontology and textual fragments, we calculate semantic similarity relations between the concepts of this hyperbook and a reference hyperbook. We finally rebuild the document by involving the retrieved hyperlinks. The aim is to show that the integration process also works without a highly detailed ontological structure of the source document.

Categories and Subject Descriptors
H 5 4 [**Hypertext/Hypermedia**]: Architectures

General Terms
Algorithms, Experimentation, Human Factors

Keywords
Hyperbook, digital library, ontology, ontology matching, virtual document, link inference

1. INTRODUCTION
The last years, we have been working on the automatic integration of hyperbooks into digital libraries. The aim was not only to put the content of a book into a DL, but also to have a true semantic integration that augments a hyperbook with relevant information fragments found in other books of the DL [2]. The idea is to provide the readers with additional information they need, for instance additional examples, term definitions, more detailed or more general information, etc.

The hyperbooks in question are based on the virtual document principle. They are made of an information layer of logically and narratively linked information fragments and a conceptual layer of semantically interconnected concepts [5]. The two layers are connected through concepts-to-fragments links. They can optionally be typed to reflect the role played by a fragment with respect to a concept (definition, example, description), much in the spirit of [1]. The augmentation technique we propose is based on the creation of 'conceptual equivalence' links between the hyperbooks' conceptual layers.

In this contribution, we focus on establishing links between a "small-scale" hyperbook that was derived from a web site and a more complete comprehensive hyperbook. Concretely, such a "small-scale" hyperbook is built from a domain ontology, which contains the most important concepts representing the hierarchical structure of the Web site. Links between the concepts are established by "is a" or "part of" links. Most of the concepts could be easily extracted, for instance by inspecting tables of contents, indexes, or glossaries. Fragments represent sections and subsections of the hypertext and are linked to the concepts. The fragment structure is stored in a XML-tree that enables the reconstruction of the Web pages after the integration process. In this way, we obtain a hyperbook from a Web site in very short time.

As example, we take a Web site including information about different aspects of Norwegian agriculture. Agriculture politics stands today in conjunction with rules of the World Trade Organization (WTO). The Web site mentions the WTO in a general context, but there is no specific information provided about the WTO. People interested in the specific functions of the WTO in conjunction with agriculture are obliged to find this information on other sites. Our aim is to create a simple hyperbook out of the mentioned Web site (the "source hyperbook") that provides information of another hyperbook (the "target hyperbook") integrated directly in the Web pages.

In recent years, we have developed a hyperbook about the WTO that includes different aspects of the organization, also agriculture. It will serve here as the target hyperbook and includes a domain ontology as well as a fragment repository. In contrast to the source hyperbook, fragments here are small texts explaining aspects of a specific concept.

Another aim is to show that the integration process is working even if there is no highly detailed structure in the ontology and the fragment repository of the source hyperbook. We take only some concepts and referred fragments that were derived from the Web pages. This approach is adequate to generate semantic relations between two hyperbooks. The next section explains our hyperbook comparison algorithm and section 3 shows how we generate augmented hypertext.

2. CONCEPTUAL COMPARISON
Figure 1 shows a graphical representation of the source hyperbook built from the mentioned Web site. It contains 11 concepts of which 9 are annotated by a fragment (in gray).

HT'05, September 6–9, 2005, Salzburg, Austria.

The target hyperbook is a part of our WTO hyperbook formed around different barriers of trade. Figure 2 shows all sub-concepts about 'domestic support'. The ontology consists of 38 concepts and 26 fragments. Fragments were annotated to 22 concepts (gray boxes).

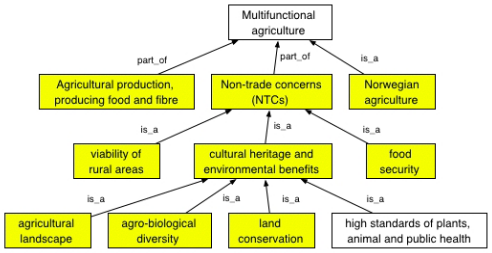

Figure 1. The ontology of the source hyperbook

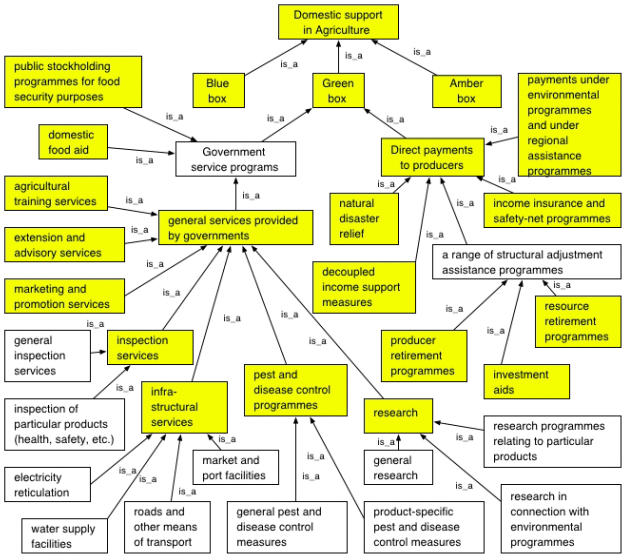

Figure 2. The ontology of the target hyperbook

We were looking for matching algorithms that include conceptual semantics. Rodríguez and Egenhofer [6] described a comparison algorithm considering the level where the concepts stand in the ontological hierarchy (expressed through α) by the following formula for similarity measures:

$$\alpha(a^p, b^q) = \begin{cases} \dfrac{depth(a^p)}{depth(a^p) + depth(b^q)} & depth(a^p) \le depth(b^q) \\ 1 - \dfrac{depth(a^p)}{depth(a^p) + depth(b^q)} & depth(a^p) > depth(b^q) \end{cases}$$

$$S(a,b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a,b)|A/B| + (1-\alpha(a,b))|B/A|},$$

They apply this similarity function to word matching, feature matching and semantic-neighborhood matching. Word matching means that the words of each concept of the source ontology p were compared with the words of each concept of the target ontology q. Feature matching means to compare attributes, roles, functions or other features that describe a concept more precisely. But it seems quite hard for authors to describe concepts by features [4]. Our hyperbook approach including textual fragments should help to avoid that users have to model their ontologies according to complex methodologies. It seems less difficult to write fragments than attaching features to ontologies. We propose to apply feature matching to the textual fragments ("fragment matching").

Using semantic-neighborhood matching might be evident, but it assumes to find a radius indicating how many nodes of the neighborhood have to be involved. Such a value is difficult to determine for a small ontology on the one side and a bigger, more hierarchical structured ontology on the other side.

We apply the approach to the above-described example by considering only word and fragment matching (418 comparisons). Running word matching by applying stopword filters sort out 12 positive values (Column 3 of Table 1).

If we consider only word matching, we immediately detect that word matching gives high values in some comparison where human beings don't expect to find a similarity relation (marked in gray in Table 1). One of the highest retrieved values is between the concepts 'Agricultural Landscape' and 'Agricultural Training Services' (0.375). It shows that even though there is no word ambiguity problem, word matching is not sufficient to evaluate the concepts' similarity.

Table 1. Results of the comparison process

Source concept	Target concept	WM	FM
food security	public stockholding programmes for food security	0.416667	0.596379
Norwegian agriculture	Domestic support in agriculture	0.375	0.266805
agricultural landscape	agricultural training services	0.375	0
food security	domestic food aid	0.37037	0.520896
Multifunctional agriculture	Domestic support in agriculture	0.357143	
Agricultural production	agricultural training services	0.294118	0
Agricultural production	domestic food aid	0.290323	0.0413793
cultural heritage and environmental benefits	research in environmental programmes	0.25	
cultural heritage and environmental benefits	payments for environmental and regional assistanc	0.185185	0.237037
Agricultural production	public stockholding programmes for food security	0.173077	0.111901
high standards (plants, animal, public health)	public stockholding programmes for food security	0.166667	
high standards (plants, animal, public health)	inspection of particular products	0.149425	
food security	Domestic support in agriculture	0	0.593243
food security	Resource retirement programmes	0	0.46801
Non-trade concerns (NTCs)	Domestic support in agriculture	0	0.463183
food security	Amber box	0	0.388739

Running fragment matching means that the words of each fragment linked to a concept of the source hyperbook are compared with words of each fragment linked to a concept of the target hyperbook. The measure is done in a similar way as in word matching by taking the above-described formula and by applying stopword filters (Column 4 of Table 1).

We immediately detect that fragment matching reinforces word matching. If word matching gives unexpectedly high values, feature matching sorts the value 0, for instance for the above-mentioned comparison (in the third row of Table 1).

But we can't conclude that fragment matching as such is a real indication for semantic similarity between two concepts. In fact, we found a total of 149 comparisons with positive values. This seems evident since we are working with big textual fragments, so there is a higher chance that some words of the source fragment match with words of the target fragment. Only few comparisons (53) give 0 for fragment matching.

One fragment in the source hyperbook was smaller compared with the others ('viability of rural areas'). Corresponding fragment matching values were more often 0. Using small fragments will get less positive results, but a better indication about the precision of the matches. But for this kind of application, we have assumed that people should not have to divide text resources more precisely because of the high workload this would imply. We want to show that our approach also works with bigger fragments. As we can't use the presented comparison algorithms in an

isolated manner, we tried to find a way to combine the two matching procedures.

Rodríguez and Egenhofer weight the different matching to get an overall comparison value, but it seems very difficult to indicate general applicalable weights [3]. When considering only our matching, it means determining a weight α for word matching and a weight β for fragment matching. At the end, we also have to determine a threshold x above which we consider that we have a good indication for semantic similarity:

$$\alpha \text{ WordMatching} + \beta \text{ FragmentMatching} > x$$

Taking into consideration the above analysis of word and fragment matching according to their precision and recall behavior and considering nearly impossible to indicate general rules for the parameters of the above formula, we propose the following simple combination:

$$\text{Similarity } s = \text{WordMatching FragmentMatching}$$

The similarity s between two concepts is taken into consideration if s is positive. But in our example, we found only six relations. In the next section, we explain how we handle the problem of few semantic similarity relations and how we generate the enriched Web pages by using the hierarchical structure of the hyperbook ontologies.

3. GENERATING VIRTUAL DOCUMENTS

As we have stored the fragment structure in a XML-tree, we can reconstruct the original Web page in a simple manner and extend it by fragments of the target hyperbook in the case where we found a semantic relation between the concepts. The fragments can be placed between the original fragments of the page or in a separated frame, for instance on the side of the initial content. The result is still a HTML or XML file that can be modified by the author. The result is similar to a virtual document that consists of a set of information fragments associated with filtering, organization and assembling mechanisms. Different representations might be possible for the integrated fragments. If the fragment is small, we could place its full text directly into the Web page. If the fragment contains a longer text, we could place so-called expand-in-place links such that the user first has to click on a hyperlink before a box with the content of the fragment will open. For instance, we can extend the Web page about 'food security' with the fragments 'public stockholding programmes for food security purposes' and 'domestic food aid' (Figure 3).

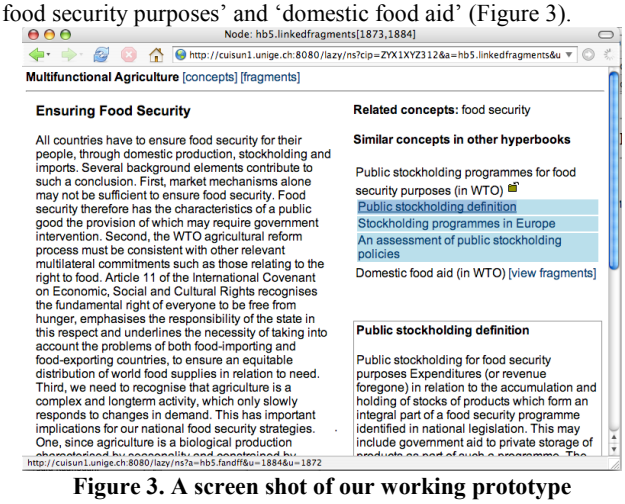

Figure 3. A screen shot of our working prototype

If there is no direct semantic similarity link between two concepts, we try to infer links by passing via concepts that owns a semantic similarity relation. In this case, we do not focus on the direct integration of fragments into the existing Web page. We further represent a part of the ontology of both hyperbooks and indicate the existence of fragments and semantic similarity links of sub- or super-concepts.

4. CONCLUSION

We extended a Web site with information of a reference hyperbook. The approach is mostly based on ontology matching and on generating virtual documents. Major problems of word matching-based approaches were resolved by considering word and fragment matching. The crucial problem of combining the different results was resolved by choosing a restrictive combination formula that sorts out only few, but evident semantic similarity relations. Finally, a virtual document was created by including the hierarchical structure of the ontology of the target hyperbook and by applying link inference mechanisms.

We also focused on extending the target hyperbook with information provided by the source hyperbook. It is easy to place the fragments into the target hyperbook, but we have to handle big pieces of text in such an inverse process. As the target hyperbook has already a fine graded structure, we consider it as a pivot hyperbook. The comparison of concepts of a third hyperbook will not take into consideration the fragments with other origin than of this pivot hyperbook, but they could serve to enrich the virtual document. In this way, a Digital Library will be created around this pivot hyperbook.

Another question is how to consider the level where concepts stand in the ontology. We assume that the source hyperbook has a flat structure. If we find a semantic similarity relation between a concept of the source hyperbook and a concept on a high level in the hierarchical structure of the target hyperbook, the semantic of this relation might be low. As sub-concepts inherit all attributes of their super-concepts, we could involve fragments of sub-concepts into the comparison process especially if there is no directly linked fragment.

5. REFERENCES

[1] Brusilovsky, P., Rizzo, R. Map-based Horizontal Navigation in Educational Hypertext. In *Proc. of Hypertext'02* (College Park, ML, 2002). ACM Press, New York, 2002.

[2] Falquet, G., Mottaz-Jiang, C.-L., Ziswiler, J.-C. Ontology Based Interfaces to Access a Library of Virtual Hyperbooks. In *Proc. of the 8th European Conference on Digital Libraries (ECDL 2004)* (Bath, UK, 2004). LNCS, vol. 3232, Springer, Berlin, Germany, 2004.

[3] Falquet, G., Nerima, L., Ziswiler, J.-C. Integration of Hyperbooks into the Semantic Web. In *Proc. of the SW-EL workshop at the 3rd International Semantic Web Conference (ISWC2004)* (Hiroshima, Japan, 2004).

[4] Marshall, C. C., Shipman, F. M. Which Semantic Web? In *Proc. of Hypertext'03* (Nottingham, UK, 2003). ACM Press, New York, 2003.

[5] Nanard, J., Nanard, M. Should Anchors be Typed Too? An Experiment with MacWeb. In *Proc. of Hypertext'93* (Seattle, WA, 1993). ACM Press, New York, 1993.

[6] Rodríguez, M. A., Egenhofer, M. J. Determining Semantic Similarity among Entity Classes from Different Ontologies, *IEEE Transactions on Knowledge and data engineering 16, 2* (2003), 442-456.