Human-machine Collaboration for Enriching Semantic Wikis using Formal Concept Analysis

Alexandre Blansché, Hala Skaf-Molli, Pascal Molli, and Amedeo Napoli

LORIA
Nancy, France
{firstname.lastname}@loria.fr

Abstract. Semantic wikis are new generation of collaborative tools. They allow to embed semantic annotations in the wiki content. These annotations allow to better organize and structure the wiki contents. It is then possible for users to build knowledge understandable by humans and computers. By this way, machines are allowed to produce or update semantic wiki pages as humans can do. In this paper, we propose a new smart agent based on Formal Concept Analysis. This smart agent can compute automatically category trees based on defined semantic properties. In order to reduce human-machine collaboration problems, humans just validate changes proposed by the smart agent. A distributed version of wiki is used to ensure consistency of the content during the validation process.

Keywords. Formal Concept Analysis, Semantic Wiki, Human-Machine Collaboration

1 Introduction

Semantic wikis are new generation of collaborative tools [1,2,3,4]. They allow to embed semantic annotations in the wiki content. These annotations allow to better organize and structure the wiki contents. Semantic wikis allow mass collaboration for creating and emerging ontological resources. They guide the users from informal knowledge contained in documents to more formal structures.

Semantic wikis allow users to build knowledge understandable by humans and computers. By this way, they also allow machines to produce or update semantic wiki pages as humans can do. This opens the opportunity to consider machines as new member of communities to produce and maintain knowledge. Consequently, such "smart agents" can reduce significantly the overhead of communities in the process of continuously knowledge building and correct humans errors.

In [5], authors coupled a case-based reasoner with a semantic wiki. The case-based reasoner can enrich the wikis with new semantic pages and thus can be considered as a smart agent. As pointed out in [5], human-machine collaboration can lead to unstable system if not managed. For example, if humans change the category tree used by the case-based reasoner, the case-based reasoner can produce incorrect results from the point of view of humans users.

In this paper, we propose a new smart agent based on Formal Concept Analysis (FCA) [6]. This smart agent can compute automatically category trees based on defined semantic properties. By this way, the FCA smart agent leverages humans from these tasks. In order to reduce human-machine collaboration problems, humans just validate changes proposed by the FCA smart agent. This is achieved using the DSMW [7] semantic mediawiki extension.

The paper is organized as follows. Section 2 introduces the FCA framework. Section 3 shows how the FCA smart agent is used to enrich the wiki. Section 4 details the validation process. The last section concludes and points future works.

2 Formal Concept Analysis

In this paper, we present a smart agent that enrich a wiki based on a classification method. Actually, any classification methods might be used. We choose Formal Concept Analysis (FCA) because it extracts concepts organized into a lattice, which is interesting for the navigation into the wiki. In this section, we briefly introduce FCA.

Formal Concept Analysis [6] is a classification method allowing to build a concept lattice where concepts are composed of an intent, a maximal set of attributes, and an extent, a maximal set of objects sharing the attributes.

A context K relies on a set of objects G, a set of attributes M and a relation between objects of attributes $I \subseteq G \times M$. Considering an object $g \in G$ and an attribute $m \in M$, $(g, m) \in I$ means that g has the attribute m.

A context can be visualized as a binary table. Table 1 shows a (simple) example of context about animals. There are five attributes that describe animals. Animals may have hair, feather, wings. They might breath in air or water. Objects are animals: bat, bird, cat and fish. In the table, a cross in one cell indicate the animal has the corresponding attribute.

	Has hair	Has feather	Has wings	Breathe in air	Breathe in water
Bat	×		×	×	
Bird		×	×	×	
Cat	×			×	
Fish					×

Table 1. Example of context (animals)

FCA allows to build concepts organized into a lattice. A concept $C_1 = (A_1, B_1)$ is defined by an extent A_1 (a set of objects) and an intent B_2 (a set of attributes that define the concept). If $C_2 = (A_2, B_2)$ is a subconcept of C_1 (denoted by $C_2 \sqsubseteq C_1$), then $A_2 \subseteq A_1$ and $B_1 \subseteq B_2$. The top concept $\top$ contains all the objects and usually its intent is empty (unless an attribute is present in each object). The bottom concept $\perp$ is defined by all attributes but usually contains no objects (unless an object has all attributes).

On figure 1 is shown the concept lattice of the context of table 1. On the graph, every node is a concept. A link between two nodes indicates a subsumption relation (a concept is a subconcept of another concept). The intent of a concept is written on a gray background, the extent on white background.

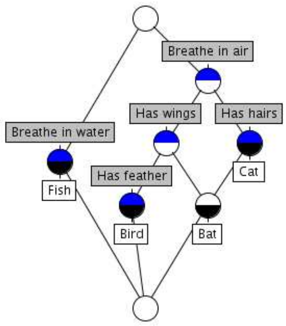

Fig. 1. Galois lattice based on the context from table 1

3 Wiki Enrichment

3.1 Principles

We developed a method that reorganizes the categories of the wiki according to the result of FCA. A new wiki will be created with the same pages and properties, but different categories, based on the lattice of concepts.

The new categories will be created based on the previous ones, and on semantic links between pages. Useful categories human users did not create might be discovered. It is even possible to start a wiki without creating any categories but only semantic links between pages, and then let the smart agent build the categories, based on the semantic links. The new categories facilitate the navigation in the wiki and provide an explicit and complete organization of the pages.

A mapping between original categories and lattice concepts is performed. Each category maps one (and only one) concept: the most general concept containing the category in its intent (the attribute concept). Each concept maps zero, one or several categories. If a concept maps a single category the category will be preserved. If a concept maps two categories or more, it means these categories are identical and should be merged (however this case is very unlikely). If a concept does not map any category, a new category will be created.

Currently, the enrichment is performed by a Java application that access the content of the wiki and create an enriched version of it.

3.2 Case study

The method presented in this paper will be illustrated by a wiki concerning academics. Here we present the initial content of the wiki. We have the following (user-defined) categories:

- `Category:Professor`;
- `Category:Topic`;
- `Category:Course`;
- `Category:Level` which contains two subcategories: `Category:Master 1 Level` and `Category:Master 2 Level`.

We also defined two properties:

- `Property:isTaughtBy`, the domain is a course, the range a professor;
- `Property:isAbout`, the domain is a course, the range a topic.

Finally, we added pages in the wiki:

- `Prof. Smith` and `Prof. Jones` in the `Professor` category;
- `Artificial Intelligence`, `Software Engineering` and `Networks` in the `Topic` category;
- `Knowledge Discovery`, in the `Course` and `Master 1 Level` categories, this page has two semantic links `isAbout:Artificial Intelligence` and `isTaughtBy:Prof. Smith`;
- `Semantic Wiki`, in the `Course` and `Master 2 Level` categories, this page has two semantic links `isAbout:Artificial Intelligence` and `isTaughtBy:Prof. Smith`;
- `Semantic Web`, in the `Course`, `Master 1 Level` and `Master 2 Level` categories, this page has two semantic links `isAbout:Artificial Intelligence` and `isTaughtBy:Prof. Smith`;
- `Design Patterns`, in the `Course` and `Master 1 Level` categories, this page has two semantic links `isAbout:Software Engineering` and `isTaughtBy:Prof. Jones`;
- `Network Administration`, in the `Course` and `Master 1 Level` categories, this page has two semantic links `isAbout:Networks` and `isTaughtBy:Prof. Jones`;
- `IPv6 Protocol`, in the `Course` and `Master 2 Level` categories, this page has two semantic links `isAbout:Networks` and `isTaughtBy:Prof. Jones`;

3.3 Formal concept analysis applied on the wiki

FCA can be applied on the content of the wiki. Objects to be classified by the FCA algorithm are the standard pages of the wiki.

The description of a page is composed of two parts: the categories it belongs to and the semantic properties it has (in our first prototype, we only considered wiki properties of type “Page”). Each of these two parts allow to build a context. We can combine these two context by apposition.

Based on the content of the wiki, as described above, we can create the context shown on table 2. When applied to this context, FCA returns the lattice shown on figure 2.

Table 2. Context based on the wiki

	Professor	Topic	Course	Level	Master 1 Level	Master 2 Level	isTaughtBy:Prof. Smith	isTaughtBy:Prof. Jones	isAbout:Artificial Intelligence	isAbout:Software Engineering	isAbout:Networks
Prof. Smith	×										
Prof. Jones	×										
Artificial Intelligence		×									
Networks		×									
Software Engineering		×									
Knowledge Discovery			×	×	×		×		×		
Semantic Web			×	×		×	×		×		
Semantic Wiki			×	×	×	×	×		×		
Design Patterns			×	×	×			×		×	
IPv6 Protocol			×	×		×		×			×
Network Administration			×	×	×			×			×

In the case study, as one can see on figure 2, four concepts match one category: `Professor`, `Topic`, `Master 1 Level`, and `Master 2 Level`. One concept matches two categories: `Course` and `Level`. All the other concepts do not match any category at all.

How to create the new categories depends on the number of categories matched by each concept. Depending on that number different methods are used. However, no categories are created for the two concepts $\top$ and $\bot$, as $\top$ always contains all pages and $\bot$ does not contain any page.

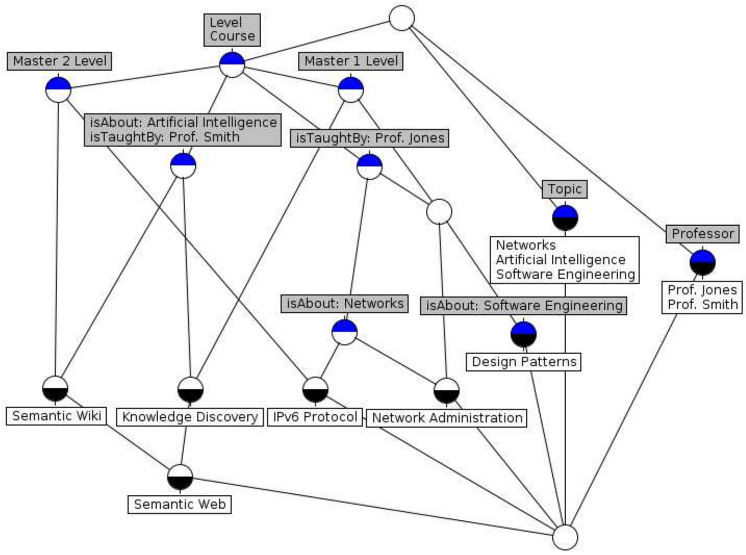

Fig. 2. Galois lattice based on the context from table 2

3.4 Preserving of an original category

If a concept matches one and only one category, this category will simply be preserved in the enriched wiki. This is the case of the category `Topic`, for instance.

Actually, in most cases, all the original categories are preserved.

3.5 Category merging

If a concept matches two categories or more, a new category is created. This new category will merge the content of the original matching categories: text of each pages are concatenated together. A default title is given to the category.

Category merging should be rare. It only happens if two or more categories always appear in the exact same pages. This would happen if several users use different terms for the same concept. Bit by bit, after a number of wiki edition, these different categories will appear in all the same pages and then will be merged by the FCA.

This is the case of the two categories `Course` and `Level`. Having these two categories is due to a naming problem. The enriched wiki has now only one category for this concept.

3.6 New categories

If a concept matches no category, a new one is created, with a default title.

This might happens in two (non-exclusive) cases:

– a page belongs to two categories or more;
– several pages having some identical properties.

A category about courses on software engineering has been created, based on the semantic relation in the page `Design Patterns`. Also, a category about courses available for both Master 1 and Master 2 students has been created, `Semantic Web` is a page of this category.

3.7 Category enrichment

Whatever the creation method of a category, all the new categories are enriched with new text content, based on properties. Sentences like "The pages belonging to this category seems to have relation T with the page P." would be appended in the page. This will help human users to understand the meaning of the category.

For instance, the category of courses about software engineering will contain the sentence "The pages belonging to this category seems to have relation `Property:isAbout` with the page `Software Engineering`.", as a description of the category.

4 Validation

4.1 Validation by human users

After the enrichment, new categories need to be validated by human users. Some merged categories might be spit, some new categories removed. Also, human users should edit all the categories: default titles should be changed into more relevant ones, text should be refined. We will present three examples of validation.

The first one concerns the two categories `Course` and `Level` that have been merged. Having this two categories was a mistake. Human users will acknowledge that and rename the merged category `Course`. They will also rename two of the subcategories `Master 1 Course` and `Master 2 Course` to make them more intelligible.

Another example concerns a new category that has been created based on the semantic relation in the page `Design Patterns` with a default name (`Category:New Category 42`, for instance). As explained in previously, the new category will contain a text describing some properties of the concept. A human user will understand that this category contains courses about software engineering and will rename it consequently. The same thing will be done for the category about courses taught by Prof. Jones.

The last example concerns a subcategory of `Master 1 Course` and `Prof. Jones' Course`. One might consider this category to be irrelevant, or at least not useful. A human user would decide to remove this category from the wiki and update the hierarchal links consequently.

4.2 Distributed wiki organization

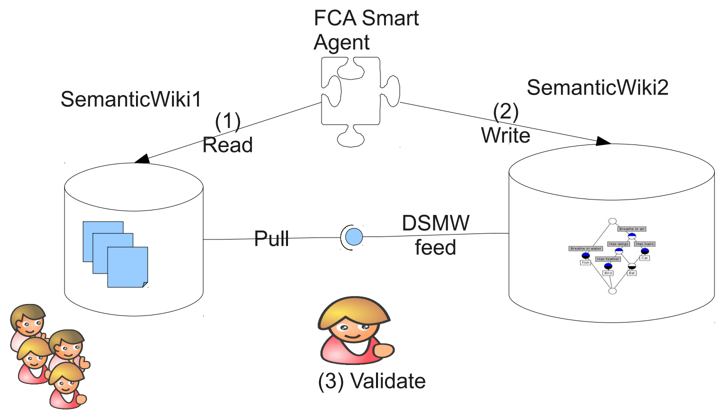

Fig. 3. Man-machine collaboration process

In order to ensure consistency of the data, we used a distributed wiki. Two semantic mediawiki sites are synchronized with the DSMW extension[1] [7] (see figure 3).

- The first one is the "SemanticWiki1" wiki. Humans access this wiki as usual.
- From this "SemanticWiki1", the FCA smart agent creates the lattice in the "SemanticWiki2" site.
- Human users will then check the content of this second wiki site, correct and refine the content.
- Next, they can push the content of "SemanticWiki2" on a push feed.
- Finally, administrator of "SemanticWiki1" can pull validated modifications from "SemanticWiki2" into "SemanticWiki1".

This scenario demonstrates how the DSMW extension can be used to implement processes. In this case, a simple process allows validation of changes produced by the FCA smart agent and avoids the problem of instability of human-machine collaboration.

4.3 Enriched wiki content

After validation, here is the content of the enriched wiki (SemanticWiki1 in figure 3) in the case study:

- `Category:Professor`, contains pages about Prof. Smith and Prof. Jones;

[1] `http://dsmw.org`

- `Category:Topic`, contains pages about Networks, Artificial Intelligence and Software Engineering;
- `Category:Course`;
- `Category:Master 1 Course`, a subcategory of `Category:Course`;
- `Category:Master 2 Course`, a subcategory of `Category:Course`;
- `Category:Artificial Intelligence Course`, a subcategory of `Category:Course`, the page indicates that Prof. Smith is teaching all the courses in this category;
- `Category:Prof. Jones' Course`, a subcategory of `Category:Course`;
- `Category:Master 1 Artificial Intelligence Course`, a subcategory of `Category:Master 1 Course` and `Category:Artificial Intelligence Course`, contains the page about Knowledge Discovery;
- `Category:Master 2 Artificial Intelligence Course`, a subcategory of `Category:Master 2 Course` and `Category:Artificial Intelligence Course`, contains the page about Semantic Wiki;
- `Category:Master 1 and 2 Artificial Intelligence Course`, a subcategory of `Category:Master 1 Artificial Intelligence Course` and `Category:Master 2 Artificial Intelligence Course`, contains the page about Semantic Web;
- `Category:Networks Course`, a subcategory of `Category:Prof. Jones' Course`;
- `Category:Software Engineering Course`, a subcategory of `Category:Prof. Jones' Course` and `Category:Master 1 Course`, contains the page about Design Patterns;
- `Category:Master 1 Networks Course`, a subcategory of `Category:Master 1 Course` and `Category:Networks Course`, contains the page about Network Administration;
- `Category:Master 2 Networks Course`, a subcategory of `Category:Master 2 Course` and `Category:Networks Course`, contains the page about IPv6 Protocol.

5 Conclusion and future work

Semantic wikis allow users to build knowledge understandable by humans and computers. By this way, they also allow machines to produce or update semantic wiki pages as humans can do. This opens the opportunity to consider machines as new member of communities to produce and maintain knowledge. Consequently, such "smart agents" can reduce significantly the overhead of communities in the process of continuously knowledge building and correct humans errors.

In this paper, we proposed a new smart agent based on Formal Concept Analysis. This smart agent allows to reorganize the wiki: new categories are computed and pages are placed into these new categories. This allows a better organization of the content and facilitate the navigation in the wiki.

The refactoring process needs to be validated by human users. Consistency of the wiki is ensured by the use of DSMW: a second wiki site is used to store the result of the smart agent and is pulled back to the main wiki after human validation.

This paper presented an early work, and more research have to be done in the future. Clearly, if applied on a real wiki, a method such as FCA would produce a large amount of concepts, and it would by impossible for human users to validate any one of them. Some filtering methods should be used to prevent irrelevant categories to be added, based on the number of instances in a category or other criteria.

Using Relational Concept Analysis instead of FCA should provide interesting results. Other clustering methods will also be considered.

In the current version of our method, human users have a feedback from the smart agent, they will take into consideration the new categories that have been created. However, the smart agent does not have a feedback from the human users: if a category has been rejected during the validation process, the smart agent will create it again when the process will be reiterated. To avoid this problem, the smart agent has to be "history-aware" and use the information of the modification by human users during the validation process.

6 Acknowledgments

This research was part of the CyWiki project, funded by the Université Henri Poincaré of Nancy.

References

1. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic wikipedia. In: WWW '06: Proceedings of the 15th international conference on World Wide Web. (2006) 585–594
2. Schaffert, S.: IkeWiki: A semantic wiki for collaborative knowledge management. In: 1st International Workshop on Semantic Technologies in Collaborative Applications (STICA06), Manchester, UK. (2006)
3. Buffa, M., Ereteo, G., Faron-Zucker, C., Gandon, F., Sander, P.: SweetWiki: A semantic wiki. Journal of Web Semantics, special issue on Web 2.0 and the Semantic Web **6**(1) (2008)
4. Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic wikipedia. Journal of Web Semantic **5**(4) (2007) 251–261
5. Cordier, A., Lieber, J., Molli, P., Nauer, E., Skaf-Molli, H., Toussaint, Y.: Wikitaaable: A semantic wiki as a blackboard for a textual case-based reasoning system. In: 4th Workshop on Semantic Wikis (SemWiki2009), held in the 6th European Semantic Web Conference. (2009) 18–32
6. Ganter, B., Wille, R.: Formal Concept Analysis, Mathematical Foundation. Springer (1999)
7. Rahhal, C., Skaf-Molli, H., Molli, P., Weiss, S.: Multi-synchronous collaborative semantic wikis. In: 10th International Conference on Web Information Systems-Wise 2009. Volume 5802 of Lecture Notes in Computer Science. (2009)