Mining for Reengineering: an Application to Semantic Wikis using Formal and Relational Concept Analysis

Lian Shi, Yannick Toussaint, Amedeo Napoli, Alexandre Blansché

Mining for Reengineering: an Application to Semantic Wikis using Formal and Relational Concept Analysis

Lian Shi[1], Yannick Toussaint[1], Amedeo Napoli[1], and Alexandre Blansché[2]

[1] LORIA CNRS – INRIA Nancy Grand-Est – Nancy Université, équipe Orpailleur, BP 70239, F-54506 Vandœuvre-lès-Nancy
{firstname.lastname}@loria.fr
[2] Laboratoire LITA, Université Paul Verlaine, Île du Saulcy F-57000 Metz
alexandre.blansche@univ-metz.fr

Abstract. Semantic wikis enable collaboration between human agents for creating knowledge systems. In this way, data embedded in semantic wikis can be mined and the resulting knowledge patterns can be reused to extend and improve the structure of wikis. This paper proposes a method for guiding the reengineering and improving the structure of a semantic wiki. This method suggests the creation of categories and relations between categories using Formal Concept Analysis (FCA) and Relational Concept Analysis (RCA). FCA allows the design of a concept lattice while RCA provides relational attributes completing the content of formal concepts. The originality of the approach is to consider the wiki content from FCA and RCA points of view and to extract knowledge units from this content allowing a factorization and a reengineering of the wiki structure. This method is general and does not depend on any domain and can be generalized to every kind of semantic wiki. Examples are studied throughout the paper and experiments show the substantial results.

1 Introduction

Wikis provide friendly environments to create, modify and update a website, where different topics or pages are linked by hyperlinks, forming a large page network [9]. In the trend of semantic web, taking advantage of knowledge representation and ontology technologies, semantic wikis extend the capabilities of wikis by allowing annotations attached to elements in a wiki page [6]. Annotations refer to the introduction of a category for "typing" a page or a relation between an element of the page and another element in another page. Knowledge units in semantic wikis are usually represented within RDF Schema and OWL constructions, and can be queried on the fly using SPARQL for example. Therefore, semantic wikis enable communities to collaboratively produce both a large set of textual documents that human agents can easily browse thanks to semantic links and a formalized knowledge base readable by software agents. Moreover, a

semantic wiki can be considered as a wide blackboard where human agents interact with software agents [3] for producing and completing knowledge. However, the collaborative and multi-user aspects introduce different perceptions of a domain and thus differences in knowledge organization. The incremental building over the time may also introduce lacks or over-definitions in the knowledge base.

Accordingly, learning techniques can be used to solve these kinds of problems by reengineering, i.e. a semantic wiki is considered as a base for discovering and organizing existing and new knowledge units. Furthermore, semantic data embedded in the semantic wikis are rarely used to enrich and improve the quality of semantic wikis themselves. There is a large body of potential knowledge units hidden in the content of a wiki, and knowledge discovery techniques are candidate for making explicit these units. The objective of the present work is to use knowledge discovery techniques –based on Formal Concept Analysis and Relational Concept Analysis– for learning new knowledge units such as categories and links between objects in pages for enriching the content and the organization of a semantic wiki. Thus, the present work aims at reengineering a semantic wiki for ensuring a well-founded description and organization of domain objects and categories, as well as setting relations between objects at the most appropriate level of description.

Reengineering improves the quality of a semantic wiki by allowing stability and optimal factorization of the category hierarchy, by identifying similar categories, by creating new categories, and by detecting inaccuracy or omissions. The knowledge discovery techniques used in this reengineering approach are based on Formal Concept Analysis (FCA) [5] and Relational Concept Analysis (RCA) [10]. FCA is a mathematical approach for designing a concept lattice from a binary context composed of a set of objects described by attributes. RCA extends FCA by taking into account relations between objects and introducing relational attributes within formal concepts, i.e. reifying relations between objects at the concept level. FCA and RCA are powerful techniques that allow a user to answer a set of questions related to the quality of organization and content of semantic wiki contents. The originality of this approach is to consider the semantic wiki content as a starting point for knowledge discovery and reengineering, applying FCA and RCA for extracting knowledge units from this content and allowing a completion and a factorization of the wiki structure (a first attempt in this direction can be found in [1]). The present approach is general and does not either depend on any domain or require customized rules or queries, thus can be generalized to every semantic wikis.

After defining some terminology about semantic wikis in Section 2, we introduce basics elements on Formal and Relational Concept Analysis and in Section 3. Section 4 gives details on the proposed approach for reengineering a semantic wiki. In Section 5, we propose an evaluation of the method based on experimental results, and we discuss the results and related issues (Section 6). After a brief review of related work, we conclude with a summary in Section 7.


```
[[Category:Film]]This movie was directed by [[directedBy::Chris Columbus]]
and produced by [[producedBy::David Heyman]] and [[producedBy::J. K.
Rowling]]. It was starred by [[Starring::Daniel Radcliffe]]
and [[Starring::Emma Watson]]. As a [[hasGenre::Fantasy]] movie, it's loved
by most of Children.
```

Fig. 1. A wiki page titled "Harry Potter and the Philosopher's Stone". The upper half shows the content of the wiki page while the lower half is the underlying annotated form.

2 Problem Setting and Wiki Terminology

Throughout this paper, we use *wiki*(*s*) to refer to *semantic wiki*(*s*). Each wiki page is considered as a "source ontological element", including classes and properties [8]. Annotations in the page provide statements about this source element. For example, the page entitled "Harry Potter and the Philosopher's Stone" describes a movie and a set of annotations attached to this page (Figure 1).

Editors annotate an object represented by a wiki page with categories, data types, and relations, i.e. an object can be linked to other objects through relations. A category allows a user to classify pages and categories can be organized into a hierarchy. For example, the annotation `[[Category:Film]]` states that the page about "Harry Potter and the Philosopher's Stone" belongs to the category `Film`. The category `Fantasy` is a subcategory of `Film` as soon as the annotation `[[Category:Film]]` is inserted in the `Fantasy` page. Binary relationships are introduced between pages. For example, the annotation `[[Directed_by::Chris Colombus]]` is inserted in the "Harry Potter..." page for making explicit the `Directed_by` relation between "Harry Potter..." page and "Chris Colombus" page.

Some attributes are allowed to assign "values": they specify a relationship from a page to a data type such as numbers. Then `[[Duration::152min]]` give the the duration of the film "Harry Potter...".

Basically, all categories in a wiki are manually created by various editors, possibly introducing several sources of inconsistency and redundancy. The fact that the number of pages is continuously growing and that new categories are introduced is a major challenge for managing the category hierarchy construction. For keeping efficient the browsing and navigation within the wiki, the category hierarchy has to be periodically updated. Thus, a tool for automatically managing the category hierarchy in a wiki is of first importance. In the following, we

show how this can be done using FCA and RCA, which are both detailed in the next section.

3 Introducing Formal Concept Analysis (FCA) and Relational Concept Analysis (RCA)

3.1 Formal Concept Analysis

The basics of FCA are introduced in [5]. Data are encoded in a formal context $\mathcal{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$, i.e. a binary table where $\mathtt{G}$ is a set of objects, $\mathtt{M}$ a set of attributes, and $\mathtt{I} \subseteq \mathtt{G} \times \mathtt{M}$ an incidence relation. Two derivation operators, both denoted by $'$, formalize the sharing of attributes for objects, and, in a dual way, the sharing of objects for attributes:

$' : \wp(\mathtt{G}) \longrightarrow \wp(\mathtt{M})$ with $\mathtt{X}' = \{\mathtt{m} \in \mathtt{M} \mid \forall \mathtt{g} \in \mathtt{X}, \mathtt{gIm}\}$

$' : \wp(\mathtt{M}) \longrightarrow \wp(\mathtt{G})$ with $\mathtt{Y}' = \{\mathtt{g} \in \mathtt{G} \mid \forall \mathtt{m} \in \mathtt{Y}, \mathtt{gIm}\}$

$\wp(\mathtt{G})$ and $\wp(\mathtt{M})$ respectively denote the powersets of $\mathtt{G}$ and $\mathtt{M}$; $\mathtt{gIm}$ states that object $\mathtt{g} \in \mathtt{G}$ is owning attribute $\mathtt{m} \in \mathtt{M}$. The two derivation operators $'$ form a *Galois connection* between $\wp(\mathtt{G})$ and $\wp(\mathtt{M})$. Maximal sets of objects related to maximal set of attributes correspond to closed sets of the composition of both operators $'$ (denoted $''$), for $\wp(\mathtt{G})$ and $\wp(\mathtt{M})$ respectively. A pair $(\mathtt{X}, \mathtt{Y}) \in \wp(\mathtt{G}) \times \wp(\mathtt{M})$, where $\mathtt{X} = \mathtt{Y}'$ and $\mathtt{Y} = \mathtt{X}'$, is a *formal concept*, $\mathtt{X}$ being the *extent* and $\mathtt{Y}$ being the *intent* of the concept. The set $\mathcal{C}_\mathcal{K}$ of all concepts from $\mathcal{K}$ is ordered by extent inclusion, denoted by $\leq_\mathcal{K}$. Then, $\mathcal{L}_\mathcal{K} = \langle \mathcal{C}_\mathcal{K}, \leq_\mathcal{K} \rangle$ forms the *concept lattice* of $\mathcal{K}$.

For example, let us consider the two formal contexts $\mathcal{K}_{Films}$ and $\mathcal{K}_{Actors}$ given in Table 1. The context on the left provides descriptions for films while the context on the right is for actors. Each film or actor is introduced by its name and has a set of attributes. The two corresponding lattices, $\mathcal{L}_{InitFilms}$ and $\mathcal{L}_{InitActors}$ are given on Figure 2 and Figure 3.

A reduced labelling is used in the drawing of lattice: attributes are inherited from high level to low levels while objects are shared form low levels to high

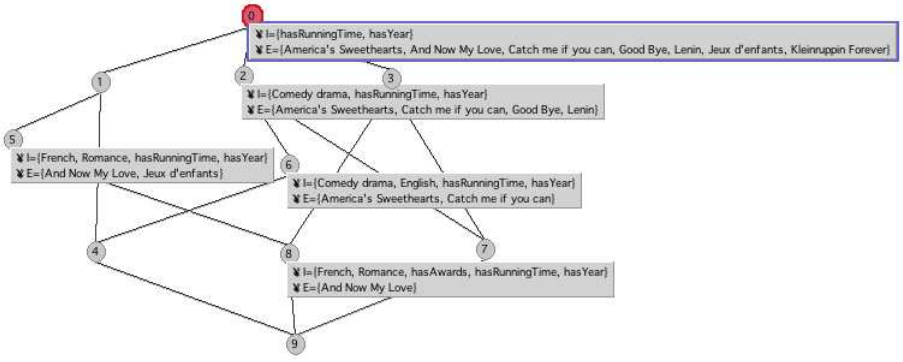

Fig. 2. The initial lattice of films $\mathcal{L}_{InitFilms}$.

	French	English	Germany	Romance	ComedyDrama	hasAwards	hasRunningTime	hasYear	American
Jeux d'enfants	x			x			x	x	
Good Bye, Lenin			x		x	x	x	x	
Catch me if you can		x			x		x	x	x
And now my love	x			x		x	x	x	
America's Sweethearts		x		x	x		x	x	x
Kleinruppin Forever				x			x	x	

	hasAwards	Female	Male	Age20	Age30
Guillaume Canet	x		x		x
Daniel Brühl	x		x		x
Leonardo DiCaprio	x		x	x	
Marthe Keller		x			
Tolias Schenke		x			
Julia Roberts	x	x			
Catherine Zeta Jones		x			
Anna Maria Muhe		x		x	

Table 1. The two binary contexts of films $\mathcal{K}_{Films}$ (left) and actors $\mathcal{K}_{Actors}$ (right).

levels in a lattice. For example, concept `c5` in $\mathcal{L}_{InitActors}$ has for intent the set of attributes {`Female, hasAward`} (respectively from `c1` and `c3`). By contrast, concept `c3` in $\mathcal{L}_{InitActors}$ has for extent the set of individuals {`Julia Roberts, Leonardo DiCaprio, Daniel Brül, Guillaume Canet`} (respectively from `c5`, `c6`, and `c9`). When attributes are mentioned following reduced labelling, they will be said *local attributes*, otherwise *inherited attributes*. Attributes obey the following rules: when there are at least two local attributes $\mathtt{a}_1$ and $\mathtt{a}_2$ in the same intent, these attributes are equivalent, *i.e.* $\mathtt{a}_1$ appears as soon as $\mathtt{a}_2$ and reciprocally. For example, `hasRunninigTime` and `hasYear` are equivalent in $\mathcal{L}_{InitFilms}$ (see Figure 2). Moreover, local attributes imply inherited attributes. For example, `ComedyDrama` implies `hasRunninigTime` and `hasYear`.

3.2 Relational Concept Analysis

RCA was introduced and detailed in [10]. Data is described by a *relational context family* (RCF), composed of a set of contexts $\mathbf{K} = \{\mathcal{K}_\mathtt{i}\}$ and a set of binary relations $\mathbf{R} = \{\mathtt{r_k}\}$. A relation $\mathtt{r_k} \subseteq \mathtt{G_j} \times \mathtt{G}_\ell$ connects two object sets, a *domain* $\mathtt{G_j}$, i.e. $\mathtt{dom}(\mathtt{r_k}) = \mathtt{G_j}$, and a *range* $\mathtt{G}_\ell$, i.e. $\mathtt{ran}(\mathtt{r_k}) = \mathtt{G}_\ell$. For example, the RCF corresponding to the current example is composed of the contexts $\mathcal{K}_{Films}$ and $\mathcal{K}_{Actors}$. The context $\mathcal{K}_{Starring}$ represents the relation `Starring` between films and actors (a film is starring an actor).

Hereafter, we briefly introduce the mechanism of RCA necessary for understanding the following (other details are given in [10]). RCA is based on a *relational scaling* mechanism that transforms a relation $\mathtt{r_k}$ into a set of *re-*

	Guillaume	Daniel Brühl	Leonardo DiCaprio	Marthe Keller	Tolias Schenke	Julia Roberts	Catherine Zeta Jones	Anna Maria Muhe
Jeux d'enfants	x							x
Good Bye, Lenin		x						
Catch me if you can			x					
And now my love				x		x		
America's Sweethearts						x	x	
Kleinruppin Forever				x	x			

Table 2. The context $\mathcal{K}_{Starring}$ of the relation `Starring` between films and actors (films and actors are objects in the contexts $\mathcal{K}_{Films}$ and $\mathcal{K}_{Actors}$).

lational attributes that are added to complete the "initial context" describing the object set $\mathtt{G_j} = \mathtt{dom}(\mathtt{r_k})$. For each relation $\mathtt{r_k} \subseteq \mathtt{G_j} \times \mathtt{G}_\ell$, there is an *initial lattice* for each object set, i.e. $\mathcal{L}_\mathtt{j}$ for $\mathtt{G_j}$ and $\mathcal{L}_\ell$ for $\mathtt{G}_\ell$. For example, the two initial lattices for the relation `Starring` are $\mathcal{L}_{InitFilms}$ (Figure 2) and $\mathcal{L}_{InitActors}$ (Figure 3).

Given the relation $\mathtt{r_k} \subseteq \mathtt{G_j} \times \mathtt{G}_\ell$, the RCA mechanism starts from two initial lattices, $\mathcal{L}_\mathtt{j}$ and $\mathcal{L}_\ell$, and builds a series of intermediate lattices by gradually completing the initial context $\mathtt{G_j} = \mathtt{dom}(\mathtt{r_k})$ with new "relational attributes". For that, relational scaling follows the DL semantics of role restrictions. Given the relation $\mathtt{r_k} \subseteq \mathtt{G_j} \times \mathtt{G}_\ell$, a relational attribute $\exists \mathtt{r_k} : \mathtt{c}$ –$\mathtt{c}$ being a concept and $\exists$ the existential quantifier– is associated to an object $\mathtt{g} \in \mathtt{G_j}$ whenever $\mathtt{r_k}(\mathtt{g}) \cap \mathtt{extent}(\mathtt{c}) \neq \emptyset$ (other quantifiers are available, see [10]). For example, let us consider the concept `c1` whose intent is $\mathtt{Starring} : \mathtt{c3}$ in $\mathcal{L}_{FinalFilms}$, i.e. the final lattice of films on Figure 4. This means that all films in the extent of `c1` are related to (at least one or more) actors in the extent of concept `c3` in $\mathcal{L}_{FinalActors}$, i.e. the final lattice of actors, through the relation `Starring` (actors in the extent of `c3` are characterized by the `hasAward` attribute).

The series of intermediate lattices converges toward a "fixpoint" or "final lattice" and the RCA mechanism is terminated. This is why there is one initial and one final lattice for each context of the considered RCF. Here, $\mathcal{L}_{InitActors}$ is identical to $\mathcal{L}_{FinalActors}$ (Figure 3), and there are two different lattices for films, namely the initial $\mathcal{L}_{InitFilms}$ (Figure 2) and the final $\mathcal{L}_{FinalFilms}$ (Figure 4).

4 Methodology

In this section, we give details on the knowledge discovery approach used for wiki reengineering. We first explain how data are retrieved and how the different formal contexts and associated concept lattices are built. Then, we analyze the concepts and propose a new organization for the category hierarchy in the wiki.

4.1 Construction of the Relational Context Family

A complete RDF data export of a given wiki can be obtained by crawling the individual RDF export of each wiki page. According to the schema defined by Semantic MediaWiki[1] (SMW), a category is defined as a `owl:Class` and the object described in a page is exported as an instance defined by SWIVT ontology [7], which provides a basis for interpreting the semantic data exported by SMW.

The construction of a relational context family from RDF data export is based on the following correspondence. Objects in SMW are considered as objects in FCA, categories and datatype attributes in SMW are considered as attributes in FCA, and finally relations in SMW are considered as relations between objects in RCA.

We conduct SPARQL queries on the RDF dump to obtain a set of objects represented by pages (`O`), a set of categories (`C`), a set of datatype attributes (`A`)

[1] `http://ontoworld.org/wiki/SMW`

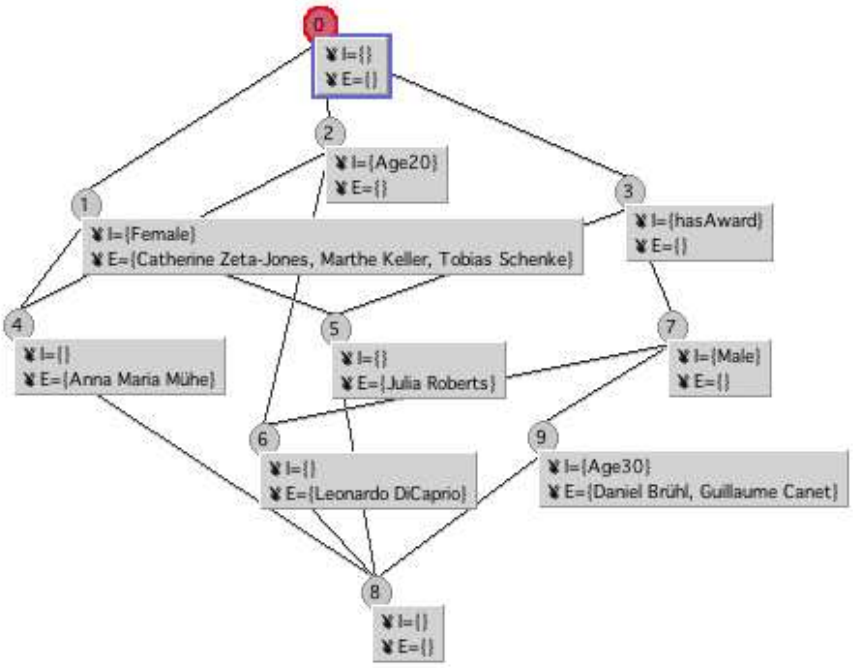

Fig. 3. The initial lattice of actors $\mathcal{L}_{InitActors}$.

and a set of relations (R). Unlike the previous example in Section 3, here each RCF has only one set of objects G (i.e. all objects represented by wiki pages) and each relation $\mathtt{r_i} \in \mathtt{R}$ ($\mathtt{0} \leq \mathtt{i} \leq \mathtt{n}$) is defined on $\mathtt{G} \times \mathtt{G}$, where

- G consists of all objects obtained from O.
- M is defined as $\mathtt{M} = \mathtt{C} \cup \mathtt{A}$.
- $\mathtt{I} \subseteq \mathtt{G} \times \mathtt{M}$. A pair $(\mathtt{g}, \mathtt{m}) \in \mathtt{I}$ iff $\mathtt{m} \in \mathtt{C}$ and object g belongs to this category, or $\mathtt{m} \in \mathtt{A}$ and g is annotated by this attribute. Additionally, the transitivity has to be explicitly stated in the context, i.e. $(\mathtt{g}, \mathtt{m}') \in \mathtt{I}$ iff $\mathtt{m}' \in \mathtt{C}$, $(\mathtt{g}, \mathtt{m}) \in \mathtt{I}$ and $\mathtt{m} \subseteq \mathtt{m}'$.
- n is the number of relations from R.

By doing this, abstraction will be maximized and objects will be classified into formal concepts without any prior knowledge but RDF data,

4.2 Analyzing Formal Concepts of the Concept Lattice

Based on the two binary contexts $\mathcal{K}_{Films}$ and $\mathcal{K}_{Actors}$, and on the relational context family $\mathcal{K}_{Starring}$, we obtain a relational lattice of films shown in Figure 4. In this lattice, the intent of each concept can be divided into a set of categories, a

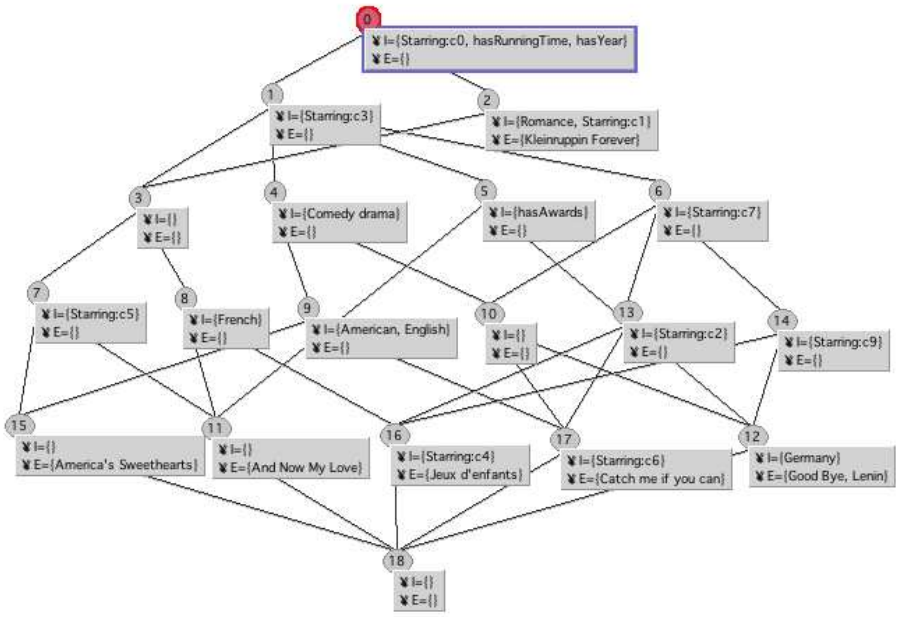

Fig. 4. The final relational lattice of films $\mathcal{L}_{FinalFilms}$.

set of attributes and a set of relational attributes. The analysis of formal concepts is driven by the following questions:

Question 1: Identifying equivalence between categories. Actually, categories appear as attributes and local attributes in a formal concept are equivalent [5]. For instance, the intent of concept `c9` in the lattice $\mathcal{L}_{FinalFilms}$ makes categories `American` and `English` equivalent, meaning that American and English movies are English speaking movies. This kind of redundancy is often due to the fact that a user may introduce a new category into the wiki without being well informed of existing ones.

Question 2: Creating new categories. A formal concept with no categories in its intent means that there is no category in the wiki for characterizing the set of objects in its extent. Therefore, a new category may be defined to capture this concept. For instance, in the lattice $\mathcal{L}_{FinalFilms}$, concept `c5` has the attribute `hasAward` that can be used for defining a new category, say "`Awarded`", which can classify objects having awards. Similarly, concept `c6` in lattice $\mathcal{L}_{FinalFilms}$ could be associated to a new category "movies starring male actors", because it has no category in its local intent but has the relational attribute `Starring:c7` and concept `c7` in the lattice $\mathcal{L}_{InitActors}$ has category `Male` as its local intent.

Question 3: Detecting category subsumption. Subsumption relations in the lattice infer subsumptions between existing categories and discovered categories. As a result, more optimized hierarchies can be defined and the organization of the wiki can be improved. In the lattice $\mathcal{L}_{FinalFilms}$, we assume that categories `English` and `American` are subcategories of `ComedyDrama` by observing concepts `c4` and `c9`. This sounds strange but is mainly due to the Closed-World Assumption of RCA, which collides with the openness of a semantic wiki and the reduced size of our running example (see Section 6).

Question 4: Defining categories. Definitions are quite rare in SMW despite they are essential. Nevertheless, definitions can help humans understanding of the purposes of categories and can be used for automatic classification by introducing necessary and sufficient conditions for an individual to belong to a category. As seen in question 1, elements in the local intent are substantially equivalent. Therefore, if a formal concept contains a category and one or more attributes in its local intent, then these attributes can be considered as a definition of that category. Moreover, any new introduced object to that category should be assigned these attributes. The case of equivalence between a category and a relational attribute is similar. For instance, concept `c2` has the category `RomanceMovie` in its intent. This category can be defined by the relational attribute `Starring:c1` where the intent of `c1` is `Female` (see lattice $\mathcal{L}_{InitActors}$). Then a romance movie would involve a female actor, and any new object in this concept should be related to at least one actress.

The result of all these is an RDF model that defines an OWL ontology containing both a TBox (new class definitions, subsumptions and equivalences) and an ABox (new instantiations). The final step is to contribute back with new knowledge to the original wiki. Our method acts as a wiki user suggesting updates. These changes, as any change from any other user, can be undone. Even if all the results are correct and consistent, some of them may be useless in practice. Therefore, in this approach it is the responsibility of a human expert or the wiki community to evaluate the reengineering proposals following the spirit of collaborative editing work for wikis.

5 Experimental Results

We applied our method to several semantic wikis and defined criteria for evaluating the experimental results.

5.1 From Wikis to Lattices

RDF dumps from a number of semantic wikis were obtained using an *ad hoc* crawler. Seven wikis were selected for the experiments due to their representative characteristics, as summarized in Table 3. The selected wikis[2] include Bioclipse, Hackerspace, Open Geospatial Encyclopedia, Referata, Sharing Buttons,

[2] Retrieved by Dec 2, 2010, from `http://wiki.bioclipse.net/`, `http://hackerspaces.be/`, `http://referata.com/`, `http://geospatial.referata.com/`,

Semantic Wiki	AP	CP	UCP	CAT	SUBCAT	DP	OP	CATSIZE	DPS/CP	OPS/CF
Bioclipse	573	373	220	74	9	3	17	4.49	0.12	0.73
Hackerspace	821	564	312	11	0	0	14	26.00	0.00	2.56
Open Geospatial	131	120	14	4	0	0	8	26.50	0.00	4.97
Referata	316	155	121	13	0	0	48	2.85	0.00	0.88
Sharing Buttons	121	54	3	2	0	7	7	25.50	4.46	3.98
TLC	371	331	23	38	14	25	9	10.16	3.56	0.89
Virtual Skipper	820	226	43	109	106	41	7	4.17	2.42	0.28

Table 3. Wiki characteristics in terms of the total number of pages (AP), the number of content pages (CP), the number of uncategorized content pages (UCP), the number of categories (CAT), the number of subsumptions (SUBCAT), the number of datatype attributes (DP), the number of relations (OP), the average cardinality of categories (CATSIZE), the average number of datatype attributes in content pages (DPS/CP) and the average number of relations in content pages (OPS/CF)

Toyota Land Cruiser (TLC) and Virtual Skipper (VSK). The characteristics obtained by querying the RDF dumps using SPARQL differ from the statistics contained in the `Special:Statistics` page. These divergences were caused by the slightly different definitions for "content page" and the incompleteness of the RDF exports of some wikis.

Some of these wikis have a dense categorization (e.g., VSK has 109 categories for 226 concepts content pages, then roughly a 2:1 ratio of content pages to categories), while others are subject to inexistent or lightweight category hierarchies (e.g., Hackerspace ratio is 50:1). Ideally, we would expect to categorize some of the uncategorized pages listed under the UCP column in the table. Unfortunately, this is hampered by the fact that almost all of these unclassified pages are also lacking attributes or relations with other pages (for instance, in the Bioclipse wiki, only one out of 220 has attributes). Therefore, FCA/RCA is unable to discover categories because of inadequate information.

All objects, wiki categories, datatype attributes and relations were obtained by using Jena[3]. A custom Java script transformed each RDF model into an RCF described in XML. All lattices were produced by the Java-based Galicia[4] platform and exported to XML documents.

5.2 Results

Table 4 shows the topological characteristics of the lattices of all wikis. The number of formal concepts defines the size of the lattice. Apparently, this number is not always proportional to the size of the wiki. For instance, in spite of Bioclipse wiki being smaller than Hackerspace wiki in terms of pages, the lattice

`http://www.sharingbuttons.org/`, `http://tlcwiki.com/` and `http://vsk.wikia.com/`, respectively.

[3] `http://jena.sourceforge.net/`

[4] `http://sourceforge.net/projects/galicia/`

Semantic Wiki	Concepts	Edges	Depth	Connectivity	Width
Bioclipse	170	319	8	1.88	21.25
Hackerspace	21	34	5	1.62	4.20
Open Geospatial	67	120	5	1.79	13.40
Referata	24	139	4	1.63	6.00
Sharing Buttons	70	130	6	1.86	11.67
TLC	520	1059	9	2.04	57.78
Virtual Skipper	148	294	12	1.99	12.33

Table 4. Characteristics of the computed lattices.

of Bioclipse has more concepts than Hackperspace one. In the lattice, each edge represents a subsumption relationship between concepts. Moreover, the *depth* of the lattice is defined by the longest path from the top concept down to the bottom concept, knowing that there are no cycles. The higher it is, the deeper the concept hierarchy is.

The connectivity of the lattice is defined as the average number of edges per concept. It is noteworthy that all lattices have a similar connectivity in a narrow range between 1.62 and 2.05. It seems that the characteristics of the wikis do not have a strong influence in the connectedness of the lattice. Finally, the last column gives the average number of concepts per level in the lattice. This value indicates the *width* of the lattice and it correlates to the size of the lattice. Consequently, the *shape* of a lattice is determined by both its depth and width.

Galicia produces XML files that represent the lattices as graphs, where concepts are labeled with their intent and extent. Using another custom Java application, we interpret these files and transform them into OWL/RDF graphs. Specifically, our application processes all concepts, and:

- generates `owl:equivalentClass` assertions for categories appearing as elements in the same intent of a concept,
- generates `rdf:type` assertions to express a membership of instances in the extent of a concept to the categories in the intent of the concept,
- generates `rdfs:subClassOf` assertions to subsumption relations between concepts,
- introduces new `owl:Class`es if the intent of a concept does not contain any category, and defines a new category by using the attributes in the intent.

After substracting the model given by FCA/RCA and removing trivial new categories, an RDF model is generated that contains concrete findings for reengineering the wiki. In Table 5 we report some metrics about the findings. It should be noticed that, in the case of VSK, the number of equivalences between originally existing categories (CAT-EQ) rises quickly due to the combinatorial effect and the symmetry of the equivalence (i.e. $A \equiv B$ and $B \equiv A$ count for two entries). Although some content pages are classified (CAT-CP), the lattice fails

Semantic Wiki	CAT-CP	SUB-CAT'	CAT-EQ	NEW-CAT-NT
Bioclipse	781	194	8	73
Hackerspace	597	19	2	5
Open Geospatial	85	29	0	25
Referata	20	19	0	5
Sharing Buttons	156	26	0	18
TLC	1521	375	0	191
Virtual Skipper	181	161	58	72

Table 5. The result of analyzing the lattices in terms of the number of new proposed memberships of content pages to categories (CAT-CP), the number of proposed sub-categorizations (SUB-CAT'), the number of category equivalences between originally existing categories (CAT-EQ) and the number of proposed non-trivial categories (NEW-CAT-NT)

to classify originally uncategorized pages. The prime reason is that these pages often lack any attribute that can be used to derive a categorization.

Figure 5 compares the category size histogram before and after the reengineering of VSK wiki. The shaded area amounts for the number of new discovered categories. The histogram clearly shows that most of the discovered categories are *small* in terms of their sizes.

The number of discovered subsumption relationships (SUB-CAT') seems to be more related to the number of discovered new categories than to the number of pre-existing ones. This indicates that in general the new categories are refinements of other ones; in other words, they have a "place" in the hierarchy.

Two of the studied wikis (Hackerspace and Referata) lead to only a few new categories. By looking into these two wikis, we found that they are already well organized and therefore provide less opportunities for reengineering, combined with the fact that these wikis do not use datatype properties.

Finally, we provide some of the generated findings to illustrate the kind of knowledge that can be obtained. For instance, in TLC wiki, the category `Has_specialty-is-Service+garage` is among the 191 proposed ones (the name is automatically generated and probably is not optimal). This new category is defined in the resulting OWL model with the Description Logic (DL) formula: $\texttt{Vendor} \sqcap \exists \texttt{has\_specialty}.\{\texttt{Service\_garage}\}$. Semantically, it can be interpreted as a subcategory of `Vendor` which aggregates those that have this precise specialty (actually, 89 out of 109 vendors).

Subsumption relations are also discovered among pre-existing categories. For instance, in the Bioclipse wiki, a rearrangement of categories into a hierarchy is proposed, with subsumptions such as $\texttt{Repositories\_maintained\_by\_Stefan\_Kuhn} \sqsubseteq \texttt{Repository}$. For some reason, this obvious subsumption link was not introduced by the wiki editors, and can be reconstructed afterwards by our method.

The discovered category equivalences and new restrictions based on attributes lead to definitions for the existing categories. Consider the discovered equiv-

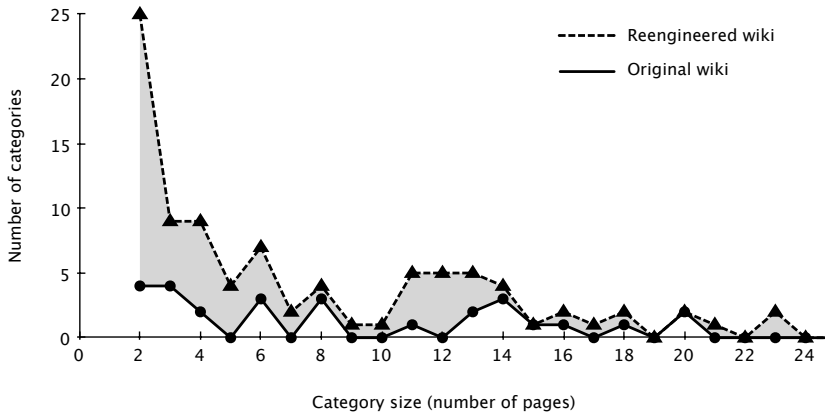

Fig. 5. Category size histogram of VSK wiki before and after FCA/RCA. Shaded area represents new proposed categories.

alence regarding TLC wiki: `Diesel_Engines` $\equiv \exists$`fuel`.{`diesel`}. In the original wiki, only `Diesel_Engines` $\sqsubseteq$ `Engine` is present. Therefore, the combination of both formulae provides a precise definition of the existing category, i.e. `Diesel_Engines` $\equiv$ `Engine` $\sqcap \exists$`fuel`.{`diesel`}.

6 Discussion

The experimental results show that our proposed method conduces to reengineering proposals that can go beyond what is obtained by DL-reasoning and querying on the original wiki. It is noteworthy to say that we do not compute the closure of our resulting model, which would produce an enlargement of the values in Table 5 but with little practical effect on the quality of the feedback provided to the wiki.

The method is suitable for any semantic wiki regardless of its topic or language. However, the computations are not linear with the size of the wiki (in terms of the number of pages). Precisely, the maximum size of the lattice is $2^{min(|G|,|M|)}$ with respect to FCA and $2^{min(|G|,2*|G|)}$ with respect to RCA. Therefore, processing large wikis can be a computational challenge.

FCA/RCA operates under the Closed-World Assumption (CWA), which diverges from the Open-World Assumption (OWA) of OWL reasoning. More importantly, CWA collides with the open nature of wikis. As a consequence, some of the results are counter-intuitive when they are translated back to the wiki, as it was exemplified in the previous section. However, the results are always consistent with the current data in the wiki, and the method can be repeated over time if the wiki changes. Recall that the process is "semi-automatic" and that an analysis is required.

A feedback analysis remains to be done. An approach for such an analysis is to provide results to human experts (e.g., wiki editors), who may evaluate the quality of the results based on their knowledge and experience. The quality can be measured in terms of correctness and usefulness. The latter will produce a subjective indication of the "insights" of the results, i.e., how much they go beyond the "trivial" and "irrelevant" proposals for reengineering.

7 Related Works and Conclusion

Krötzsch and co-authors [6] underlined that semantic wikis manage to disseminate semantic technologies among a broad audience and many of the emerging semantic wikis resemble small semantic webs. The knowledge model of a semantic wiki often corresponds to a small fragment of OWL-DL, but this fragment differs from one wiki to another as illustrated with SMW and IkeWiki [11]. Some extensions of SMW (like Halo Extension[5]) enable to introduce the domain and the range of a relation that are needed by FCA/RCA in order to abstract properties and relations between objects at the category level.

There exist several attempts to combine DL-based knowledge representation and FCA. One of the main works is the thesis of B. Sertkaya (see an overview in [12]). The objective is to use FCA for building a conceptualization of the world that could be encoded in an expressive DL. An "extended bottom-up approach" (computing Least Common Subsumers, Good Common Subsumers) is proposed for a bottom-up construction of a knowledge base, even including an existing knowledge base. However, SMW does not provide either disjunction or negation. Furthermore, concept lattices are mapped to the so called $\mathcal{FL}$ description logic in [10]. Then, the FCA/RCA combination provides substantial capabilities for working with wikis.

Among several domains of experiment, reengineering or refactoring UML models [4] is quite similar to the purpose of the present paper, i.e. wiki reengineering. The goal was to build abstractions in UML models.

Chernov et al. [2] defined a method for introducing semantics in "non-semantic" wikis. They attempt to define relations between categories looking at links between individuals. The method is essentially based on statistics: observing the number of links between pages in categories and computing Connectivity Ratio in order to suggest semantic connections between categories. Although there is no measure involved in the their approach, we are currently working at using numerical measures to deal with noise or omissions in the data.

Contrasting the previous work, the approach presented in [1] uses FCA on the semantic information embedded in wikis, however, authors did not distinguish attributes, relations and categories. In the present work, we go beyond by distinguishing semantical elements, in particular, datatype attributes, relations and categories, and we complete the work of FCA with the work of RCA on relations, giving better results on the reengineering of wikis.

[5] The Halo Project: `http://semanticweb.org/wiki/Project_Halo`

In this paper, we proposed an approach for reengineering semantic wikis based on Formal and Relational Concept Analysis. Our approach alleviates the human effort required to detect category redundancy, discover potential categories and identify membership between pages and category, subsumption and equivalence between categories. These objectives are achieved by analyzing formal concepts of lattices built on the semantic data contained in wikis. We argue that the use of FCA and RCA helps to build a well-organized category hierarchy. Our experiments show that the proposed method is adaptable and effective for reengineering semantic wikis. Moreover, our findings pose several open problems for future study.

References

1. A. Blansché, H. Skaf-Molli, P. Molli, and A. Napoli. Human-machine collaboration for enriching semantic wikis using formal concept analysis. In C. Lange, J. Reutelshoefer, S. Schaffert, and H. Skaf-Molli, editors, *Fifth Workshop on Semantic Wikis – Linking Data and People (SemWiki-2010)*. CEUR Workshop Proceedings Vol-632, 2010.
2. S. Chernov, T. Iofciu, W. Nejdl, , and X. Zhou. Extracting semantic relationships between wikipedia categories. In *1st International Workshop SemWiki2006 - From Wiki to Semantics, co-located with the ESWC 2006*, Budva, 2006.
3. Amélie Cordier, Jean Lieber, Pascal Molli, Emmanuel Nauer, Hala Skaf-Molli, and Yannick Toussaint. 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*, May 2009.
4. Michel Dao, Marianne Huchard, Mohamed Rouane Hacene, Cyril Roume, and Petko Valtchev. Improving generalization level in uml models iterative cross generalization in practice. In *ICCS*, pages 346–360, 2004.
5. B. Ganter and R. Wille. *Formal Concept Analysis*. Springer, Berlin, 1999.
6. M. Krötzsch, S. Schaffert, and D. Vrandečić. Reasoning in semantic wikis. In *Reasoning web 2007*, volume 4636 of *Lecture Notes in Computer Science*, pages 310 – 329. Springer, 2007.
7. M. Krözsch and D. Vrandečić. Swivt ontology specification. `http://semantic-mediawiki.org/swivt/`.
8. M. Krözsch, D. Vrandečić, M. Kolkel, H. Haller, and R. Studer. Semantic wikipedia. *J. Web Sem*, pages 251–261, 2007.
9. B. Leuf and W. Cunningham. *The Wiki Way: Quick Collaboration on the Web*. Addison-Wesley Longman, 2001.
10. M. Rouane-Hacene, M. Huchard, A. Napoli, and P. Valtchev. A proposal for combining formal concept analysis and description logics for mining relational data. In S.O. Kuznetsov and S. Schmidt, editors, *Proceedings of ICFCA 2007*, Lecture Notes in Artificial Intelligence 4390, pages 51–65. Springer, Berlin, 2007.
11. Sebastian Schaffert. Ikewiki: A semantic wiki for collaborative knowledge management. In *1st International Workshop on Semantic Technologies in Collaborative Applications (STICA'06)*, Manchester, UK, 2006.
12. B. Sertkaya. *Formal Concept Analysis Methods for Descriptions Logics*. PhD thesis, Dresden university, 2008.