Ontological Analysis of the Wikipedia Category System

Alexander Kirillovich¹ and Olga Nevzorova¹,²

¹Kazan Federal University, Kazan, Russia
²Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia

Keywords: Wikipedia, Categorization, Ontology Validation, OntoClean, Ontological Dependence.

Abstract: We analyse violations of the transitivity principle of the Wikipedia category system, i.e. the situations where articles from a subcategory doesn't logically belong to its parent category. The causes of the violation have been analysed on the base of ontological modelling methodologies such as OntoClean. We propose a new approach to automatically eliminating the violations. This approach is based on analysis of the relation of ontological dependence between categories. As a theoretical foundation of such analysis we propose a new deflationistic interpretation of the essential account of ontological dependence. The proof of concept has been evaluated on the category C:Mathematics. We are going to apply the proposed approach to derive a new large-scale domains hierarchy from the Wikipedia category system, and use it to provide BabelNet and DBpedia with fine-grained domain annotations.

1 INTRODUCTION

Wikipedia is one of the largest knowledge bases on the Web. Wikipedia data are used in such tasks as word sense disambiguation, text categorization, calculation of semantic similarity, and machine translation. To automatically process information from Wikipedia, means of data structuring are needed.

The category system is the main mean for structuring information in Wikipedia. The categories specify the thematic classification of articles.

There are two types of categories:

- Set categories, for example C:Cities category that contains articles about various cities (New York, Moscow, Kazan, Seville, etc).
- Topic categories, for example C:City category that contains articles on city-related topics (Urban planning, Urbanization, History of cities, Urban culture, etc).

Each category can include subcategories and can be included in a parent category. Thus the category system is organized as an directed acyclic graph.

Categories can be grouped using meta-categories, for example C:Writers → C:Writers_by_nationality → C:Russian_writers.

Thus it is important that the category system demonstrate transitivity and nested subcategories be relevant to the parent category. However, building a chain of nested categories does not always satisfy this requirement. This fact causes problems while searching for articles. Therefore it is necessary to analyze the causes of violations of transitivity which leads to a discrepancy between the category and the article that is located in one of the nested subcategories.

We understand relevance as rating a category-subcategory pair in terms of preserving the volume of common and distinctive features. We assume that for relevant categories, the volume of common attributes exceeds the volume of differences, and in this sense one can speak of the transitivity property of categories.

However, the requirement of transitivity is often violated. For example, the category C:Arithmetic contains irrelevant subcategories such as C: Colombianspeople_executed_by_firing_squad (see example 1).

Example 1: C:Arithmetic → C:Ratios → C:Rates → C:Temporal_rates → C:Acceleration → C:Force → C:Motion_(physics) → C:Flight → C:Ballistics → ... → C:Projectile_weapons → C:Firearms → C:People_associated_with_firearms → C:Shooting_victims → ... → C: Colombianspeople_executed_by_firing_squad.
The objective of this paper is to analyze the problem of transitivity violations in the Wikipedia category system and to propose an approach to solving it.

The paper is organized as follows. In Section 2, we briefly enumerate other projects of structuring Wikipedia data and note its limitations. In Section 3, we analyze Wikipedia category system on base of ontological modeling methodologies and reveal causes of transitivity violations. In Section 4, we propose an approach that corrects the category system by removing intransitive chains. Section 5 describes the direction of future work.

2 RELATED WORKS

There are many projects for extracting structured data from Wikipedia (Medelyan et al., 2009; Hovy et al., 2013), such as DBpedia (Auer et al., 2007; Bizer et al., 2009; Lehmann et al., 2015), YAGO (Suchanek et al., 2007; Hoffart et al., 2013; Mahdisoltani et al., 2015), WikiTaxonomy (Ponzetto and Strube, 2007; Ponzetto and Strube, 2011; Zrn et al., 2008), WikiNet (Nastase and Strube, 2008; Nastase et al., 2010; Nastase and Strube, 2013), ORA (Gangemi et al., 2012; Nuzzolese et al., 2013), WiBi (Flati et al., 2014; Flati et al., 2016), MENTA (de Melo and Weikum, 2010), BabelNet (Navigli and Ponzetto, 2010; Navigli and Ponzetto, 2012), WikiNet (Moro and Navigli, 2012), KOG (Wu and Weld, 2008), as well as projects for alignment of Wikipedia and WordNet (Ruiz-Casado et al., 2005; Toral et al., 2008; Niemann and Gurevych, 2011; Ponzetto and Navigli, 2009; Gella et al., 2014; Titze et al., 2014).

However, these projects cannot completely substitute the Wikipedia category system in the case of subject retrieval, i.e. retrieval of articles on the particular subject (for example, all the articles related to Ancient Greece).

3 ANALYSIS OF TRANSITIVITY VIOLATIONS

We have analyzed the category system using the methodology of ontological modeling and have identified the causes of transitivity violations.

The category system can be considered as a thesaurus (ANSI, 2005; ISO 2011; Loukachevitch, 2011). In this case the categories will correspond to thesaurus concepts, and the relationships between the category and its subcategory will be viewed as standard ontological relationships.

Here are examples of such relations:

- **Generic relation:**
  - C:Cities_in_Europe → C:Capital_in_Europe;
  - C:Software → ... → C:Operating_systems;
  - C:Mathematical_axioms → C:Proof_of_Godel’s_incompleteness_theorem;
  - C:Machines → C:Engines;
  - C:Wars → ... → C:Wars_involving_the_Soviet_Union;
  - C:Fiction_books → ... → C:Dystopian_novels.

- **Instance relation:**
  - C:Capital_in_Europe → C:Moscow;
  - C:Intergovernmental_organizations → C:United_Nations;
  - C:Universities_and Colleges_in_Connecticut → C:Yale_University;
  - C:Operating_systems → C:Unix;
  - C:Fields_of_mathematics → C:Algebra;
  - C:Axioms_of_set_theory → C:Axiom_of_choice;
  - C:Abstract_strategy_games → C:Chess;
  - C:Engines → C:Internal_combustion_engine;
  - C:Wars_involving_the_Soviet_Union → C:World_War_II;
  - C:Dystopian_novels → C:Nineteen_Eighty-Four;
  - C:Organs → C:Brain;
  - C:Space_stations → C:International_Space_Station.

- **Part-whole relation:**
  - C:Moscow → C:Cities_and_towns_under_jurisdiction_of_Moscow → C:Zelenograd;
  - C:Yale_University → C:Yale_University_Library;
  - C:United_Nations → C:International_Atomic_Energy_Agency;
  - C:World_War_II → ... → C:Attack_on_Pearl_Harbor;
  - C:Central_nervous_system → C:Brain;
  - C:Unix → C:Network_socket;
Errors due to the application of ambiguous categories:
- Record charts is included in the Diagrams category: C:Record_charts → C:Charts: In one case Charts is interpreted as diagrams and in other as music charts.
- Ship construction is included in the Real estate category: C:Ship_construction: In one case Libraries is interpreted as buildings type, and in another as the social institute.

Using a concept in different senses:
- Digital library Lib.ru is included in the Buildings:
  C:Lib.ru → C:Digital_libraries → C:Libru
- Nihilism is included in the Biology:
  C:Nihilism → C:Philosophy_of_life: In one case Nihilism is interpreted as biological process, and in another as a social process.
- Snow category is included in the Liquids:
  C:Snow → C:Forms_of_water → C:Water: In one case Water is interpreted as a chemical substance.
as such, and in another as a chemical substance in liquid state.

- Incompatible identity criteria:
  - The Kaaba (a Muslim sacred building) is included in the Mathematical objects:
    - C:Mathematical_objects
    - C:Geometry
    - C:Elementary_shapes
    - C:Cubes
    - C:Cubic_buildings
    - C:Kaaba. The error is located in the chain: C:Cubes → C:Cubic_buildings → C:Kaaba.
  - Incompatible identity criteria:
    - C:Shooting_victims
    - C:People_associated_with_firearms
    - C:Ballistics
    - C:Flight
    - C:Force
    - C:Temporal_rates
    - C:Arithmetic
    - C:Occupational_safety_and_health
    - C:Labour_relations
    - C:Toxicology
    - C:Biological_weapons
    - C:Illegal_drugs
    - C:Morphine
    - C:Analgesic. The reason for the error is that the Illegal drugs is not a type. It is a role, and it should not contain type categories.
  - Incompatible identity criteria:
    - C:Optical_bombshights
    - C:Office_equipment
    - C:Computers
    - C:Analog_computers
    - C:Office_equipment
    - C:Optical_bombshights.
  - Poison berry Sambucus is included in the Foods category:
    - C:Foods → C:Berries → C:Sambucus.

In the following cases, transitivity is violated not because of an error, but because of the very principle of organization of the Wikipedia categories system:

- Nontransitivity of the class-instance relation:
  - The Santa Maria ship is included in the Ship types category:
    - C:Ship_types
    - C:Exploration_ships
    - C:Santa_Maria_(ship).
  - The ‘Ode’ poem is included in the Literary genres category:
    - C:Literary_genres
    - C:Poetry → ... → Ode_(poem). This case is complicated by the real lexical homonymy (the ode as a literary genre).

- Nontransitivity of associative relation:
  - Zelenograd, a city under Moscow jurisdiction is included in the Capitals in Europe category:
    - C:Capitals_in_Europe
    - C:Moscow → ... → C:Zelenograd.
  - Blondi the dog is included in the Nazi leaders category:
    - C:Nazi_leaders
    - C:Nazi_leaders
    - C:Adolf_Hitler → Blondi.
  - Fictional Galactic Empire is included in the Northern American countries category:
    - C:Northern_American_countries
    - C:United_States
    - C:American_people → ... → C:George_Lucas → C:Star_Wars → Galactic_Empire_(Star_Wars).
  - Languages of Djibouti is included in the Statistics category:
    - C:Statistics
    - C:Statistical_data_sets
    - C:Demographics_by_country → C:Demographics_of_Djibouti → C:Languages_of_Djibouti.
  - Biological weapons is included in the Labour law category:
    - C:Labour_law
    - C:Labour_relations
    - C:Occupational_safety_and_health
    - C:Toxicology → C:Biological_weapons.
  - Colombian people executed by firing squad is included in the Arithmetic category:
    - C:Arithmetic
    - C:Ratios
    - C:Rates → C:Temporal_rates → C:Acceleration
    - C:Force → C:Motion_(physics)
    - C:Flight → C:Ballistics → ... → C:Projectile_weapons → C:Firearms → C:People_associated_with_firearms → C:Shooting_victims.
Thus, the transitivity violation in the Wikipedia category system is caused by two groups of reasons. The first group includes the reasons related to the violation of rules for constructing a hierarchy of concepts in an ontology. These violations can be eliminated by the authors of Wikipedia. The second group includes the reasons related to the very principle of organizing the system of Wikipedia categories, the main one of which is the nontransitivity of the associative relationship.

4 APPROACH TO ELIMINATION OF TRANSITIVITY VIOLATIONS

In this section, we propose a method for elimination of intransitive category-subcategory chains.

4.1 The Basic Idea

As we have demonstrated, one of the main causes of transitivity violations is the associative relation, which is not transitive.

Up-to-date methods of extracting transitive hierarchical structure from the category system (such as YAGO and WikiTaxonomy) detect associative relationships between categories and eliminate them all. A disadvantage of these methods is elimination of potentially relevant relationships, that don’t violate transitivity. As a result, there is a need for a method, that eliminates associative relationships violating transitivity (e.g. Statistics → Demography), but keeps relationships not violating transitivity (e.g. Education → Teacher).

A proposed method is based on approach to establishing associative relationships in RuThes thesaurus (Loukachevitch, 2011; Loukachevitch and Dobrov, 2014; Loukachevitch et al., 2014; Loukachevitch and Dobrov, 2004a; Loukachevitch and Dobrov, 2004b). According to this approach associative relationship between two concepts doesn’t violate transitivity if ontological dependence relation holds between these concepts.

In RuThes the notion of ontological dependence (Tahko and Lowe, 2016; Correia, 2008; Koslicki, 2012, Koslicki, 2013) is formalized according to modal-existential account: an object A is ontologically depends on an object B iff necessary if A exists then B exists (Simons, 1987, chap. 8, pp. 290–323; Thomasson, 1999, chap. 2, pp. 24–34).

The modal-existential account has several advantages, including simplicity and mathematical rigorous. A disadvantage is that its application requires a human participation. In this regard, it is not appropriate for the given problem.

Additionally, modal-existential account has been criticized on pure ontological grounds. Kit Fine has demonstrated, that this account is very rough approximation to the notion of ontological dependence and has counterexamples. As an alternative, Fine proposed an essential account. According to this account, A depends on B iff B is a constituent of the essence of A. Essence of object is defined as a collection of propositions that are true in virtue of the identity of this object. These propositions, in turn, constitute real definition of the object (Fine, 1994; Fine, 1995).

The notion to be true in virtue of the identity of calls for clarification. According to Fine’s own interpretation, based on meta-ontological realism, this notion designates unanalyzed relation between an object and a proposition. We propose another interpretation, based on deflationist approach of Amie L. Thomasson. According to Thomasson, existence and identity criteria for an object A are application/co-application conditions for the term “A” (Thomasson, 2008; Thomasson, 2009). According to our interpretation, a proposition is true in virtue of identity of the object A if it is a part of application/co-application conditions for the term “A”. And, accordingly, an object A is ontologically depended on an object B iff application/co-application conditions for the term “A” contain application/co-application conditions for term “B”. Our interpretation demonstrates that the essential account is independent from meta-ontological assumptions.

So, employing the essential account we obtain the following criteria of ontological dependence: A ontologically depends on B iff B is ineliminably involved in the definition of A. This criterion suits better for automatic application.

As an approximation of a definition of an object, represented by any given Wikipedia category, we take the lead section of the main article of this category as well as of other language versions of this article. Involvement of an object in the definition of another object is roughly approximated by existence of a hyperlink between the two definitions.

4.2 Proof of the Concept

The proof of concept of the proposed method has been realized as follows:

- Check whether the relationship between the cate-
gory and its subcategory is associative. Assume that the relationship is associative if one of its participants is a topic category. The type of the category is detected by the method from WikiTaxonomy project.

- If the relationship is associative, then check whether the ontological dependence relation exists between the corresponding concepts on the base of the criterion described above. If the ontological dependence holds, then keep the relationship between the categories, and eliminate it otherwise.
- If the relationship is not associative, but taxonomic, then refer to YAGO that contains refined taxonomic relationships. If the corresponding relationship is contained in YAGO, then keep it, and eliminate it otherwise.

4.3 Evaluation

The proof of concept has been evaluated with category C:Mathematics. The choice of this category was motivated by the practical task of interlinking OntoMathPro ontology (Nevzorova et al., 2014; Elizarov et al., 2014; Elizarov et al., 2016; Elizarov et al., 2017) with DBpedia.

For the task of evaluation we recursively formed a list of subcategories of C:Mathematics category. Then we applied our method to eliminate supposedly irrelevant subcategories. The lists of the eliminated and the kept categories were analyzed manually. The assessor’s task was to check whether the kept categories were relevant, and the eliminated categories were irrelevant. The assessment result is represented in Table 1.

Table 1: The result of the preliminary assessment of the proposed method with C:Mathematics.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>4281</td>
</tr>
<tr>
<td><strong>True positives</strong></td>
<td>2136</td>
</tr>
<tr>
<td><strong>True negatives</strong></td>
<td>650</td>
</tr>
<tr>
<td><strong>False positives</strong></td>
<td>1010</td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td>485</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.814956</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>0.678957</td>
</tr>
<tr>
<td><strong>F1 score</strong></td>
<td>0.740766</td>
</tr>
</tbody>
</table>

5 CONCLUSION

We analyzed the causes for transitivity violations in the Wikipedia category system and proposed an approach to their elimination.

The main contributions of this paper are:

1. an ontological analysis of the Wikipedia category system;
2. a new approach to eliminating the violations of transitivity in this category system, based on analysis of the relation of ontological dependence between categories; and
3. a new deflationistic interpretation of the essential account of ontological dependence, obtained by application of the meta-ontological framework of Amie Thomasson to the ontological dependence account of Kit Fine.

However, our work is on early stage and will be continued in the following directions:

1. **Ontological Analysis of the Wikipedia Category System.** We are going to conduct an ontological analysis in a more systematic way. In particular, it is supposed to find a numerical distribution of the category-subcategory relationships as well as of the causes for transitivity violations.

2. **Aboutness Analysis of the Wikipedia Category System.** We are going to complement our ontological analysis of the Wikipedia category system with an analysis based on the notion of aboutness. Aboutness is the relation between a meaningful item (such as a document or a proposition) and its subject matter. Subject matter is a central concept in Library and information science (Hjørland, 2016; Hjørland, 1992) and has been formalized in formal semantics and mathematical logic (Hawke, 2017). In our analysis, we would rely on the formalization of Yablo (2014), according to which subject matter is conceived as a partition of logical space and the part-whole relation between two subjects is defined as refinement relation between two partitions. With this formalisation in hands, we consider a category-subcategory chain as transitivity-preserving if subject matter, associated with the subcategory is a part of the subject matter, associated with the category.

Additionally, we are going to investigate correspondence between the notions of aboutness and ontological dependence. Our working hypothesis to be proved or disproved is that $A$ is ontologically depended on $B$ iff the subject matter associated with $A$ is a part of the subject matter associated with $B$.

3. **Formalization of the Deflationistic Interpretation of the Essential Account of Ontological Dependence.** In this paper we proposed a new deflationistic interpretation of the essential account of ontological dependence, where this
relation has been explained in terms of application/co-application conditions. However, the notion of application/co-application conditions itself requires clarification and formalization. We are going to formalize this notion on base of the situation semantics for relevant logic (Mares, 2004).

4. Refining of the proposed Approach. In particular, it is supposed to analyze other articles in the target category, as well as to use the context of the link within the definition.

5. Fine-grained Domain Annotation of DBpedia and BabelNet. We are going to apply the developed approach to derive a new large-scale domains hierarchy from the Wikipedia category system. In contrast to relatively small domains hierarchies of WordNet, BabelNet and DBpedia, the extracted hierarchy will contain a large number of fine-grained domains, such as Ancient Greece, Modal logic, Object-oriented programming or Star Wars.

This hierarchy will be integrated into RuThes Cloud, a multilevel multilingual resource of the Linguistic Linked Open Data (Kirlilovich et al., 2017; Galieva et al., 2017), and can be naturally applied to carry out fine-grained domain annotation of the resources, automatically extracted from Wikipedia, first of all, DBpedia and BabelNet.

ACKNOWLEDGEMENTS

This work was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, grant agreement no. 1.2368.2017, and by the subsidy of the Russian Government to support the Program of Competitive Growth of Kazan Federal University.

REFERENCES


