Understanding How Users Edit Ontologies: Comparing Hypotheses About Four Real-World Projects

Simon Walk[1(✉)], Philipp Singer[2], Lisette Espín Noboa[2], Tania Tudorache[3], Mark A. Musen[3], and Markus Strohmaier[2,4]

[1] Graz University of Technology, Graz, Austria
simon.walk@tugraz.at
[2] GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
{philipp.singer,Lisette.Noboa}@gesis.org
[3] Stanford Center for Biomedical Informatics Research, Stanford, CA, USA
{Tudorache,Musen}@stanford.edu
[4] University of Koblenz-Landau, Mainz, Germany
Markus.Strohmaier@gesis.org, Strohmaier@uni-koblenz.de

Abstract. Ontologies are complex intellectual artifacts and creating them requires significant expertise and effort. While existing ontology-editing tools and methodologies propose ways of building ontologies in a normative way, empirical investigations of how experts *actually* construct ontologies "in the wild" are rare. Yet, understanding actual user behavior can play an important role in the design of effective tool support. Although previous empirical investigations have produced a series of interesting insights, they were exploratory in nature and aimed at gauging the problem space only. In this work, we aim to advance the state of knowledge in this domain by systematically defining and comparing a set of hypotheses about how users edit ontologies. Towards that end, we study the user editing trails of four real-world ontology-engineering projects. Using a coherent research framework, called HypTrails, we derive formal definitions of hypotheses from the literature, and systematically compare them with each other. Our findings suggest that the *hierarchical structure* of an ontology exercises the strongest influence on user editing behavior, followed by the *entity similarity*, and the *semantic distance* of classes in the ontology. Moreover, these findings are strikingly consistent across all ontology-engineering projects in our study, with only minor exceptions for one of the smaller datasets. We believe that our results are important for ontology tools builders and for project managers, who can potentially leverage this information to create user interfaces and processes that better support the observed editing patterns of users.

1 Introduction

Large real-world ontologies are intellectual artifacts that are inherently complex and hard to build. Most such ontologies are found in the biomedical domain. For

M. Arenas et al. (Eds.): ISWC 2015, Part I, LNCS 9366, pp. 551–568, 2015.
DOI: 10.1007/978-3-319-25007-6_32

example, SNOMED-CT,[1] a comprehensive clinical health terminology, has over 300,000 classes, the National Cancer Institute Thesaurus (NCIT)[2] has more than 100,000 classes, and the 11th revision of the International Classification of Diseases (ICD-11)[3] has over 50,000 classes. The development of such large ontologies usually takes place in distributed teams, and requires a significant effort both in the ontological modeling and coordination of the entire process.

One of the biggest challenges in developing large real-world ontologies is proper tool support. While existing ontology-editing tools and methodologies prescribe certain ways of building ontologies, there is very little research on how users actually use these tools. Empirical analyses of how users develop ontologies "in the wild" are very rare. We address this gap with this paper, by aiming to broaden our understanding of editing behaviors in large ontology-engineering projects. It is the ultimate vision of our work to lay a more solid foundation for creating tools that better support ontology authors based on their actual authoring behavior.

We define a *sequential edit trail* as a chronologically sorted list of all actions a user takes while editing an ontology. We derive such editing trails from the change logs recorded by the ontology-editing tools. In previous work, we have conducted exploratory empirical analyses of various types of edit trails in several ontology-engineering projects [21,22], and we have discussed our findings and potential implications [23]. In these works, we have been able to explore different editing patterns and potential explanations via manual inspection and qualitative interpretation. For example, we have speculated that users edit ontologies in a top-down fashion or that users navigate along similar concepts. However, it is still unclear how such hypotheses can best be expressed formally, or how they can be systematically compared with each other in order to explain the production of edit trails, and hence an ontology, at hand.

Thus, in this paper, we systematically investigate previous, mostly exploratory, results using HypTrails [11]—a generic methodology for comparing hypotheses about human trails in ontology-engineering projects. This allows us to (i) formally define, (ii) systematically study, and (iii) rank different hypotheses about ontology-editing behavior within a coherent research framework. By using HypTrails, we approach this problem by modeling edit trails as first-order Markov chains (see Section 3.2) and hypotheses as priors. From our analyses, we find that the *hierarchical structure* of an ontology exercises the strongest influence on observed user behaviors, followed by the *similarity* of entities, and the *distance* of classes in the ontology. These findings are strikingly consistent across the four real-world ontology-engineering projects used in our study, with only minor exceptions for one of the smaller datasets. We believe that our results are important for ontology tools builders and for project managers, who can potentially leverage this information to create user interfaces and processes that better support the observed editing patterns of users.

[1] http://www.ihtsdo.org/snomed-ct
[2] http://ncit.nci.nih.gov
[3] http://who.int/classifications/icd/revision/en/

The main research contributions of this work are:

- A formal way to define hypotheses about how users edit an ontology (e.g., top-down vs. bottom-up editing strategies).
- A detailed systematic comparison of such hypotheses across four real-world ontology-engineering projects.
- A ranking of all investigated hypotheses according to their relative plausibility for each dataset by adopting a coherent research approach.

The remainder of the paper is structured as follows: In Section 2, we discuss the related work. The methodology and datasets are described in Section 3, followed by a detailed formal description of all investigated hypotheses in Section 4. We present the results of our analysis in Section 5, discuss implications and limitations of our findings in Section 6 and conclude our work and discuss opportunities for future work in Section 7.

2 Related Work

The related work relevant for this paper is covered by two different research fields: *Human Trails on the Web* and *Analysis of Ontology Editing Behavior*.

2.1 Human Trails on the Web

Previous research has studied human trails on the Web in various settings. Modeling trails has received a lot of attention [3,12], as well as the detection of regularities, patterns and strategies in trails of interest [6,25]. Most prominently, researchers have focused on studying human navigational trails on the Web—capturing the subsequent websites that humans navigate to [6,12,25]. This research on navigational trails has inspired other works in the effort to improve the Web, e.g., better website design (usability) [4], identifying related links [18] or constructing an e-learning Semantic Web [2]. Researchers have also investigated other kinds of human trails, e.g., search trails [13,26], diffusion trails [1] or song listening trails [11]. Our work directly connects to these studies as we are interested in shedding more light on the production of human trails on the Web; however, in our case, we look at human edit trails in ontology-engineering projects by using the approach presented in [11].

2.2 Analysis of Ontology Editing Behavior

In this line of research, a large part of the literature has focused on analyzing the editing behavior or identifying editing patterns in collaborative ontology-engineering. To perform these types of analyses, researchers have used the change logs recorded by the different ontology-editing environments, similar to our approach.

Strohmaier et al. [14] conducted an empirical analysis to investigate the hidden social dynamics that take place when editors develop an ontology, and

provided new metrics to quantify various aspects of the engineering processes. Falconer et al. [5] did a change-log analysis of different ontology-engineering projects, showing that contributors exhibit specific roles, which can be used to group and classify these users. Pesquita and Couto [9] analyzed the influence of the location and specific structural features to determine if and where the next change will be conducted in the Gene Ontology[4]. The work by Wang et al. [24] presents an analysis of user editing patterns derived from change logs of several real-world ontology-engineering projects utilizing association-rule mining. The results suggest that users tend to edit in a vertical way, i.e., users edit the same properties for different classes in a sequential way. Rospocher et al [10] analyzed the change logs for two different Web-based collaborative ontology-editing tools and found similar collaboration and editing patterns. For example, they found that users tend to edit in the local neighborhood of an entity. Van Laere et al. [19] analyzed behavior-based user profiles in collaborative ontology-engineering projects using K-means clustering to group similar users.

In contrast to our previous research [21–23], this work represents a systematic and comparative study of different hypotheses in a coherent mathematical research framework, whereas our previous analyses have mostly been exploratory. We can thereby—for the first time—make relative, empirically grounded statements about the plausibility of different hypotheses given data.

3 Materials and Methodology

We present the four datasets used in our research (Section 3.1), and the HypTrails framework (Section 3.2) that forms the basis of the methodology used in this work.

3.1 Datasets

We used the change logs of four real-world ontology-engineering projects to conduct the analyses presented in this work. These projects use WebProtégé [17] as the editing platform, a Web-based generic ontology-editing tool, which records a log of all changes performed by each user. Each change record stores metadata about the change, such as the user who performed the change, a textual description of the change, the timestamp, and the entity on which the change occurred.

To extract the editing trails from the change logs, we performed a preprocessing step in which we merged consecutive changes on the same entity by the same user (i.e., *self-loops*) into one change. Such changes occurred when users would edit different properties of the same entity. For the purpose of this work, we have not been interested in such changes, but rather in the ones which occurred on different entities. Further, we have limited all our analyses on *isA* relationships and removed equivalence links. However, multiple *isA* inheritances have been kept "as-is". We provide a brief description of the four datasets used in our research below.

[4] http://www.geneontology.org

The International Classification of Diseases (ICD),[5] developed by the World Health Organization (WHO), is the international standard for diagnostic classification used to encode information relevant to epidemiology, health management, and clinical use in over one hundred United Nations countries. WHO regularly publishes new revisions of the classifications. The 11th revision of the classification, **ICD-11**,[6] is currently in progress, and is planned to be finalized in 2017. In contrast to previous revisions, ICD-11 is developed as a rich OWL ontology [16]. Over 100 domain experts are using a customized version of WebProtégé to author the ontology collaboratively.

The International Classification of Traditional Medicine (ICTM)[7] is a WHO-led project that aimed to produce an international standard terminology and classification for diagnoses and interventions in Traditional Medicine. ICTM was developed collaboratively as an OWL ontology with the goal to unify the knowledge from the traditional medicine practices from China, Japan and Korea. Its content is authored in 4 languages: English, Chinese, Japanese and Korean. More than 20 domain experts from the three countries developed ICTM using a customized version of WebProtégé. The development of ICTM ended in 2012.

The Biomedical Resource Ontology (BRO) [15] was developed as part of the Biositemaps project. Biositemaps is a mechanism for researchers working in biomedicine to publish metadata about biomedical data, tools, and services. Applications can then aggregate this information for tasks such as semantic search. BRO is the enabling technology used in Biositemaps; a controlled terminology for describing the resource types, areas of research, and activity of a biomedical related resource. A small group of editors authored BRO using WebProtégé to modify the ontology and to carry out discussions.

The Ontology for Parasite Lifecycle (OPL) models the life cycle of the *T.cruzi*, a protozoan parasite, which is responsible for a number of human diseases [8]. OPL uses expressive OWL (SHOIF) to represent its knowledge base, and extends several other OWL ontologies. Several users from different institutions collaborate on OPL development using WebProtégé as a collaborative platform.

Table 1 provides some characteristics about each of the datasets used in our analysis. The average trail length ranges from 1,637.13 transitions for ICD-11 to 136.60 transitions for BRO. Trails refer to the number of different human edit trails per dataset, where each trail represents a chronologically ordered list of all the classes a user has edited. Users with less than 2 distinct changes have been removed from our analysis.

[5] http://who.int/classifications/icd/en/
[6] http://who.int/classifications/icd/ICDRevision/
[7] http://who.int/mediacentre/news/notes/2010/trad_medicine_20101207/en/

Table 1. Characteristics of the four datasets.

	ICD-11	ICTM	BRO	OPL
Classes	48,771	1,506	528	393
Changes	439,229	67,522	2,507	1,993
Users	109	27	5	3
Trails	102	26	5	3
Avrg. trail length	1,637.13	673.54	136.60	152.00
Transitions	361,491	66,708	2,388	2,668
Self-Loops	194,504	49,196	1,705	2,212
First change	18.11.2009	02.02.2011	12.02.2010	09.06.2011
Last change	29.08.2013	17.7.2013	06.03.2010	23.09.2011
Period (ca.)	4 years	2.5 years	1 month	3 months

3.2 Methodology

By and large, HypTrails [11] is an approach that allows us to compare hypotheses about human trails. In our case, we are interested in studying: (i) the human edit trails in ontology-engineering projects, and (ii) the relative plausibility of hypotheses about the production of these trails that have been manifested in previous studies. In Section 1, we used the hypothesis that users edit ontologies in a top-down manner as an example. Using HypTrails, we are able to compare this hypothesis to other such hypotheses, and determine which one is more plausible to describe the production of the corresponding editing trails, and hence the ontology at hand. Section 4 provides a formal description of all hypotheses that we have compared as part of this research. Figure 1 shows a graphical representation of the editing patterns represented by each hypothesis. Next, we introduce the core concepts of HypTrails; for a more thorough introduction please refer to [11].

Technically, HypTrails models trails with first-order Markov chain models, and compares hypotheses using Bayesian inference, and more specifically, the *marginal likelihood* which can also be referred to as the *evidence* (we use both terms throughout this work synonymously). The marginal likelihood $P(H|D)$ describes the probability of a hypothesis H (e.g., uniform hypothesis) given the data (trails). For expressing generic hypotheses and being able to compare them, HypTrails uses the sensitivity of the marginal likelihood on the prior. Thus, hypotheses are expresses as different priors—in case of a Markov chain model the conjugate prior is the *Dirichlet distribution*. The hyperparameters of Dirichlet distributions can be interpreted as *pseudo counts*. Thus, simply put, higher pseudo counts refer to higher beliefs in corresponding transition for a given hypothesis.

Consequently, we have to provide HypTrails with matrices that capture our generic hypotheses and corresponding beliefs in transitions (see Section 4). Based on these matrices, HypTrails internally elicits proper Dirichlet priors for given hypotheses by setting the pseudo counts accordingly, based on a parameter k which steers the total number of pseudo counts assigned. Basically, the higher we

set k, the stronger we believe in a given hypothesis. Analogously, this means that with higher k, we expect to see less transitions contradicting the corresponding hypothesis (e.g., only transitions from higher level classes to lower level classes in the top-down hypothesis). For fairness, we always want to compare hypotheses with each other for the same values of k.

Finally, by using different priors for different hypotheses, we get different marginal likelihoods when combined with empirical trail data. Based on these evidences, we can compare the relative plausibility of hypotheses—higher evidences indicate higher plausibility. In theory, we need to further calculate *Bayes factors* [7] between the marginal likelihoods of two hypotheses, so that we would be able to judge the strength of the evidence for one hypothesis over the other. However, as all Bayes factors are decisive, we resort from presenting them individually throughout this paper. Thus, we can produce a partial ordering of hypotheses based on their relative plausibility by ranking their marginal likelihoods from largest to smallest for single values of k.

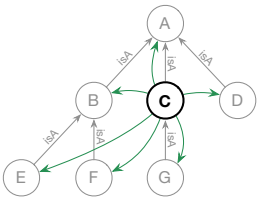
4 Hypotheses

HypTrails allows us to compare hypotheses about the production of human edit trails in ontology-engineering projects, and helps us to understand how an ontology is produced in an ontology-development tool. Hypotheses are *beliefs about transitions* (see Figures 1(a)–1(h)) opposed to actual empirical transitional observations (see Figure 1(i)). With HypTrails, we express these transitional beliefs as our assumptions about Markov chain transitions. In detail, we specify hypotheses as matrices that reflect our assumptions about transitions between states where higher values correspond to higher beliefs.

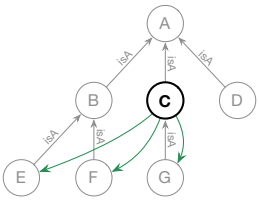
Thus, for each hypothesis, we need to specify the *hypothesis matrix* Q with elements $q_{i,j}$ that represent the belief in the transition between states s_i and s_j. A *state* corresponds to a class in the ontology that users are editing. A *transition* between states s_i and s_j corresponds to a two sequential user edit: first of the class represented by s_i, and then of the class represented by s_j. In order to express our hypotheses as beliefs in Markov transitions, and to have a better interpretation capability, we directly set $q_{i,j}$ as row probabilities $P(s_j|s_i)$. Thus, for each row i of Q it holds that $\sum_j q_{i,j} = 1$.

For example, Figure 1(e) depicts the *hierarchy-based hypothesis*, which postulates the belief that users are likelier to edit classes along the hierarchical (*isA*) structure of the ontology and the shortest distance. In this example, if a user has just previously changed class C, this hypothesis believes that the user is most likely to change class A (the *parent*) or G (the *child*) next. Classes B and D are both *siblings* (and two steps away) of C, which is why this hypothesis expresses a smaller belief in these transitions. Other hierarchical transitions, *ancestors*, *descendants* and *cousins*, follow analogously with less belief (i.e., lower proabability; not depicted in Figure 1(e)).

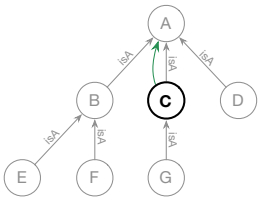
Figure 2 shows an exemplary illustration of the transition graph and the corresponding matrix for the *top-down* hypothesis, which believes that users



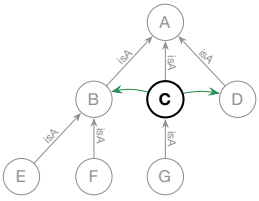
(a) Uniform hypothesis: all transitions are equally likely



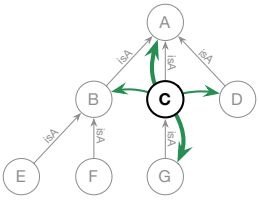
(b) Top-down hypothesis: transitions to lower classes are most likely



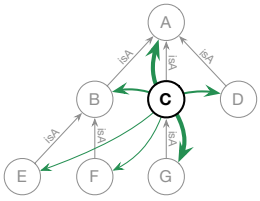
(c) Bottom-up hypothesis: transitions to higher classes are most likely



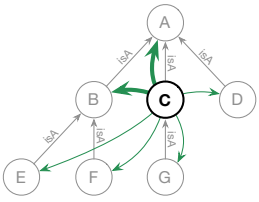
(d) Breadth-first hypothesis: transitions to same level classes are most likely



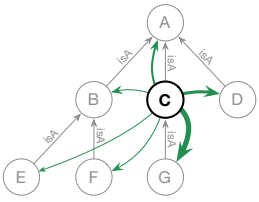
(e) Hierarchy hypothesis: transitions along hierarchical relations (parent, child, sibling, cousin) are most likely



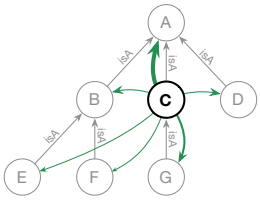
(f) Shortest path hypothesis: transitions to close classes are most likely



(g) Connectivity hypothesis: transitions to popular classes are most likely

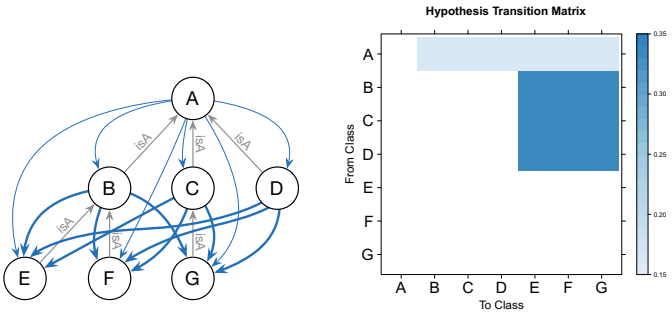


(h) Similarity hypothesis: transitions to similar (title and definition) classes are most likely



(i) Empirical transitions: obtained from real world data

Fig. 1. Sample-Hypotheses. This figure depicts eight hypotheses about how humans consecutively edit classes in ontology-engineering projects derived from our previous research (a-h), as well as empirical observations (i). The curved arrows represent transitions we believe in for a given hypothesis (a-h), or observed transition probabilities from data (i). The thicker an arrow, the higher our belief in the corresponding transition for a given hypothesis (a-h), or the higher the number of transitions we observed in the data (i). For simplicity, we always only visualize the transitions for class C; all other classes follow analogously.



(a) Top-down hypothesis: transitions to classes on lower levels are most likely.

(b) Hypothesis matrix: the transition matrix representing the top-down hypothesis.

Fig. 2. Top-down hypothesis. This figure depicts (a) the top-down hypothesis and (b) its corresponding hypothesis matrix Q that is generated from its formal definition. Darker transitions between classes represent a strong belief in these transitions, while white transitions represent a disbelief in a transition. Note that the matrix is normalized per row, hence the sum of all beliefs for each row is 1.

consecutively edit classes at deeper levels in the hierarchy. In this example, our state space consists of seven classes $S = \{A, B, C, D, E, F, G\}$. The beliefs in the transitions between states are shown in Figure 2(a). As this hypothesis has stronger beliefs in top-down transitions, the graph and matrix will only contain beliefs in transitions from higher-level classes to lower-level classes, such as, from C to E, F and G. Figure 2(b) shows the corresponding representation of the beliefs in the hypothesis matrix. For example, for the row corresponding to the transitions from class C, we may set $q_{C,E} = 1/3$, $q_{C,F} = 1/3$ and $q_{C,G} = 1/3$. For all other classes, we can proceed analogously.

In the remainder of this section, we thoroughly describe the hypotheses used in this research, and provide formal descriptions of how we built the corresponding hypothesis matrices Q. Note that for each hypothesis and equation, we always calculate $q_{i,j}$, for all i and j. We set the diagonal of each hypothesis matrix Q to 0 as we do not consider self-loops in our data. As it is not always possible to express our beliefs with direct probabilities, we additionally normalize each row of Q using the ℓ_1-norm.

Figure 1 shows a graphical representation of the hypotheses investigated in our research. The *top-down*, *bottom-up*, *breadth-first* and *hierarchy hypotheses* resulted as part of our prior research from a manual inspection of Markov chains of different orders [21–23]. Additionally, we are also considering the *shortest path*, *connectivity*, and *similarity* hypotheses to also investigate further "strategies" of how users edit an ontology that could provide plausible explanations for the underlying data.

Uniform Hypothesis. This hypothesis believes that each transition from one state to any other state is equally likely (cf. Figure 1(a)). Thus, it assumes that humans edit ontologies at random. We can see this hypothesis as a baseline. If other hypotheses are not more plausible than this uniform one, we cannot expect them to provide good explanations about the production of the trails (and the ontology) at hand. The elements of matrix Q for this hypothesis are defined as follows:

$$q_{i,j} = \frac{1}{|S-1|} \tag{1}$$

Top-down Hypothesis. For the top-down hypothesis, we express the belief that classes that are deeper in the hierarchy (further away from the root class) than the previously edited class, are likelier to be changed next. For expressing this hypothesis, we measure the depth level of each class (the distance to the root); classes deeper in the hierarchy have larger depth levels. In this hypothesis, we have stronger beliefs in transitions to classes that have a *larger depth level* than the current class (cf. Figure 1(b)). We express this hypothesis according to the following definition with $depth_i$ and $depth_j$ representing the depth-levels of the corresponding classes s_i and s_j.

$$q_{i,j} = \begin{cases} 1, & \text{if } depth_i < depth_j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Bottom-up Hypothesis. Analogously to the top-down hypothesis, this hypothesis believes that classes that are closer to the root class (i.e., they have lower depth levels) than the previously edited class, are likelier to be changed next (cf. Figure 1(c)).

$$q_{i,j} = \begin{cases} 1, & \text{if } depth_i > depth_j, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Breadth-first Hypothesis. Similar to the top-down and bottom-up hypotheses, we express the belief that classes are likelier to be changed next, if they are on the same depth levels (cf. Figure 1(d)).

$$q_{i,j} = \begin{cases} 1, & \text{if } depth_i = depth_j, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Shortest Path Hypothesis. With this hypothesis, we express the belief that users consecutively edit classes in an ontology that are close to each other in the class hierarchy (cf. Figure 1(f)). In detail, we look at the shortest path distances $d(i, j)$ between pairs of classes—the shorter the distance, the stronger we believe

in the corresponding transition. To invert the shortest path length, we subtract it from the diameter $\max_{x,y}(d(x,y))$ of the whole hierarchy.

$$q_{i,j} = \max_{x,y}(d(x,y)) - d(i,j) \tag{5}$$

Hierarchy Hypothesis. The hierarchy hypothesis represents our belief that users edit classes along the hierarchical structure of the ontology (i.e., *isA* links). In particular, the next edit operation is likelier to occur on close relatives than on relatives that are further away (cf. Figure 1(e)). This hypothesis has the following weight initialization of our belief matrix:

$$q_{i,j} = \begin{cases} 4, & \text{if } d(i,j) = 1 \text{ and } depth_i \neq depth_j, \\ 3, & \text{if } d(i,j) = 2 \text{ and } depth_i = depth_j \text{ and } check\_siblings(i,j) > 0, \\ 2, & \text{if } d(i,j) = 4 \text{ and } depth_i = depth_j \text{ and } check\_cousins(i,j) > 0, \\ 1, & \text{if } sp(i,j) = |depth_i - depth_j|, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Where $sp(i,j)$ is the shortest path between pairs (i,j). It holds that $check\_siblings(i,j) = |parents(i) \cap parents(j)|$ and $check\_cousins(i,j) = |grandparents(i) \cap grandparents(j)|$. Hence, both functions are larger than zero, if classes i and j share at least one parent or grandparent, respectively.

Connectivity Hypothesis. In this hypothesis, we believe that the next edit operation will likelier occur on a class that is better connected in the class hierarchy. We define the *connectivity level* of a class as the number of *isA* relationships a class has to and from other classes. We represent the connectivity level of class j as k_j. The higher the connectivity level of a class, the higher our belief in a given transition (cf. Figure 1(g)). Note that for this hypothesis, each row of Q is the same—it can be seen as a zero-order Markov chain hypothesis that is weighted by the connectivity of nodes.

$$q_{i,j} = k_j \tag{7}$$

Similarity Hypothesis. In this hypothesis, we believe that transitions between similar classes are likelier to occur than between less similar classes (cf. Figure 1(h)). To calculate the similarity between classes i and j, we first generate *tf-idf* vectors, v_i and v_j, consisting of the values of the annotation properties corresponding to the label of a class, and the textual definition. Using these *tf-idf* vectors, we compute the cosine similarity between classes.

$$q_{i,j} = cos\_sim(v_i, v_j) \tag{8}$$

$cos\_sim(v_i, v_j)$ is the cosine similarity between the *tf-idf* vectors of the property values corresponding to the labels and textual definitions of classes i and j.

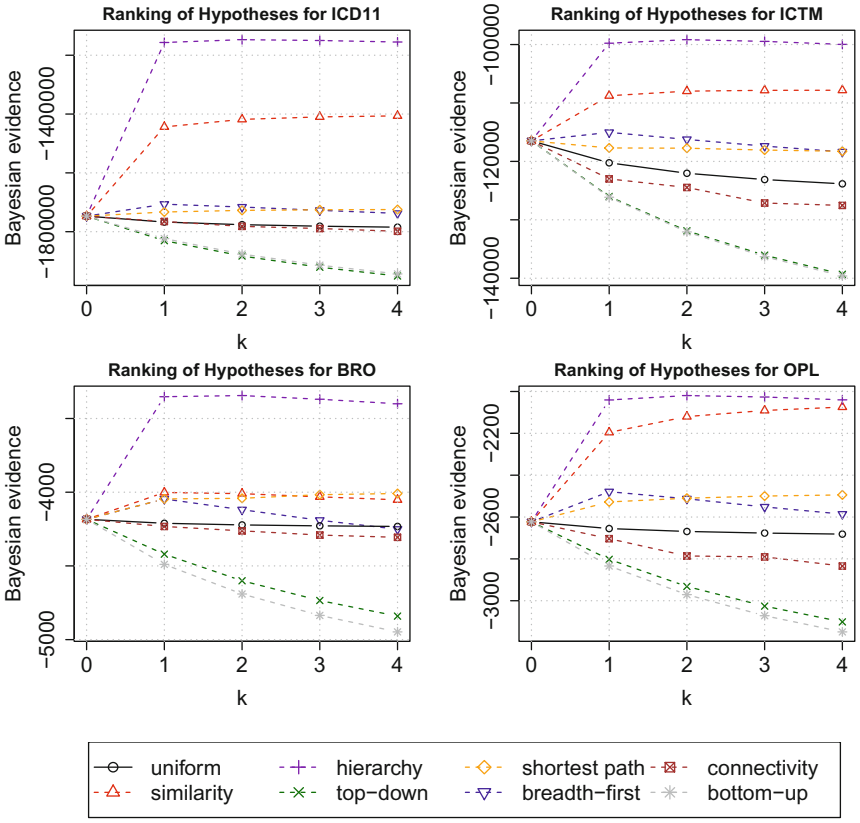


Fig. 3. Hypotheses ranking. Results for comparing hypotheses for the four datasets using HypTrails. The x-axes represent the hypothesis weighting factor k representing the "strength" of our belief in a hypothesis. In general, the stronger we believe in a hypothesis (i.e., the higher we set k), the less we expect to see transitions opposing the parametric beliefs of the corresponding hypothesis. The y-axes depict the Bayesian evidences. The higher the evidence for a given hypothesis, the better it is suited for describing the production of the extracted human edit trails (see Section 3).

5 Results

By applying HypTrails, we are able to gain insights into the relative plausibility of the hypotheses of interest based on the empirical data at hand. We illustrate the results in Figure 3. As mentioned in Section 3, we can compare the plausibility of hypotheses by comparing their marginal likelihoods—the higher, the more plausible. The hypothesis weighting factor k describes the "strength" of our belief in a given hypothesis; for fairness, we compare the plausibility of hypotheses by comparing their Bayesian evidences for the same values of k. For tractability, we report and interpret results for $0 <= k <= 4$; for higher values of k the

results might slightly vary. Next, we highlight the main results for each ontology-engineering project (see Table 2 for a comparison of all hypotheses and datasets). We thoroughly discuss the results in Section 6.

International Classification of Diseases (ICD-11). The results for ICD-11, our biggest dataset, are depicted in the top-left part of Figure 3. The top-down and bottom-up hypotheses indicate lower evidences than the uniform hypothesis, suggesting that users are likelier to randomly change classes in the ontology than strictly follow a top-down or bottom-up approach. The connectivity hypothesis starts out to be nearly as plausible as the uniform hypothesis, but looses in Bayesian evidence faster with increasing k. The breadth-first and shortest-path hypotheses indicate higher evidences than the uniform hypothesis for our $k > 0$ at interest and thus, seem to be plausible explanations for the creation of the given human edit trails. Clearly, for ICD-11, the hierarchy hypothesis represents the most plausible explanation for the production of the trails, and thus the ontology at hand, followed by the similarity hypothesis.

International Classification of Traditional Medicine (ICTM). Similarly to ICD-11, the top-down, bottom-up and connectivity hypotheses exhibit lower evidences than the uniform hypothesis for all analyzed values of $k > 0$ (see top-right part of Figure 3). According to our experiments, the most plausible hypothesis for explaining the production of the edit trails of ICTM is the hierarchy hypothesis as it exhibits the highest Bayesian evidences for all $k > 0$. Further, the similarity hypothesis, as well as the breadth-first and shortest path hypotheses, are also better suited for describing the production of the human edit trails in ontology-engineering projects than the uniform hypothesis. For $k > 2$, we can also observe that the shortest-path hypothesis is increasing in plausibility and takes over the breadth-first hypothesis at $k = 4$.

Table 2. Results. The table depicts the relative ranking of each hypothesis for the corresponding datasets at $k = 4$. The best performing hypotheses are highlighted boldface. If a hypothesis is less likely to explain the production of the corresponding edit trails than the uniform hypothesis, we have marked them with "-" for the corresponding dataset.

	ICD-11	ICTM	BRO	OPL
Hierarchy Hypothesis	**1**	**1**	**1**	**1**
Similarity Hypothesis	2	2	3	2
Shortest Path Hypothesis	3	3	2	3
Breadth-First Hypothesis	4	4	-	4
Uniform Hypothesis	5	5	4	5
Connectivity Hypothesis	-	-	-	-
Bottom-Up Hypothesis	-	-	-	-
Top-Down Hypothesis	-	-	-	-

Biomedical Resource Ontology (BRO). For BRO, the hypothesis with the highest Bayesian evidences for $k > 0$ is, again, the hierarchy hypothesis. Similarly to ICTM, the connectivity, top-down and bottom-up hypotheses are less plausible for explaining the production of the human edit trails in ontology-engineering projects than the uniform hypothesis. In contrast to ICD-11 and ICTM, the similarity hypothesis is less likely to be a plausible explanation for the trails than the shortest path hypotheses. Further, the shortest path hypothesis gains evidence with growing k, while the breadth-first hypothesis drops below the uniform hypothesis at $k = 4$.

Ontology for Parasite Lifecycle (OPL). Similarly to all other projects, the most plausible hypothesis for explaining the production of the trails at hand for OPL is the hierarchy hypothesis, followed by the similarity hypothesis (especially for higher k). The top-down, bottom-up and connectivity hypotheses are again, less plausible than the uniform hypothesis at $k > 0$. Analogously to ICTM, the breadth-first and shortest path hypotheses are more plausible for explaining the creation of the human edit trails than the uniform hypothesis, and switch ranks with growing k.

6 Discussions

The results of comparing the different hypotheses for the four datasets with HypTrails are surprisingly consistent. In all of the four ontology-engineering projects, the hierarchy hypothesis represents the most plausible hypothesis to explain the production of the human edit trails in ontology-engineering projects, and therefore the corresponding ontology at hand. The similarity hypothesis is the second most plausible hypothesis for explaining the production of the human edit trails in ontology-engineering projects for ICD-11, ICTM and OPL (at $k = 4$). The reason for the high Bayesian evidences of the similarity hypothesis is most probably due to the fact that (semantically) similar classes are usually grouped into the same parts of an ontology, hence the similarity calculations are likely to reflect our beliefs of the hierarchy hypothesis. For example, in a biomedical ontology, similar classes are grouped together as siblings or cousins, sharing at least one common parent or grandparent among them. Hence, additional adaptions to further distinguish the similarity hypothesis from the hierarchy hypothesis are warranted. In particular, we plan on investigating correlation between the similarity of classes and existing hierarchical links in future work.

In Walk et al. [23], we have been arguing that users are editing the ontology in a combined top-down and breadth-first fashion. The results of our analysis confirm the results from our exploratory analysis. In particular, the hierarchy hypothesis emphasizes transitions along top-down and breadth-first hierarchical relations (i.e., children, siblings and cousins opposed to uncles and aunts). This finding is also supported by the empirical research conducted by Vigo et al. [20], which shows that the class hierarchy is the central focus of user activity in an ontology-editing session. Users spend more than 45% of their time navigating

or editing the class hierarchy, which serves as an index and external memory of the ontology. The authors have identified the class hierarchy as the central component of the user interface, which also explains very well our findings.

Thus, these observations reinforce our initial belief that the ontological hierarchy influences the selection of which class to edit next. Among other potential scenarios, this information can be leveraged by ontology-engineering tools creators to minimize the efforts required by users to create new, or edit existing content in an ontology. For example, ontology-editing tools may visually highlight the corresponding classes in the user interface, and provide keyboard shortcuts that allow for quicker and more productive editing sessions. Vigo et al. [20] also make the recommendation to place editing features close to the class hierarchy to better support the users in their editing patterns.

In our investigations, we have also identified hypotheses that were weak, and potentially not useful for the purpose of improving the user interface or editing process: the top-down, bottom-up and connectivity hypotheses are less plausible than the uniform hypothesis, meaning that randomly selecting classes to work on is likelier to produce the corresponding edit trails than specifically editing highly connected classes, or editing classes in a top-down or bottom-up fashion.

Our study also has limitations, for example, all investigated ontologies are authored with the same tool, WebProtégé (or its customizations), which may biases some of our findings. However, we believe that the bias is attenuated by the fact that the projects are completely different efforts by different teams, and they also use different customizations of the user interface. Furthermore, Rospocher et al. [10], who have analyzed the change logs of two different ontology-editing platforms (WebProtégé and a Wiki system), have come to the conclusion that users tend to edit around the hierarchy, indifferent of the tool that they used. One difficulty in overcoming this limitation is the fact that obtaining change logs for real-world projects from different platforms is almost impossible. Another limitation is the fact that HypTrails focuses on comparing the relative plausibility of hypotheses. Hence, we can say that the hierarchy hypothesis is the most plausible one for explaining the production of the edit trails at hand. However, we do not know if another hypothesis, other than the ones compared, is more plausible than the hierarchy hypothesis. For example, calculating the actual transition probabilities directly from the trails yields highest Bayesian evidences. However, understanding and interpreting this empirical "hypothesis" is very hard. Also, to be able to conduct an analysis using HypTrails, we need to have detailed change-tracking information, which WebProtégé provides, but might not be as easily obtained for other projects and tools.

7 Conclusions

In this paper, we have formally defined several hypotheses of how users edit an ontology, and systematically investigated, analyzed, and ranked these hypotheses according to their relative plausibility for describing edit trails of four real-world ontology-engineering projects using HypTrails, a coherent research approach.

We have found that the hierarchical structure of an ontology exercises the strongest influence on the observed user behavior, followed by the similarity of concepts. These findings are remarkably consistent across four different real-world projects, with some minor exception for the BRO dataset. We have also discussed how these findings may be used to improve ontology-editing tools. We think that our findings represent an advancement of the empirical research on how ontologies are created, which is a field that has been chronically lacking in our community.

We believe that the insights, uncovered in this paper, into how users *actually* edit real-world ontologies, represent a great opportunity for ontology-tools builders and for project managers, who can potentially leverage this information to create user interfaces and processes that better support the editing patterns of the users.

For future work, we plan to extend our set of formally defined hypotheses by including theories on how users edit properties (current work only considers class-based trails) and include different types of relationships for the analyses presented in this paper. In particular, studying individual (clustered) user behavior to automatically detect subsets of users that behave differently to other subsets of users represents a very promising opportunity for future work. On the longer term, we would like to create a recommendation module for ontology-editing tools, which would be informed by the editing patterns that we identify through our empirical research. We believe that the recommendation module and an adapted user interface will vastly improve the editing experience of the users.

Acknowledgments. This work is supported in part by grants GM086587 and GM103316 from NIH, and grant STR1191/2-1 from the German Research Foundation (DFG).

References

1. An, J., Quercia, D., Crowcroft, J.: Partisan sharing: facebook evidence and societal consequences. In: Conference on Online Social Networks, pp. 13–24. ACM (2014)
2. Beydoun, G.: Formal concept analysis for an e-learning semantic web. Expert Systems with Applications **36**(8), 10952–10961 (2009)
3. Borges, J., Levene, M.: Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions. IEEE Transactions on Knowledge and Data Engineering **19**(4), 441–452 (2007). http://dx.doi.org/10.1109/TKDE.2007.1012
4. Chi, E.H., Pirolli, P., Pitkow, J.: The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a web site. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 161–168. ACM (2000)
5. Falconer, S.M., Tudorache, T., Noy, N.F.: An analysis of collaborative patterns in large-scale ontology development projects. In: K-CAP, pp. 25–32. ACM (2011)
6. Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E., Lukose, R.M.: Strong Regularities in World Wide Web Surfing. Science **280**(5360), 95–97 (1998). http://www.sciencemag.org/content/280/5360/95.abstract

7. Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association **90**(430), 773–795 (1995)
8. Parikh, P., Zheng, J., Logan-Klumpler, F.J., Stoeckert Jr., C.J., Louis, C., Topalis, P., Protasio, A.V., Sheth, A.P., Carrington, M., Berriman, M., et al.: The Ontology for Parasite Lifecycle (OPL): towards a consistent vocabulary of lifecycle stages in parasitic organisms. J. Biomedical Semantics **3**, 5 (2012)
9. Pesquita, C., Couto, F.M.: Predicting the Extension of Biomedical Ontologies. PLoS Comput. Biol. **8**(9), e1002630 (2012)
10. Rospocher, M., Tudorache, T., Musen, M.A.: Investigating collaboration dynamics in different ontology development environments. In: Buchmann, R., Kifor, C.V., Yu, J. (eds.) KSEM 2014. LNCS, vol. 8793, pp. 302–313. Springer, Heidelberg (2014)
11. Singer, P., Helic, D., Hotho, A., Strohmaier, M.: HypTrails: a bayesian approach for comparing hypotheses about human trails on the web. In: International Conference on World Wide Web (2015)
12. Singer, P., Helic, D., Taraghi, B., Strohmaier, M.: Detecting Memory and Structure in Human Navigation Patterns Using Markov Chain Models of Varying Order. PloS one **9**(7), e102070 (2014)
13. Singla, A., White, R., Huang, J.: Studying trailfinding algorithms for enhanced web search. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 443–450. ACM (2010)
14. Strohmaier, M., Walk, S., Pöschko, J., Lamprecht, D., Tudorache, T., Nyulas, C., Musen, M.A., Noy, N.F.: How Ontologies are Made: Studying the Hidden Social Dynamics Behind Collaborative Ontology Engineering Projects. Web Semantics: Science, Services and Agents on the World Wide Web **20** (2013)
15. Tenenbaum, J.D., Whetzel, P.L., Anderson, K., Borromeo, C.D., Dinov, I.D., Gabriel, D., Kirschner, B.A., Mirel, B., Morris, T.D., Noy, N.F., Nyulas, C., Rubenson, D., Saxman, P.R., Singh, H., Whelan, N., Wright, Z., Athey, B.D., Becich, M.J., Ginsburg, G.S., Musen, M.A., Smith, K.A., Tarantal, A.F., Rubin, D.L., Lyster, P.: The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. Journal of Biomedical Informatics **44**(1), 137–145 (2011)
16. Tudorache, T., Falconer, S., Nyulas, C., Noy, N.F., Musen, M.A.: Will semantic web technologies work for the development of ICD-11? In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part II. LNCS, vol. 6497, pp. 257–272. Springer, Heidelberg (2010)
17. Tudorache, T., Nyulas, C., Noy, N.F., Musen, M.A.: WebProtégé: A Distributed Ontology Editor and Knowledge Acquisition Tool for the Web. Semantic Web Journal, 11–165 (2011)
18. Tufts, P.: Use of web usage trail data to identify related links. US Patent 6, 691, 163, February 10, 2004. https://www.google.com/patents/US6691163
19. Van Laere, S., Buyl, R., Nyssen, M.: A method for detecting behavior-based user profiles in collaborative ontology engineering. In: Meersman, R., Panetto, H., Dillon, T., Missikoff, M., Liu, L., Pastor, O., Cuzzocrea, A., Sellis, T. (eds.) OTM 2014. LNCS, vol. 8841, pp. 657–673. Springer, Heidelberg (2014)
20. Vigo, M., Jay, C., Stevens, R.: Constructing conceptual knowledge artefacts: activity patterns in the ontology authoring process. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, pp. 3385–3394 (2015)

21. Walk, S., Singer, P., Strohmaier, M.: Sequential action patterns in collaborative ontology-engineering projects: a case-study in the biomedical domain. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Managemen, pp. 1349–1358. ACM (2014)
22. Walk, S., Singer, P., Strohmaier, M., Helic, D., Noy, N.F., Musen, M.A.: Sequential Usage Patterns in Collaborative Ontology-Engineering Projects (2014). arXiv preprint arXiv:1403.1070
23. Walk, S., Singer, P., Strohmaier, M., Tudorache, T., Musen, M.A., Noy, N.F.: Discovering beaten paths in collaborative ontology-engineering projects using markov chains. Journal of Biomedical Informatics **51**, 254–271 (2014)
24. Wang, H., Tudorache, T., Dou, D., Noy, N.F., Musen, M.A.: Analysis of user editing patterns in ontology development projects. In: Meersman, R., Panetto, H., Dillon, T., Eder, J., Bellahsene, Z., Ritter, N., De Leenheer, P., Dou, D. (eds.) ODBASE 2013. LNCS, vol. 8185, pp. 470–487. Springer, Heidelberg (2013)
25. West, R., Leskovec, J.: Human wayfinding in information networks. In: International Conference on World Wide Web, pp. 619–628. ACM (2012). http://doi.acm.org/10.1145/2187836.2187920
26. White, R.W., Huang, J.: Assessing the scenic route: measuring the value of search trails in web logs. In: Conference on Research and Development in Information Retrieval, pp. 587–594. ACM (2010)