Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Inverse Dynamical Inheritance in Stack Exchange Taxonomies

César Ojeda, Kostadin Cvejoski, Rafet Sifa, Christian Bauckhage
Fraunhofer IAIS, Bonn, Germany

Abstract

Question Answering websites are popular repositories of expert knowledge and cover areas as diverse as linguistics, computer science, or mathematics. Knowledge is commonly organized via user defined tags which implicitly create population folksonomies. However, the interplay between latent knowledge structures and the answering behavior of users has not been fully explored yet. Here, we propose a model of a dynamical tagging process guided by taxonomies, devise a robust algorithm that allow us to uncover hidden topic hierarchies, apply our method to analyze several Stack Exchange websites. Our results show that the dynamics of the system strongly correlate with uncovered taxonomies.

1 Introduction

Questions answering (QA) sites are Web platforms where millions of users create and organize content. The established protocol requires the use of tags or keywords to characterize incoming questions. On sites such a MathOverflow, StackOverflow, or others from the Stack Exchange family, these tags usually refer to scientific disciplines which with known hierarchical relationships. The organization of knowledge on QA sites therefore depends on the interplay between the arrival of questions and the latent organizational structure defined by the co-occurrence of tags. Although algorithms which uncover ontologies from tags have been developed before, most work on QA sites so far has focused on the relationships between tags and users using unstructured graph models, ignoring the taxonomic structure.

In the case of QA web sites, we can conduct detailed empirical studies of the dynamical behavior of a complex interactive system that is implicitly hierarchically organized. In this paper, we first introduce an algorithm that can uncover knowledge taxonomies from collections of tags. We then we characterize these taxonomies through adequate metrics such as branch size and level distribution and connect these structural metrics to the behavior of the users of these sites. In particular, we observe a phenomenon of *dynamical inheritance* where user activity related to a given tag correlates with activity related to tags on lower levels of the hierarchy.

Table 1: Statistics of the Stack Exchange data studied here.

site	# questions	# answers	# users	# tags	# n-tuples
Biology	8958	11628	10631	642	5643
English	57112	147517	91621	947	19381
Finance	4098	6416	8155	495	3186
Math	63161	102447	46379	2602	28639
Physics	39355	63579	39117	824	24095
Statistics	42921	47755	40324	1032	28232

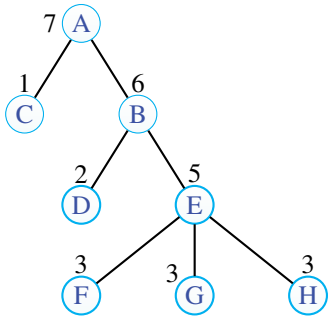
2 Stack Exchange Data

The Stack Exchange family of question answering sites covers a wide range of topics and allows subscribed users to post questions to the community. Answers are submitted and rated by the community and can be deemed acceptable by the user who posted the question. A major incentive for users to answer questions is a reputation building feature where a user's reputation grows with favorable ratings and increasing numbers of accepted answers. Indeed, reputation scores may nowadays boost careers; *Stackoverflow*, the main site of the Stack Exchange network, is known to be used as a recruitment platform where companies are looking for knowledgeable talent and experts.

In this paper, we focus on the analysis of the 5 different sub communities in Tab. 1.

Our goal is to uncover emergent knowledge structures from sets of questions and to investigate how they relate to user behavior. While individual questions only refer to a limited set of topics, the aggregated behavior of users reveals whole knowledge taxonomies (Bhat et al. 2014). Since prior work on tag dynamics (Halpin, Robu, and Shepherd 2007; Ramage et al. 2009) indicates that tags are a reliable proxy for content classification, we analyze n-tuples of tags assigned to questions rather than the content of questions.

Table 1 summarizes statistics as to the data we consider in this paper (observation periods, numbers of questions, answers, users, and tags crawled, and number of observed n-tuples). Clearly the numbers of observed n-tuples exceeds the number of observed tags. Table 2 lists ratios between the number of observed n-tuples and the numbers of questions and users. As these ratios indicate a repeated occurrence of tuples, these statistics suggests that users do not randomly select tags but are guided by some latent relationship among them.



(a) a hidden taxonomy

$\alpha_1 = \{B,D\}$
$\alpha_2 = \{E,H\}$
$\alpha_3 = \{A,C\}$
$\alpha_4 = \{A,C\}$
$\alpha_5 = \{A\}$
$\alpha_6 = \{E,H\}$
$\alpha_7 = \{A,B\}$
$\alpha_8 = \{A,D\}$
$\alpha_9 = \{A,E\}$
$\alpha_{10} = \{B,E\}$
$\alpha_{11} = \{B,E\}$
$\alpha_{12} = \{B,A,F\}$
$\alpha_{13} = \{E,A,F\}$
$\alpha_{14} = \{A\}$
$\alpha_{15} = \{E,G\}$
$\alpha_{16} = \{A,C\}$
$\alpha_{17} = \{B,A,G\}$
$\alpha_{18} = \{E,B,F\}$
$\alpha_{19} = \{E,F\}$
$\alpha_{20} = \{A,D\}$
$\alpha_{21} = \{B,F\}$
$\alpha_{22} = \{B,E\}$
$\alpha_{23} = \{A,B,H\}$
$\alpha_{24} = \{A,C\}$

(b) training sample of tuples of tags

t	2	3
A	7	{C,B,E,D,G,F,H}
B	6	{A,E,D,G,F,H}
E	5	{A,H,B,G,F}
H	3	{A,B,E}
F	3	{A,B,E}
G	3	{A,B,E}
D	2	{A,B}
C	1	{A}

(c) data structure

Figure 1: Example of a taxonomy tree and a training set of n-tuples of tags resulting from a normal tagging process guided by this tree. In the tree, letters represent tags and numbers indicate how many tags a tag co-occurs with in the training set. The rightmost panel shows the data structure for our taxonomy learning algorithm; it can be obtained directly from the training set.

Site	Ratio #NT/#Q	Ratio #NT/#U
Biology	0.63	0.53
English	0.34	0.21
Finance	0.78	0.39
MathOverflow	0.45	0.62
Physics	0.61	0.62
Statistics	0.66	0.70

Table 2: Main statistics for the Stack Exchange Sites studied

The main empirical motivation behind our work is the striking similarity we observed between tag frequency distributions and the distribution of the size of sets of co-occurring tags. For both empirical distributions, we found the Log Normal distribution to provide very good fits. This corroborates earlier results (Halpin, Robu, and Shepherd 2007) and complies with our hypothesis as to the existence of a hierarchical structure modulating the tagging process: if the number of tags which co-occur with a given tag is proportional to the number of leafs in a taxonomic sub-tree emanating from it, the number of tags can be thought of to result from successive multiplications of random variables and will therefore be log-normally distributed (Gallager 2012).

3 Tagging Process Model

Our task at hand is to identify a taxonomy or tree structure of topics from observed co-occurrences of tags. This is an unsupervised learning problem where we need to learn a model (taxonomy) from a set of data points (the n-tuples of tags). In order for this learning task not to be ill-posed, we shall inform it with prior expectations.

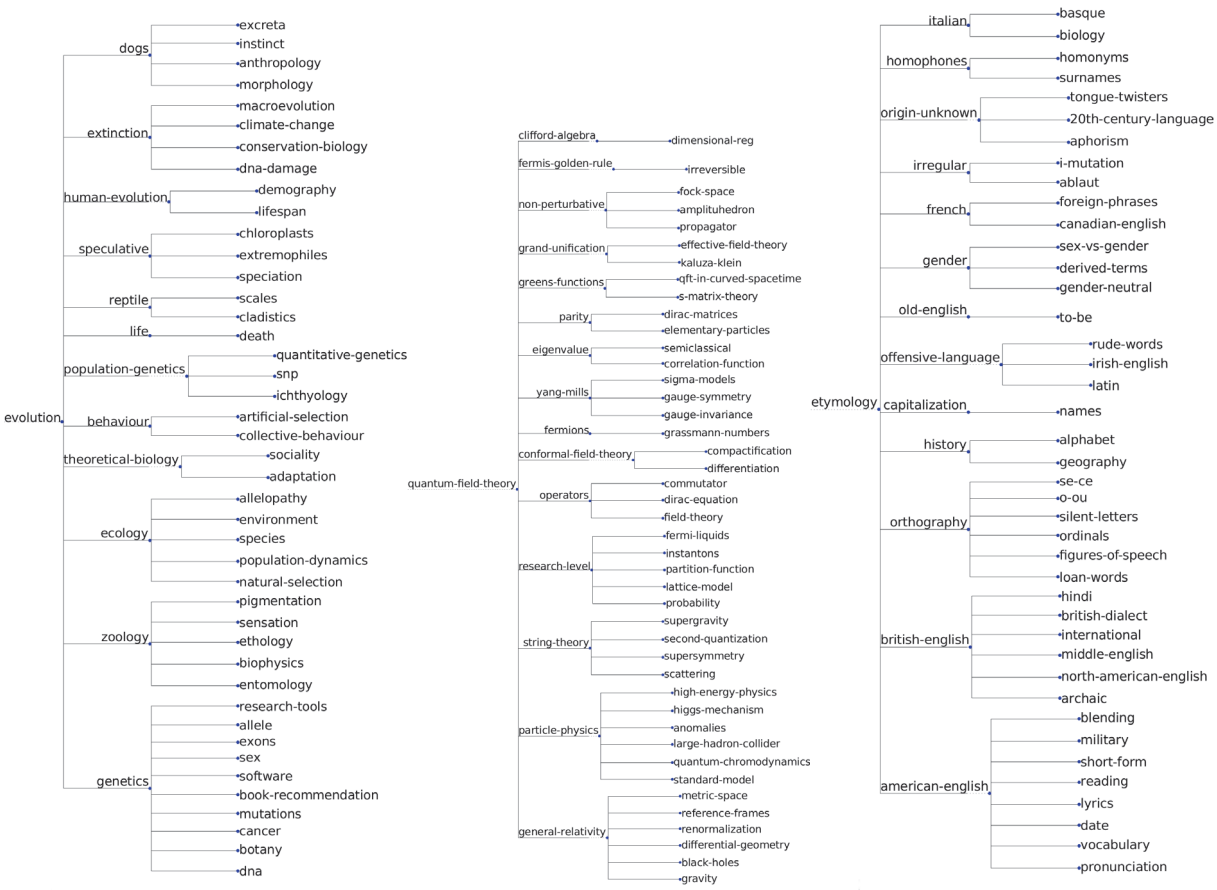
To this end, we assume that there are universal hidden hierarchies which define relations between different tags and are, to some extend, known to the users. For instance, for the biology Stack Exchange site, for example, *human anatomy* will be a part of *human biology*. We also assume that the way users assign tags to questions is conditioned on these knoeledge trees. One of the tags assigned to a question will reside on a level less or equal than all other tags. We define this tag or node as the subject node. If we follow the branch of all ancestor nodes of the tag up to the root node of the tree, we encounter all the knowledge areas to which the given question pertains. To describe the content of the question, the user is thus assumed to randomly select tags from this branch. Under this model, the creation of the question is a local process; the user only knows about the branch to which the question pertains. The overall tree, on the other hand, is a global construct; the hierarchical structure we want to uncover emerges from the cumulative process of collective question answering.

There is a possibility for anomalous tagging behavior which is not guided by latent taxonomies but instead arises from random user behavior. For example, a user might pose a question related human-biology *and* botany. Questions like these do not provide information as the latent hierarchy. To be able to argue formally, we define different uninformative tagging processes:

1. Children Tagging: once the subject node is selected (conditioned on the latent tree), the user selects the upper branch as well as several children; this will create co-occurrences between tags on the same level of the tree.
2. Horizontal Tagging: the user selects a subject node, its upper branch as well as random tags on the same level.
3. Random Tagging: the user decides for a subject node and then randomly selects from all nodes on levels above the given one.

The learning algorithm which we describe next[1], requires the data set to meet the following three conditions. Firstly, there must exist at least one n-tuple where t_a and t_b co-occur. Secondly, anomalous tagging is assumed to be rare. Thirdly, t_a must have at least two children. Fig. 1 shows a didactic example to clarify the procedure and the stated principles. The number attached to each node of the tree in Fig. 1(a) indicates the size of the co-occurrence set of

[1]All code and implementation details are available at https://github.com/cesarali/Tag2Hierarchy



(a) Biology (b) Physics (c) English language

Figure 2: Taxonomies learned form the Stack Exchange data. For better readability, we only show three levels per tree.

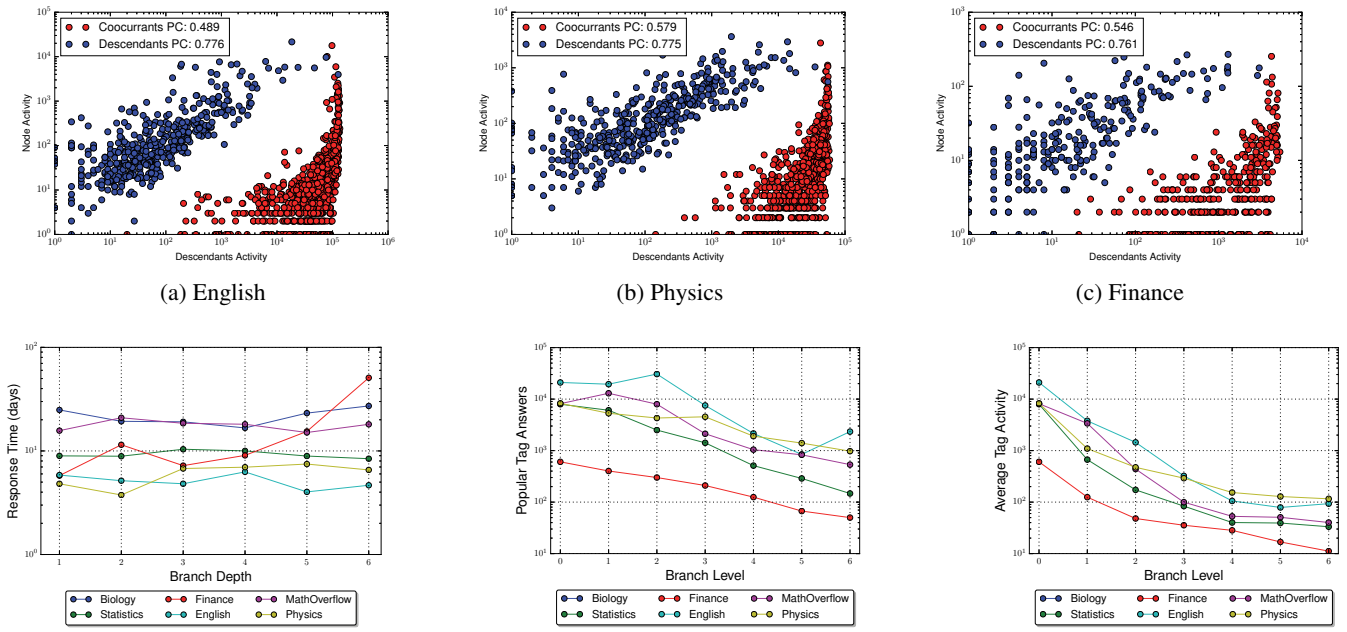
the corresponding tag. The training samples in Fig. 1(b) are sampled from that tree. In order to efficiently obtain the tree from the sampled tuples, we define a data structure $\mathcal{J}$ which orders the concurrence sets $O(t_i)$ (according to size) for every tag t_i which is shown in Fig. 1(c). We perform one bottom up pass over $\mathcal{J}$ to uncover pairwise relations between parent t_p and child t_c. This is given when $t_c \in O(t_p)$. Also, we examine the fraction of the descendants of t_c which also pertain to $O(t_p)$. Among all possible parents we select the one with higher number of descendants also in $O(t_p)$.

In order to quantitatively evaluate the taxonomy learning algorithm, we performed taxonomy inference on synthetic data. Given artificially created taxonomies, we sampled n-tuples of tags using the above process model. We then applied our algorithm to infer a taxonomy from the data and compare the results to the ground truth. This allowed us to test the algorithm under different anomalous tagging behaviors as well as under different taxonomy structures. To quantify the quality of taxonomies inferred from data samples, we applied a variant of the concept of dendogram purity (Heller and Ghahramani 2005) and obtained over 0.8 purity.

4 Stack Exchange Results

We performed taxonomy inference for all Stack Exchange sites in our data set. Figure 2 shows excerpts of the trees we obtained for biology, physics, and english. Notice that these hierarchies result from from the tagging behavior of the Stack Exchange communities. For instance, in Fig. 2b, *general relativity* appears as sub-field of *quantum field theory*. Although this would not be the case in a typical physics taxonomy, an inspection of the siblings of *general relativity* shows that *research level* is a sibling indicating that, as a research field, *general relativity* is a sub-field of quantum field theory. Given the current state of physics research on a unified field theory, this is indeed an acceptable classification.

According to our model, there is a difference between the dynamics of the tagging process and the intrinsic knowledge hierarchy. That is, the probability of selecting a given node as the *subject node* is independent of the node's location in the hierarchy. Yet, this location usually depends on the current interest of the population in the corresponding topic. Nonetheless, another node might be selected trough the branch dependencies. We refer to this process as **Inverse**



(a) English (b) Physics (c) Finance

(d) response times per level (e) maximum tag activity per level (f) average tag activity per level

Figure 3: Dynamical dependance between taxonomy structures and community behaviors; (a)–(c) shown via correlations of the activities related to a parent node and activities related to its descendants (blue); as a reference, activities are also shown for tags and their co-occurrence sets (red); (d)–(f) show the average response times per level, the number of answers for the most popular tag per level, and the average number of answers per level.

Dynamical Inheritance. The structure of the hierarchy affects the dynamics in a backwards manner. Parent nodes get activated through activities of their descendants. Quantitative results as to this phenomenon can be seen in Fig. 3. Here, we plot the activity of a node against the activity of all its descendants. Activity is given by the number of answers to questions where a particular tag occurs. In order to remove spurious correlations due to using the same set of answers, we remove all the answers which where also part of the descendants. For all inferred taxonomy trees, we obtained a Pearson correlation of 0.7 or above for the descendants activities. For the co-occurrence activities, we found a correlation of 0.5.

Figure 3d displays the average response time per level. Interestingly, fluctuations are small and response times are almost invariant. This is unexpected given the dependance on descendants and the fact that fewer descendants could have been thought to imply less activity, since fewer users may specialize in the corresponding topics. Yet, this might be attributed to the reward mechanism of the Stack Exchange sites. Independent of the taxonomic level, users responds as soon as possible to gain reputation in their community. Figure 3e shows the number of answers for the most popular tag at different levels. As expected, there is a decreasing trend since deeper levels indicate higher degrees of specialization. Yet, the fluctuations in these curves indicate that tag related activity is not solely level driven. Tag in deeper levels can show more activity than their parents. Finally, average activities per tag per level are displayed in Figure 3f and show a more pronounced decreasing trend.

5 Conclusion

We presented an algorithm to infer knowledge taxonomies from n-tuples of tags assigned to questions on Stack Exchange sites. It was based on a probabilistic model of the tagging process and our results show that the automatically uncovered taxonomies can account for the popularity of certain tags. In future work, we intend to study how a user's reputation is related to his or her level of expertise as reflected by where in the taxonomy their answers are located.

References

Bhat, V.; Gokhale, A.; Jadhav, R.; Pudipeddi, J.; and Akoglu, L. 2014. Min(e)d Your Tags: Analysis of Question Response Time in Stackoverflow. In *Proc. ASONAM*. IEEE.

Gallager, R. 2012. *Discrete Stochastic Processes*. Springer.

Halpin, H.; Robu, V.; and Shepherd, H. 2007. The Complex Dynamics of Collaborative Tagging. In *Proc. WWW*. ACM.

Heller, K., and Ghahramani, Z. 2005. Bayesian Hierarchical Clustering. In *Proc. ICML*.

Ramage, D.; Heymann, P.; Manning, C. D.; and Garcia-Molina, H. 2009. Clustering the Tagged Web. In *Proc. WSDM*. ACM.