Social Network Analysis of Ontology Edit Logs

Nenad Tomašev and Dunja Mladenić

Jožef Stefan Institute, Ljubljana, Slovenia

This paper presents an approach applying social network analysis on collaborative edit log data. Semantic Web Wiki and FAO ontologies are given as case studies. A number of users that are editing the same ontology or the same pages can be viewed as a social network of people interacting via the ontology. We propose to represent the edit log files as a graph either of users that are connected if they are editing the same ontology concepts or of concepts that are connected if edited by the same users. We apply social network analysis on such graphs in order to provide some insights into activity of the Wiki/ontology editors. Finally, a plugin was developed which provides a comfortable GUI to some of the used analysis techniques, so that the people interested in monitoring the editing activity can perform that analysis and visualization on their own.

Keywords: social network analysis, wiki, semantic web, collaborative editing, NEON project

1. Introduction

Social network analysis [3] enables analysis of connections between actors in a social network. Strictly speaking, the connections between the actors are based on some social interactions, but they can also be of different nature.

In this paper we address a problem of indirect interactions between users when they are collaborating on editing a set of pages representing ontology concepts – both when taking into account the available ontological information and when disregarding it.

The data that will be used in this paper to demonstrate how these social network analysis techniques can help in unveiling the underlying dynamics of the community collaborating on ontology editing, comes from two different sources.

The first source is the history of changes made on wiki pages of Semantic Web Wiki. These change logs are publicly available from the Wiki [2] under changes history. Most of the described analysis was presented in [5].

The second data source was a small subsample of the edit logs of AGROVOC Thesaurus of FAO (Food and Agriculture Organization of United Nations) [6]. It is a vocabulary designed to cover most of the terms used in agriculture, forestry, fisheries, food and the related domains.

Our goal was to demonstrate how the proposed approach can bring insights into the details of the collaborative editing process by representing the change logs as a set of graphs. Even though the data we were using could have been observed as is customary in Web mining [4], we decided to approach the problem from a different perspective. We create a graph of users, where two users are connected if they have been making changes to the same page or ontology concept. We also make a graph of concepts, where the concepts are connected if they have been edited by some common user or a group of users. Apart from that, we make an alternative view of the data as a bipartite graph of users and concepts, as well as similar graphs for countries that users come from and ontologies that the concepts belong to, if such information is provided. The resulting visual representations of the change logs clearly show several dimensions of the users' activity including the most active users, grouping of users based on accessing the same pages, the most central users and the most frequently edited pages.

The rest of this paper is structured as follows. Section 2 gives data description, Section 3 provides analysis of the data, Section 4 describes the constructed NeOn [7] plugin while Section 5 concludes the paper by discussion.

2. Data Description and Preprocessing

An overview of the data analyzed in this paper is given below, separately for the two used data sources.

2.1. Semantic Web Wiki

The data with change log of semantic Wiki pages as of the end of 2008 were given in an XML form, presented as a sequence of revisions for each page in the Wiki, containing a timestamp and either a username with an internal ID code or an IP of the person performing the page revision. Consecutive changes in page content were also included in the file, but we have discarded them in the pre-processing phase as they are not used in our further analysis. There was a total of 36,078 page edits. Most of those edits (75.5%) were made by the registered users. We have identified 617 registered users and 2,512 different IPs of the anonymous (non-registered) users. For the purposes of further analysis, each unique IP was considered a separate user giving us a total of 3,129 users. We have also performed some merging of user IDs in the process of data cleaning, since there were some cases of people using several usernames when logging in, for instance *MaxVölkel* and *Max Völkel* most likely refer to the same person, considering the fact that name collisions of such sort seem pretty unlikely in such a small user community.

The most "collaborative" users can be seen in Figure 1. By collaborative we mean that they are frequently (at least 5 times) editing the same pages as (at least 3) other users. In our visualization, size of the circle reflects the number of changes the user has made on all the pages, while color shows if the user is registered (red) or anonymous (blue). We can see that the most active users are registered (the right top part in Figure 1), a few of them are exceptionally active compared to the other users (e.g, the most active user has made 6980 page edits, the second most active 2982, while the 10th most active made 589 page edits).

After a brief inspection, it was determined that there were many very similar pages in the Wiki, for instance: *WikiSpammer, WikiSpammers, Talk:WikiSpammer*. Since it was conjectured that a same user or a group of users of similar knowledge and skills is likely to be editing all of such very similar pages, it was deemed useful to
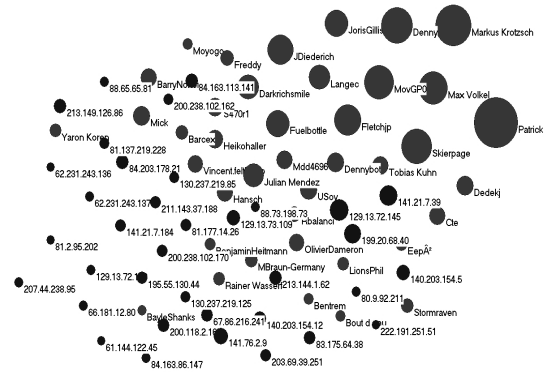

Figure 1. A subset of users that have edited the same pages as at least three other users (min. vertex degree = 3) and share at least five pages with them (min. edge weight = 5).

first group together such pages into one concept and form an aggregate concepts representation of the data.

When combining the pages we used additional information that we have on the data, namely that the pages represent concepts of the ontology on semantic web represented as Semantic Wiki. However, in the analysis itself, no information about the structure of the original concept ontology was used. The concepts were grouped based on syntactical similarity of their names i.e. page titles and the shortest title was used to label the aggregated concept. Short words that were too common in the concept names had been added to the exception list in order to contain a reasonable maximum group size. Some of these words include: *Property, Talk, Category, Template*, etc. The number of concepts was thus reduced from the original 7,369 to 5,500.

2.2. AGROVOC Sample

The sample that was observed comprised 702 actions performed by the users in the period 17.11.2008. – 11.12.2008. For each action, the following information was provided: username, full name of the user in case it was a registered user, term name, country of the user, ontology name where the concept belongs to, as well as action type.

Thirty different action types were given. Out of 702 edits, 40.5% were of the type term-create, 14.8% of the type term-relationship-add, 9.8% of the type concept-create and 7.9% of the type term-edit. The rest of the action types are less frequent. Figure 2 shows which ontologies are
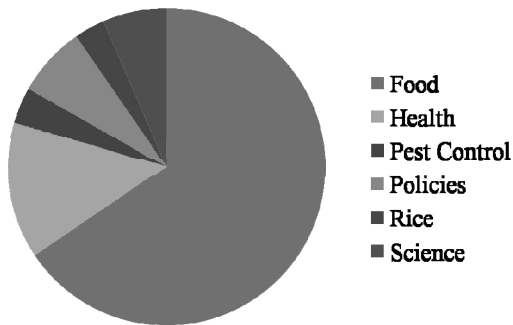
Figure 2. Proportion of actions performed in the sample on present ontologies.

most represented in the sample and it is clear that the Food ontology is the predominant one, as well as Health ontology.

Anonymous edits are allowed, as was the case with Semantic Web Wiki. On the other hand, in the data that we had access to, no IP was given, so all anonymous edits were simply marked by username *Guest*. There were 23 registered users editing the terms in the sample and a total of 280 terms.

AGROVOC is a multilingual thesaurus, which means that concepts are given in various languages. This is precisely why there is a difference between terms and concepts — term is basically a concept expressed in some specific language. Unlike with Semantic Web Wiki, here the analysis was performed on the level of terms. Hence, only term-create, term-edit and term-delete actions were taken into account, the total of 358 actions.

3. Searching for Relations

In this section, we present the results of the analysis on two data sources separately, focusing much more on Semantic Web Wiki, because of the difference in size of the edit logs, which allows for more information to be extracted from the Wiki edit logs than the small subsample of AGROVOC ontology editing. Redundancy in the presented analysis will be avoided, so for the second case only those things which are different than in the first case will be discussed.

3.1. Graphs Constructed from Semantic Web Wiki Edit Logs

After the preprocessing, three different graphs were extracted from the data: a graph of users,

a graph of concepts and a bipartite graph of concepts and users. These were given as input to social network analysis system Pajek [1] that was used here for data analysis including graph visualizations provided in Figure 1. The system includes several social network analysis methods and is one of the commonly used systems in research on graph visualization and social network analysis.

3.1.1. User Graph

It had already been stated that the initial dilemma regarding the mapping of users declared in the XML schema to the graph nodes in the network model of the community was related to the phenomenon of anonymous revisions. Such revisions accounted for 24.5% of the total number of page edits, so it would not have been beneficial to simply disregard them. Even though the average number of revisions per anonymous user was only 3.5, the maximum number was 251, which was not negligible. The overall most active user was *Patrick* accounting for 19.3% of the total number, by performing 6,980 page edits in the Wiki.

User graph was formed by connecting the users who contributed to the same aggregated concepts, i.e. who had edited pages belonging to the same group of pages. The edges were associated with weights, corresponding to the number of different concept groups that were edited by the same two users.

The user graph has 86,401 edges, which is less than 1% of the maximum possible number of edges. Hence, it can be considered a sparse graph. Only about 5% of these edges have a weight of more than one. About 9.7% of the graph consists of isolated nodes. The pages that were edited by these users were not edited by any other user. Most of them were pages about the users themselves (so, most of these users didn't participate in building the Wiki any more than making an entry about themselves or their coworkers). However, there were also some exceptions, such as concepts *Ontology learning, RELAX SEO Services* and *Category: Czech person*. These were only edited by users *Dmanzano, Webmissile* and *Tom*, respectively.

In the user graph, there are 30 connected components having at least two nodes, but the biggest component comprises most of the users (84%).

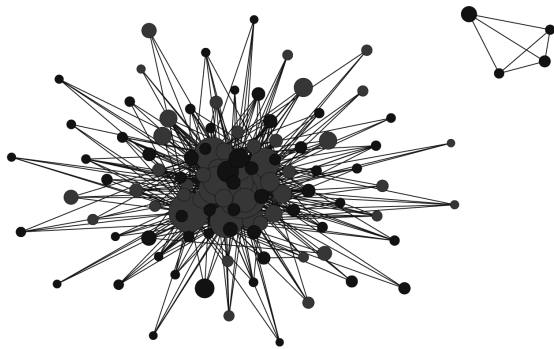Average distance among reachable pairs in the graph is 2.23.

Figure 3. Graph of the most collaborative users in Semantic Web Wiki (min. edge weight=4, min. Vertex degree=3) showing an isolated group of four anonymous users actively working on a group of 9 Wiki pages (top right).

Centrality was calculated for all the users and is presented in Figure 4. It is immediately apparent that there is a small number of very active users in the center of the network and a lot of less well centered users.
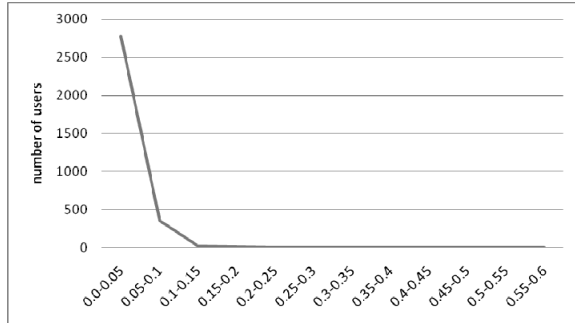

Figure 4. Distribution of degree centrality among the Semantic Web Wiki users.

The betweenness centralization achieved in the user network is 0.15. It is interesting to compare the top five users according to these different criteria. Such a comparison is given in Table 1. The users *Patrick*, *Markus Krotzsch* and *Denny* occur at the top in all three lists, with some difference in their ranking. As can be seen in Table 1, while *Patrick* has the highest number of revisions, *Markus Krotzsch* has the highest degree of centrality and betweenness. Even though these user lists (not just the top 5) are somewhat similar, they do show that an activity of a contributor does not directly reflect his/her collaborative potential.

Number of revisions	Degree centrality	Betweenness
1. Patrick	1. Markus Krötzsch	1. Markus Krötzsch
2. Markus Krötzsch	2. Patrick	2. Patrick
3. Denny	3. MovGPO	3. Denny
4. Skierpage	4. Denny	4. MovGPO
5. Joris Gillis	5. Knud	5. WikiSym

Table 1. Top five users according to some importance criteria from social network analysis.

3.1.2. Concept Graph

In the same way in which we have observed the similarities between various users, it is possible to determine similar concepts, or rather in this case – similar groups of concepts. Hence, a weighed concept graph was constructed, each aggregate concept represented as a node, edge weights corresponding to the number of common users that have edited concepts belonging to both concept groups.

Even though the approach is the same as in the case of user activity analysis, the concept graph displays some different properties. Unlike the user graph, concept graph is dense. It has 5600 vertices, but approximately 2,5 million edges, which is 15.95% of the number of edges in a complete graph with the same number of vertices. Average distance between reachable pairs is thus smaller than the one of user graph, though diameter of the largest component is the same, six. There are 47 connected components that contain at least two vertices, but the larger component comprises 91.6% of the total number of vertices.

By observing strongly related concepts, it becomes quite apparent that there aren't that many concepts that are both being edited by a larger group of users. After making a cut in the graph (see Figure 5) by removing all those aggregate concepts that do not have at least 10 common revisers, we have only 35 concepts remaining, most of which are some pretty general concepts (such as, *Math, Time, Help, Relation, Programming, Germany, Sandbox*). Concept *Sandbox*, for instance, represents a page pointing to a test
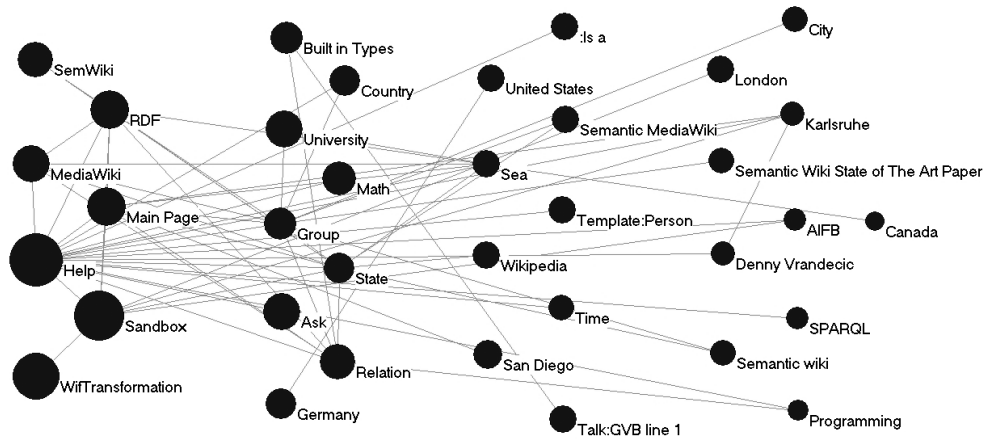
Figure 5. Concept graph showing concepts having at least 10 mutual revising users with other concepts.

server where users can test their queries and test the appearance of their content prior to adding it to the main Wiki. This might indicate a lack of cooperation between larger user groups, since we have already seen that there are pairs of users exhibiting great similarity in their Wiki editing effort. On the other hand, it might also be a sign of good partitioning of work, because there is no need for many people to participate in an editing of the same page, under the assumption that the editing is done by domain experts who do not need others to provide corrections of their contributions to the Wiki.

Another approach in search for strong relations between concepts is to determine line islands in the concept graph. Line island is a subgraph induced by a set of vertices such that the relations (edge weights) between the vertices in an island are stronger than the ones between those vertices and the rest of the vertices in the original graph. As it turns out, the graph of Semantic Web Wiki aggregate concepts has 9 such islands comprising 1.2% of the entire graph. Three most interesting islands are shown in Figure 6. In the bottom left corner one can see the following concepts: *Category:AlkaliEarthMetalsGroup, Cat-*
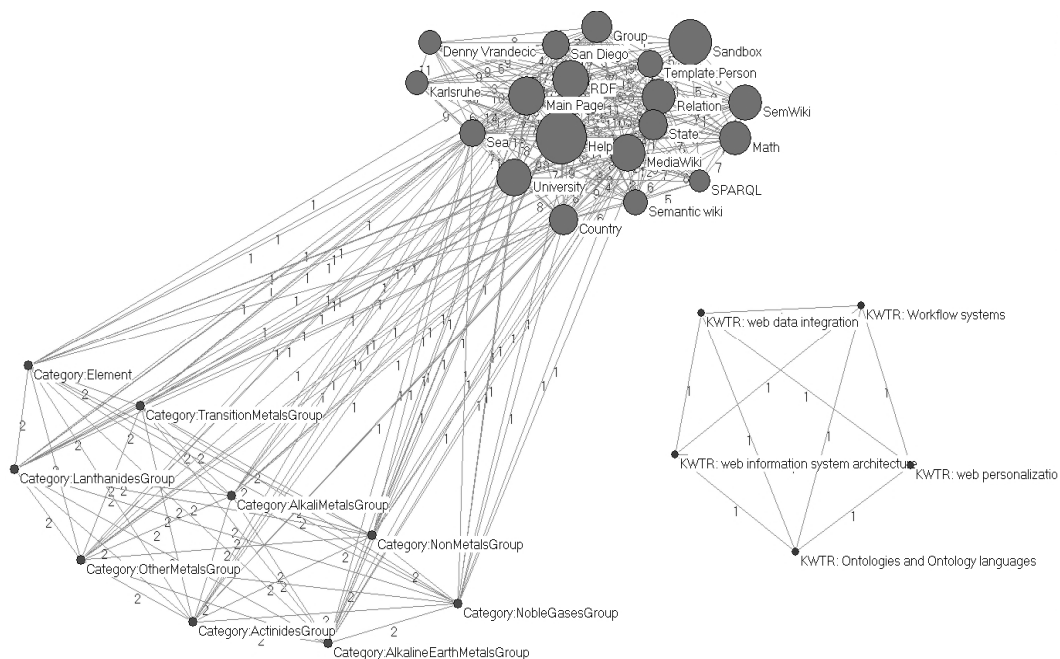

Figure 6. Concept graph showing some strongly correlated concept groups in Semantic Web Wiki editing.

egory:TransitionMetalsGroup, Category:Acti-nides, Category:NonMetalsGroup, Category: AlkaliMetalsGroup and a few more chemistry-related concepts. They form an island of their own, though they are apparently connected with the largest island composed of the most common aggregate concepts: *Help, Semantic Wiki, Media Wiki, Main_page,Sandbox,SPARQL, Template: Person*, etc. Five isolated concepts on the right are related to KWTR (Knowledge Web Technology Roadmap), specifically: *KWTR: web_data_integration, KWTR:Workflow_systems, KWTR:web information_system_architecture, KWTR:web_ personalization, KWTR:Ontolo-gies_andOntology_language.*

Not only does the examination of the most frequently edited concepts and strongly correlated ones provide us with the information about trends in Wiki page editing, but also does the analysis of rarely edited concepts.

Surprisingly, 4400 aggregate concepts (78.6% of merged concept graph) have been edited less than 6 times during Wiki development. This means that most of the concepts in the ontology have not been edited many times, nor have they been edited by many people (no more than 5, certainly much less on average). In this reduced graph, the following concepts have the greatest centrality: *Property:Rights, Property:Source, Template:Esoteric, Rdfs:subClassOf*. The number of components (not counting isolated vertices) in this graph is significantly larger than in the previous case (101 components), since commonly edited concepts had been removed and they connect many of the rarely edited concepts into larger components.

3.1.3. Users and Concepts

Bipartite graph of the users and the aggregated concepts they are editing (see Figure 7) show the most active users and the concepts they are editing. For instance, *Markus Krotzsch* has edited *MainPage* 100 times, *TemplatePerson* 67 times, *Help* 152 times and *MediaWiki* 125 times.

Checking all the concepts and users shows that the two most active users *Patrick* and *Skierpage* have edited 359 same concepts. One of the more illustrative cores of this bipartite graph is shown in Figure 8.

3.1.4. Using Non-aggregated Concepts

Even though we have opted for merging concepts prior to collaboration analysis, based on the assumption that the concepts within the formed groups are likely to be similar regarding the editing user groups, there is certainly some merit in taking a closer look at the graphs produced without the mentioned merge, most of all at the concept graph obtained in such a manner. The reason for this lies in the fact that the performed merge disregarded any ontological information about concept relations and was based on syntactical similarities instead. Naturally, not all similarity relations could have been captured without some semantic context. Some of those lost similarities can be found in this concept graph and outlining them might provide us with yet another insight into the wiki editing process.
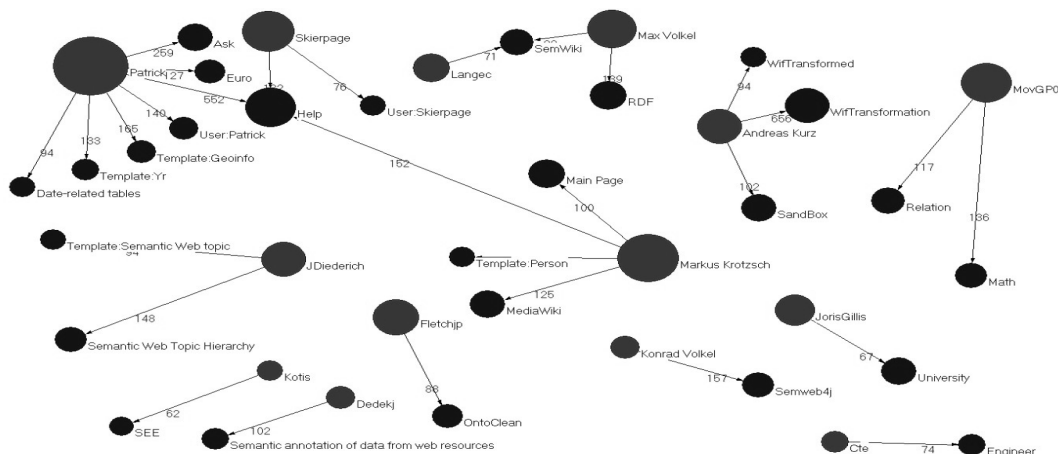

Figure 7. Graph of users and the most frequently edited concepts (over 60 edits per concept).
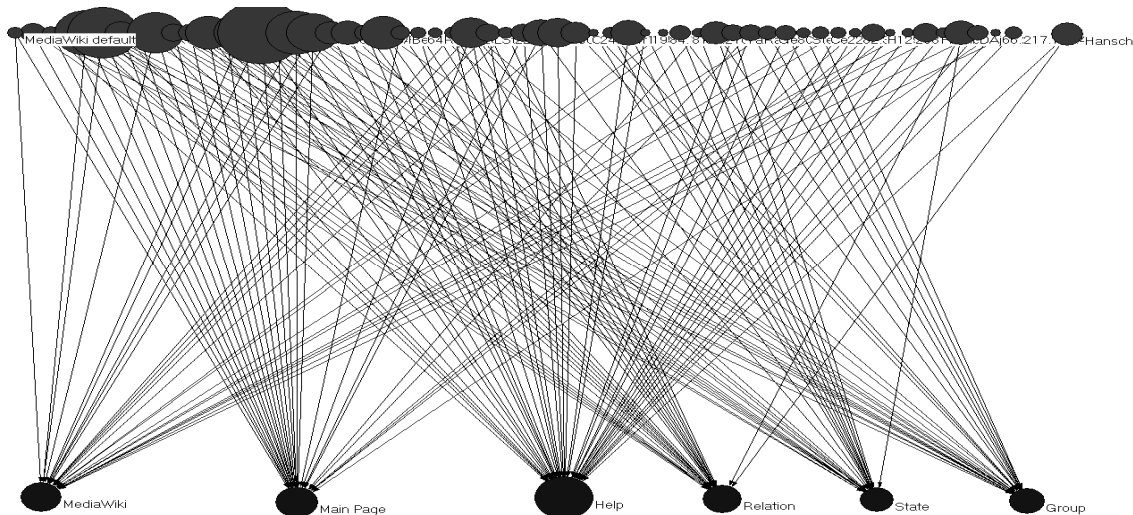
Figure 8. Core (2, 24) of bipartite user-concept graph clearly showing the most frequently edited aggregate concepts: MediaWiki, Main Page, Help, Relation, State, Group.

This graph of basic, unmerged concepts, consists of 7369 vertices and it's density is 14,4% – which is quite similar to the above discussed graph of merged concepts. There are more connected components of at least two vertices in this graph (81, to be more precise), which was to be expected since some of the isolated nodes in the merged graph were basically expanded into one or more separate connected components here.

One of the interesting correlations between concepts was discovered for northern European countries, namely Norway, Sweden and Denmark, as shown in Figure 9. Even though all three edges have the weight 7, it were not the same 7 users that have edited all three pages, so it is just a matter of coincidence. On the other hand, there are some users present in all three relations, namely *Patrick*, *Skierpage* and also some anonymous users (given only by their IPs).
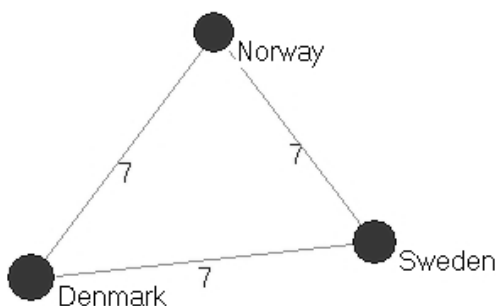
Further inspection of the graph also suggests the connection between pages *United States of America* and *Germany*, as well as within a group of similar concepts: *Semantic Wiki, Semantic Wiki mailing list, Semantic Wiki interest group, Semantic Wiki State of the Art, Semantic Wiki State of the Art Paper*. This last group of connections points to the fact that the concept grouping based on syntactic similarity wasn't completely illogical.

This graph also has some interesting cores, for example one containing: *Neuron, Neuronal stem cell, Electrically responsive cell, Electrically signaling cell, Electrically active cell, Molecule*, but also *Math, Minimum Spanning Tree* and some other concepts not related to neural activity. This type of relations shows that some of the contributors to the Wiki want to contribute to various parts of the ontology.

3.2. Graphs Constructed from AGROVOC Edit Logs

We will not go into details of the graphs of terms and users, since that perspective has already been covered in the previous chapters. Instead, we will give the basic properties and describe other graphs.

The term graph has 280 nodes, and 16.5% density. There are 21 connected components, the largest of which contains 110 nodes, and there are two isolated nodes.

Figure 9. Example of concepts correlated based on editing by the same user.

User graph is interesting because it only has 3 edges! This is certainly due to the fact that AGROVOC is a multilingual thesaurus, so people only edit concepts in the languages they speak, mostly the languages of the countries they live in – as well as the fact that this was a really small sample. Therefore, there is nothing more of any importance to say about user graph.

The terms were edited by users from 14 different countries. As it turns out, within the sample, there is no overlap between the countries in term editing, so there are no edges in this graph, one can just observe the relative sizes corresponding to the editing activity, for which one doesn't really need a graph, a pie chart would be quite sufficient. On the other hand, this is probably not true for the entire data set, but we didn't have access to a larger sample to verify this.

In the case of AGROVOC data, the most informative views of editing were when observing the editing on the ontological level, instead of the level of simple terms. The graph given in Figure 10 shows the country-ontology editing graph. It is basically an aggregation of the graph of user ontology editing by collapsing all the user nodes from the same country into a country node. The countries in Figure 10 are given in yellow coloring, while the ontologies are displayed green. We see that most edges go into the Food ontology, which we already concluded was the most edited one. Note that the country-ontology editing graph doesn't get much bigger even when there is more data, so

it is quite readable. On the other hand, in larger data samples, some subgraphs would have to be observed in the case of the user-ontology graph to perceive relationships more easily.

4. Analysis and Visualization Plugin

The presented results of analysis were obtained using Pajek [1] tool. However, our goal was to make a lightweight, easy-to-integrate tool which could help people gain a quick overview of the current state of editing process from the collaborative perspective.

In this particular case, the idea was to enable observing the progress of ontology editing and running some basic analysis.

The application was developed as a plugin to NeOn Toolkit [7]. The goal of the NeOn project is to provide support for managing of semantic applications. The toolkit is an open source ontology-engineering environment based on the Eclipse platform.

The application runs in Java and uses JGraph libraries [9], which are the most common Swing graph visualization libraries used. The displayed graph is editable in a way that the users can move and resize the nodes with the mouse.

We created a GUI over a set of graph analysis functions. It has been tailored for importing
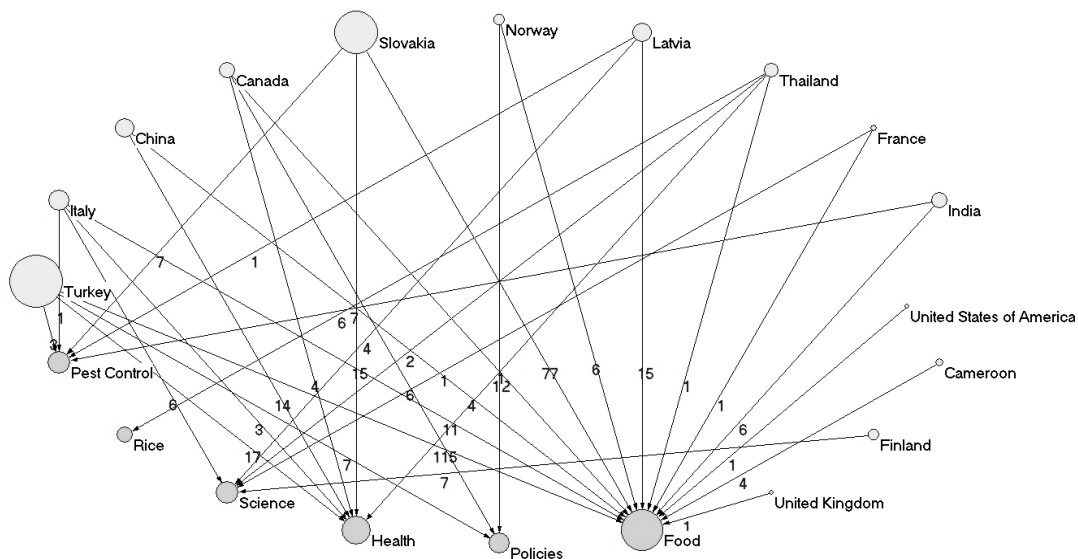

Figure 10. Country-ontology editing graph, displayed in a circular fashion.

AGROVOC logs, but some other data could also be imported. It accepts ARFF file format [8] on input, which is basically a coma separated file with attribute descriptions in the header. It is also possible to load files in format for Pajek tool and to export graphs in that data format, as well.

When ARFF file containing edits for AGROVOC ontologies is loaded, all the graphs are generated automatically in the target directory and the user can load them into the application for further processing and visualization one by one, in arbitrary order.

When visualizing a graph, initial vertex coordinates are calculated automatically. Two options are supported — the barycentric method [10] and Fruchterman-Reingold method [11], and the user can choose which one to use. There is a progress bar which indicates when the drawing is going to end in case of very large graphs with lots of vertices.

After the graph has been drawn, as displayed in Figure 11, there is an option to export the visualization as a JPG image file. It is also possible to write into a file the sorted list of nodes according to their degree and closeness centrality.

It is also possible to perform graph cuts, by selecting that all edge values and vertex degrees be at least some minimal values. It is also possible to get the connected components for a graph.

5. Conclusions

Both the Semantic Web Wiki (or any Wiki, for that matter) and FAO ontologies are a product of combined user effort and collaboration. Some aspects of that collaboration can be glimpsed by analyzing edit logs via application of social network analysis and graph visualization techniques.

Several graphs have been constructed to reflect both the relationship between actual participants – users, as well as conceptual objects, terms and ontologies, which are the products of user actions. We have demonstrated how these graphs can be constructed both when using or disregarding available ontological background information.

In the case of Semantic Web Wiki, it was determined that there exist a number of users which make unique contributions to the Wiki. Also, there is a number of very active users, a small group responsible for most of the current page edits. The importance of each of those users was determined according to several criteria. Areas of interest/expertise for such users have been detected and displayed in visual form (as graph cuts). Anonymous users (those for which only the IP was known) have been compared to the ones having a proper username. This comparison has revealed that even though the named users are on average much more active, there
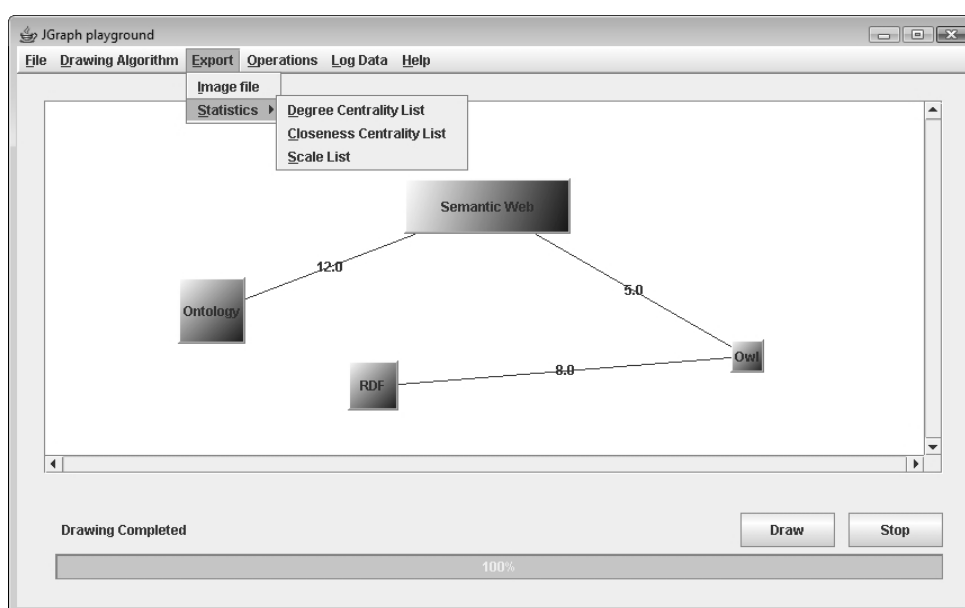

Figure 11. A view of the constructed plugin when visualizing a graph.

has been an amount of editing by the anonymous users which is not negligible and that there are some IPs displaying a higher activity than many users with a username. Both most frequently edited and least edited representative concepts have been determined, as well as groups of similar concepts based on several criteria (islands, cores, cuts). The most edited group of concepts is *Help*. User and concept graphs display some similarities, but also a significant difference in density.

A small sample of AGROVOC ontology editing data has been analyzed. The results clearly demonstrate the difference between the usual and multilingual setting. It was shown that it was possible to visualize data on the ontological level as well. Also, the information about users' countries might show the aggregate interest of local communities within those countries with respect to ontology editing. More data would be required to get some precise results for this dataset.

A plugin for NeOn Toolkit was developed in a form of a GUI supporting several graph analysis functions, as well as graph visualization. The plugin could also be used to analyze some other edit logs, but was crafted specifically for processing AGROVOC ontology editing logs. It supports a set of fast, lightweight methods for a quick overview of the current progress of collaborative ontology editing.

Acknowledgments

References

[1] V. BATAGELJ, A. MRVAR PAJEK, Program for Analysis and Evaluation of Large Networks. 2003.

[2] Semantic Web: http://semanticweb.org/wiki/Main_Page.

[3] S. WASSERMAN, K. FAUST, Social Network Analysis: Methods and Applications. United Kingdom, Cambridge University Press, 1994.

[4] SOUMEN CHAKRABARTI, Mining the Web: Discovering Knowledge from Hypertext Data. Morgan-Kaufmann Publishers, 2002.

[5] N. TOMAŠEV, D. MLADENIĆ, Semantic Web Wiki: Social Network Analysis of Page Editing. In *Proceedings of 31st International Conference on Information Technology Interfaces ITI2009*, 2009.

[6] AGROVOC: http://aims.fao.org/website/AGROVOC-Thesaurus/sub.

[7] NeOn project: http://www.neon-project.org/nw/Welcome_to_the_NeOn_ProProj.

[8] Attribute-Relation File Format: http://www.cs.waikato.ac.nz/~ml/weka/arff.html.

[9] JGraph: http://www.jgraph.com/.

[10] WILLIAM T. TUTTE, How to draw a graph. In *Proceedings of the London Mathematical Society*, 13(52), 743–768, 1963.

[11] T. FRUCHTERMAN, E. REINGOLD, Graph drawing by force-directed placement. *Software Practice and Experience*, 21(11), 1129–1164, 1991.

Contact addresses:
Nenad Tomašev
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: nenad.tomasev@ijs.si

Dunja Mladenić
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: dunja.mladenic@ijs.si

NENAD TOMAŠEV is a Computer Science PhD student at the Jožef Stefan International Postgraduate School in Slovenia, as well as being a mathematics student at the University of Novi Sad. He is also an associate at Petnica Science Center. His research interests are currently in the area of intelligent data analysis, ranging from simple numerical data to images and social network analysis. He first attended ITI conference in 2009 presenting a paper by Nenad Tomašev and Dunja Mladenić: Semantic Web Wiki: Social network analysis of page editing.

DUNJA MLADENIĆ is an expert on study and development of Machine Learning, Data Mining and Text Mining techniques and their application on real-world problems. Her current research focuses on data analysis, with particular interest in learning from Text/Web and Semantic Technologies. She is associated with the Department of Knowledge Technologies of the J. Stefan Institute since 1992. She got her MSc and PhD in Computer Science at University of Ljubljana in 1995 and 1998 respectively. She was a visiting researcher at School of Computer Science, Carnegie Mellon University, USA in 1996–1997 and in 2000–2001. She first attended ITI in 1993 presenting a paper by Dunja Mladenić, Ivan Bratko, Ray J. Paul and Marko Grobelnik: Using machine learning techniques to interpret results from discrete event simulation.