Analysis of Community Structure in Wikipedia (Poster)

Dmitry Lizorkin
Inst. for System Programming,
Russian Academy of Sciences
lizorkin@ispras.ru

Olena Medelyan
Computer Science Dept.
Univ. of Waikato, New Zealand
olena@cs.waikato.ac.nz

Maria Grineva
Inst. for System Programming,
Russian Academy of Sciences
rekouts@ispras.ru

ABSTRACT

We present the results of a community detection analysis of the Wikipedia graph. Distinct communities in Wikipedia contain semantically closely related articles. The central topic of a community can be identified using PageRank. Extracted communities can be organized hierarchically similar to manually created Wikipedia category structure.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation**]: Semantic Networks; H.3.3 [**Information Search and Retrieval**]: Clustering

General Terms

Experimentation

Keywords

Graph analysis, community detection, Wikipedia

1. INTRODUCTION

The category structure in Wikipedia is created by human contributors to thematically organize articles.[1] However, the information it reveals about the semantics of Wikipedia is not complete and not always reliable. For instance, it contains inaccuracies like *Domestic Pig* $\subset$*Pork* and cycles, like *The Beatles* $\subset$*Apple Records artists* $\subset$*Apple Records* $\subset$*Apple Corps* $\subset$*The Beatles*, where "$\subset$" means *belongs to category*.

It seems that Wikipedia is becoming too large for human contributors to grasp its knowledge organization without machine-aided tools. We propose a new approach to automatic hierarchical organization of Wikipedia articles. We apply a state-of-the-art community detection algorithm to Wikipedia graph to identify individual communities. These contain semantically related articles with one central topic computed using PageRank. Re-applying this analysis to large communities creates a meaningful hierarchy, which reflects the underlying link structure of the encyclopedia.

2. COMMUNITIES IN WIKIPEDIA

The notion of a *community structure* builds on a common characteristic of many complex real-world networks, like computer networks and social networks. Their graphs commonly contain densely-connected groups of nodes called *communities*, which are in turn sparsely connected with each other. Community detection algorithms analyze node connections in a graph and divide it into a number of such dense subgraphs, or communities. We use the popular Girvan-Newman algorithm [2] that creates graph partitions based on *modularity*: a measure comparing connectivity in a community division to that of a graph with random connections between nodes. Modularity values ranging from 0.3 to 0.7 indicate a distinguishable community structure.

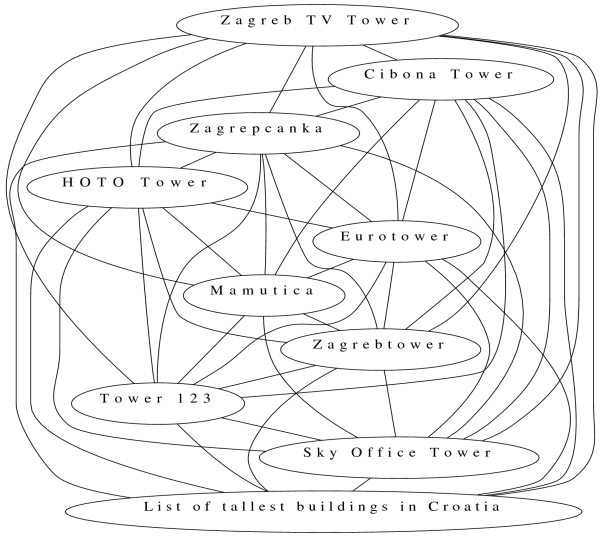

Figure 1: Connectivity in a sample small community

While other network analysis methods, like PageRank and HITS algorithms, were applied to the Wikipedia graph before [1], this is the first experiment analyzing its community structure. Given the original graph covering 4.1M nodes of the English Wikipedia,[2] we first, for computational reasons, restrict it to meaningful links: *see also*, *category links*, *links to main page*, *mutual links*, *links between articles from the same category*. This produces several intra-connected components, the largest with 1.1M nodes and 4.6M edges. Running the community detection algorithm on this component took over four days and produced 2,038 communities with a high modularity value of 0.63.

Although the largest community contains over 300K nodes,

[1]http://en.wikipedia.org/wiki/Wikipedia:Categorization

[2]http://en.wikipedia.org, snapshot of August 2008

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

	Community with 50 nodes		Community with 640 nodes	
Top PageRank	American football strategy	0.1542	List of railroad-related periodicals	0.1028
	Category:American football formations	0.0599	Category:Rail transport magazines	0.0542
	Trick play	0.0552	M250 series	0.0432
Middle PageRank	Swinging Gate (American football)	0.0133	Suwanotaira Station	0.0023
	Flea-flicker	0.0132	Category:Stations of Enoshima Electric Railway	0.0023
	Notre Dame Box	0.0130	Category:Railway stations of Japan by company	0.0023
Lowest PageRank	A formation	0.0030	Category:Enoshima Electric Railway Line	0.0002
	Fly (American football)	0.0030	Kawaguchi Station	0.0002
	Zone coverage	0.0030	Maruoka Station	0.0002

Table 1: Wikipedia article titles with top, middle and lowest PageRank scores in two sample communities

99% of communities consist of less than 2K nodes each. The distribution of their sizes is close to the power law, with a few large and a long tail of small communities,–a consequence of Wikipedia being a scale-free graph.

Individual communities, particularly small ones, contain heavily inter-connected nodes, which all relate to a common topic, see Figure 1. Evaluation of semantic similarity [3] of Wikipedia articles shows that average similarity of randomly chosen nodes is close to zero. However, nodes in communities up to 50 and 1000 nodes are on average 35% and 25% related, respectively.

3. MAIN TOPICS IN COMMUNITIES

To identify the central node, or main topic, in a semantically cohesive community, we apply the PageRank algorithm, where the highest PageRank score indicates centrality. For example, the top-scoring node in Figure 1 is the term *List of tallest buildings in Croatia*, which also corresponds to the human intuition.

Table 1 shows PageRank statistics for two communities with 50 and 640 nodes sampled at three levels depending on their scores. Terms in the 50-nodes community are all related to strategies and formation building in *American football*. The two top-scoring terms summarize the main subject as *American football strategy* and *Category:American football formations*. Terms with average and low PageRank values refer to specific strategies (*Notre Dame Box*) or terminology used to described these (*Zone Coverage*).

Analysis of the 640-nodes community is less straightforward. Here, two subjects appear to be mixed up: *railroad-related periodicals* and *Japanese railway stations*, with a common general topic *railway*. We conclude that top-ranked terms in smaller communities can be treated as their main topics, but larger communities first need to be further subdivided into more coherent subtopics.

4. INFERRING THE HIERARCHY

Following the above findings, we re-apply the community detection algorithm to large communities to derive second-level communities. The largest first-level community has been divided into 349 second-level communities. The modularity values of these divisions are again very high, confirming that such groupings are semantically meaningful. The size distribution again follows the power law.

We now recursively divide each large community into sub-communities using modularity values as quality indicators. This creates a hierarchy of communities as plotted in Figure 2. Each circle placed within another circle denotes a sub-community of a parent community, with circle sizes roughly illustrating the relative community sizes. Interestingly, in each recursive subdivision, the largest sub-community takes up approximately 25% of its parent community. These properties of hierarchical community structure in the Wikipedia graph are typical for scale-free graphs, and the extracted communities inherit these properties.

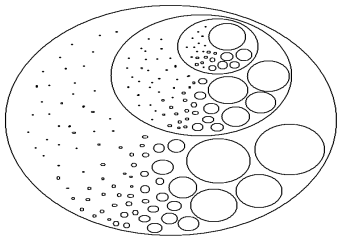

Figure 2: Hierarchically organized communities

5. CONCLUSIONS

The entire Wikipedia graph can be automatically organized into a hierarchy of communities comprising thematically related Wikipedia articles. Combined with the PageRank analysis to identify their central topics, we can automatically produce an ontological structure similar to the existing Wikipedia category tree. Evaluation of the accuracy of such structure will be a part of our future work, however the initial experiments demonstrate the potential of our method.

The community-detection analysis is fully language-independent. Thus, it will be particular useful for Wikipedias in languages, where a category structure is not as well developed as in the English Wikipedia. Furthermore, community detection analysis could be used to improve existing categories, created by humans without the knowledge of the integral hyperlink organization, or to augment Wikipedia search results with same-community terms.

6. REFERENCES

[1] F. Bellomi and R. Bonato. Network analysis for Wikipedia. In *Wikimania*, 2005.

[2] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[3] D. Milne and I. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Wikipedia and AI workshop at the AAAI*, 2008.