Graph-Theoretic Analysis of Collaborative Knowledge Bases in Natural Language Processing

Konstantina Garoufi Torsten Zesch Iryna Gurevych

{garoufi, zesch, gurevych}@tk.informatik.tu-darmstadt.de
Ubiquitous Knowledge Processing Lab
Computer Science Department
Technische Universität Darmstadt
Hochschulstr. 10, D-64289 Darmstadt, Germany

ABSTRACT

We present a graph-theoretic analysis of the topological structures underlying the collaborative knowledge bases Wikipedia and Wiktionary, which are promising uprising resources in Natural Language Processing. We contrastively compare them to a conventional linguistic knowledge base, and address the issue of how these Social Web knowledge repositories can be best exploited within the Social-Semantic Web.

1. INTRODUCTION

The Social-Semantic Web endeavor pledges to combine the expressibility and formal reasoning capabilities of the Semantic Web with the large amounts of human knowledge collaboratively constructed via the community-oriented techniques of Web 2.0. This would enable a new class of applications, in which the semantic relations latently existing in web-accessible data are automatically identified and aggregated in a network of structured knowledge. The Natural Language Processing (NLP) community has been making moves towards this vision, experiencing a perceptible shift from classical Linguistic Knowledge Bases (LKBs) like wordnets and ontologies to Collaborative Knowledge Bases (CKBs) as the background knowledge in applications. The latter have evolved by collective contributions of users participating in the Social Web and are constructed in a bottom-up rather than top-down manner, thus posing challenges due to their semi-structured and occasionally noisy knowledge representation model. It is not yet well-studied how traditional knowledge processing techniques can be suitably applied to CKBs. To address this issue, we examine graph-theoretic properties of the CKBs Wikipedia and Wiktionary[1], and compare them to a LKB, the German lexical semantic wordnet GermaNet [1].

[1] `http://www.{wikipedia|wiktionary}.org`

2. NLP KNOWLEDGE BASES AS GRAPHS

The knowledge bases examined consist of separate interconnected substructures that reflect different types of semantic relations. The structures of our interest are (i) the network of semantically related terms formed from Wikipedia's articles, (ii) the user-generated taxonomy (a.k.a. folksonomy) of the categories tagging Wikipedia's articles, (iii) the network of Wiktionary's entries, and (iv) as an instance of a LKB, the taxonomy of GermaNet's concepts.[2] We abstract over these different types of semantic networks with the formal notion of a directed **graph** G defined as a pair (V, E), where V is a finite set of elements called vertices or **nodes**, and E is a set of ordered pairs of nodes called **edges**.

Each Wikipedia article is viewed as a node of a graph (**WAG**). Each *hyperlink* between two articles is a directed edge between the article nodes, while articles redirecting to each other are represented as a single node. In Wikipedia's category graph (**WCG**), a node represents a category, whereas a directed edge between two nodes exists in case the two corresponding categories are connected by means of a *subcategory* relation. The directed graph modeling the structure of Wiktionary's entries (**WiktG**) represents each concept in Wiktionary as a node, whereby a concept is specified by surface form, language and part of speech. As nodes we include exclusively concepts defined by German language words. Two such nodes are connected with a directed edge if the corresponding concepts stand in one of selected semantic relations: *hypernymy, hyponymy, meronymy, holonymy, synonymy, antonymy, troponymy, coordination and see-also*. Finally, GermaNet is modeled as a directed graph (**GNG**) by representing each synset (i.e. set of synonyms) as a distinct node. The edges are given by the *hyponymy* relation. Representing the knowledge bases as graphs enables the direct application of graph-theoretic and social network analysis tools in order to characterize their topological structures.

The main elements that define the topology of a graph are its nodes and edges. In a directed graph $G = (V, E)$, the **out-degree** of a node is the number of edges leaving it, and its **in-degree** is the number of edges entering it. The sum of the node's out- and in-degree is its **degree** k. A **scale-**

[2] We use a snapshot of the German edition of Wikipedia from February 6, 2007 and a snapshot of the German edition of Wiktionary from October 9, 2007. The version of GermaNet we employ is 5.0, released in May 2006.

free graph has the property that the degree distribution of its nodes follows a power law $P(k) \sim k^{-\gamma}$, where the probability $P(k)$ that a certain node connects with k other nodes is roughly proportional to $k^{-\gamma}$ for some **power law exponent** γ. In a scale-free network, therefore, a small number of nodes have many connections, whereas most nodes have only a few. A **path** of length n from a node v to a node u is a sequence $(v_1, v_2, \ldots, v_n)$, where $(v_i, v_{i+1}) \in E$ for $i = 1, 2, \ldots, n-1$, $v = v_1$ and $u = v_n$. The length of the shortest path between nodes v and u is their **distance** $d(v, u)$. The maximum distance from v to any other node is its **eccentricity** $\varepsilon_v = max\{d(v, u) \,|\, u \in V\}$. The **diameter** of the graph is then defined as $D_G = max\{\varepsilon_v \,|\, v \in V\}$. The weakly **connected components** (CCs) are the equivalence classes of nodes under the *is-reachable-from* relation, whereby reachability between nodes is established by the existence of a path connecting them. The **largest connected component** (LCC) is the CC that is the largest in node size. We denote the **second largest CC** as LCC2.

The average distance between pairs of nodes in a connected graph is referred to as the graph's **characteristic path length** L_G. The clustering coefficient C_v of a node v is the fraction of the allowable edges between v's neighbors that actually exist, while the **clustering coefficient** C_G of G is the average of C_v over all nodes v in V. **Small-world** networks [4] have relatively low values of characteristic path lengths, comparable to the ones of random graphs, but much higher values of clustering coefficients than the ones expected by random graphs: $L_G \succeq L_{random}$ whereas $C_G \gg C_{random}$. The values for the corresponding random graphs are approximated as $L_{random} \approx \frac{\ln |V|}{\ln(\overline{k})}$ and $C_{random} \approx \frac{\overline{k}}{|V|}$, where $\overline{k}$ is the **average degree** over the nodes in V. Thus, in a small-world network most nodes are not each other's neighbors, yet are reachable from each other by relatively few hops.

3. ANALYSIS

The analysis focuses on LCCs, as connectivity is particularly important for several NLP tasks, e.g. the computation of semantic relatedness. The results are presented in Table 1. The connectivity analysis suggests the existence of a large portion of concepts in Wiktionary with few or no semantic connections. With almost half of its concepts practically not having been semantically related to any other concept, as the size of the LCC2 indicates, it is clear that the knowledge base is still at a premature stage of development. The highest connectivity appears in GermaNet. All graphs are found to be sparse, i.e. with an actual number of edges much lower than the possible number of edges that would correspond to a fully connected graph. Moreover, average degrees are low, ranging between approx 4 and 6. This indicates that all four knowledge bases encode semantic relations among their concepts only in a limited, selective way. With a relatively higher average degree, Wiktionary appears to be richer in the encoding of explicit semantic relations than the other knowledge bases. On all graphs, the distributions of the number of semantic connections between the concepts follow a power law, denoting that the graphs are scale-free. The diameters are small, indicating that the largest number of nodes having to be traversed in order to navigate between two concepts cannot be more than 28, even in a knowledge base of almost 40,000 concepts. The comparison of characteristic path lengths and clustering coefficients against corresponding random graphs demonstrates that the graphs indeed have the dynamics of small-world networks.

	WAG*	**WCG**	**WiktG**	**GNG**
$\|V\|$	38,594	38,057	20,011	42,129
$\|E\|$	80,567	74,975	33,650	99,130
#CCs	1	48	8,214	355
$\|V_{LCC}\| \,/\, \|V\|$	1.00	0.99	0.57	0.67
$\|V_{LCC2}\| \,/\, \|V\|$	-	< 0.01	< 0.01	0.21
$\overline{k}$ in LCC	4.18	3.94	5.80	3.82
γ in LCC	1.98	1.89	2.28	1.96
D_{LCC}	28	20	17	25
L_{LCC}	5.0014	6.9390	5.0290	8.7668
L_{random}	7.3897	7.6852	5.3074	7.6480
C_{LCC}	0.0120	0.0134	0.0822	0.0155
C_{random}	0.0001	0.0001	0.0005	0.0001

Table 1: Results of the graph-theoretic analysis. *The figures for the WAG correspond to a sample of size 7% of the original graph, created following [2].

These distinctive topological features shared by CKBs and LKBs (high degree of sparsity, a single CC containing the vast majority of concepts, small-world characteristics and scale-free pattern of connectivity) have also been found in many other biological, social or man-made networks, such as the WWW, and the principles of their large-scale structures are extensively analyzed in [3]. Models of semantic processing should be sensitive to these principles and adapt to the semantic structures of the knowledge bases, accounting for similarities but also differences. For example, the computation of concept relatedness using a path-based measure would not perform optimally on Wiktionary, which suffers from particularly low network connectivity. Further observations are made in ongoing work extending to centrality and link analysis, as well as content analysis of the networks.

4. ACKNOWLEDGMENTS

This work has been supported by the German Research Foundation (DFG) under the grants No. GU 798/3-1 and GU 798/1-3, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806.

5. REFERENCES

[1] C. Kunze. *Lexikalisch-semantische Wortnetze*, chapter Computerlinguistik und Sprachtechnologie, pages 423–431. Spektrum Akademischer Verlag, 2004.

[2] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, New York, NY, USA, 2006. ACM.

[3] M. Steyvers and J. B. Tenenbaum. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29:41–78, 2005.

[4] D. J. Watts and S. H. Strogatz. Collective Dynamics of Small-World Networks. *Nature*, 393:440–442, 1998.