Cleansing Wikipedia Categories using Centrality

Paolo Boldi
Dipartimento di Informatica
Università degli Studi di Milano
Italy
paolo.boldi@unimi.it

Corrado Monti
Dipartimento di Informatica
Università degli Studi di Milano
Italy
corrado.monti@unimi.it

ABSTRACT

We propose a novel general technique aimed at pruning and cleansing the Wikipedia category hierarchy, with a tunable level of aggregation. Our approach is endogenous, since it does not use any information coming from Wikipedia articles, but it is based solely on the user-generated (noisy) Wikipedia category folksonomy itself. We show how the proposed techniques can help reduce the level of noise in the hierarchy and discuss how alternative centrality measures can differently impact on the result.

General Terms

Networks; Categorization

1. INTRODUCTION

Folksonomies are collaborative attempts to categorize items of some type, with the aim of helping users in their searches (e.g., to have information on related items or to cluster items that are similar under some viewpoint). Wikipedia, the largest free-access collaborative Internet encyclopedia, is itself endowed with a folksonomy, that takes the form of a category hierarchy: each Wikipedia article is tagged with one or more categories, that are themselves structured in a collaborative hierarchical framework.

Under this point of view, Wikipedia can be seen as a knowledge graph with an explicit, human-authored form of article classification. Users interested in mining data from Wikipedia can naturally rely on categories as a further, precious information source (e.g., [2, 20, 16]).

Nonetheless, using Wikipedia categories without filtering is problematic, at best: the category hierarchy is extremely sparse and noisy, it contains duplications, errors and oversights, and it is more often than not too fine-grained to be directly employed.

In this paper, we propose an easy, tunable, endogeneous technique to cleanse and prune the category hierarchy. After briefly discussing the problems that the hierarchy exhibits, we focus on the usage of centrality measures to identify important categories and show how harmonic centrality [4] outperforms other alternative measures. The method we propose can be used fruitfully as a preprocessing phase in every algorithm that wants to exploit categories in mining Wikipedia data.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2891111.

2. RELATED WORK

Wikipedia has attracted an ever-growing academic interest as a crowd-sourced, openly-investigable source of information. For example, it is considered a good mean to study complex social behavior—e.g., to test the ability of centrality measures in capturing the stability of edits [17], or to analyze how online conflicts evolve [11]. Since it is the largest open encyclopedia in the world, it has been helpful in creating countless knowledge bases (e.g., DBPedia [2], YAGO [19]).

Another extremely valuable use of Wikipedia is enriching text processing with semantic information (e.g., [7, 16, 9]). Within this area, many researchers obtained good results in measuring semantic relations of concepts through information stemmed from Wikipedia [20, 18, 8]).

These works have shown how valuable category tagging and the Wikipedia Category Graph can be. The former is the bipartite graph where Wikipedia articles are tagged by ("belong to") one or more categories; the latter is the hierarchy specifying how categories are organized in subcategories (i.e., there should be a a link between x and y whenever "x is a subcategory of y"). Both of these graphs are completely user-generated and in continuous evolution.

Indeed, the category graph is far from perfect: since the very notion of "subcategory" is fuzzy and no universal policy is strictly enforced, the resulting hierarchy is not a forest, and not even a directed acyclic graph, as [11] pointed out. The category graph has been described instead as a thesaurus that combines collaborative tagging and hierarchical subject indexing by [22] and as an overlay between different trees by [15].

In fact, although many works heavily employed categories and the graph they form, all authors had to cope with the extrem level of noise one can find in them. In particular, the fact that the subcategory graph actually contains cycles forces users to take a cleansing step into careful consideration. Diverse techniques have been tried to do that. Common examples include considering only a tree based in a root category (from a global root [9] or a local one [24, 15]); others have arbitrarily removed cycles [7]. However, to the best of our knowledge, a work focusing on this *denoising* operation, comparing different techniques, was still missing.

Problem solving → Artificial intelligence → Cybernetics → Applied mathematics → Mathematical problem solving → Problem solving

Table 1: A cycle in the category pseudo-forest.

3. WIKIPEDIA CATEGORIES

Wikipedia articles are endowed with *categories*, intended to group together articles related to similar subjects; categories serve many purposes, like enabling users to browse sets of related articles, or enhancing the automated production of inclusion and navigation boxes. Categories are organized into a structure called *category hierarchy* that reflects the notion of "being a subcategory of"; according to the guidelines, the category hierarchy should correspond to a partial order[1], and should therefore be acyclic.

Nonetheless, like the rest of the Wikipedia effort, categories are created and edited collaboratively by users: as a result, the categorization process in Wikipedia is quite noisy and, in fact, a continuous work-in-progress: most importantly, the absence of cycles is neither enforced nor guaranteed. In fact, the *category pseudo-forest* (the directed graph whose nodes are the Wikipedia categories and with an arc from x to y whenever x is a subcategory of y) does contain cycles. Most of them are either consequence of a factual error by the Wikipedia editors or, more commonly, of the fact that the very notion of "being a subcategory" is not precisely defined. An example of a cycle[2] is shown in Table 1.

While a very large majority of categories (1 125 823, amounting to 99.22%) lie in a strongly connected component (SCC) of their own, there are some non-trivial SCCs, the largest one counting 6 833 categories. On the other hand, the graph itself is reasonably connected in the weak sense, with 952 833 (83.97%) nodes belonging to one single weakly connected component (WCC); most of the remaining categories (171 889, amounting to 15.15%) are isolated nodes (the 45% of these nodes are related to specific years, like "1833 births" or "Populated places established in 1864").

4. CLEANSING THE CATEGORY HIERARCHY

The presence of cycles is not the only form of noise that we find in the Wikipedia category pseudo-forest. Duplications, misplaced eponymous categories, excessive fragmentation are other problems that make a direct use of the hierarchy difficult at best. This was also highlighted in many previous works [6, 7].

It is natural to try to employ the Wikipedia articles belonging to each category as a mean to obtain a cleaner category hierarchy; for example, one may forget about the hierarchy structure altogether and try to reconstruct (a less noisy version of) it from the sets of articles belonging to each category. This approach, besides throwing away a large amount of human-cured data, has a quite serious (theoretical and practical) limitation: if one wants to use the category hierarchy to enrich the information on articles, using articles to get the category hierarchy is by all means a catch-22.

Our cleansing technique, instead, is completely endogenous (i.e., it uses only the information contained in the category pseudo-forest) and it consists of three phases.

Phase 1: Milestones determination.

In the first phase, we select the C topmost categories that we want to preserve (called *milestones*). While C is a parameter that determines the granularity of the output hierarchy, and it is set by the user, categories are ranked according to their centrality in the category pseudo-forest. The choice of which centrality measure [4] should be adopted will be discussed below. While selecting the C topmost categories, we expunged utility categories such as *"Categories by country"* and *"Main topic classifications"*.

Note that the milestones determined in this way have a natural hierarchy, that is the one induced from the original pseudo-forest, throwing away the hierarchical arrows that do not match the centrality score chosen (a supercategory cannot be less important than its subcategories). In other words, given two milestone categories x and y, we postulate x is a subcategory of y iff it was marked as a subcategory in the original pseudo-forest and the centrality score of x is smaller than that of y. This reconstruction process guarantees that the resulting hierarchy is acyclic.

Phase 2: Category remapping.

Once the milestones categories have been determined, each category is mapped to the closest reachable milestone (i.e., to the milestone category that is the more specific generalization of the category under examination). Categories for which no milestone is reachable are *orphan*, and they in fact disappear. The (partial) map from categories to milestones $\iota(-)$ is called the *category remapping*.

Phase 3: Article categorization.

At this point, each Wikipedia article is re-assigned to the remapped categories it belongs; more precisely, if an article was marked as belonging to categories $c_1, \ldots, c_k$ it is now mapped to the milestone categories $\iota(c_i)$ $(i = 1, \ldots, k)$ for which $\iota(-)$ is defined. If all the categories it belonged to are orphan, the article itself remains an orphan. This phase is not strictly part of the category-hierarchy cleansing, but it is necessary to use the cleansed category hierarchy as a folksonomy for Wikipedia articles.

5. CENTRALITY RANKING FOR WIKIPEDIA CATEGORIES

The most important steps in our cleansing procedure, that crucially determine its output, are the choice of C and the selection of the centrality score to be used.

[1]"Categories are organized as overlapping "trees", formed by creating links between inter-related categories (in mathematics or computer science this structure is called a partially ordered set)." [In `https://en.wikipedia.org/wiki/Wikipedia:Categorization`]

[2]All the experiments shown in this paper are based on the `enwiki` snapshot of February 3, 2014 (`enwiki-20140203-pages-articles` according to the Wikipedia naming scheme). This dump consists of 4 514 662 Wikipedia articles (with 110 699 703 links), each belonging to one or more categories; the associated category pseudo-forest contains 1 134 715 categories and 2 215 353 arcs.

5.1 Centrality Measures for the Wikipedia Pseudotree

In recent years, there has been an ever-increasing research activity in the study of real-world complex networks [23]; these are typically graphs generated directly or indirectly by human activity and interaction (and therefore dubbed "social"), and appear in a large variety of contexts and often exhibit a surprisingly similar structure. One of the most important notions that researchers have been trying to capture in such networks is "node centrality": ideally, every node in the graph (in our intended application: every category in the pseudo-forest) has some degree of importance within te network under consideration, and one expects such importance to surface in the structure of the social network; centrality is a quantitative measure that aims at revealing the importance of a node.

As explained in [4], the most used centrality measures can be broadly categorized into geometric measures (e.g., closeness centrality [3], Lin's index [12] or harmonic centrality [4]), path-based measures (e.g., betweenness [1]) and spectral measures (e.g., PageRank [14] or Katz's index [10]).

Note that we are here sticking to our principle of endogeneity and want therefore to avoid selecting important categories based on notions that are not internal to the category hiearchy itself.

In this paper, we took into consideration the network centrality measures that more widely used and appear to be better-behaved for these kinds of problems [4], namely, indegree (number of incoming arcs), closeness centrality, Lin's index, harmonic centrality, PageRank (with damping factor $\alpha = 0.85$), Katz's index (with parameter $\beta = (2\lambda)^{-1}$ where λ is the spectral radius). For comparison, we also considered the category popularity (i.e., number of articles belonging to that category). To compute geometric centralities in an efficient way, we employed HyperBall [5].

A further, widely used, endogeneous measure that was sometimes proposed for Wikipedia categories is their distance from the root category ("Articles") [9]. Albeit simple and natural, this measure has a number of drawbacks: first of all, it has a very limited granularity and a huge number of ties (see Figure 1). Moreover, the distance from the root is easy to spam and not very robust: one single misplaced subcategory link is enough to make a category more (or less) important than it should be. While the latter observation is probably irrelevant for the very top levels of the hierarchy (where errors and inconsistencies are easily spotted and corrected), the more crowded lower levels are certainly problematic.

The availability of many different ways to rank the categories immediately raises two problems: the first is whether (and how much) those ranking techniques differ in choosing the topmost categories; the second, in case the rankings are significantly different, is to understand which one is more suitable for our needs.

In order to answer the first question, we compared the various rankings using a variant of the classical Kendall's τ [21] that treats differently discordances depending on whether they happen at the top or at the bottom of the rankings, still handling ties in a proper way.

As Table 2 shows, there is a group of measures (Lin's index, harmonic centrality, PageRank, indegree and Katz's index) that strongly correlate to one another especially at the top, but appear to be much uncorrelated to the "Distance from the root" and, even more evidently, from "Closeness" and "Popularity".

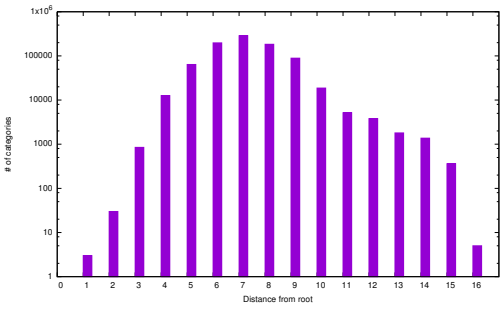


Figure 1: The number of categories depending on the distance from the root category "Article". 56% **of the categories is at distance** 6 **or** 7**.**

These differences make it urgent to answer the second question raised above, that is, which measure seems more "correct"; to answer this question, though, we need some ground truth on which categories are "relevant" in a broad sense.

Following [15] we decided to use an expert-curated bibliography classification. We decided to make use of the Library of Congress Classification[3], (LoCC) through the outline (main classes and subclasses) available from within Wikipedia. This choice allowed us to employ Wikipedia itself[4] to map LoCC classes to Wikipedia categories in the following way.

For each listed LoCC class (e.g., "Philosophy"), we followed the hyperlink (if any) to the related Wikipedia article (dropping the word "Outline" if needed; e.g., for the category tag "Philosophy" we ended up to the "Philosophy" article of Wikipedia); then, we joined all the categories of the article[5] ("Philosophy", "Academic disciplines", "Humanities" etc.). This process resulted in a set of 682 golden-truth categories assumed to be high-ranked, and we computed for each ranking the average precision (AP), the Normalized Discounted Cumulative Gain (NDCG), and the Area Under ROC curve (AUC) [13] in retrieving these categories; the results are displayed in Table 3, showing that Lin's index and harmonic centrality appear to be the best techniques under this viewpoint.

From the discussion so far we can conclude that harmonic centrality and Lin's index are the best centrality measures to identify milestones categories; we hereby preferred the former over the latter because harmonic centrality is more natural and it enjoys better theoretical properties [4].

5.2 Choice of C

The choice of C depends on the level of granularity we desire in the final output. Depending on the application, we may want different levels of aggregation among categories.

[3] `http://www.loc.gov/aba/cataloging/classification/`

[4] `https://en.wikipedia.org/wiki/Library_of_Congress_Classification`

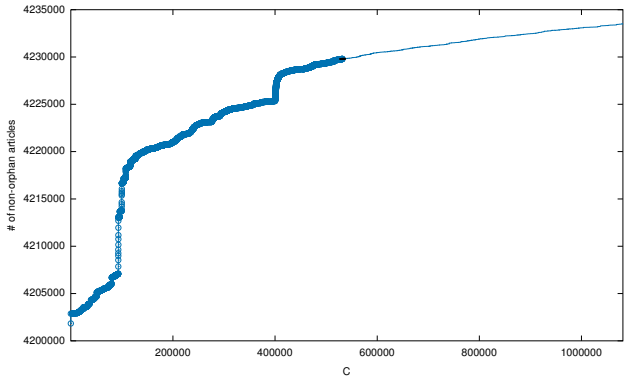
[5] As a more restrictive alternative, we only considered the category whose name matched that of the article, provided that it was present; this alternative approach produced a much smaller, less noisy, set of categories (205 instead of 682) but yielded essentially the same results.

	Lin	Harmonic	PageRank	Katz	Indegree	Distance from root	Closeness	Popularity
Lin	1.0000	0.9924	0.9539	0.9653	0.9497	0.6102	0.4010	0.1356
Harmonic	0.9924	1.0000	0.9545	0.9679	0.9490	0.6657	0.3665	0.1127
PageRank	0.9539	0.9545	1.0000	0.9547	0.9392	0.5345	0.4575	0.1683
Katz	0.9653	0.9679	0.9547	1.0000	0.9924	0.6200	0.4222	-0.0786
Indegree	0.9497	0.9490	0.9392	0.9924	1.0000	0.5646	0.4461	-0.0770
Distance from root	0.6102	0.6657	0.5345	0.6200	0.5646	1.0000	0.5810	0.3014
Closeness	0.4010	0.3665	0.4575	0.4222	0.4461	0.5810	1.0000	0.3842
Popularity	0.1356	0.1127	0.1683	-0.0786	-0.0770	0.3014	0.3842	1.0000

Table 2: The weighted Kendall's τ [21] between the centrality measures under comparison.

Centrality	AP	NDCG	AUC
Lin's index	**0.14641**	**0.71230**	0.94324
Harmonic centrality	**0.13914**	**0.70149**	**0.94444**
PageRank	0.07411	0.64503	**0.95001**
Distance from the root	0.05339	0.60720	0.92983
Katz's index	0.01606	0.53651	0.92708
Indegree	0.00917	0.50636	0.91532
Category popularity	0.00491	0.47994	0.90082
Closeness centrality	0.00083	0.40392	0.65134

Table 3: Comparison of various rankings in retrieving the golden-truth categories from the Library of Congress Classification, highlighting the best two rankings for each metric. We repeated the computation with many different shuffles of the tied scores, to see how ties influenced the result: it turned out that ties do not impact on the relative ordering of the measures; in fact the variance of the accuracy scores computed is quite limited, except for "Distance from root" (where, for example, the average precision ranged from 0.050686 **to** 0.056097**).**



Clearly, one further important factor in the choice of C is the number of categories (and articles) that are non-orphan if only C milestones are selected; while the former value increases steadily and smoothly, the latter is bursty and irregular (Figure 2): this is explained by the very poor correlation between centrality and popularity of categories (number of articles that belong to that category). The corresponding weighted Kendall's τ is only 0.113! It should be noted, however, that even extremely small values of C already give the 99% of article coverage.

5.3 Results with Harmonic Centrality

As justified above, harmonic centrality is apparently the most powerful among all centrality measures we tried. To give an idea about the effectiveness of harmonic centrality in capturing the generality of categories, we report in Tables 4 and 5 the first and the last categories on our list, when $C = 20\,000$. In Table 6 we show some examples of the induced category remapping: on average, each article belongs to 4 categories. As the examples show, this cleansing process yields very clean labels. Figure 3 shows the rank-size distribution of the number of categories remapped to each milestone category for various choices of C.

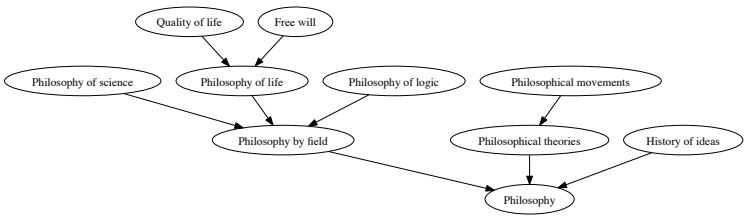
We also tested the cycle-removing procedure that we explained in Section 4: by combining the original pseudo-forest with the total order given by our rank, we can obtain a Directed Acyclic Graph. In Figure 4 we show how many arcs are discarded by our approach, with respect to the number C of categories to preserve. The plot indicates that with $C < 10\,000$ categories, the fraction of discarded arcs is approximately between 30% and 40%; increasing C, we discard fewer and fewer arcs. In other words, for the top thousands of categories removing cycles means discarding a significa-

t

C5.4

Original category c	**Substitution milestone** $\iota(c)$
Southern Tang poets	Poets by nationality
Antsiranana Province	Country subdivisions of Africa
Fellows of Magdalen College, Oxford	University of Oxford
Actresses from Greater Manchester	Greater Manchester
Guyanese slaves	History of South America
Swiss manuscripts	Swiss culture
Wilson Pickett songs	Songs by artist
Baroque architecture in Austria	Baroque architecture by country
Eastern Collegiate Roller Hockey Association	∄
Art schools in Washington (state)	Washington (state) culture
Rivers of Kostroma Oblast	Rivers by country
Flamenco compositions	Spanish music
Oil fields of Gabon	Geology of Africa
Basketball teams in Georgia (U.S. state)	Basketball teams in the United States by state
2004 in Australian motorsport	2004 in sports
Populated places established in 1821	∄
Elections in Southwark	Local government in London
Permanent Representatives of Norway to NATO	Ambassadors of Norway
Basketball in Turkey	Basketball by country
Balli Kombëtar	∄

Table 6: An excerpt of the category remapping process (using harmonic centrality with $C = 20\,000$ milestones). We write ∄ if there is no milestone category reachable from c.



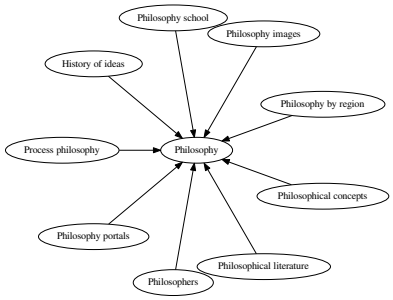
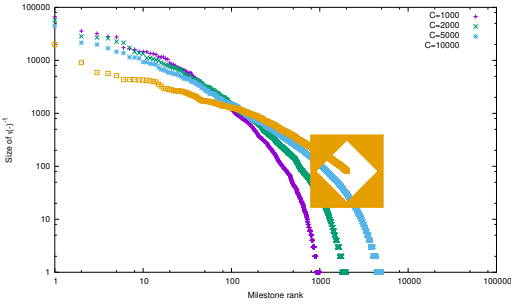


Figure 5: Sample from the cleansed category hierarchy (on the left, with harmonic centrality; on the right, for comparison, using "distance from the root").



909
9085
908
9075
907
9065
9055
905

Figure 4:

Figure 3: The mil

Rank	Category
1	Countries
2	Society
3	Nationality
4	Political geography
5	Culture
6	Humans
7	Social sciences
8	Structure
9	Human–geographic territorial entities
10	Contents
11	Geographic taxonomies
12	Fields of history
13	Places
14	Humanities
15	Continents
16	Political concepts
17	Human geography
18	Subfields of political science
19	Articles
20	Subfields by academic discipline

Table 4: Topmost twenty Wikipedia categories according to their harmonic centrality in the Wikipedia category pseudo-forest.

Rank	Category
19981	Maldives
19982	Gov. buildings on the Nat. Register of Historic Places
19983	Illinois waterways
19984	Bodies of water of Illinois
19985	2002 in association football
19986	Electronica albums by British artists
19987	Visitor attractions in Arkansas by county
19988	Years of the 20th century in Europe
19989	Commonwealth Games events
19990	Albums by English artists by genre
19991	American football in Pennsylvania
19992	Ethnic groups in Poland
19993	Card games
19994	Central African people
19995	Deaths by period
19996	Visitor attractions in Vermont
19997	Ancient roads and tracks
19998	People in finance by nationality
19999	Populated places in Greater St. Louis
20000	Religion in Poland

Table 5: Bottommost twenty Wikipedia categories (with $C = 20\,000$) according to their harmonic centrality in the Wikipedia category pseudo-forest.

[5] Paolo Boldi and Sebastiano Vigna. In-core computation of geometric centralities with HyperBall: A hundred billion nodes and beyond. In *Proc. of 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW 2013)*. IEEE, 2013.

[6] MS Fabian, K Gjergji, and W Gerhard. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International World Wide Web Conference, WWW*, pages 697–706, 2007.

[7] S. Faralli, G. Stilo, and P. Velardi. What women like: A gendered analysis of twitter users' interests based on a twixonomy. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[8] P. Jain, P.Z. Yeh, K. Verma, R.G. Vasquez, M. Damova, P. Hitzler, and A.P. Sheth. Contextual ontology alignment of lod with an upper ontology: A case study with proton. In *The Semantic Web: Research and Applications*, pages 80–92. Springer, 2011.

[9] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth. User interests identification on twitter using a hierarchical knowledge base. In *The Semantic Web: Trends and Challenges*, pages 99–113. Springer, 2014.

[10] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[11] A. Kittur, E.H. Chi, and B. Suh. What's in wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proc. SIGCHI conference on human factors in computing systems*, pages 1509–1512. ACM, 2009.

[12] Nan Lin. *Foundations of Social Research*. McGraw-Hill, New York, 1976.

[13] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[14] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library Technologies Project, Stanford University, 1998.

[15] A.A. Salah, C. Gao, K. Suchecki, and A. Scharnhorst. Need to categorize: A comparative look at the categories of universal decimal classification system and wikipedia. *Leonardo*, 45(1):84–85, 2012.

[16] P. Schonhofen. Identifying document topics using the wikipedia category network. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 456–462. IEEE, 2006.

[17] M. Sicilia, N.Th. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[18] M. Strube and S.P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424, 2006.

[19] F.M. Suchanek, G. Kasneci, and G. Weikum. Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203 – 217, 2008.

[20] Z.S. Syed, T. Finin, and A. Joshi. Wikipedia as an ontology for describing documents. In *ICWSM*, 2008.

[21] Sebastiano Vigna. A weighted correlation index for rankings with ties. In *Proceedings of the 24th international conference on World Wide Web*. ACM, 2015.

[22] Jakob Voss. Collaborative thesaurus tagging the wikipedia way. *arXiv preprint cs/0604036*, 2006.

[23] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Cambridge Univ Press, 1994.

[24] T. Zesch and I. Gurevych. Analysis of the wikipedia category graph for nlp applications. In *Proc. TextGraphs-2 Workshop (NAACL-HLT 2007)*, pages 1–8, 2007.