Visualizing Wikipedia for Interactive Exploration

Ron Bekkerman
University of Haifa
Haifa, Israel
ronb@univ.haifa.ac.il

Olga Donin
University of Haifa
Haifa, Israel
olgad@univ.haifa.ac.il

ABSTRACT

We aim to visualize (almost) the entire Wikipedia as a two-level coarse-grained / fine-grained graph representation of Wikipedia categories, for which we customize a hierarchy. We face the challenge of visualizing large scale-free graphs and propose an effective method for edge elimination that preserves the topical locality property of the original graph. The resulting visualization is sensible, traversable, and therefore actionable. It is a big step towards establishing comprehensiveness of Wikipedia as the collective memory of our and future generations.

KEYWORDS

Data Visualization, Interactive Exploration, Wikipedia

ACM Reference format:
Ron Bekkerman and Olga Donin. 2017. Visualizing Wikipedia for Interactive Exploration. In *Proceedings of KDD 2017 Workshop on Interactive Data Exploration and Analytics (IDEA'17), Halifax, Nova Scotia, Canada, August 14th, 2017 (IDEA'17),* 7 pages.
DOI:

INTRODUCTION

Wikipedia has de-facto become the collective memory of our generation [10, 21]. Our ancestors did not have the luxury of accessing a comprehensive memory bank. Over generations, people were considered intellectuals if they remembered a variety of facts and had a mental ability to integrate them into a compelling story [16]. We no longer need to develop a strong declarative memory. The classic model of human intellect [8] is to be adjusted to the new reality when the memory retention operation is effectively "outsourced" to the Web, and to Wikipedia specifically. Facts are – from now on – always at the tips of our fingers. And, remarkably, the content of our new outsourced memory is roughly the same for everyone. It is safe to say that the humankind is developing a collective intellect as our cognition is now based on the common, shared memory source.

There are many advantages of the collective memory as represented in Wikipedia. First, it never fails on us (as soon as, naturally, the Wikipedia website is accessible). We can always retrieve a missing fact, provided that we remember what to search for. Admittedly, Wikipedia is being constantly changed, some pages deleted while new pages added, the content of others updated. However, Wikipedia is never fading as human memory is. We can retrieve the same fact twice, many years apart, and chances are good that the fact will not change, regardless of our physical and mental wellbeing. Moreover, an argument can be made that Wikipedia is updating more slowly than the human memory is fading. For all practical purposes, our new collective memory is pretty static.

Second, in contrast to our biological memory that always plays tricks on us, Wikipedia is not changing inadvertently. Wikipedia pages are being added and deleted for a good reason, which is to always improve the content quality. Wikipedia is known for tending to objectiveness – opinionated reasoning is being aggressively fought against. Actually, the notion of objectiveness is very new in the context of human memory – we are never objective in our choice of facts to remember, nor we are able to keep our memories unaffected by our attitude towards them. Wikipedia, however, is widely considered unbiased [18], and facts presented in Wikipedia are perceived as correct. Indeed, they are verified by a community of highly qualified editors. While pure objectiveness cannot be possibly achieved, Wikipedia might be the most objective source of information that the humanity has ever had access to.

Third, and probably foremost, there is nothing mysterious about Wikipedia. While human memory has not been fully researched and some biological processes in our brains are yet to be understood, Wikipedia is just a few (million) pages in the Web that are – conceptually – trivial to grasp. Wikipedia pages hyperlink each other so its underlying structure is a graph [28], which we – computer scientists – are intimately familiar with. And whoever believes that a graph with a few million nodes is too large should not forget that they carry a graph of about 100 billion neurons to the north of their neck.

Being a conceptually simple notion, Wikipedia as our new digital memory allows answering questions that would sound completely outrageous were they asked about the human memory. One of the most exciting questions is comprehensiveness: does Wikipedia contain all the world's knowledge? Needless to say, asking such a question would make no sense in the context of human memory – no one would doubt its selectivity. A skeptical reader would argue that the lack of comprehensiveness characterizes Wikipedia just as well as the human memory. The proof might be straightforward: it is enough to come up with an example of a piece of knowledge that Wikipedia lacks. We, however, would like to offer two counterarguments. First, not every piece of knowledge has to be included in the world's collective memory. In fact, Wikipedia editors meticulously assess the value of each piece of knowledge to be presented on Wikipedia pages. Information that might not be in the general public interest is cold-bloodedly erased. This does not necessarily jeopardize the comprehensiveness of Wikipedia as knowledge can be effectively summarized to obfuscate auxiliary details.

Our second counterargument is: given a specific piece of knowledge, how does one know that this knowledge is not already in

IDEA'17, Halifax, Nova Scotia, Canada

DOI:

Wikipedia? We are used to applying keyword search to document repositories such as Wikipedia, but knowledge is not always easy to describe in a few keywords. Even if we applied many searches of many keywords, and did not find anything, would this mean that the knowledge we were looking for is not in Wikipedia or our search methodology is just not good enough? Based on the two arguments above, we may conclude that Wikipedia comprehensiveness is not that easy to contradict. Apparently, it is not easy to prove either.

Some work has been done on assessing comprehensiveness of several topics in Wikipedia, by mapping topical pages on a set of books published on the topic [9]. An attempt was made of assessing comprehensiveness of the entire Wikipedia [23], however the proposed methodology did not go far beyond word frequency computations.

To claim comprehensiveness or lack of comprehensiveness, one needs to first understand what it is out there in Wikipedia. What does our collective memory actually contain? When referred to memory, this question sounds both lunatic and thrilling at the same time. On the one hand, no one has dared to overview the entire content of memory. This would be impossible in the context of human memory which is an ever-changing, intrinsically complex, only partly studied medium. Even in the context of (English only) Wikipedia, this question is hard to answer. On the other hand, once answered, this question may lead to a breakthrough in a global understanding of our intellectual and cultural heritage which we (and our children) are substituting for our long-term memory. So, what does Wikipedia know? That is the question that we aim to answer in this paper.

METHODOLOGY

We will show how to climb 30,000 feet and view (almost) the entire content of Wikipedia in a digestible and actionable format. After exploring the content, we will be able to make decisions about which topics are missing in Wikipedia, which are underrepresented, and how our efforts need to be allocated to make Wikipedia the ultimate source of truth in all areas of human interest.

At the time of our bulk download (September 9, 2016), English-language Wikipedia contained 16,857,586 pages out of which 7,785,959 were redirect pages, 162,236 were disambiguation pages, and 3,830,032 were auxiliary pages, such as pages of Wikipedia categories, files, templates etc. After removing redirect, disambiguation, and auxiliary pages, we ended up with 5,079,359 content pages. We are on a quest to summarize five million Wikipedia pages.

From the classical Text Classification perspective [17], summarizing five million pages is not too hard: each page can be automatically categorized to one of N categories. Once all the pages are categorized, we would be able to summarize the entire Wikipedia as a ranked list of categories sorted by their frequency: say, N_1 pages on the topic of chemistry, N_2 pages on politics, N_3 pages on arts, etc. There are a number of deficiencies in this approach: (a) categories have to be chosen beforehand and might not directly correspond to the topics covered in the data; (b) text classification is error-prone – some pages will be misclassified; (c) choosing too few categories will lead to coarse-grained, imprecise categorization, while choosing too many categories will overcomplicate the categorization algorithm which would result in a large amount of misclassifications.

Fortunately, most Wikipedia pages are already categorized by their contributors: at the time of creating a Wikipedia page, a set of relevant categories has to be provided. Out of the 5,079,359 content pages, 4,913,089 pages belong to at least one category. Unfortunately, the entire number of Wikipedia categories is 1,303,021 which is only four times less than the number of content pages. Overviewing those categories would be as tedious as overviewing Wikipedia pages themselves. Nevertheless, Wikipedia categories hold the aggregation property such that content pages can be overviewed in groups whenever the corresponding categories are considered.

Creating a ranked list of Wikipedia categories is not an ideal way of overviewing Wikipedia. A one-dimensional interface of the ranked list – while being intimately familiar to us from our everyday interactions with search engine results – suboptimally exploits the area of the computer screen, and misses the advantages of using visual primitives such as color and shape [7]. A two-dimensional, graph-based representation would be more plausible for overview and exploration purposes.

We build a graph of Wikipedia categories, with nodes being the categories themselves and the edges being the (weighted) semantic connections between the categories as captured on Wikipedia pages: in 83% cases, a Wikipedia page belongs to more than one category. The more pages belong both to category A and category B, the stronger the connection between A and B is. When spread over a two-dimensional surface, the graph of Wikipedia categories naturally holds the topical locality property [5]: similar categories will be shown close to each other – which will allow easy exploration. At this point, it appears that all we are left to deal with is the graph's enormous size.

Since the times Wikipedia first got measured [27], attempts were made to visualize Wikipedia. Holloway et al. [11] generated an image with 79 thousand Wikipedia categories – which at that time was the overall number of categories. Needless to say, given such an enormous number of represented categories, this visualization is not appropriate for exploration. Moreover, the number of categories has increased 16.5 times since then, which makes the visualization of all Wikipedia categories no longer feasible. Pang and Biuk-Aghai [20] proposed a Wikipedia visualization in the style of a geographical map, while not attempting to achieve the visualization comprehensiveness. Silva et al. [24] visualized small graphs of Wikipedia pages hyperlinking each other. Some previous works dealt with visualizing Wikipedia dynamics: Brandes et al. [3] visualized Wikipedia's edit network, Kimmerle et al. [14] visualized knowledge evolution in Wikipedia.

Wikipedia categories are power-law distributed over the pages, with a long tail of categories each covering very few pages. In fact, 64% Wikipedia categories cover 90% of Wikipedia content pages. We decided to ignore the long tail and to visualize only categories covering 90% of Wikipedia pages, however those categories are still too many to visualize. Literature offers a variety of methods for visualizing large graphs, by using techniques such as edge clustering [4], edge bundling [12], and edge compression [6]. We did not adapt those techniques due to their imprecision, high complexity, and low scalability. Instead, we got inspired by the wealth of research on visualization of hierarchical information (e.g. [22]). If we

impose a hierarchy on the Wikipedia categories, we could visualize the graph of top-level categories each covering a large number of pages, while each top-level category could in turn be visualized as a graph of second-level categories.

Let us emphasize the fact that we need to visualize only the top two levels of Wikipedia category hierarchy – because the overall number of categories to visualize is under one million. If the hierarchy is carefully designed, e.g. second-level categories are uniformly distributed among the top-level categories, at any time we may show a graph with under $\sqrt{1,000,000} = 1000$ nodes. This number is manageable in a visualization – both in terms of layout and explorability. In a real-world situation, however, the uniform distribution is too much to require. Nevertheless, the number of categories is not expected to grow fast beyond a million, so the two-level hierarchy design will hold water years from now.

As a matter of fact, Wikipedia already offers a category hierarchy: most category pages are themselves listing one or more categories. However, Wikipedia category hierarchy is extremely noisy and not appropriate for visualization. Consider, for example, category *"BioShock"* which is a first-person shooter video game series. Traversing one path of the Wikipedia category hierarchy from *"BioShock"* upwards, we can see the following categories: *"BioShock"* → *"Dieselpunk"* → *"Retro Style"* → *"Nostalgia"* → *"Melancholia"* → *"Romanticism"* → *"German Idealism"* → *"Rationalism"* → *"A priori"* → *"Latin Logical Phrases"* → *"Latin Philosophical Phrases"* → *"Latin Words and Phrases"* → *"Ancient Rome in Art and Culture"* → *"Culture in Rome"* → *"Tourism in Rome"* → *"Rome"* → *"Renaissance Architecture in Lazio"* → *"Italian Renaissance"*. Apparently, the Wikipedia hierarchy is not a hierarchy but rather a network of associations. The longest path we could detect in this graph is of the length of 881. Besides, we detected 32,678 cycles in the graph, the shortest being of length 2, the longest – 829.

It is clear that we need to construct the category hierarchy of our own. We consulted with Kittur et al. [15] who mapped all 277 thousand Wikipedia categories of that time to 26 top-level categories, and then used the top-level categories to overview the content of Wikipedia. While Kittur et al.'s result is the closest to ours, we find it too coarse-grained, not explorable, and therefore not actionable. Milne and Witten [19] present a visual tool for analyzing Wikipedia which is, in contrast, suitable for exploration but too fine-grained: it does not provide an overview of Wikipedia. Suchecki et al. [25] investigated the evolution of Wikipedia category structure and concluded that it is quite stable, which implies that our results are unlikely to become obsolete any time soon.

We preprocessed the set of Wikipedia categories by first removing "technical" categories (that auxiliary pages belong to), such as categories containing the following phrases: *"Archived"*, *"COI-Bot"*, *"Created"*, *"Defunct"*, *"Deprecated Parameters"*, *"Did You Know"*, *"Disambiguation"*, *"Draft"*, *"DYK"*, *"Infobox"*, *"Lists of"*, *"Missing"*, *"Navigational Boxes"*, *"Nominations"*, *"Redirects"*, *"Requests"*, *"Templates"*, *"Uncertain"*, *"Unknown"*, *"Wikipedia"*, and *"Wikiproject"*. We also removed 70 noisy categories (categories in foreign languages, personal names, etc). Examples of noisy categories are: *"Living People"*[1], *"Births"*, *"Deaths"*, *"Nacional"*, and *"Michael"*. We mapped all plural words onto their singular forms. We then manually added 9 aggregation rules for all categories belonging to *"US States"*, *"UK Counties"*, *"Canada Provinces"*, *"India States"*, *"Countries"*, *"Towns"*, *"Years"*, *"Centuries"*, and *"National"* (into the latter, we aggregated nationality categories, such as *"German"*, *"Brazilian"*).

We are now ready to build the hierarchy of Wikipedia categories. For a category A, we denote $W(A)$ the set of words in the category's name. We create the category hierarchy as follows: category A is included in a more general category A' if $W(A')$ is a proper subset of $W(A)$. The resulting hierarchy is a DAG – circles are not allowed by definition. The depth of the constructed hierarchy is 6. An example of a depth-6 hierarchy is: *"College of Charleston Cougars Women's Basketball Players"* → *"College of Charleston Cougars Women's Basketball"* → *"College of Charleston Cougars Basketball"* → *"College of Charleston Cougars"* → *"College of Charleston"* → *"College"*.[2]

The top level of the category hierarchy contains 441 largest categories covering 90% of the entire Wikipedia. Those categories will be the nodes in our top-level graph representation. If two categories appear together on at least one Wikipedia page, we connect them with an edge. We end up having 68,764 edges in the top-level graph – the number that is way beyond the boundaries of aesthetic appeal. Besides the problem of the enormous number of edges, we face another problem: the graph is scale-free.

Visualization of scale-free graphs is difficult. In the majority of cases, the graph looks like an image of an explosion whose epicenter is a tangled bundle of edges with many separate branches sticking out of it in all possible directions. The larger the epicenter is, the messier the graph appears. To our surprise, literature on visualizing scale-free graphs is very sparse (see e.g., [13, 26]). Accepted approaches are mostly related to stochastic edge sampling, which does not really solve the aesthetics problem if the sample is large, while breaking the graph to disconnected parts if the sample is small. We propose a different technique for eliminating unnecessary edges.

As a preprocessing step, we need to eliminate low-weight edges (edges between categories that rarely appear together on Wikipedia pages). Unfortunately, in a scale-free graph of categories, the two endpoints of an edge might have dramatically different coverage, such that the number of pages on which they appear together can be negligible for one and substantial for another. A standard approach of using a universal threshold to filter out low-weight edges is therefore not applicable in this case. We eliminate an edge between two categories if they appear together on less than 5% of pages covered by either of them. The motivation for this choice is that the eliminated edge needs to be negligible for both its endpoints. For example, let us say that category A covers 100 pages, and category B covers 10,000 pages. Say, A and B appear together on 4 pages, which is 4% of A's coverage, and 0.04% of B's coverage. We eliminate the edge between A and B because it is negligible for both nodes. For the top-level category graph, applying this heuristic led to eliminating 93% of edges. Still, the remaining 4815 edges are too many for an aesthetic visualization.

We noticed that both the top-level graph and second-level graphs contain many triangles. Triangles tangle nodes while creating extra

[1] *"Living People"* is the largest category in Wikipedia, covering over 786 thousand pages. It is simply too common to be meaningful.

[2] Despite its apparent superiority over the existing Wikipedia category hierarchy, our hierarchy is not 100% error-proof. For example, the category *"Ambassadors of the United Kingdom to the Ottoman Empire"* was identified as a subcategory of *"Ambassadors of the Ottoman Empire"*.

ties between them. If we break each triangle by eliminating one of its edges, the distance between two previously adjacent nodes will then be 2, which will still preserve the topical locality property. The remaining question is which edge out of the three edges of a triangle we need to eliminate. The power-law distribution of node degrees in a scale-free graph naturally splits up to the head, body, and tail. Nodes from the distribution's head are connected to many others, while nodes from the tail are connected to very few, with the body nodes staying in between. For simplicity, the sets of nodes belonging to the head, body, and tail of the degree distribution will be called the first layer, second layer, and third layer, respectively. Inspired by the Hamiltonian ball model of Asratian and Oksimets [1], we propose the following algorithm for eliminating triangles in scale-free graphs:

(1) Eliminate edges that connected nodes of the same layer.
(2) Eliminate edges between nodes of the first and third layer.
(3) If the process above resulted in isolating nodes, restore one (arbitrary) edge per such node.

The logic behind this algorithm is in taking into account only connections between the first and the second layers, as well as between the second and the third layers. All the other edges would not matter: nodes of the second layer are likely to be connected to each other through the nodes of the first layer, while each node from the third layer is likely to be connected at least one node from the second layer (and if not, a connection will be kept to one node from the first or third layer).

Theorem 0.1. *The algorithm proposed above eliminates all triangles in the graph.*

Proof. Assume a triangle remained in the resulting graph. According to step 1 of the algorithm, there cannot be two nodes of the triangle that belong to the same layer. Thus, the only option for the triangle to exist would be when each of its nodes belongs to a different layer. However, according to step 2 of the algorithm, the resulting graph does not contain edges drawn from layer 1 to layer 3, which means that the triangle with nodes at each of the three layers is not possible. Edges restored at step 3 of the algorithm increase node degrees from 0 to 1, which implies that those nodes cannot participate in any triangle. □

Figure 1 is an example of a subgraph from the top-level graph before and after applying the triangle elimination graph – clearly, the resulting graph is more comprehensible. After applying the algorithm to the top-level graph, we eliminated 60% edges – and all 19,412 triangles. The distance between two previously adjacent nodes became 2.1 on average (that is, the topical locality of the graph is almost fully preserved).

RESULTS

The resulting visualization of the top-level graph is in Figure 2. All visualizations are obtained using the Gephi graph visualization tool with Fruchterman-Reingold rendering preprocessed by Force Atlas [2]. Larger nodes correspond to categories with higher coverage. As can be seen in Figure 2, the top-level categories split to four large groups: *Science and Society* (including history, religion, and technology), *Arts and Culture* (including films and television), *Places and Nature* (including flora and fauna), and *Sports*, while some ambiguous categories are referred to as *Other*. Percentage-wise, *Science and Society* covers 32.7% of Wikipedia, *Arts and Culture* 25.6%, *Places and Nature* 76.7%, *Sports* 16.0%, and *Other* 24.4% (obviously enough, these topics heavily overlap). It is not a surprise that *Places and Nature* covers more than 3/4 of Wikipedia – the majority of Wikipedia pages are location-bound. What is more of a surprise is that as much as 1/6 of Wikipedia deals with sports.

Figure 3 shows four examples of visualizing the top-level categories as graphs of their subcategories. Analogously to the top-level, in the second-level visualizations we decided to present only the largest subcategories covering together at least 90% of the category's pages. The top graphs in Figure 3 show two large categories (*"Districts"* and *"Descent"*) with over a thousand subcategories each, while the bottom graphs show two small categories (*"Models"* and *"Gold"*) with under a hundred subcategories each.

Our edge elimination methodology (low-weight edge elimination + triangle elimination) split the *"Districts"* graph to many small subgraphs, each representing a separate type of a district. Many such subgraphs look like flowers – those often correspond to a specific country and its districts (the central category is global, such as *"Districts of India"*, while the peripheral categories cover local districts). In the case of *"Descent"*, the vast majority of categories shown are quite homogeneous in their meaning: they cover pages of people of a certain descent. In this situation, separation of the graph to smaller subgraphs is infeasible. Our edge elimination methodology can, however, substantially detangle the complex network of connections between people of various descents. In the resulting visualization, areas can be clearly identified that correspond to people of European, Asian, Hispanic, and Middle Eastern descent. *"Models"* is an ambiguous category that got split in our visualization to two main subgraphs: scientific models and fashion models, with the latter being significantly larger in size. Category *"Gold"* was split to many more subgraphs, the largest of which is related to gold medals in sports. The average number of nodes in the second-level visualizations is 321, the average coverage is 94%. The original number of edges (before edge elimination) is 7942 on average, it goes down to 1321 after the low-weight edge filtering, and down to 609 after applying our triangle elimination algorithm. Each edge eliminated by the algorithm became a path of length 2.4 on average.

The website[3] presents the interactive system where each top-level node from Figure 2 is clickable and once clicked it unfolds into the second-level visualization, examples of which are shown in Figure 3. We invite encyclopedians, library and information scientists, philosophers, and subject matter experts to use our visualization for assessing the comprehensiveness of Wikipedia. Underrepresented topics can now be identified, and additional content may be created. Content creation in overrepresented topics might be slowed down. If rebuilt periodically, our visualization can capture the dynamics of content creation, which may lead to defining the general strategy of maintaining Wikipedia as our main source of factual knowledge.

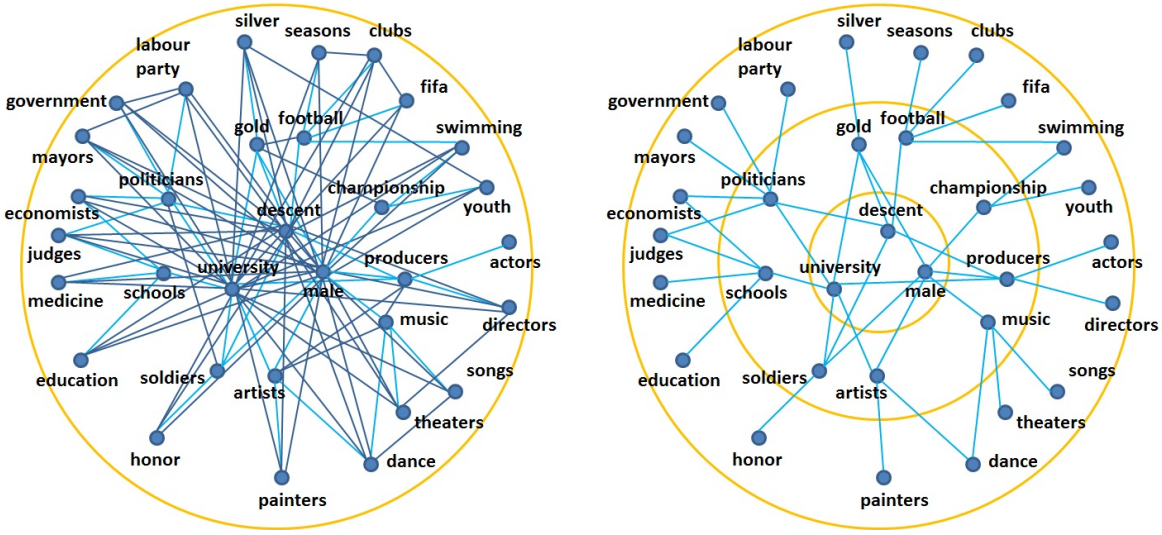
ACKNOWLEDGMENTS

We thank Dr. Roi Krakovski and Prof. Asher Koriat for fruitful discussions.

[3] http://the-wikipedia-viz-project.s3-website-eu-west-1.amazonaws.com/

Figure 1: A subgraph of the top-level Wikipedia category graph before (left) and after (right) applying the triangle elimination algorithm. Circles represent node layers.

-1in0in



REFERENCES

[1] A Asratian and N Oksimets. 1998. Graphs with Hamiltonian balls. *Australasian Journal of Combinatorics* 17 (1998), 185–198.

[2] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the International Conference on Web and Social Media*. 361–362.

[3] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. 2009. Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th international conference on World Wide Web*. 731–740.

[4] Weiwei Cui, Hong Zhou, Huamin Qu, Pak Chung Wong, and Xiaoming Li. 2008. Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1277–1284.

[5] Brian D Davison. 2000. Topical locality in the Web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and Development in Information Retrieval*. 272–279.

[6] Tim Dwyer, Nathalie Henry Riche, Kim Marriott, and Christopher Mears. 2013. Edge compression techniques for visualization of dense directed graphs. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2596–2605.

[7] Wilbert O Galitz. 2007. *The essential guide to user interface design: an introduction to GUI design principles and techniques*. John Wiley & Sons.

[8] Joy Paul Guilford. 1956. The structure of intellect. *Psychological bulletin* 53, 4 (1956), 267.

[9] Alexander Halavais and Derek Lackaff. 2008. An analysis of topical coverage of Wikipedia. *Journal of Computer-Mediated Communication* 13, 2 (2008), 429–440.

[10] Maurice Halbwachs. 1992. *On collective memory, edited and translated by Lewis Coser*. University of Chicago Press.

[11] Todd Holloway, Miran Bozicevic, and Katy Börner. 2007. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity* 12, 3 (2007), 30–40.

[12] Danny Holten and Jarke J Van Wijk. 2009. Force-Directed Edge Bundling for Graph Visualization. In *Computer graphics forum*, Vol. 28. 983–990.

[13] Yuntao Jia, Jared Hoberock, Michael Garland, and John Hart. 2008. On the visualization of social and other scale-free networks. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1285–1292.

[14] Joachim Kimmerle, Johannes Moskaliuk, Andreas Harrer, and Ulrike Cress. 2010. Visualizing co-evolution of individual and collective knowledge. *Information, Communication & Society* 13, 8 (2010), 1099–1121.

[15] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2009. What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1509–1512.

[16] Patrick C Kyllonen and Raymond E Christal. 1990. Reasoning ability is (little more than) working-memory capacity?! *Intelligence* 14, 4 (1990), 389–433.

[17] David D Lewis, Robert E Schapire, James P Callan, and Ron Papka. 1996. Training algorithms for linear text classifiers. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 298–306.

[18] Sorin Adam Matei and Caius Dobrescu. 2010. Wikipedia's fiNeutral Point of Viewfi: Settling Conflict through Ambiguity. *The Information Society* 27, 1 (2010), 40–51.

[19] David Milne and Ian H Witten. 2013. An open-source toolkit for mining Wikipedia. *Artificial Intelligence* 194 (2013), 222–239.

[20] Cheong-Iao Pang and Robert P Biuk-Aghai. 2011. Wikipedia world map: method and application of map-like wiki visualization. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. 124–133.

[21] Christian Pentzold. 2009. Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place. *Memory Studies* 2, 2 (2009), 255–272.

[22] George G Robertson, Jock D Mackinlay, and Stuart K Card. 1991. Cone trees: animated 3D visualizations of hierarchical information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 189–194.

[23] Cindy Royal and Deepina Kapila. 2009. What's on Wikipedia, and what's not...? Assessing completeness of information. *Social Science Computer Review* 27, 1 (2009), 138–148.

[24] Filipi Nascimento Silva, Matheus Palhares Viana, Bruno Augusto Nassif Travençolo, and L da F Costa. 2011. Investigating relationships within and between category networks in Wikipedia. *Journal of Informetrics* 5, 3 (2011), 431–438.

[25] Krzysztof Suchecki, Alkim Almila Akdag Salah, Cheng Gao, and Andrea Scharnhorst. Evolution of wikipedia's category structure. *Advances in Complex Systems* 15 (????).

[26] Tatiana Von Landesberger, Arjan Kuijper, Tobias Schreck, Jörn Kohlhammer, Jarke J van Wijk, J-D Fekete, and Dieter W Fellner. 2011. Visual analysis of large graphs: state-of-the-art and future research challenges. In *Computer graphics forum*, Vol. 30. 1719–1749.

[27] Jakob Voß. 2005. Measuring wikipedia. In *Proceedings of 10th International Conference of the International Society for Scientometrics and Informetrics*.

[28] Vinko Zlatić, Miran Božičević, Hrvoje Štefančić, and Mladen Domazet. 2006. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E* 74, 1 (2006).

Figure 2: Top-level Wikipedia category graph.

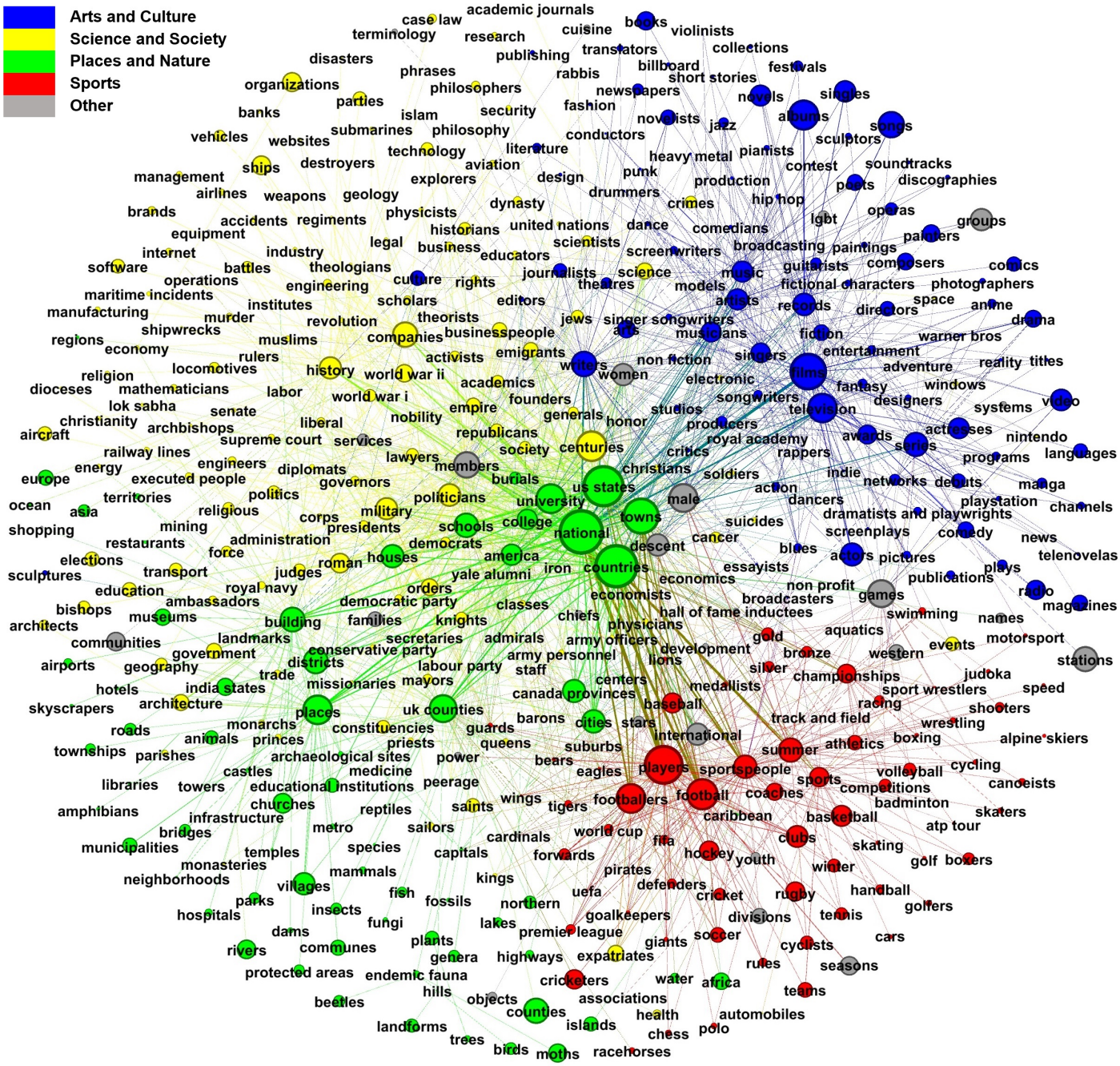


Figure 3: Examples of second-level visualization (of top-level categories *"Districts"*, *"Descent"*, *"Models"*, and *"Gold"*).