All Paths Lead to Philosophy

Dmitriy Brezhnev, Stephen Trushiem, Vikas Yendluri
{brezhnev, trusheim, vikasuy}@stanford.edu (Group 26)

***Abstract*—Have you ever tried to recursively click the first link of Wikipedia articles to see where it takes you? If you have, chances are you ended up at ”Philosophy.” Based on this observation, we hypothesized that articles on Wikipedia can be automatically categorized by generating a graph of Wikipedia pages based on their first link. Unlike human-curated categories of pages, which provide multiple possible parents for a page, these singular “is-a” relationships derived from natural text represent the most important (conceptual) parent of each Wikipedia article.**

We generated three “is-a” graphs over all English-language Wikipedia articles based on three different methods of determining the “first link:” one based on the naive approach, and two based on NLP techniques. We then automatically categorized pages based on their parent node in each tree, and found that the resulting categories had high precision into existing human-curated Wikipedia categories. This shows that automatic categorization of Wikipedia using the first link results in a useful set of categories.

I. Introduction

A. Background and Purpose

Wikipedia is a free online encyclopedia that contains over four million articles in English and over 30 million articles in total. [2] Like the Web, each article contains links to other articles; and in 2011, the online comic XKCD revealed an interesting property about the graph: if you keep traveling from article to article by following the first link, you will eventually end up at the article ‘Philosophy.’ [1]

This property is likely caused by a Wikipedia style convention that the first sentence of each article should “tell the non-specialist reader what (or who) the subject is.” [15] The XKCD result, then, indicates that every concept on Wikipedia derives from philosophy – an interesting proof of a epistemological concept. More formal research on this property showed that as many as 95% of all articles derive from Philosophy [10], and various websites have been created that visualize the path. [4]

Wikipedia currently relies solely on human-curated categorization for each article to induce an ontology of concepts. The average article has 4.2 human-curated categories, confounding efforts to determine what should be the “parent” category of a page is. For instance, the article “Barack Obama” has over 40 categories, and the most obvious categorization, “President of the United States,” appears 26th. However, the first sentence clearly states: “Barack Obama is the 44th and current President of the United States.” Based on these observations, we propose that the first-link property is more than just an interesting trick. In fact, the first link of each article is the conceptual parent of the article (X “is a” Y); and following all of those parent articles induces a tree, rooted at Philosophy, that categorizes almost all Wikipedia articles.

In this study, we develop algorithms that automatically categorize Wikipedia based on 3 different ”is-a” graphs: one graph based on the naive first link found in the natural text of the article, and two more graphs based on natural language processing of the first sentence of each article. We compare these automatic categorizations against human-curated Wikipedia categorizations, and show that automatic “is-a” categorization has high precision into human-curated categories, indicating its utility. We conclude by discussing the implications of automatic categorization and automated tools to investigate the different category trees we have generated.

B. Properties and definitions of Wikipedia graphs

Formally, we define the term “is-a” (in the context X “is-a” Y) to be equivalent to the term “hypernym.” For instance, the sentence “Barack Obama is the 44th and current President of the United States” indicates that Barack Obama “is-a” President of the United States. X is known as the hypernym to Y because the definition of Y sacrifices subtleties in the definition for a broader version of the definition of X (e.g. ”Red” is broader than “Ruby” or “Scarlet”). [16]

We define the “first sentence” of a Wikipedia article to be the first sentence of the main article text, which is technically unstructured language but, as previously discussed, typically follows a style convention resulting in extractable “is-a” relationships. We do not consider any structured information in this study (e.g. infoboxes, image captions, templates).

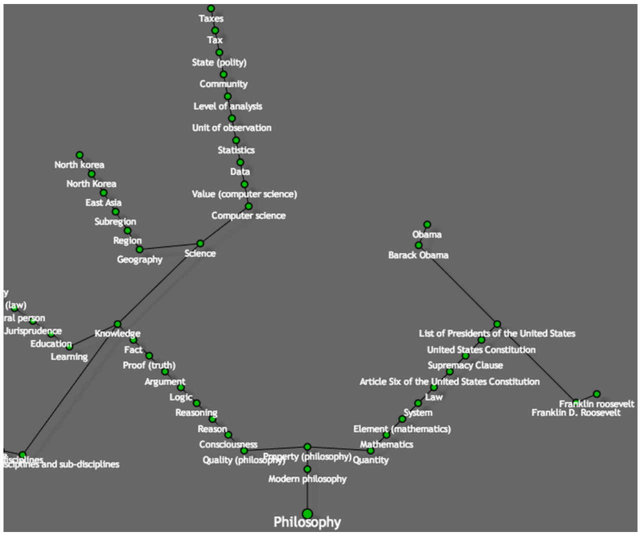


Fig. 1. Sample paths to ‘Philosophy’ [4]

The extraction of the general term in the “is-a” relationship relies on accurate classification of words into parts of speech. We used a maximum entropy classifier. The classifier assigns probability distributions to each part of speech combination and pick the one with the biggest entropy. This is a standard approach outlined in MaxEnt and is implemented by many standard packages. We used the Python NLTK package [8].

II. Prior Work

A. Categorizing Wikipedia

Categorizing Wikipedia has traditionally meant trying to label a fresh Wikipedia article with a set of new labels. In the past, researchers have approached the categorization of Wikipedia from three main angles: (1) assigning labels to articles based on NLP feature vectors; (2) deriving the labels after calculating article neighborhood properties (like PageRank/Hub scores); or (3) a combination of those two.

The NLP feature-vector approach measures relative importance of words that appear in the article to pull out the terms with the highest term frequency - inverse document frequency (tf-idf) scores. The terms get passed into a feature set and the documents are clustered based on their similarity. Extensions of this type of work combine features to come up with category labels: for example, see Gantner et al. [5]

The network structure / network feature approach has also shown promise, categorizing Wikipedia articles using PageRank-like scoring of network structure. Colgrove, Neidert and Chakoumakos showed that Wikipedia articles could be categorized with high precision using logistic regression on features such as percentage of out-links and in-links that belong to the given category [9]. While this method achieves very high precision, it requires developing a classifier for each category or subcategory on Wikipedia.

Our work expands upon these classifiers by making a combination NLP & network structure classifier that is easily scalable to the entire Wikipedia graph.

B. Linking to ‘Philosophy’

Previous work analyzing the “all roads lead to Philosophy” observation has been limited to finding basic statistics about linking to the ‘Philosophy’ article. Karonen showed that approximately 95% of pages on Wikipedia lead to the Philosophy article [10]. Karonen deduced this by extracting the first link of each page, skipping links inside of image captions, language translation information, “hatnote” sections, comments, and other wikipedia/mediawiki templates. Skipping this information allows for modeling the ”is a” relationship best. Kelcey was able to replicate these results, showing that of the pages on Wikipedia, only approximately 100,000 do not end at the Philosophy article. [11]. Our work extends the state-of-the art by using the “first link” observation to make useful observations about the conceptual structure of pages on Wikipedia.

C. Topic Extraction

Previous work on generating single categories for each page is typically called “topic extraction” and focuses on categorizing articles’ content into “topics” uncorrelated with the Wikipedia category tree. Previous work has successfully utilized the manually entered categories for extracting underlying concepts in documents that come from any domain. For example, [6] shows that getting Wikipedia titles and the associated categories is enough to produce high quality summarization data. The basic idea is to take an article and to retreat all document titles words of which are fully contained within the given document. The category probability is directly proportional to the tf-idf over Wikipedia titles (that is, inversely proportional to how frequently the words in the title appear across Wikipedia). The method utilized only article titles, yet showed promising results.

Similarly, the authors of [13] demonstrate how to successfully use an outside metric to improve ambiguities in human categories. They show how to use WordNet, a dictionary package that groups English words into groups of synonyms (synsets) and describes the semantic connections between synsets. The authors use WordNet to unambiguously map to category associated with a word using a distance metric, skipping the look-up of the article titles.

Our work extends this field by finding a new set of Wikipedia categories that are similarly contained within the natural text of the article, but also compares these categories to existing Wikipedia categorizations as a success metric.

III. Methods & Algorithms

A. Overview

Our study had four major components:

- *Category graph generation.* We generated a graph of the human categorizations for each page, and a graph on the ”first-category” of each page. These category graphs are our baseline to compare our automatic category generation against.
- *Is-a graph generation.* We generated 3 different ”is-a” graphs. One is based off the first link in an article: the *first-link graph*. Two others are based on natural-language processing of the first sentence of each article: the *NLP first-noun graph* and the *NLP first-noun-link graph*.
- *Automatic categorization of nodes in all graphs.* We automatically categorized each node (i.e., article) based on its parent node in each of the 3 is-a graphs.
- *Measuring category overlap.* We measured the precision, recall, F-score, and Jaccard similarity of each set of automatically generated categories to all human-curated categories. Taking advantage of the structure of the statistics, we implemented an all-pairs comparison that runs in approximately $O(5 * num\_nodes)$ time instead of the naive $O(num\_categories^2)$ approach.

Because our study relies on different analysis than is typically seen in the Wikipedia dataset, no acceptable dataset existed. We created appropriate datasets through this study, along with all of the supporting code to regenerate the dataset in the future.

Our study started with the full August 5, 2013 revision of the English language Wikipedia dataset (44GB; 14M pages). We split the Wikipedia dataset into 64 shards (approximately 700 MB / shard) prior to processing, and wrote custom batch-processing code to split processor-intensive tasks across three 64-core Xeon machines. Generally, our code uses Python 2.7.3, mwlib 0.15.12 [17], Snap.py 0.8.4 [19], and Snappyer [20].

We found that the August 5, 2013 revision of Wikipedia contained 4,436,660 articles (the remaining 9,525,687 pages were user pages, category pages, redirects, or other "meta" pages).

Together, our study provides a full toolchain from raw Wikipedia data to "is-a" graphs, automatic categorization, and analysis.

B. Category graph generation

We generated 2 baseline category graphs, based on the human categorization of Wikipedia.

1) Full human-curated category graph: To establish ground truth for article categorization, we generated a graph summarizing all the human-curated categorizations of every article. We created this graph by finding all instances of a Wikipedia category link occurring at the end of the article, following Wikipedia style conventions for categorization of an article.

The human-category tree contains 809,208 categories, each with an average of 15.72 articles (in-links) after eliminating outliers in the tail. The in-degree distribution of human-curated categories appears to follow a power law, with one category ("Living People") containing over 600,000 articles. A plot is available in Appendix A.

Each article has an average of 4.2 categories (out-links). The out-degree distribution, interestingly, does not appear to follow a power law.

2) First-category baseline graph: As a baseline to comparing metrics of any scheme that results in a single "most important" category, we generated a graph that limits each page into one human-curated category. For this graph, we assume that the first listed human-curated category is the parent category. Note that, as discussed in the introduction, this is a naive approach; nonetheless, it provides a baseline of maximum likely success. While it may be possible to perform better than baseline, it is very unlikely, as the baseline has knowledge of actual human-curated categories that will become the later benchmark.

C. Is-a graph generation: overview

We generated three different "is-a" graphs based on the contents of each Wikipedia article. Table 1 shows how each algorithm would extract node links from a sample sentence. Note that the phrase 'Classical Era' is actually the only url-link on Wikipedia.

Algorithm	**Extracted link**
First-link	'Classical era'
NLP first-noun	'prolific and influential composer' or 'composer'
NLP first-noun-link	none: *first noun not a link*

TABLE I

HOW THE DIFFERENT GRAPH ALGORITHMS PARSE THIS SAMPLE SENTENCE: *Wolfgang Amadeus Mozart, baptised as Johannes Chrysostomus Wolfgangus Theophilus Mozart, was a prolific and influential composer of the Classical era.*

D. Is-a graph generation: First-link graph

We generated a graph of first links of Wikipedia pages, based on the hypothesis that the first link out of a wikipedia article indicates the article's parent topic (If the first link on page X is to page Y, then X is a Y). Nodes in the graph are articles on Wikipedia. We filter out meta-pages and redirect pages in this graph, so that our graph only consists of nodes corresponding to true Wikipedia articles. A node X has a directed edge to node Y if the first link on the wikipedia page for X is page Y. We also filter out redirects in the graph. For instance, the Wikipedia pages for Definition of Philosophy and Philosopher both link to the main page for Philosophy. So if we have a page whose first link is to "Definition of Philosophy," we insert an edge to the node for Philosophy instead, and omit the node for Definition of Philosophy.

Note that we only consider internal links in our first link calculation. That is, links to other wikipedia articles. We use a Python library called article_parser.py to extract the first link; like previous research, this library ignores text in headers and sidebars on wikipedia pages so that the first link extracted is the first link in the article body [21].

E. Is-a graph generation: NLP first-noun graph

In this graph, we ignore the actual links to other Wikipedia pages. Instead, to establish connections, we find articles whose first sentence or paragraph follows the "is-a" pattern (e.g., "Barack Obama is the 44th President of the United States.") We extracted the noun following a recognized conjugation of the verb "to be." The algorithm then matched the noun to a Wikipedia page under the same title as that word.

In many cases, the noun was surrounded by adjectives and nouns that narrowed the breadth of the term, e.g. "44th President" rather than "President". To account for this, our algorithm attempted to match the specific term (including all adjectives and the noun); if that page could not be found, it would then attempt to match the general term (noun only). This behavior is shown in Table 1.

F. Is-a graph generation: NLP first-noun-link graph

This graph is a subset of the *NLP first-noun graph*. Article A is connected to a parent article if article A follows the "is-a" model and the extracted noun is also a link to some Wikipedia page. If these conditions hold, article A would be connected to the page under the title of the full link. Note that the link does not have to be the first link in the sentence, but rather is the first link following an "is-a" construction.

G. Is-a graph generation: Summary

Below is a summary of the 3 generated is-a graphs. Note that the number of edges is equal to the number of nodes with an ”is-a” relationship to another node. At most, each node has 1 ”is-a” classification, so the max out-degree of any node in the graph is 1.

Is-A Graph	Nodes	Edges
First-Link	4,033,513	4,016,239
NLP First Noun	3,705,897	3,687,466
NLP First Noun-Link	1,927,191	1,793,505

We represented each of these graphs in the format required by Snap.py to allow us to calculate weakly-connected-components and distribution of in-degrees [19]. The outcome of this analysis is listed in our results section.

H. Automatic categorization of articles in is-a graphs

For each generated ”is-a” graph, we automatically categorized each article based on its immediate parent in the graph. For instance, if the node for “Barack Obama” in the generated graph links to (has an is-a edge to) “President of the United States,” then Barack Obama is in the category President of the United States. This approach is naive, but we hypothesized that it would be sufficient to generate useful categorization.

Results from the automatic categorization step are provided in the results section.

I. Statistical testing of automatic categories

To validate whether the automatic categorizations were useful, we had to determine whether they contained substantially the same set of nodes as any human-curated category. This requires an all-pairs comparison between the human-curated categories (900k) and the automatic categorizations induced by each graph (1.2-2.5M). A naive approach would require $O(num\_human\_categories * num\_auto\_categories)$ comparisons, or approximately 1.2-3.0 trillion comparisons per graph, an intractable problem.

We engineered an all-pairs comparison that was optimized for sparse overlaps and runs in $O(num\_nodes * average\_categories\_per\_node) \approx O(5 * num\_nodes)$, requiring less than 30 million iterations per graph to complete an all-pairs comparison. The algorithm trades memory for time by moving node-wise across the graph and generating overlap counts for every pair of categories that actually overlaps. All other counts are zero, and a second iteration of the algorithm can easily compute metrics on existing elements of the overlaps dictionary. We tested the algorithm on four toy graphs, and verified that the results matched expected values.

Using this sparse-optimized all-pairs comparison, we computed the precision, recall, F-score, and Jaccard index of automatic categories induced by each graph, compared to the ”ground truth” human categories. We selected these metrics because they are standard success metrics for information-retrieval problems. [22]

We determined the maximum value of each metric for each automatic category; this maximum value corresponds to the human category with the highest overlap, which we assume to be the ”corresponding” human-curated category. We report these maximal scores in the Results section. We do not report Jaccard index because it does not add new information to the other metrics.

IV. Results & Analysis

A. Automatic category size distribution matches human-curated category size distribution, and follows a power law

We compared category size distribution (i.e., in-degree distribution) of our three generated is-a graphs to the real Wikipedia graph, and found that each plot obeys a power law with similar alpha.

We plotted in-degree distribution for all human-curated Wikipedia categories (as a baseline). In this graph, a node N has in-degree X if X articles have category N. This plot obeyed a power law with alpha = 1.64 (calculated using maximum likelihood estimate, or MLE). This makes sense because there a few, general categories with many articles, and many, specialized categories with few articles. The power law on this plot thus corresponds to the hierarchy of Wikipedia categories.

Interestingly, when we plotted in-degree distribution for our 3 generated is-a graphs, the plots also obeyed a power law (with alpha ranging from 1.87 to 2.57, estimated using MLE). Note that in each of these graphs, a node N has in-degree X iff X articles have an ”is-a” relationship to that node. The power-law here indicates that we have a few, general pages that are hypernyms to many pages, and many, specialized pages that are hypernyms to few pages.

The correspondance in power law alpha between our 3 is-a graphs and the real Wikipedia category graph thus shows that examining automatic categorization is merited. The few, general pages with high in-degree in our generated graphs may correspond to the few-general categories in the human-curated category graph.

A plot of these 4 graphs is available in Appendix A.

B. First-links graph confirms: all paths lead to Philosophy

We examined the largest weakly-connected components (WCCs) in each of our 3 is-a graphs. For the largest WCCs, we ran a breadth-first-search (BFS) out of every node in the WCC to see which nodes had the largest resulting BFS trees.

In the first-link graph, the largest tree came out of the node corresponding to the article for ”Philosophy”. This tree reaches 76% of the total nodes in the graph – that is, 3,065,038 out of 4,033,513 nodes in the first-link graph. This tree has depth 178. This result is consistent with others’ findings that approximately 95% of Wikipedia is rooted in ”Philosophy,” and we suspect the difference in absolute value is a result of our more careful elimination of redirect pages.

In the NLP first-noun graph, the largest trees came out of the pages for ”Collection”, ”Element”, ”Study”, ”Type”, and ”Differentation.” However, each of these trees reached only between 2 to 13% of the total nodes in the graph. We hypothesize that this is because many Wikipedia articles begins with ”X is a collection of Y” or ”X is a type of Y,” indicating a need for better parsing of this common idiom as well as a lack of common pages other than these general terms.

In the NLP first-noun-link graph, the largest trees cae out of the pages for ”Statistics” and ”Team sport.” However, these trees reach at most 3% of the NLP first-noun-link graph, and are thus insignificant. Because the total number of nodes in this graph (approximately 1 million) is only a quarter of the total pages in Wikipedia, it is likely that many nodes are missing edges between each other, hampering the potential of there being a large BFS tree out of a single node, like there was in the first link graph.

We also plotted the WCC size distribution of our 3 generated is-a graphs and compared it to the WCC size distribution on the baseline first-category graph. These plots obeyed a power law, with alpha ranging from 1.08 to 1.10 (estimated using MLE). These plots are available in Appendix B.

C. First-link automatic categories have high precision with human-curated categories

Figure 2 shows that 64.84% of the automatic categories generated from the first-link graph had precision of 1.0 with a human-curated category. That is to say, approximately 65% of the automatic categories induced by the first-link graph contained only articles that were also contained in one human-curated category. Further, 75.04% of the automatic categories generated from the first-link graph have precision of 0.5 or higher; our analysis showed that these cases typically occur when one page is miscategorized in a category of three pages.

Figure 3 shows some of the high-precision matches; our informal review confirms that they make sense as valid “is-a” relationships.

Importantly, the first-link automatic categories also cover over 90% of Wikipedia articles. The only pages that cannot be categorized occur when a page has no out-links in the first paragraph.

D. Automatic categories are “is-a” subsets of human-curated categories

Figure 2 also shows that automatic categories, generally, have low recall on human-curated categories. For example, only 16.32% of automatic categories generated from the first-link graph have a recall score of 0.5 or higher, approximately half of the baseline recall score. As a result, the F-scores of automatic categories are low; only 13.53% of automatic categories have a F-score of 0.5 or higher.

These results indicate that automatic categories are predominantly subsets of human categories for which the “is-a” relationship holds. For instance, the automatic category for “pickup truck” includes all the expected pages (models of pickup trucks such as Ford F-150 and Chevy Silverado), but the human-curated category “pickup truck” also includes hundreds of other pages – for instance, pickup truck transmissions, pickup truck manufacturing corporations, and a pickup truck conference. This leads to high precision, but low recall and thus low F-score.

Our analysis has also showed that human-curated categories frequently categorize broadly, lowering recall when we compare our “most-important” category to the whole human-curated category graph. The low baseline recall score of 0.4 indicates that any metric of “most-important” categorization misses pages that cross clean category boundaries: for instance, many hybrid truck/SUVs are described as “crossovers” in their article text, but are also human-categorized as both “Pickup truck” and “Sport utility vehicle.”

E. NLP first-noun-link graph performed best, but has limited coverage

Figure 2 shows that the “NLP first-noun-link” graph had the best performance of all graphs tested. Intuitively, this makes sense; if a Wikipedia editor wrote a sentence stating “Barack Obama is the President of the United States,” it is a sure sign that article has an “is-a” relationship with the linked entity, and that the linked entity is a specific concept. As a result, 78.13% of the categories created in the “NLP first-noun-link” graph have precision of 1.0, and over 84% have precision of 0.5 or higher. As shown in Figure 2, these statistics are all better than the “naive” first-link graph.

Despite these positive results, the NLP first-noun-link graph is less useful because it only covers 40.42% of Wikipedia articles, less than half of the first-link graph. This limited coverage indicates that this method must be combined with a back-up method for the majority of the graph where the natural-language processor cannot interpret a more precise “is-a” relationship.

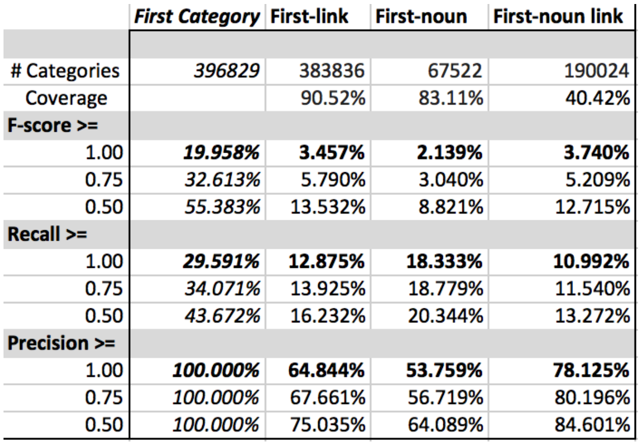
F. NLP first-noun categorization performed poorly

Figure 2 shows that the “NLP first-noun” graph resulted in precision and F-scores lower than the first-link graph. Combined with the difficulty seen in creating a useful tree from the data, we believe that this method is not a viable way to create “is-a” relationships from the Wikipedia. Our analysis has shown that this is likely due to complicated sentence structure that must be parsed through more advanced techniques, not simply finding nouns after a conjugation of “[to be].”

V. Discussion & Conclusion

This study has created a useful categorization of Wikipedia articles based on the “is-a” relationships found in the first paragraph of natural text of each article. We have compared the “naive” first-link approach (generating a graph based on the first link that appears in the first paragraph) with two more advanced approaches based on natural-language processing, and found that the naive approach resulted in the best combination of coverage of the Wikipedia graph and accuracy of categorization.

The categories we automatically generated from the first-link graph had high precision (64% had precision of 1.0) and low recall (16.23% had recall of 0.5 or less) when



	First Category	First-link	First-noun	First-noun link
# Categories	*396829*	383836	67522	190024
Coverage		90.52%	83.11%	40.42%
F-score >=				
1.00	***19.958%***	**3.457%**	**2.139%**	**3.740%**
0.75	*32.613%*	5.790%	3.040%	5.209%
0.50	*55.383%*	13.532%	8.821%	12.715%
Recall >=				
1.00	***29.591%***	**12.875%**	**18.333%**	**10.992%**
0.75	*34.071%*	13.925%	18.779%	11.540%
0.50	*43.672%*	16.232%	20.344%	13.272%
Precision >=				
1.00	***100.000%***	**64.844%**	**53.759%**	**78.125%**
0.75	*100.000%*	67.661%	56.719%	80.196%
0.50	*100.000%*	75.035%	64.089%	84.601%

Fig. 2. Meric scores from each graph (baseline in italic). Note that the baseline selects one human-curated category from the known list for each page, which is an upper limit on expected performance. More detailed plots are shown in Appendix C.

compared to the ground truth, human-curated categories. These scores indicate that automatic categories are typically subsets of human-curated categories. These metrics indicate that automatic categorization provides accurate article → category mapping, but may not result in a useful view of all pages in a category. However, as stated in [9], it is most important to create precise categorizations; mis-categorizing something is worse than not categorizing it at all, because many research applications depend on the accuracy of labels.

Automatic categorization based on the first link of an article provides a useful new metric for analyzing the Wikipedia graph because it allows research into conceptual "is-a" relationships based on only natural text for almost all of Wikipedia's four million articles. Our automatic categories encompass over 90% of Wikipedia articles, and most map precisely onto one human-curated category to allow further exploration.

During this study, we also found two interesting results: first, we confirmed previous studies indicating that most articles in Wikipedia are part of a first-link tree anchored at "Philosophy;" however, we found that the Philosophy connected component only contains 76% of Wikipedia articles, not 95%, after carefully eliminating redirect links as part of the article tree. And finally, we found that category size, regardless of the construction of categories (including human categorization, first-link categorization, and more advanced NLP categorization) follows a power law distribution with alpha in the range of 1.8 to 2.5.

Finally, to help others explore "is-a" relationships in Wikipedia, we have created a graph explorer that shows the different paths to root over each of our four generated trees.

VI. Future Work

Future research informed by our study centers around improving NLP classifiers to increase the utility of the first-noun-link graph. We are confident that further research to determine more edge cases could build upon our simple parser to dramatically improve the coverage of that graph (currently less than 45%) while keeping its high precision (currently over 78%). For instance, phrases that use terms such as "element of [important concept]" are currently linked to the word "element," where a more natural "is-a" classification would list them as part of "[important concept]."

References

[1] XKCD. "An Extended Mind." http://xkcd.com/903/

[2] *Wikipedia:Database download* . Found at http://en.wikipedia.org/wiki/Wikipedia:Database_download

[3] http://weblab.infosci.cornell.edu/papers/Bank2008.pdf

[4] Xefer philosophy tree visualization http://www.xefer.com/wikipedia

[5] Gantner, Zeno and Schmidt-Thie Lars. "Automatic content-based categorization of Wikipedia articles."

[6] Schonhofen, Peter. Identifying document topics using the Wikipedia category network. *ICWI, 2006*.

[7] Berger A., Della Pietra V., and Della Pietra S. A Maximum Entropy Approach to Natural Language Processing. http://acl.ldc.upenn.edu/J/J96/J96-1002.pdf

[8] Natural Language Toolkit. https://code.google.com/p/nltk/

[9] Colgrove, Neidert, and Chakoumakos. "Using Network Structure to Learn Category Classifcation in Wikipedia" http://snap.stanford.edu/class/cs224w-2011/proj/colgrove_Finalwriteup_v1.pdf

[10] Karonen, Ilmari. First link research. http://en.wikipedia.org/wiki/User:Ilmari_Karonen/First_link

[11] Kelcey, Mat. Wikipedia parser research. http://matpalm.com/blog/2011/08/13/wikipedia-philosophy/

[12] Krishnan Ramanathan, Yogesh Sankarasubramaniam, Nidhi Mathur, Ajay Gupta. 'Document summarization using Wikipedia'. 2009. http://www.hpl.hp.com/techreports/2009/HPL-2009-39.pdf

[13] Toral, A., Ferrandez, O., Agirre, E., and Munoz, R. 'A study on Linking Wikipedia categories to Wordnet synsets using text similarity.' RANLP, 2009. http://www.computing.dcu.ie/~atoral/publications/2009_ranlp_sem_sim_paper.pdf

[14] http://en.wikipedia.org/wiki/Jaccard_index

[15] Wikipedia Style Guide. http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section#First_sentence

[16] Wikipedia: Hypernym http://en.wikipedia.org/wiki/Hypernym

[17] MWLib Library, hosted on pypi. https://pypi.python.org/pypi/mwlib

[18] MWParser Library, hosted on pypi. https://pypi.python.org/pypi/mwparser

[19] Snap.py 0.8.4. http://snap.stanford.edu/snap/snap.py.html

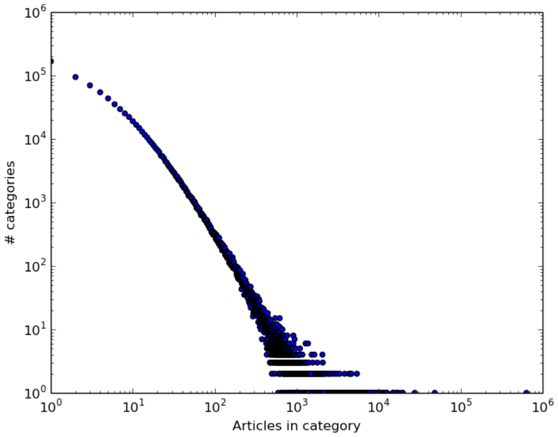
[20] Snappyer library, hosted on GitHub. https://github.com/trusheim/snappyer

[21] ArticleParser.py, hosted on GitHub https://github.com/matpalm/wikipediaPhilosophy/blob/master/article_parser.py

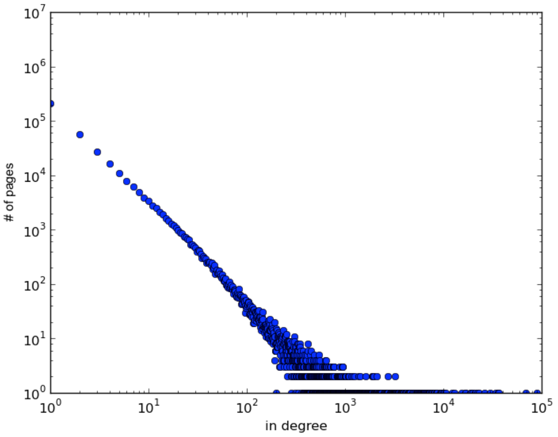
[22] Manning, Christopher, et al. Introduction to Information Retrieval: Evaluation of unranked retrieval sets. http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-unranked-retrieval-sets-1.html

Appendix A
In-Degree Distribution

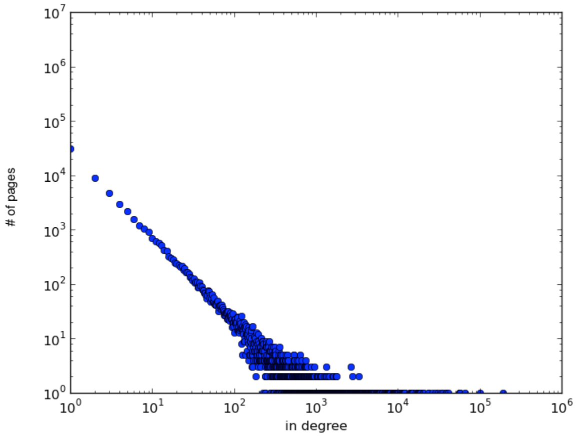
We plotted in-degree distribution for the graph of human-curated wikipedia categories as well as our 3 generated ”is-a” graphs. In each case, the distribution obeyed a power law.



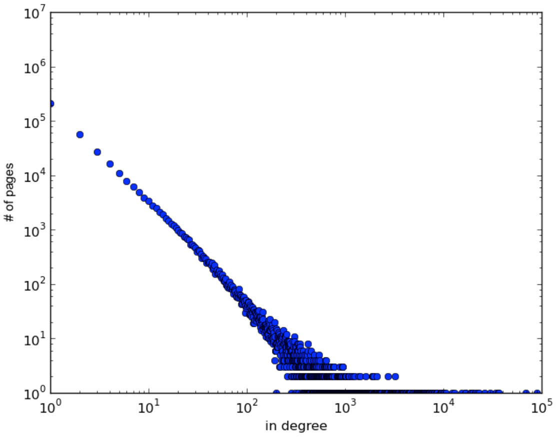
(a) Human-curated Wikipedia categories (baseline). Alpha is estimated to be 1.64, using MLE



(b) First-link graph. Alpha is estimated to be 2.33, using MLE



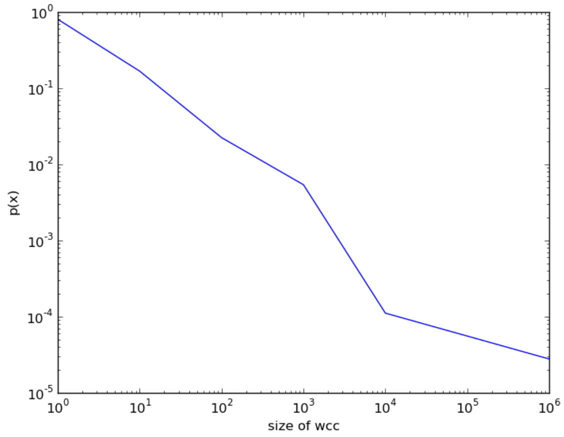
(c) NLP first-noun graph. Alpha is estimated to be 1.87, using MLE



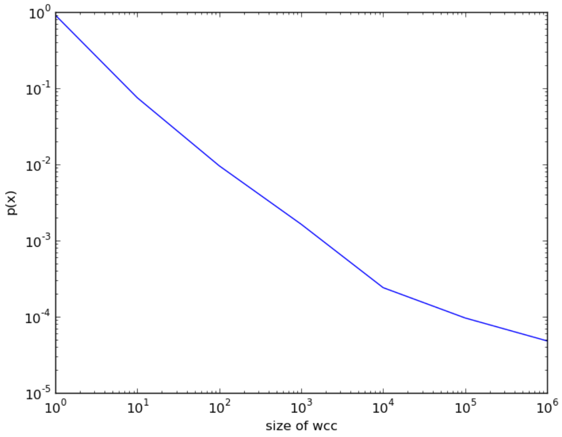
(d) NLP first noun-link graph. Alpha is estimated to be 2.57, using MLE

Appendix B
WCC Size Distribution

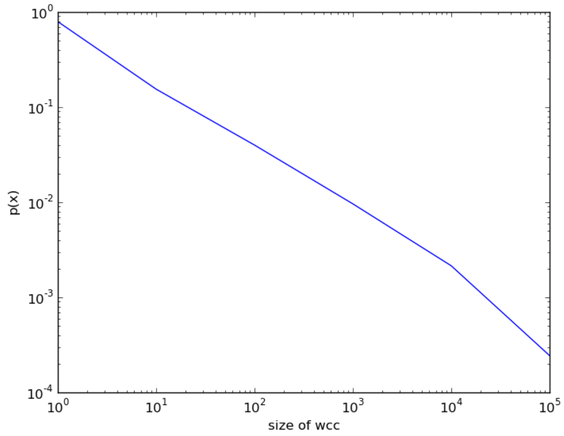
We plotted weakly-connected component (WCC) size distribution for the first-category graph as well as our 3 generated ”is-a” graphs. In each case, the distribution obeyed a power law.



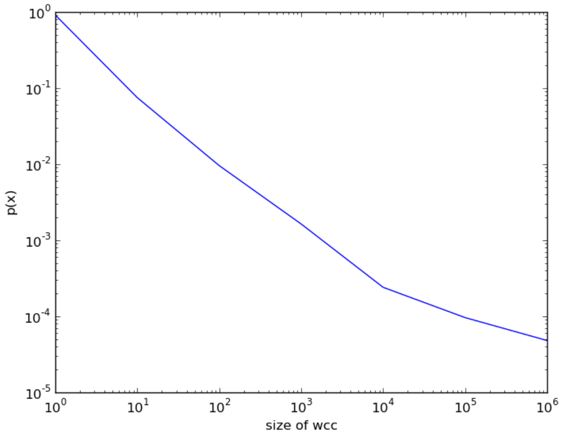
(e) First-category graph (baseline). Alpha is estimated to be 1.101, using MLE



(f) First-link graph. Alpha is estimated to be 1.103, using MLE



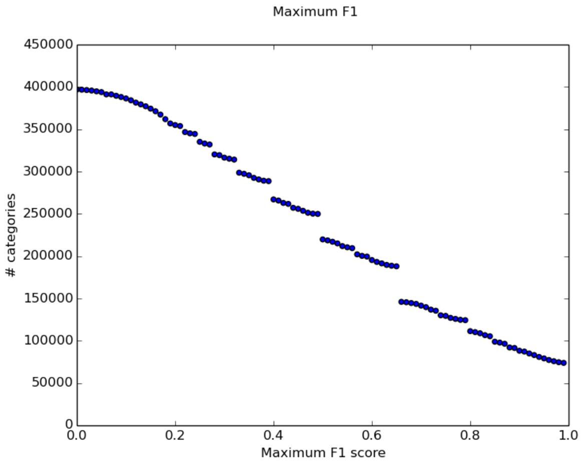
(g) NLP first-noun graph. Alpha is estimated to be 1.107, using MLE



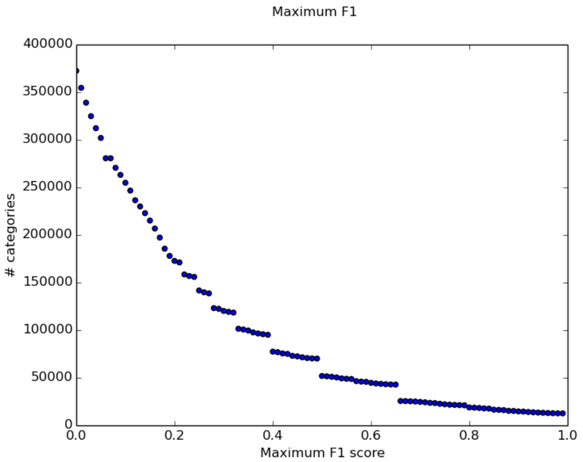
(h) NLP first noun-link graph. Alpha is estimated to be 1.08, using MLE

Appendix C
Graph Success Metrics

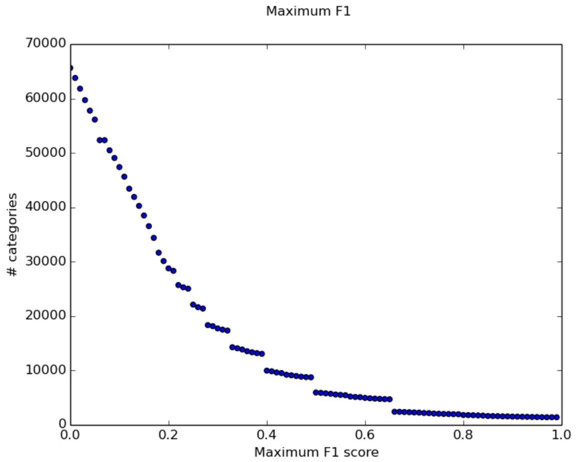
The precision, recall, and F-score of each method is shown here. The graphs are discontinuous at values corresponding to common division products (e.g. 1/3, 1/2, 2/3...). All values are rounded to their nearest 0.01 before plotting.



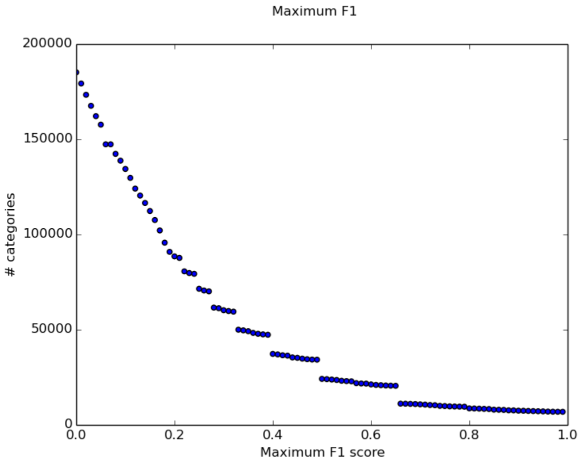
(i) Baseline first-human-category, cumulative F1 score



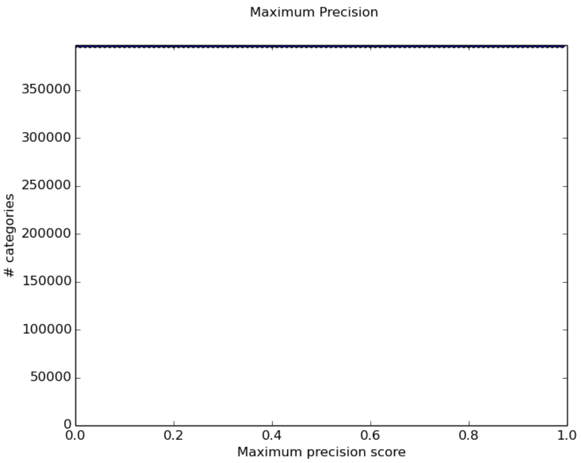
(j) First-link graph, cumulative F1 score



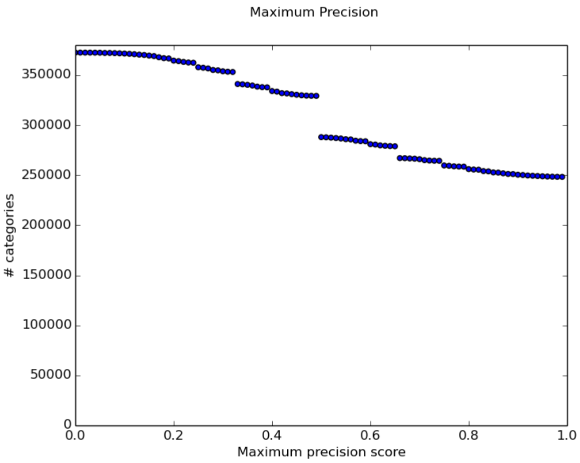
(k) First-noun graph, cumulative F1 score



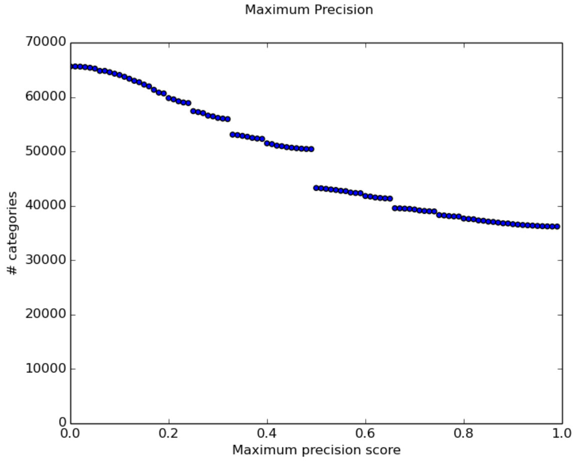
(l) First-noun-link graph, cumulative F1 score



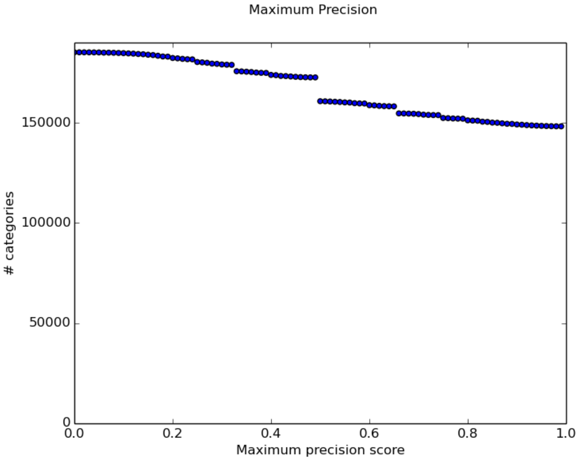
(m) Baseline first-human-category, cumulative precision score



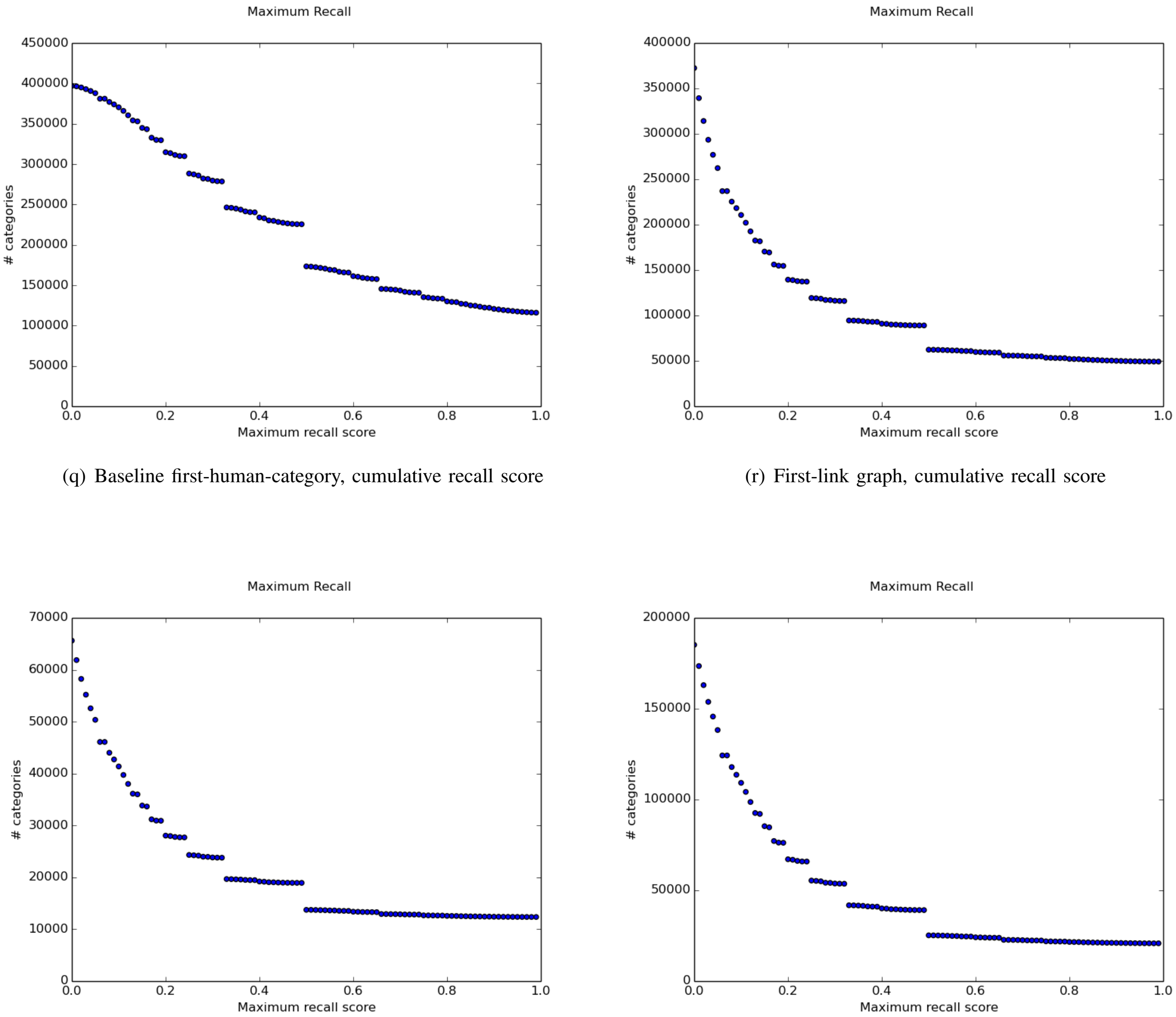
(n) First-link graph, cumulative precision score



(o) First-noun graph, cumulative precision score



(p) First-noun-link graph, cumulative precision score



(q) Baseline first-human-category, cumulative recall score

(r) First-link graph, cumulative recall score

(s) First-noun graph, cumulative recall score

(t) First-noun-link graph, cumulative recall score