What's in Wikipedia? Mapping Topics and Conflict Using Socially Annotated Category Structure

Aniket Kittur
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213 USA
nkittur@cs.cmu.edu

Ed H. Chi, Bongwon Suh
Palo Alto Research Center
3333 Coyote Hill Rd, Palo Alto, CA 94304 USA
{echi, suh}@parc.com

ABSTRACT

Wikipedia is an online encyclopedia which has undergone tremendous growth. However, this same growth has made it difficult to characterize its content and coverage. In this paper we develop measures to map Wikipedia using its socially annotated, hierarchical category structure. We introduce a mapping technique that takes advantage of socially-annotated hierarchical categories while dealing with the inconsistencies and noise inherent in the distributed way that they are generated. The technique is demonstrated through two applications: mapping the distribution of topics in Wikipedia and how they have changed over time; and mapping the degree of conflict found in each topic area. We also discuss the utility of the approach for other applications and datasets involving collaboratively annotated category hierarchies.

Author Keywords

Wikipedia, wiki, visualization, mapping, annotation, social computing, distributed collaboration, conflict.

ACM Classification Keywords

H.5.3 [Information Interfaces]: Group and Organization Interfaces – Collaborative computing, Computer-supported cooperative work, Web-based interaction

INTRODUCTION

Wikipedia is a volunteer-created online encyclopedia that includes over 2.5 million articles in the English version alone, and has become one of the most important information resources on the Web. As an information ecology, the evolution of its community and content structure has become the focus of considerable attention. Coverage of specific topic areas such as military history or the Beatles are driven in part by the passion of groups of netizens online, while the lack of coverage in specific areas may require recruitment efforts by the Wikipedia community to fill in missing details.

The graph of categories and pages that makes up Wikipedia is a rich resource for understanding the topical coverage and evolution of the system. Each page in Wikipedia can be annotated with multiple categories, organized into a loose ontology of topics. However, any user can add or change a category assignment to any article or category. The resulting category structure is noisy, ill-formed, and difficult to make sense of.

In this paper, we use a simple algorithm to infer a topic distribution from the category ontology for every article in Wikipedia and compare predictions of the algorithm to human judgments with positive results. We then demonstrate applications of the model to mapping the coverage of topics in Wikipedia as well as mapping conflict.

Our work makes three main contributions: first, we demonstrate and empirically validate how useful hierarchy information can be derived from noisy, socially annotated category data; second, we provide the first comprehensive quantitative mapping of the distribution of topics in Wikipedia; and third, we demonstrate how page-level metrics such as conflict can be similarly mapped.

RELATED WORK

There has been significant research on Wikipedia aimed at understanding its evolution, the dynamics of its users, the quality and growth of its content, and the norms and behavior of its community (see, e.g., [3][6] for reviews). Few studies however have tried to use Wikipedia's category structure to understand its topic distribution.

Holloway et al. [4] developed a semantic map of the category network of Wikipedia based on category co-occurrence to highlight patterns such as edit times and author coverage. However, they did not use quantitative measures of topic distributions, instead using a graph layout algorithm to place similar articles closer together.

Halavais and Lackaff [3] quantitatively compared the distribution of 3,000 Wikipedia articles coded into Library of Congress categories with a distribution of published books. They found substantial overlap between Wikipedia and other encyclopedias in three target topics (physics, linguistics, and poetry). However, the number of articles

CHI 2009, April 4–9, 2009, Boston, MA, USA.
Copyright 2009 ACM 978-1-60558-246-7/09/04…$5.00.

sampled and categories examined was relatively small, as extensive hand-coding was needed. Our research addresses the limits of both of the above studies, providing a method for quantitatively coding the topic distribution of an article, and that can scale to all articles.

APPROACH

Categories in Wikipedia are socially annotated, and any user can classify a page into a category simply by appending a category label to it. Categories are different from tags in social tagging applications in that they form a pseudo-hierarchical structure, as a category can be assigned to another category such that the categories form a graph. In Wikipedia as of January 2008 there were 11 top level categories, 276,834 subcategories, 666,537 category hierarchy assignments, and over 20 million category-page assignments. The nature of the MediaWiki category structure makes mapping categories difficult for a number of reasons. The very large size (200+ million links) dwarfs other social linking structures (e.g., tagging analyses so far typically have <1 million edges). There is no strict enforcement of which higher-level categories a child category can belong to; thus, the category structure is neither a tree nor a directed acyclic graph, permitting such paradoxes as a category being its own "grandparent".

Unlike social tagging, in which the problem is to generate a hierarchy from non-hierarchical data, in wikis the problem involves category data that already has hierarchical information and mapping lower-level nodes to a specified higher-level ontology. We created a distribution of topics for each article in Wikipedia by aggregating the top-level categories associated with the article's annotated categories. That is, for a category tag, we first compute a topic distribution for that tag. For an article, we take all of the category tags that have been assigned to it and calculate an aggregate topic distribution using those tags. Specifically, we transform the Wikipedia category hierarchy into a tree structure through breadth-first traversal on the category relationships, starting with the top-level categories and assigning subcategories to them using path-based semantic relatedness.

In the above approach, we determine semantic relatedness for Wikipedia category nodes through link distance metrics. The simplest path-based method of calculating semantic relatedness is edge counting [8], in which semantic distance is the length of the shortest path between two nodes. Other measures include normalizing the distance by taxonomy depth, or additionally including the depth of the least common subsumer of the two nodes [11]. After investigating the performance of more complex metrics (such as normalizing by taxonomy depth), we found little substantial difference in the results when applied to Wikipedia data. In the following analyses we use the simple shortest-path metric as our distance measure, which is also consistent with prior work [10]. In the case of equally short paths to multiple top-level categories, we split the assignment weight for both. Again, testing with variations of weight assignment resulted in no substantial differences. One minor complication is that the distribution of categories among pages is not homogeneous. While most categories are distributed approximately equally, the "People" category is an outlier in having over 2.5 times as many category assignments per page as other categories. In all subsequent analyses we control for this bias.

The Albert Einstein article provides a concrete example. This article includes 26 categories, with some categories (such as "Jewish-American scientists") associated with multiple top-level categories ("Religion", "Science", and "Society"). Overall, Einstein's topic distribution primarily falls under "People"; however, his roles as both a prominent scientist and social figure are reflected in associations with "Science", "Society", "History", "Philosophy", "Religion", and "Culture". His involvement with the Manhattan Project also leads to associations with "Technology".

EMPIRICAL EVALUATION

To test validity we compared topic distributions generated by the algorithm and by human raters. Forty-eight articles were randomly sampled from Wikipedia's "Featured article" class to ensure mature category annotations and content coverage. Raters were recruited using Amazon's Mechanical Turk market (mturk.com), and were asked to generate a topic distribution for each article by "split[ting] 100 points among the different categories that you believe this article best belongs to." Mechanisms from [5] were used to minimize invalid responses in Mechanical Turk. Users self-selected to complete one or more rating tasks, with 70 users completing 240 ratings (5 per article). Four ratings were rejected due to distributions not summing to 100. For each article the ratings were averaged to generate an aggregate topic distribution that was compared to the distribution generated by the algorithm.

To analyze the data we used a robust regression approach which adjusts the standard errors for intra-article correlation [1]. This was done as category assignments within an article may not be independent of each other. Results indicated a significant positive correlation between human- and algorithm-generated article topic distributions, $r = .67$, $p < .001$. This provides encouraging evidence that the model reasonably approximates human judgments.

MAPPING TOPICS IN WIKIPEDIA

Although Wikipedia has grown to be one of the largest encyclopedic sources of information, there is little visibility into what it contains [3][4], and the lack of a method to automatically assign articles to topics has prevented a comprehensive quantitative approach. We apply the algorithm developed and validated above to map the

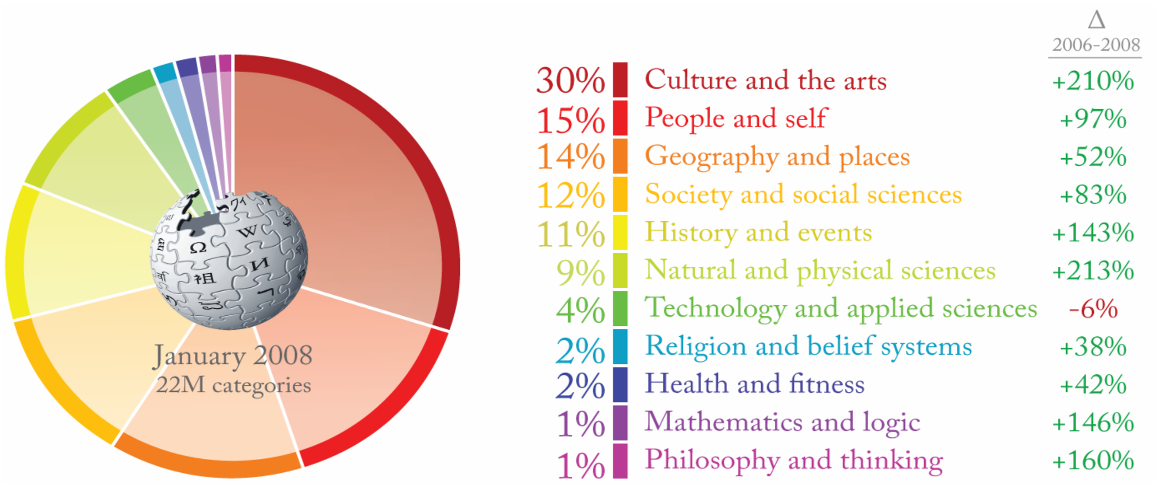

Figure 1. Distribution of topics in Wikipedia from January 2008 along with change since July 2006.

coverage of Wikipedia and examine the growth of topics over time by comparing the January 2008 and the July 2006 datasets

We use full history downloads of Wikipedia provided by the MediaWiki foundation. The data indicate tremendous growth: from July 2006 to 2008, the number of pages, categories, and category assignments more than doubled. Approximately half of all pages had at least one category assignment. The distribution of topics computed from the ~6M page-category assignments in July 2006 and ~20M in January 2008 is shown in Figure 1.[1] In order to represent a comprehensive view of topics in Wikipedia, the data shown include all page types, although using article data alone does not substantially affect the results.

As is evident from Figure 1, “Culture and the arts” and “People and self” are the most represented topics. These topics include popular subjects such as musicians and sports teams, though they also include more traditionally encyclopedic subjects such as political figures and scientists. Conversely, articles dealing with the “harder” topics, such as the natural sciences, technology, and mathematics have lower representation; accounting for only 14% of all category assignments. However, some of these topics are rapidly growing: “Natural and physical sciences” is the fastest growing area in Wikipedia at 213% growth, and “Mathematics and logic” has had 146% growth. Surprisingly, “Technology and applied sciences” is the only topic to show a decrease in size, with a 6% loss from 2006. As both pages and categories often change and are consolidated, it is difficult to pinpoint the exact cause of this decrease, but given its anomalous nature it may merit further investigation.

MAPPING CONFLICT IN WIKIPEDIA

We now extend the model to deal with aggregate article-level metrics such as quality, page views, or conflict. Here we will demonstrate its application to mapping conflict in Wikipedia, though the same technique could be used for any quantitative measure. We focus on conflict as it has been of interest to many researchers and thus methods have been developed to quantify it; additionally, little is known about the distribution of conflict in Wikipedia.

Kittur et al. [6] developed a page-level metric of conflict shown to correlate with expert ratings based on the number of times a page has been labeled as controversial. We calculated this metric for all revisions of every article in the July 2006 Wikipedia dataset, ending up with 1343 articles. For each article we split its conflict score amongst its tagged categories and passed the split score to the corresponding top level categories. If a category had two equally short paths to different top level categories, its score would be further split among them to prevent double-counting[2]. This results in the total amount of conflict per topic; however, to determine the relative degree of conflict (or “contentiousness”) per article we normalized by the number of article-category assignments in a topic. The resulting normalized conflict graph is shown in Figure 2. The categories are listed in order of their absolute (non-normalized) amount of conflict clockwise from “People”. This enables us to easily identify anomalous topics such as

[1] Some top level categories differed between datasets; however, most could be unambiguously mapped across datasets. Only one category, “Self-care”, existed in the January 2008 but not the July 2006 dataset. As this category only had 212 assignments, its impact was considered insignificant.

[2] We also investigated a number of variations including: 1) whether the quantitative conflict score was used or whether every article contributed an equal amount; 2) whether the contribution to a top level category was weighted by path length; and 3) whether the score was split amongst all tagged categories of an article or whether the full score was given to each. None of these variations or combinations thereof substantially affected the reported pattern of results.

"Religion" and "Philosophy" which stand out as highly contentious despite having relatively few articles.

CONCLUSION

We demonstrated a simple technique for determining the distribution of topics for articles in Wikipedia. This method works with collaboratively constructed, noisy and ill-formed category structures and scales to large amounts of data. Comparison with human ratings showed significant positive correlations. We applied this approach to map the distribution of topics for all articles in Wikipedia, as well as how those topics have changed over time. We also demonstrated how article-level metrics such as conflict in Wikipedia can be similarly aggregated and mapped.

Our results demonstrate one method by which category hierarchies generated through distributed means can be exploited for their rich semantics despite their noisy or ill-formed structure. They may be useful for navigating and making sense of data resulting from the growing use of distributed knowledge building. For example, thousands of wikis have been created and edited in domains ranging from enterprise knowledge sharing (e.g., socialtext) to gaming (e.g., wowwiki) to popular culture (e.g., Star Wars' wookiepedia). Many of these wikis face challenges similar to Wikipedia in having large numbers of pages (both wowwiki and wookiepedia have ~60-70,000 articles) and ill-formed, socially annotated category structures.

Our results also have implications for researchers studying wiki systems. Most existing studies of Wikipedia have been at the system level, implicitly treating it as a homogeneous entity. However, topic areas in Wikipedia are highly varied, and may have different populations of editors, different norms and rules, and different dynamics. The ability to quantify topic areas with greater granularity could enable these differences to be better captured and explored.

The correlation between independent user judgments and the predictions of the algorithm provides an empirically grounded foundation for other researchers of Wikipedia in areas ranging from understanding conflict to expertise identification to coverage analysis. For example, expertise identification has been a topic of significant research interest in both enterprise and volunteer communities, and our technique could straightforwardly be adapted to characterizing a user's editing topic distribution, providing a potentially more sensitive metric than others such as text similarity, explicit links, or co-editing patterns [2].

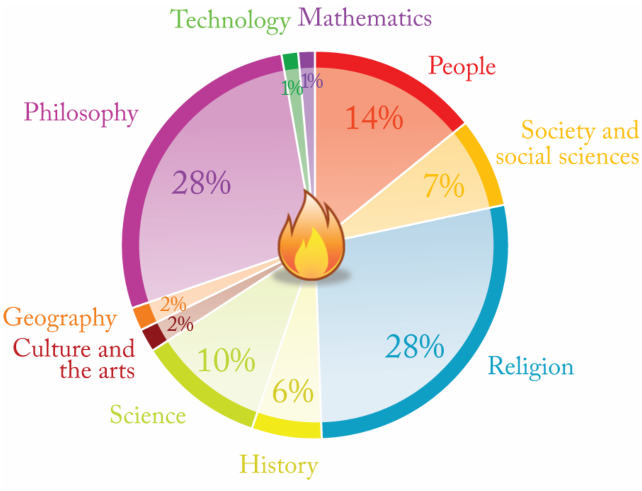

Figure 2. Distribution of conflict in Wikipedia. Sizes represent normalized conflict, while the topic order (clockwise from *People*) reflects the absolute amount.

REFERENCES

[1] Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. 2003. *Applied multiple regression/correlation analysis for the behavioral sciences.* Lawrence Erlbaum Associates, Mahwah, New Jersey.

[2] Cosley, D., Frankowski, D., Terveen, L., Riedl, J. 2007. SuggestBot: Using intelligent task routing to help people find work in Wikipedia. In *Proc. IUI*, Honolulu, HI, 32-41.

[3] Halavais, A. & Lackaff, D. 2008. An analysis of topical coverage of Wikipedia. *JCMC*, *13,* 429-440.

[4] Holloway, T., Bozicevic, M., & Börner, K. 2005. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *ArXiv Computer Science e-prints*, cs/0512085.

[5] Kittur, A., Chi, E., & Suh, B. 2008. Crowdsourcing user studies with Mechanical Turk. In *CHI 2008.*

[6] Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. 2007. He says, she says: Conflict and coordination in Wikipedia. In *CHI 2007*, San Jose, CA, 453-462.

[7] Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., Riedl, J. 2007. Creating, destroying, and restoring value in Wikipedia. In *Proc. GROUP, 2007.*

[8] Rada, R., H., Mili, E., Bicknell & M. Blettner (1989). Development and application of a metric to semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 19(1):17.30.

[9] Schonhofen, P. 2006. Identifying document topics using the Wikipedia category network. In *Proc. Intl. Conf. on Web Intelligence*, 456-462.

[10] Strube, M., & Ponzetto, S. P. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proc. of AAAI 2006*, 1419-1424.

[11] Wu, Z. & M. Palmer (1994). Verb semantics and lexical selection. In *Proc. of ACL 1994*, pp. 133-138.