Wikipedia Category Visualization Using Radial Layout

Robert P. Biuk-Aghai
Dept. of Computer and Information Science
Faculty of Science and Technology
University of Macau
Macau S.A.R., China
robertb@umac.mo

Felix Hon Hou Cheang
Dept. of Computer and Information Science
Faculty of Science and Technology
University of Macau
Macau S.A.R., China
ma86514@umac.mo

ABSTRACT

Wikipedia is a large and popular daily information source for millions of people. How are articles distributed by topic area, and what is the semantic coverage of Wikipedia? Using manual methods it is impractical to determine this. We present the design of an information visualization tool that produces overview diagrams of Wikipedia's articles distributed according to category relationships, and show examples of visualizing English Wikipedia.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*web-based services*; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*collaborative computing*; I.3.8 [**Computer Graphics**]: Applications

General Terms

Design, Experimentation

Keywords

Wikipedia, wiki, category, radial layout, information visualization

1. INTRODUCTION

Wikipedia is a public online user-contributed encyclopædia available in many languages. Most Wikipedia articles may be edited by any user. In order to organize content there is a category structure that allows authors to categorize articles corresponding to their content into as many categories as needed. Moreover, categories themselves may also belong to other categories, forming a category hierarchy. But how is Wikipedia's content distributed over its categories, i.e. what is the *semantic coverage* [2]? Which topic areas are the most popular? How does article content relate to these topic areas and to that of other articles? What kinds of patterns of distribution of content over topic areas can be observed? To address these questions we use information visualization in order to visually explore the content of an entire wiki and to visually identify patterns of content distribution. We designed and implemented a new wiki information visualization tool, using radial layout (see [1] for an overview) of wiki articles and categories, and then applied it to several language Wikipedias. For illustration we show examples of the visualization of English Wikipedia.

WikiSym 2011 October 3-5, 2011, Mountain View, California, USA
ACM 978-1-4503-0909-7/11/10.

2. DATA PRE-PROCESSING

The Wikipedia category structure is manually created over time in an incremental fashion with little coordination and no central control in which Wikipedia articles and categories may be arbitrarily inter-connected. The resulting graph structure representing a category hierarchy is neither complete nor perfect [4], containing anomalies such as loops and multiple parents. This causes problems for certain types of analysis, which is why we first simplify and pre-process the data.

2.1 Calculating Category Similarity

To avoid duplicate nodes in our visualization we require that the category graph is a tree. Converting the graph into a tree is based on the similarity between linked categories (similar to our previous method [3]), thus the first step is calculating category similarity. Our similarity calculation between a pair of categories is based on co-assignment in articles. That is, a larger number of co-occurrences of the same two categories in a set of articles implies a stronger similarity. Given parent category p and child category c, and given a root category node r, we calculate the category similarity $S_{p,c}$ as: $S_{p,c} = D_c - \frac{C_{p,c}}{k}$, where D_c is the depth of category c in the category graph, i.e. the shortest distance from the root category node r; $C_{p,c}$ is the number of co-assigned articles of categories p and c; and k is a constant that is empirically determined. Through experimentation we have found that a value of $k = 2$ produces the best results, i.e. results that agree with human intuition as to similarity of a given pair of categories. A *smaller* value of $S_{p,c}$ indicates a greater similarity (i.e. a smaller distance between the nodes). The number of co-assigned articles $C_{p,c}$ of parent category p and child category c is simply the cardinality of the intersection of their assigned article sets: $C_{p,c} = |A_p \cap A_c|$, where A_p and A_c are the sets of articles assigned to categories p and c, respectively.

2.2 Converting the Category Graph

Given the category similarity calculated above we can convert the category graph. We traverse the graph from the root category node to every child node using a breadth-first search and maintaining a list of visited nodes. Once we encounter a child node previously visited we need to keep one and eliminate the other of the vertexes. In this case, given child category c and two parent categories $p1$ and $p2$, we choose which parent link to keep according to following rules: (1) Choose the parent whose similarity value $S_{p,c}$ is lower; (2) If $S_{p1,c} = S_{p2,c}$, choose the parent whose depth D is lower; (3) If $D_{p1} = D_{p2}$, choose the parent with the larger value of $C_{p,c}$; (4) If $C_{p1,c} = C_{p2,c}$, choose the parent with the lower page ID. The last rule is provided as a fallback for those few cases when all else is equal (page ID is the unique identifier assigned to the article page in the Wikipedia database). Applying this list of rules to all cases of multiple parents, the parent links to keep can be identified and all other parent links can be eliminated. The result is a category tree without cycles or multiple parents.

2.3 Calculating Article Classification Strength

Each article in Wikipedia may be assigned to any number of categories. However, the content of an article may be more relevant to some of the categories that this article is classified under than to other categories. To enable accurate visualization we need to determine the strength of the classification of each article-category pair. We use both article-category links (i.e. article to category assignment), and article-article links (i.e. hyperlinks from one article to another) to determine classification strength. The assumption is that if a given article a has been classified under a certain category $c1$, and this article a contains many links to other articles $a1, a2, \ldots, an$ that are also classified under the same category $c1$, then the classification is stronger than for a different category $c2$ where only few or no links to articles under that category $c2$ exist. Exploration of actual data in the Wikipedia database supports this assumption. We calculate the article classification strength $A_{c,a}$ of an article a for a given category c as: $A_{c,a} = 1 + |A_a \cap A_c|$, where A_a is the set of articles linking to, or being linked to by, article a; and A_c is the set of all articles classified under category c (the classification of article a to category c is the initial number 1 on the equation's right-hand side).

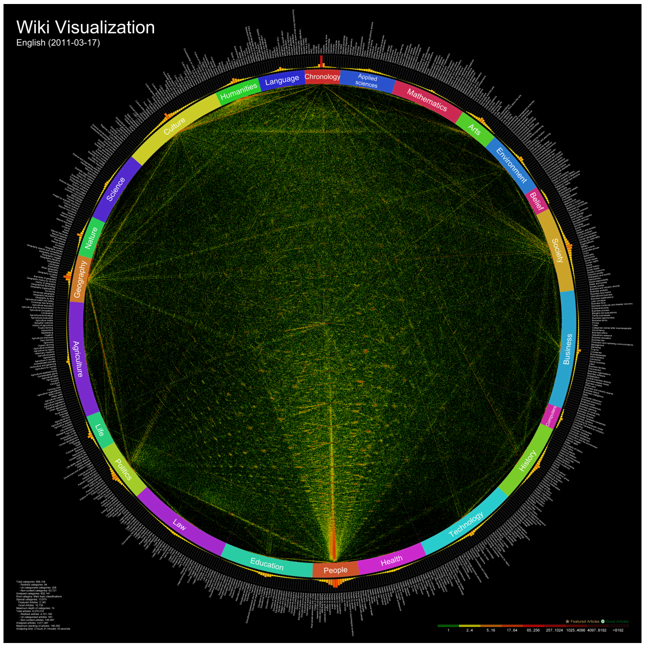

Figure 1: Radial visualization of English Wikipedia

3. VISUALIZATION DESIGN

Once the data pre-processing described above has been completed, producing the visualization itself is relatively simple. Typically a radial visualization represents nodes and edges in a graph by node labels placed on the outside of a circle and lines connecting these nodes. The nodes in our visualization are Wikipedia categories from the category tree resulting from our data preprocessing, and displaying the upper two category levels: top-level categories (such as "Science", "History", etc.) and first level subcategories (such as "Logic", "Scientists", etc. under top-level category "Science"). The entire circle is divided into arcs of different lengths that are assigned to top-level categories, with the arc length proportional to its number of sub-categories. Then arcs of top-level categories are placed on the circle, and sub-categories are placed within each top-level category arc. In addition to each subcategory label we also show a small bar that visualizes the number of articles contained in that sub-category, normalized relative to the maximum article number of all sub-categories. However, we do not display edges connecting sub-category nodes. Instead we visualize all Wikipedia articles as small icons placed on the inside of the circle at positions relative to their classification strength values respective all assigned categories. The article classification strength of each category classification of the article can be thought of as exerting a proportional pull on the article node. When articles overlap each other, a different node colour on a spectrum from green to red represents the amount of article stacking. In this way the visualization displays the distribution of articles over the entire wiki.

4. DISCUSSION

We have applied our visualization tool to five Wikipedia language editions: Danish, Chinese, Swedish, German and English. Figure 1 shows an example of the English Wikipedia visualization. Many sub-categories are connected by articles lined up exactly between them, such as can be seen in the Geography category on the left of Figure 1. These articles are only assigned to the two categories at either end and thus form an exact straight line. In other cases the influence of a third category becomes noticeable, which manifests itself by shorter lines perpendicular to the main connecting line pulling some articles sideways away from the line, cases of which are visible near the lower left, near category Politics. Featured articles are represented by a special icon, and some areas of the graph have many featured articles clustered closely together. Many areas of the graph show article nodes stacked up highly, sometimes thousands in the same spot (evident at the bottom near category People). Such clustering reflects identical category assignment and classification strength.

In our comparison of different language Wikipedia visualizations we could perceive strong local variations and many interesting visual patterns. For example, we noticed that all language editions have many articles close to the category "People", but that this was far more pronounced in the English, German and Swedish Wikipedias than in the others. This strong interest in writing people-related articles may reflect a cultural characteristic of the countries in question. In the Chinese Wikipedia, category "History" occupies a far wider space than in the other Wikipedias, perhaps reflecting the strong interest that Chinese take in their rich history.

5. REFERENCES

[1] G. Draper, Y. Livnat, and R. Riesenfeld. A survey of radial methods for information visualization. *Visualization and Computer Graphics, IEEE Transactions on*, Sept. 2009.

[2] T. Holloway, M. Bozicevic, and K. Börner. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity*, 12-3:30–40, 2006.

[3] C.-I. Pang and R. P. Biuk-Aghai. A method for category similarity calculation in wikis. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, WikiSym '10, pages 19:1–19:2. ACM, 2010.

[4] J. Yu, J. A. Thom, and A. Tam. Ontology evaluation using Wikipedia categories for browsing. In *Proceedings of the Sixteenth ACM conference on Conference on Information and Knowledge Management (CIKM '07)*, pages 223–232, 2007.