A Method for Category Similarity Calculation in Wikis

Cheong-Iao Pang
University of Macau
Faculty of Science and Technology
Av. Padre Tomás Pereira
Taipa, Macau
+853-83978638
ma76543@umac.mo

Robert P. Biuk-Aghai
University of Macau
Faculty of Science and Technology
Av. Padre Tomás Pereira
Taipa, Macau
+853-83974375
robertb@umac.mo

ABSTRACT

Wikis, such as Wikipedia, allow their authors to assign categories to articles in order to better organize related content. This paper presents a method to calculate similarities between categories, illustrated by a calculation for the top-level categories in the Simple English version of Wikipedia.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *collaborative computing*. I.5.3 [**Pattern Recognition**] Clustering – *similarity measures*.

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Wiki, category similarity.

1. INTRODUCTION

Wikis have enjoyed rapid and widespread adoption, not only on public sites such as the well-known user-contributed online encyclopaedia Wikipedia, but also in the form of many intra-organizational wikis around the world [3]. Gaining an understanding of the content of these large knowledge repositories has become the focus of numerous research efforts, leading to methods and implementations of wiki analysis and visualization, e.g. [1], [2]. Some of these efforts have concentrated on a *micro-level* of analysis, focusing on a single wiki article and exploring its evolution, relations to other articles, or other aspects. Our research is concerned with a *macro-level* of analysis, aiming at an understanding of a wiki as a whole, which could involve a huge number of articles (for example, the current count for the English Wikipedia stands at over three million articles). A challenge faced when trying to analyze and visualize such a large collection of articles is how to relate different parts of the wiki with each other; i.e. which parts of the wiki are similar to others, and thus should be visualized adjacent to each other?

Wikis typically have a category feature: an article may be assigned to a category, and categories themselves may also belong to other categories, forming a category hierarchy. Taking co-occurrence of a pair of categories within an article as an indication of a relation between the categories, we calculate the degree of similarity for each such pair. Our method is specifically designed for the MediaWiki system that Wikipedia and many other wikis run on, but should be applicable to other wiki systems.

2. METHOD OF CALCULATION

In MediaWiki, authors of an article can assign it to any number of categories by inserting a category tag in the text, and it is common for a given article to belong to multiple categories, sometimes more than a dozen. Likewise authors can create new categories and assign these to existing parent categories. As it is possible for a category to be assigned to multiple parents, and also for a given category to be both the parent and child category of another category, the whole Wikipedia hierarchy is not a simple tree structure [4]. Our proposed method for calculating category similarity includes following steps:

1. Simplify the category hierarchy
2. Calculate similarity for each co-occurring category pair
3. Aggregate similarities for a defined number of levels

Simplifying the category hierarchy: To eliminate loops and to simplify the category hierarchy for later processing we transform it into a directed rooted tree. A root is selected depending on the category hierarchy of the wiki in question; for instance, for the English Wikipedia Holloway et al. start at the Categories node [2]. From the root we traverse the tree to every child node using a breadth-first search. Moreover, we maintain a list of visited nodes. Once we encounter a child node that was previously already visited we eliminate the repeated vertex to it in our tree. Thus we can prevent cycles and multiple parents in the tree. As the breadth-first search retrieves sibling (i.e. same-level) categories in ascending order of category ID, and the order of IDs corresponds to chronological order of category creation, older categories are favoured over younger ones in cases where two or more categories at the same depth link to the same child category. Assuming that more important or significant categories were created earlier in the history of the wiki, this method thus connects subcategories to the more significant parent categories.

Calculating similarity: Co-occurrence of a pair of categories in an article is used for computing their similarity as we assume that a larger number of co-occurrences implies stronger similarity. We retrieve the number of articles citing a given category and for each pair of categories use it to calculate the cosine similarity [2]:

WikiSym '10, July 7-9, 2010, Gdańsk, Poland
Copyright © 2010 ACM 978-1-4503-0056-8/10/07... $10.00

$$cos_{i,j} = cos_{j,i} = \frac{\sum_{k=1}^{n} A_k C_{ij}}{\sqrt{\sum_{k=1}^{n} A_k C_i \sum_{k=1}^{n} A_k C_j}}$$

where $cos_{i,j}$ is the cosine similarity of categories C_i and C_j, A_kC_i is the assignment of article A_k to category C_i, and similarly for A_kC_j, and A_kC_{ij} is the co-occurrence of article A_k in categories C_i and C_j.

Aggregating similarity: To determine the similarity of a given pair of categories, we should consider not only their direct similarity but also the similarity of their subcategories, possibly several levels deep. How many levels below the given pair of categories it is suitable to include can be empirically determined, as we demonstrate in Section 3 below. The aggregate similarity is calculated as follows:

$$ac_{i,j} = w_1 cos_{i,j} + w_2 cos'_{i,j,n}$$

where $ac_{i,j}$ is the aggregate similarity between two categories C_i and C_j, $cos_{i,j}$ is the direct similarity between C_i and C_j, and $cos'_{i,j,n}$ is the average of the similarity values of the subcategories of C_i and C_j, to n levels depth. w_1 and w_2 are coefficients for adjusting the weights of above similarity values.

3. EXPERIMENT AND RESULTS

We applied our category similarity calculation method to the top-level categories in the Simple English version of Wikipedia, using the database dump of Sep. 20, 2009 which contains 77,126 articles and 12,058 categories. Under the category root "Articles" there are eight top-level categories (Everyday life, Geography, History, Knowledge, Literature, People, Religion and Science, abbreviated by their initial letters in the tables that follow). Table 1 shows the calculated values: direct cosine similarity $cos_{a,b}$, average of cosine similarity for top level and first level child categories $cos'_{a,b,n}$, and aggregate similarity $ac_{i,j}$ with different values for weights w_1 and w_2 (columns marked A, B, C). Because of space limitations only the first few rows are shown. Using direct similarity alone is not sufficient because in some cases this value is zero, as in the case of the pair Science and People. Thus including lower level categories is necessary. We found that including one level below the top categories ensured all values were greater than zero and delivered on average the maximum value of $cos'_{a,b,n}$, compared to including 2, 3, 4 or 5 levels where the similarity value rapidly dropped off as shown in Table 2.

Table 1. Cosine and aggregate similarity values

Cate-gories	$cos_{a,b}$	$cos'_{a,b,n}$	$ac_{i,j}$ (A) w_1=0.67 w_2=0.33	$ac_{i,j}$ (B) w_1=0.5 w_2=0.5	$ac_{i,j}$ (C) w_1=0.33 w_2=0.67
S - G	0.010726	0.000686	0.007413	0.005706	0.003999
S - E	0.018155	0.001288	0.012589	0.009722	0.006854
S - P	0.000000	0.000413	0.000136	0.000207	0.000277
S - R	0.009917	0.000765	0.006897	0.005341	0.003785
S - L	0.054464	0.000746	0.036737	0.027605	0.018473
S - H	0.011330	0.000572	0.007780	0.005951	0.004122
S - K	0.000000	0.001910	0.000630	0.000955	0.001280

Table 2. Value of $cos'_{a,b,n}$ by number of subcategory levels

Cate-gories	1 Level Deep	2 Levels Deep	3 Levels Deep	4 Levels Deep	5 Levels Deep
S - G	0.000686	0.000056	0.000042	0.000041	0.000038
S - E	0.001288	0.000189	0.000088	0.000050	0.000038
S - P	0.000413	0.000155	0.000108	0.000085	0.000074
S - R	0.000765	0.000088	0.000056	0.000057	0.000048
S - L	0.000746	0.000101	0.000047	0.000042	0.000038

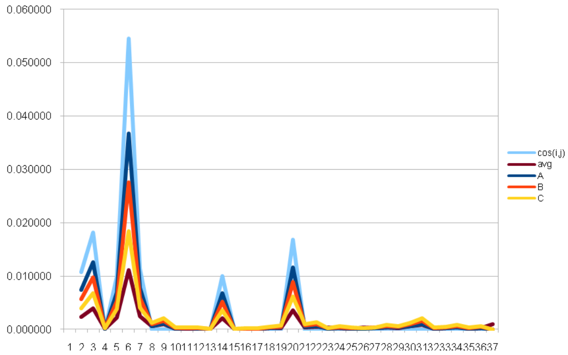

Figure 1. Comparison of different coefficients w_1 and w_2

To determine suitable values for w_1 and w_2 we visualized the similarity values as shown in Figure 1 to explore the effect of altering these weights. Firstly, this visual exploration shows that using only the cosine similarity values $cos_{i,j}$ and $cos'_{i,j,n}$ is unsatisfactory as it magnifies differences between categories and includes zero values. We found that w_1=0.33 and w_2=0.67 was the most satisfactory, producing a more balanced set of values that smoothed extreme differences between high and low, while lifting very low values. We suggest using this set of values in the further calculation of aggregated similarity.

4. CONCLUSION

We have presented a method for calculating the similarity among wiki categories, showing how to simplify the category hierarchy and how to aggregate individual similarity values to produce a similarity value of their parent categories. While our paper illustrated this calculation for the top-level categories of Simple English Wikipedia only, the method can also be applied to other levels deeper down the category hierarchy, and to other wikis.

While users are still in charge of making article category assignments, our method takes these category assignments and calculates similarity values that can be put to use in various applications of Wiki analysis and visualization, as well as in document clustering. Our current work applies these similarity values to develop whole-wiki overview visualizations.

5. REFERENCES

[1] Chan, B., Wu, L., Talbot, J., Cammarano, M., and Hanrahan, P. 2008. Vispedia: Interactive Visual Exploration of Wikipedia Data via Search-Based Integration. *IEEE Transactions on Visualization and Computer Graphics*, 14, 6, 1213-1220.

[2] Holloway, T., Bozicevic, M., and Börner, K. 2006. Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. *Complexity*, 12, 3, pp. 30-40.

[3] O'Leary, D.E. 2008. Wikis: 'From Each According to His Knowledge'. *Computer*, 41, 2 (Feb. 2008), 34-41.

[4] Yu, J., Thom, J. A., and Tam, A. 2007. Ontology Evaluation Using Wikipedia Categories for Browsing. *Conference on Information and Knowledge Management*, 223-232.

[5] Zesch, T. and Gurevych, I. 2007. Analysis of the Wikipedia Category Graph for NLP Applications. *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 1-8.