Automatically assigning Wikipedia articles to macro-categories

Jacopo Farina[‡]
jacopo1.farina@mail.polimi.it

Riccardo Tasso[‡]
riccardo.tasso@gmail.com

David Laniado[*, ‡]
david.laniado@barcelonamedia.org

‡ DEI, Politecnico di Milano * Barcelona Media – Innovation Centre

ABSTRACT

The online encyclopedia Wikipedia offers millions of articles which are organized in a hierarchical category structure, created and updated by users. In this paper we present a technique which leverages this rich and disordered graph to assign each article to one or more topics. We modify an existing approach, based on the shortest paths between categories, in order to account for the direction of the hierarchy.

Categories and Subject Descriptors

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

General Terms

Algorithms, Human Factors

Keywords

Wikipedia, category graph, topic coverage

1. INTRODUCTION

Wikipedia is an online encyclopedia whose contents are freely editable by users. Founded in 2001, it underwent a rapid growth and nowadays counts over 3 million articles in the English version. To manage the increasing amount of articles, in 2004 a system of categories was introduced. Any user can change the categories to which a page is assigned, and any category can be itself assigned to one or more categories. Nowadays, more than 500 000 categories exist in the English Wikipedia.

Most of category assignments are taxonomic and represent an "*is a*" relationship, like "Conifers" assigned to "Tree", but they may also represent other relationship types as shown in [4]; for example, "Brain" is a subcategory of "Cognitive science" as well as "Psychology". The structure can be naturally represented as a graph where nodes represent pages and categories, and edges the oriented relationship "*is assigned to*". Whereas in principle the graph represents a hierarchy of topics and subtopics, with *broader* categories assigned to *narrower* ones, nothing prevents users from assigning categories following any criterion, sometimes just a "*related to*" relationship, so also loops are possible.

Most of the attention of previous literature has focused on the extraction of ontologies from this pseudo-hierarchical structure, restricting the analysis on only taxonomic relationships. In this work we want to take into account all the richness of the category graph in order to assign one or more topics to each Wikipedia article. To this end we rely on the algorithm proposed in [2], based on the shortest path between categories and topics, and we introduce and evaluate some variations.

The most relevant work related to ours is the one presented in [1], where similarity between categories is computed according to their co-occurrence within individual articles; a map of topic coverage in Wikipedia is drawn and 8 top level categories are highlighted.

2. APPROACH

The idea on which the technique is based is simple: if two categories are connected by an edge, they are probably semantically related. The closer two categories in the category graph, the closer their semantics. We can this way estimate, given a category, the macro-category in which it fits better, as the closest one in the graph. In the case of equally short paths from a category to multiple macro-categories, these are all considered suitable for the category being evaluated.

An article is assigned to macro-categories by evaluating the categories to which it is directly assigned (labels). More precisely, the degree to which an article belongs to a given macro-category is computed as the proportion of its labels which belong to that macro-category. In case of a label belonging to more than one macro-category, its contribution is split in equal parts among the macro-categories. So, suppose for example that the article "Barack Obama" is labeled with 4 categories, two of which are assigned to "Politics" and the third one to "Arts", and the remaining one is equally close to "Law" and "People": then the article will be considered related to "Politics" with a score of 0.5, to "Arts" with a score of 0.25 and to "Law" and "People" with a score of 0.125 each.

Though the category graph is based on directed relationships linking categories to super-categories, Kittur et al. [2] considered it as an undirected graph to compute the shortest paths between each category and the macro-categories, thus loosing the information carried by the assignments' di-

HT'11, June 6–9, 2011, Eindhoven, The Netherlands.
Copyright 2011 ACM 978-1-4503-0256-2/11/06 ...$10.00.

rection. The simplest way to correct the algorithm would be to compute distances in the directed graph, considering only relationships followed according to the hierarchy direction, i.e. from the most specific, low level categories, up to the macro-categories. However, in this way many categories would remain disconnected from all the macro-categories, and many articles could not be assigned to any topic. Instead, we propose another way to improve the effectiveness of the algorithm by accounting for edge direction: while computing the shortest path between a category and a macro-category, we penalize by a factor w the edges followed in the wrong direction.

3. RESULTS AND EVALUATION

For this study we relied on a dump of the English Wikipedia dated March 12th, 2010, containing about 3.2 million articles and over 500 thousand categories. We removed all the categories which we identified as non-semantic, but project-based (e.g.: "Stubs"). While Kittur et al. [2] used 11 macro-categories, we chose to use 21, corresponding, with minor arrangements, to the current official Wikipedia top level categories[1].

We ran both the original algorithm as described in [2] and our modified version with $w = 3$, i.e. penalizing edges followed in the wrong direction in the hierarchy by a factor 3. All articles could be assigned to some macro-category, except for less than 100 pages, mostly corresponding to pages created by mistake or not yet completed when the dump was created.

The topic coverage emerging from the results of the modified algorithm are shown in Figure 1, where the percentages assigned to each macro-category over the whole wiki have been aggregated in order to estimate the importance of the different topics in terms of number of articles. The two largest macro-categories are "Geography and places" and "History and events". "Agriculture" is larger than expected; this is due to the high density of links between its subcategories, which makes it easily reachable in a few steps. Moreover, Wikipedia has a huge amount of pages about plant species. The smallest categories are "Arts" and "Computing"; this is partly due to the fact that some related low level categories are assigned to other "competitor" macro-categories, like "Culture" in the first case, and "Technology and applied sciences" in the second.

After executing the two algorithms, we evaluated them comparing the results with manually generated assignments. Assessment has been performed on 200 randomly selected articles, manually labeled by three human evaluators according to the 21 macro-categories. The cosine similarity between the assignments performed by human evaluators and the ones produced by the original algorithm is of 0.34; by accounting for edge direction we get a similarity of 0.37.

4. CONCLUSIONS

In this work we faced the problem of automatically assigning each Wikipedia article to one or more topics, leveraging the rich and messy structure of categories and subcategories created by the community. We modified the algorithm proposed by Kittur et al. [2] and based on the shortest path between a category and a macro-category. By penalizing edges followed in the wrong direction with respect to the hierarchy, we are able to account for the orientation of the categories assignments, without loosing the information brought by these connections.

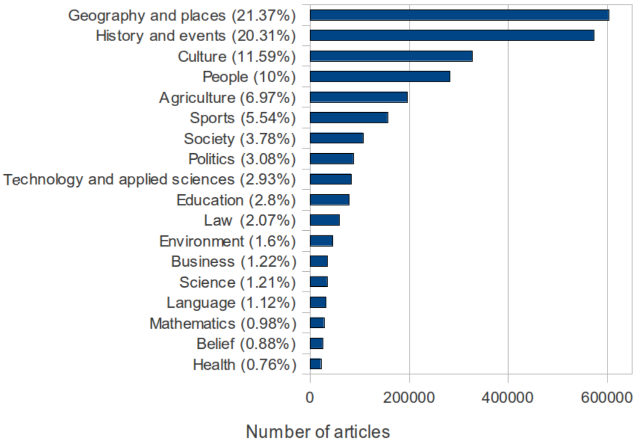

Figure 1: Size of the macro-categories, computed by aggregating the relatedness scores over all articles.

The algorithm proposed shows to outperform the original one improving the accuracy, measured as the similarity with manually generated assignments, from 0.34 to 0.37. This result is encouraging, though a more rigorous evaluation process would be needed in order better assess the statistical significance of the improvement obtained. Beyond refining the evaluation process, we plan to test and compare other algorithms, based on the probability of reaching each macro-category in the graph, starting from a given article.

The topic coverage computed here gives the same importance to pages of different sizes, and thus risks of overestimating categories containing many short pages, and in particular those automatically generated by bots. The count may be improved by considering, instead of the number of pages assigned to a macro-category, the number of edits or words in these pages, to obtain a more representative map of the wiki. Other article-level metrics, such as the number of polls, or of edits done by specific classes of users, can be aggregated by topic, to study how activity varies over different semantic areas. An analysis of the discussions in article talk pages from different macro-categories, based on the results described in this paper, can be found in [3].

5. REFERENCES

[1] T. Holloway, M. Bozicevic, and K. Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity*, 12(3):30–40, 2007.

[2] A. Kittur, E. H. Chi, and B. Suh. What's in wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proceedings of CHI*, 2009.

[3] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the wikipedians talk: network and tree structure of wikipedia discussion pages. In *Proceedings of ICWSM*, 2011.

[4] V. Nastase and M. Strube. Decoding wikipedia categories for knowledge acquisition. In *Proceedings of AAAI*, 2008.

[1] See: http://en.wikipedia.org/wiki/Category:Main_topic_classifications