Examining the “Leftness" Property of Wikipedia Categories

Karl Gyllstrom
karl.gyllstrom@cs.kuleuven.be

Marie-Francine Moens
sien.moens@cs.kuleuven.be

Department of Computer Science
Katholieke Universiteit Leuven
Leuven, Belgium

ABSTRACT

Wikipedia’s rich category structure has helped make it one of the largest semantic taxonomies in existence, a property that has been central to much recent research. However, Wikipedia’s category representation is simplistic: an article contains a single list of categories, with no data about their relative importance. We investigate the ordering of category lists to determine how a category’s position in the list correlates with its relevance to the article and overall significance. We identify a number of interesting connections between a category’s position and its persistence within the article, age, popularity, size, and descriptiveness.

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General

General Terms

Experimentation, Measurement

1 Introduction

A typical Wikipedia article contains a list of categories (e.g., “21st century actors”), which serve to place the article within the semantic, conceptual structure of Wikipedia. Categories, in turn, are implemented as special Wikipedia pages, which list all of the pages for which the category label is assigned. There is a many-to-many relationship between articles and categories, and they are intended as a way for users to navigate among articles (e.g., by selecting a category through which to find related articles). As with normal Wikipedia content, categories are added or removed to a page by edits from the community of Wikipedia authors (i.e., the general public). Hence, they reflect a form of editorial consensus; an aggregate human view of how an article fits within the conceptual space.

In this work, we examine the role played by the order of an article’s category list. As stated by Wikipedia’s editing guidelines:

The order in which categories are placed on an article is not governed by any single rule [...] Normally the most essential, significant categories appear first. [2]

We investigate the reliability of this guideline. In particular, we consider how a category’s *leftness*, or how early it appears in an article’s category list (i.e., earlier/lower list position), correlates with various attributes connected to its importance to an article and overall significance.

This work contributes an analysis of Wikipedia categories which may help guide other work that makes use of categories. The use of Wikipedia in research is widespread; in particular, categories have been central to a number of works (e.g, [3–6]). For example, Koolen and Kamps use categories to measure the semantic relationship among articles, but consider all categories equally [4]. Ponzetto and Strube derive a linguistic taxonomy using the Wikipedia category structure, also treating categories equally [5]. Our work suggests that categories are not equal and, in particular, the leftmost categories should be given significant priority. To our knowledge, no such examination yet exists.

2 Leftness examined

We used a collection of Wikipedia articles, selected as the 3000 most frequently viewed articles during an hour of August 25th, 2010, which we refer to as *Wikitop*, downloaded from a public Wikipedia page access repository [1]. For each article, we examined its entire edit history, recording when the category list was changed, and how it was changed.

2.1 Category modifications

Category changes are infrequent. Across 347,438 article modifications, only 24,398 (6.6%) included changes to the category list itself. The occurrences of different types of changes are depicted in Table 1. We observe that insertions to and deletions from the category lists dominate the number of operations. Reordering, in which at least one category changes position in the list, is relatively infrequent. Pure reorders, where the only change is a reorder (i.e., no coinciding a insertions or deletions) are even less common. Category positions appear to be relatively stable once established.

2.2 Duration and age

In this section we investigate two temporal properties of position: (1) the duration over which a category is present in

CIKM’11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

Type	#	Portion
Insert	11637	0.477
Delete	11264	0.462
Insert+Delete	941	0.039
Duplicate inserted/deleted	422	0.017
Reorder	436	0.018
Pure reorder	134	0.005

Table 1: Operations. As some operations can coincide with others, the portion sum is > 1.

an article, and (2) the age of the category (as defined by time since the category was created). Conceptually, the longer a category exists within an article, the more scrutiny it persists, and the older a category article, the more established it is within Wikipedia's category structure.

For each article in *Wikitop*, we examined its current list of categories (i.e., the list appearing in the most recent version), and measured the duration over which each category in this list was present in the article. For each category position from 0-14 (with 0 being leftmost), we recorded the mean duration for categories at that position across all articles in *Wikitop*, depicted in Figure 1[1]. In other words, the value at position 0 reflects the mean duration of categories appearing at position 0 across all articles in *Wikitop*. A category at position 0 appears, on average, within its article for about 73% of the article's lifetime. As position increases, we observe a large decline in the duration for which categories at that position are present in the article. Figure 2 depicts the mean age of categories over position. As in Figure 1 we see a decline as position increases. This indicates that older and more persistent categories are placed in lower positions.

2.3 Popularity and Size

We report the connection between position and category popularity and size. Popularity is defined by number of times web surfers accessed the category article directly (e.g., `http://en.wikipedia.org/wiki/Category:1975_births`) over a given time period[2]. The size of a category is defined by the number of articles it contains.

Figure 3 depicts the popularity of articles across position. We observe an overall general though variable decline of popularity with position[3].

Figure 4 depicts the mean size of categories across position. We observe a sharp decline among the first few positions. This could indicate that categories placed in leftmost positions are more applicable overall (as they have more articles) or are generally more favored by Wikipedia authors.

2.4 Descriptiveness

In this section we examine the relationship between position and category descriptiveness, which indicates how well a category describes a particular article. We examine this property in three ways.

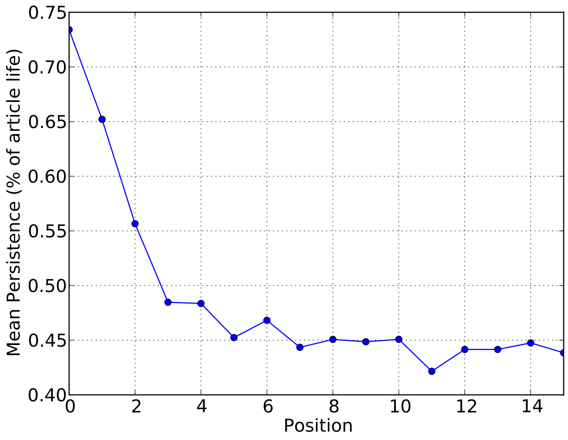

Figure 1: Duration existing in article at position.

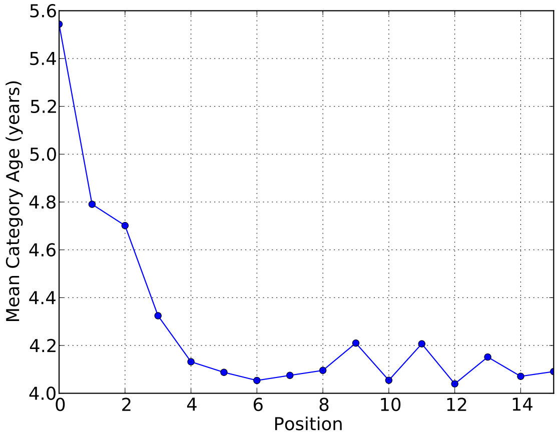

Figure 2: Category age across position.

2.4.1 *Mean category count*

First, we consider the mean number of categories per article within a category. For example, articles within the category *Zoos* have a mean of 3.72 categories per article. Our hypothesis is that this reflects the exclusivity of a category: when its articles have few other suitable categories, it is likely the category itself is a better descriptor of the article. For each article in *Wikitop*, we compute the *mean category count* (MCC) for each category by calculating the mean number of categories per article within the category. Note that the number of articles considered in this calculation (1,002,937) is much higher than the number of articles in *Wikitop*. We then aggregate MCC scores by position number. Figure 5 depicts MCC across position; we observe that as categories appear at higher positions, they are more likely to contain pages with higher number of categories, indicating that they become less effective descriptors.

2.4.2 *Mean position*

Next, we assign a positional score to categories, defined as the mean position of the category within articles in which it appears. Assuming a connection between position and

[1] Around 85% of articles have 15 or fewer categories, so we chose this as the number of category positions to consider.

[2] We aggregated access data from the repository, obtained from three one hour periods [1].

[3] In Figure 3 and Figure 4, we omit data pertaining to the *Living people* category, an exceptionally large outlier.

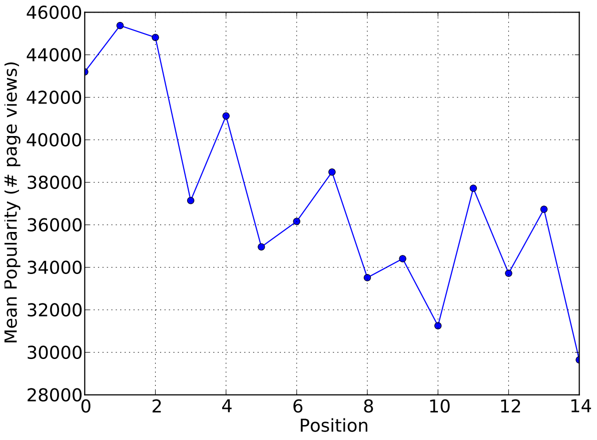

Figure 3: Category popularity across position.

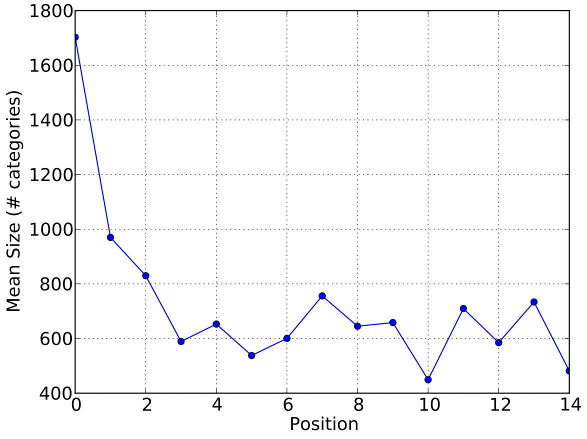

Figure 4: Category size across position.

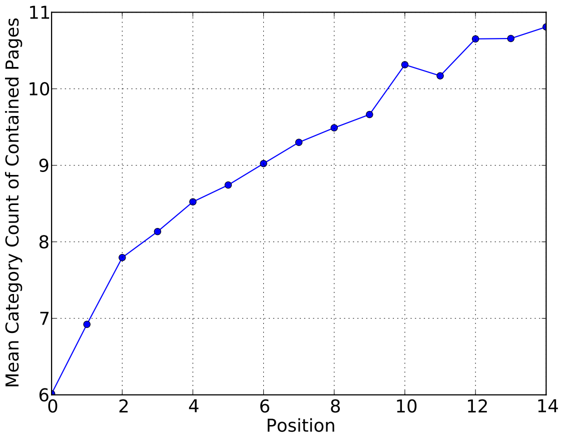

Figure 5: Mean category count across position.

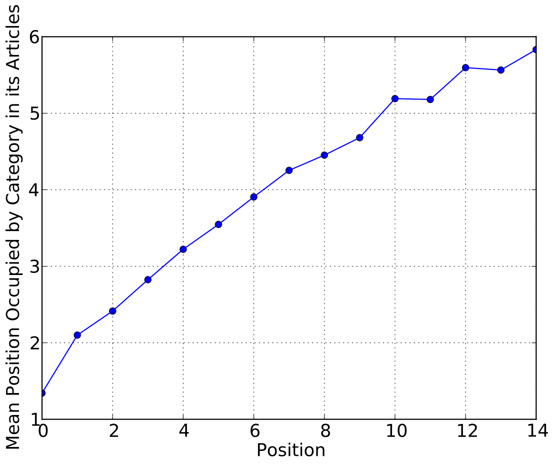

Figure 6: Mean position of category articles across position.

category importance, a category with a lower mean position across all of its articles is likely to be more significant overall. Figure 6 depicts the mean leftness of categories across position, revealing that the mean leftness of a category increases with position. In other words, categories that tend to appear at a higher position in *Wikitop* also tend to appear in higher positions in all of the articles they contain, indicating that they are generally subordinate categories.

2.4.3 *Similarity*

The next score we consider is categorical similarity. We consider a category to be a more specific descriptor if the articles with which it is associated tend to be more similar to each other. We measure this via the mean similarity of a category's articles among each other, with similarity calculated by the Jaccard coefficient:

$$similarity(P_1, P_2) = \frac{|Categories_{P_1} \cap Categories_{P_2}|}{|Categories_{P_1} \cup Categories_{P_2}|}$$

In other words, it is the proportion of category overlap among categories in a pair. For each category, we calculated the mean similarity score among each pair of articles within it, capping at 200 articles where necessary to ease computation.

Figure 7 depicts the relationship of position and categorical similarity. As position increases, the similarity among articles decreases, indicating that it is less coherent as a descriptor. One challenge to this interpretation is that the overall number of common categories remains stable across position, but as MCC increases with position, the similarity score necessarily declines. We believe that this still reflects an overall decline in similarity, as each category should be considered less individually descriptive, therefore fewer categories in common relative to total categories still achieves a reasonable approximation of overall similarity.

2.5 Parent categories

A category may have parent categories, which represent its generalizations. For example, the category *1975_births* has *1970s_births* as one of its parent categories. Figure 8 depicts the mean number of parent categories across position. We observe the 0th position containing a much higher number of parent categories, with a slight, general increase in parent category count across position. Figure 9 depicts the mean

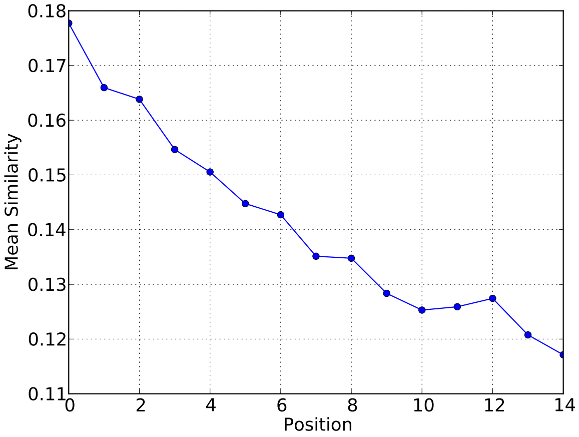

Figure 7: Mean similarity across position.

position of a category's parents within the articles they contain. This echoes the trend in Figure 6, indicating parent categories can also be distinguished by relative importance. We are uncertain about the implications of Figure 8; our intuition is that categories with many parent categories are likely to be more specific while also less independently useful. Nonetheless, the exceptional value for the 0th position in both cases should be noted.

3 Concluding remarks

This work represents a preliminary analysis of some Wikipedia category attributes, specifically with respect to their position in category lists. We identified a number of connections to position; at lower positions, categories tend to be (1) those which persisted for longer durations in the article, (2) older, (3) more popular, and (4) larger. Our analysis of descriptiveness also indicates that categories at lower positions (5) contain articles with fewer categories, hence their categories are more exclusive; (6) generally appear at lower positions across all the articles they contain; and (7) contain articles that are more similar to each other in terms of category overlap. With respect to (1), (2), (4), and (5), the drop off after position 0 is dramatic, indicating that there is increased importance on the first few listed categories, especially at the leftmost position. Though our work is at an early stage, we believe these results provide support that Wikipedia's guidelines are generally followed, and we advocate the emphasis of leftmost categories among researchers who make use of Wikipedia categories.

We plan to continue this work. First, we intend to execute a user evaluation in which category quality – especially with respect to article assignments – is assessed by human raters. Next, we seek to continue the analysis on (1) less popular and (2) lower quality articles (as indicated by improvement flags placed by editors) to find if the observations hold. Finally, we desire to apply our discoveries in a set of ranking functions, which can help determine the best category for an article based on a preference metric, such as quality, comprehensiveness, exclusivity, or descriptiveness.

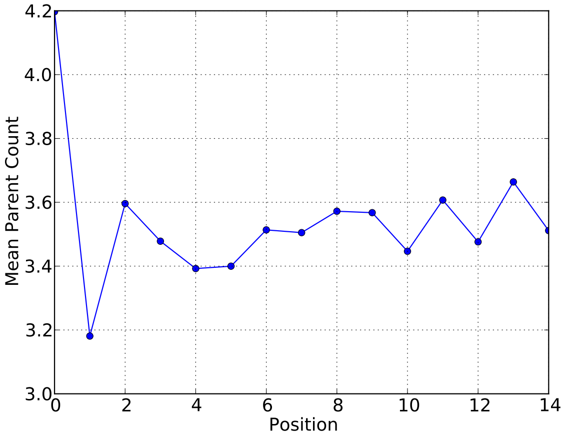

Figure 8: Mean number of parent categories across position.

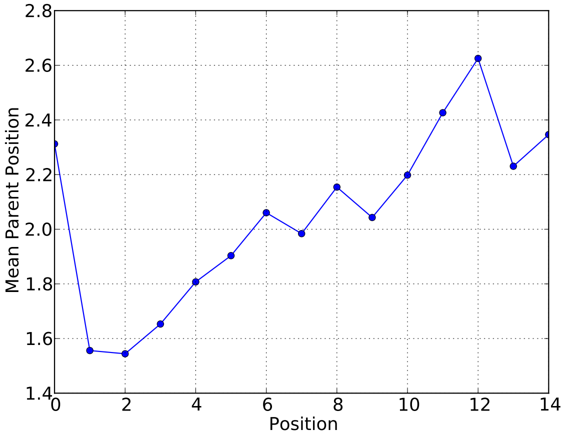

Figure 9: Mean parent position in its pages (see Figure 6).

References

[1] Wikipedia traffic. `http://dammit.lt/wikistats/`.

[2] Wikipedia:Categorization. `http://en.wikipedia.org/wiki/Wikipedia:Categorization`.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *ISWC*, pages 722–735. Springer, 2007.

[4] M. Koolen and J. Kamps. Are semantically related links more effective for retrieval? In *ECIR*, pages 92–103. Springer, 2011.

[5] S. P. Ponzetto and M. Strube. Deriving a large-scale taxonomy from Wikipedia. In *AAAI*, pages 1440–1445, 2007.

[6] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.