Wikipedia, a Social Pedia: Research Challenges and Opportunities: Papers from the 2015 ICWSM Workshop

On the Evolution of Wikipedia: Dynamics of Categories and Articles

Ramakrishna B. Bairi
IITB-Monash Research Academy
IIT Bombay
Mumbai, India, 400076

Mark Carman
Monash University
Victoria 3800, Australia

Ganesh Ramakrishnan
IIT Bombay
Mumbai, India, 400076

Abstract

We present several statistics related to English Wikipedia category and article evolution between Wikipedia 2012 (Oct) and 2014 (Jun) instances. This includes analysis of categories, articles and links creation and deletion. We also present the distribution of Wikipedia articles over 14 broad topics and compare them across 2012 and 2014 Wikipedia. We provide several statistics of Wikipedia category graph. We demonstrate that Wikipedia category hierarchy can be treated as an is-a graph locally, but over longer paths it is no more an is-a graph.

1 Introduction

Wikipedia is an online encyclopedia which has undergone tremendous growth. Many knowledge discovery and information extraction kind of applications heavily make use of Wikipedia contents. From the engineering perspective of these applications, it becomes important to know various statistics and dynamics pertaining to different aspects of Wikipedia. In this paper we present some statistics on the Wikipedia categories and articles. We analyze the category graph of Wikipedia and present the details such as graph width, depth, cycle statistics, etc. We also show that treating the category hierarchy as an is-a graph can have negative semantics over a long paths. We compare and present the statistics between two instances of English Wikipedia: (i) Oct 2012, referred to as Wiki-12 in this paper and (ii) Jun 2014, referred to as Wiki-14.

1.1 Structure of Wikipedia

In this section we briefly highlight the organization of important elements of Wikipedia that are required to understand this paper.

“Articles” are the normal pages in Wikipedia, that contain information about a subject, such as *Airbus A330*, *Gasoline*, *Albert Einstein* or *Holiday*.

“Categories” are special pages that are used to group articles together. E.g., category *Physics* is used to group articles related to physics. Every category can have multiple parent categories and multiple child categories. Parent categories are more abstract categories that group related detailed categories. E.g, categories *Physics*, *Chemistry*, *Biology* are grouped under *Science*. This creates a notion of hierarchical organization of categories.

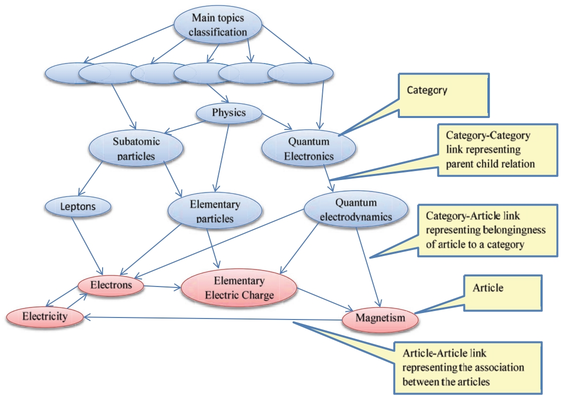

Main topics classification
Category
Physics
Subatomic particles
Quantum Electronics
Category-Category link representing parent child relation
Leptons
Elementary particles
Quantum electrodynamics
Category-Article link representing belongingness of article to a category
Electrons
Elementary Electric Charge
Article
Electricity
Magnetism
Article-Article link representing the association between the articles

Figure 1: Wikipedia article and category organization

Articles are assigned multiple categories that are related to that article. E.g, article on *Electron* is assigned categories such as *Leptons*, *Elementary particles*, *Quantum electrodynamics*, etc. Every article also links to related articles. E.g., article *Electron* associates with *Magnetism*, *Electricity*, *etc*.

Figure 1 depicts a part of Wikipedia structure.

2 Related work

Kittur et al. (Kittur, Chi, and Suh 2009) analyzed the growth of categories, and developed an algorithm to semantically map articles through its category links to the 11 top categories chosen by the research team. Krzysztof et al. (Suchecki et al. 2012) investigate the evolution of the category structure of the English Wikipedia from its birth in 2004 to 2008. They treat the category system as if it is a hierarchical Knowledge Organization System, capturing the changes in the distributions of the top categories. They investigate how the clustering of articles, defined by the category system, matches the direct link network between the articles and show how it changes over time.

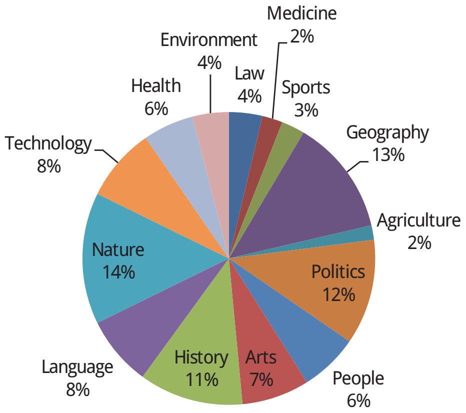


Figure 2: Topic-wise relative distribution of articles in Wiki-12

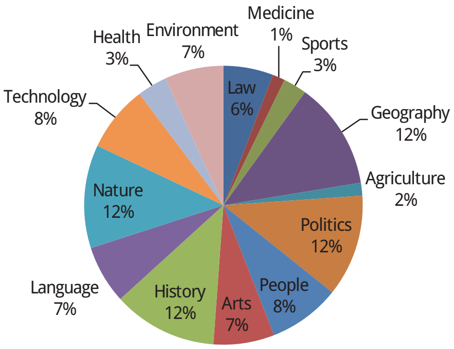


Figure 3: Topic-wise relative distribution of articles in Wiki-14

3 Evolution of Categories and Articles

We empirically studied the evolution of of Wikipedia categories and articles and present the results in this section. We present various statistics related to the categories and articles.

3.1 Category and Article Creation

Categories and articles in Wikipedia are socially created and annotated, and any user can create an article and classify into a category by simply appending a category label to it. This leads to an enormous growth of both categories and articles. Wikipedia editors also modify the existing category structure to better organize the articles. We studied changes to the categories and articles between Wiki-12 and Wiki-14. There has been an addition of 258483 number of categories in Wiki-14. This accounts for 25% increase in the number of categories. These are the categories that are present in Wiki-14 but not in Wiki-12. Note that, due to refinement of certain concepts, an existing category may be split into multiple fine-grained categories. We find that this is the major cause for increase in the number of categories.

We observed that Wiki-14 has 507746 additional articles than Wiki-12, which accounts for 12% increase.

We further analyzed the distribution of articles in Wiki-12 and Wiki-14 on 14 broad areas shown in Figure 2 and 3. In table 1, we compare the evolution of articles in each of these broad areas. Area *Environment* has seen maximum growth rate followed by areas *Law* and *People*.

Note that, Wikipedia category structure is a directed graph with every category having multiple parents and children categories. Each category is also connected to a number of articles. This gives us a notion of coverage of articles by a category. By starting at a category c, we can traverse through the descendant categories and collect all the articles that are connected to one or more descendant categories. Let $\Gamma\left(c\right)$ be the set of articles that are reachable from the category c in the above said manner. We say that, all the articles $\Gamma\left(c\right)$ belong to the category c. Alternatively, category c covers $\Gamma\left(c\right)$ articles. Although this seems to be a reasonable notion of belongingness and coverage, there is a caveat to this. Wikipedia category graph is highly interconnected and almost any article can be covered any of the 14 broad categories shown in Figure 2. This is demonstrated in Table 1. For each of the 14 broad categories, we calculate their coverage $\Gamma\left(.\right)$ at each *level* of the *breadth first traversal* starting from that category. As the level increases, we can see that more and more articles are covered and after level 10, almost all the articles in the Wikipedia are covered. That means, all the articles belong to all the broad categories, which is incorrect. As demonstrated in Section 4.3, long distance belongingness in Wikipedia category graph does not make sense. By randomly sampling and then manually inspecting 39 articles, we concluded that, coverage of a category should not look beyond 7 *levels* in a *breadth first traversal* for these 14 broad categories. Hence we reported the article coverage in Figure 2 by considering the coverage of broad categories at level 7.

3.2 Category Deletion

We also observed that, 32328 number of categories have been dropped in Wiki-14. These are the categories in Wiki-12, but not in Wiki-14. Further analysis shows that, deletion of categories are due to (1) Renaming of categories. E.g., category *Asian countries* is renamed to *Countries in Asia*; *Electronic-commerce* to *E-commerce*, *etc*. (2) Category turning empty due to the reassignment of the pages under that category to other categories. E.g., category *Hellkite Records albums* became empty when the editors reassigned the pages under this category to other music related categories.

3.3 Administrative Category Evolution

Wikipedia has many categories that are used for administrative purposes. For e.g., the category “Articles needing additional categories” is used to group articles that need more specific categories, or may need additional categories. Wiki-12 has 85195 administrative categories and Wiki-14 has 93388. We used a simple pattern matching rule from (Ponzetto and Strube 2011) to detect administrative categories. These categories cover 72% (2994814) articles in Wiki-12 and 70% (3237366) in Wiki-14.

Broad Categories	Levels														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Law	11 (+1)	8 (+5)	4 (+15)	3 (+16)	3 (+13)	3 (+10)	6 (+76)	7 (+37)	8 (+27)	8 (+16)	8 (+13)	8 (+12)	8 (+10)	7 (+10)	7 (+10)
Medicine	10 (-42)	6 (-33)	2 (-21)	2 (-16)	1 (-22)	1 (-24)	1 (-28)	4 (+25)	5 (+5)	7 (+5)	7 (+7)	7 (+7)	7 (+8)	7 (+9)	7 (+9)
Sports	6 (+15)	5 (-5)	5 (-1)	10 (+12)	6 (+15)	4 (+17)	3 (+19)	2 (+20)	2 (+23)	2 (+27)	2 (+35)	2 (+46)	3 (+92)	3 (+151)	4 (+217)
Geography	3 (+1)	8 (+16)	5 (-26)	9 (-9)	12 (-15)	13 (-10)	12 (+9)	10 (+12)	9 (+11)	8 (+10)	8 (+11)	8 (+11)	8 (+11)	8 (+11)	7 (+11)
Agriculture	8 (-2)	4 (-6)	2 (-4)	2 (+4)	2 (+10)	2 (+4)	1 (+5)	2 (+7)	2 (+6)	3 (+19)	5 (+47)	6 (+35)	7 (+19)	7 (+6)	7 (+5)
Politics	8 (+11)	11 (+7)	7 (+2)	8 (+7)	9 (+12)	12 (+9)	12 (+14)	10 (+11)	9 (+8)	8 (+9)	8 (+10)	8 (+9)	8 (+10)	8 (+10)	7 (+10)
People	5 (+55)	6 (-98)	40 (-9)	24 (+6)	14 (+17)	11 (+38)	8 (+45)	7 (+44)	7 (+40)	7 (+31)	7 (+22)	7 (+17)	7 (+14)	7 (+11)	7 (+9)
Arts	5 (+16)	8 (+11)	11 (+20)	12 (+4)	11 (+12)	8 (+11)	7 (+6)	7 (+9)	7 (+10)	8 (+10)	8 (+11)	8 (+10)	8 (+10)	8 (+10)	7 (+10)
History	6 (-24)	8 (-17)	6 (-9)	8 (-23)	18 (+73)	15 (+16)	12 (+18)	10 (+16)	9 (+16)	8 (+11)	8 (+11)	8 (+11)	8 (+10)	8 (+10)	7 (+10)
Language	5 (-12)	8 (-1)	3 (-2)	5 (+3)	5 (-8)	6 (-6)	7 (-2)	8 (+2)	9 (+7)	8 (+9)	8 (+10)	8 (+10)	8 (+10)	8 (+10)	7 (+10)
Nature	2 (+4)	2 (-11)	3 (-13)	5 (-22)	6 (-62)	10 (-38)	12 (-7)	11 (+8)	9 (+10)	9 (+11)	8 (+10)	8 (+10)	8 (+10)	8 (+10)	7 (+10)
Technology	16 (-1)	13 (+9)	6 (+15)	7 (+14)	8 (+14)	8 (+12)	8 (+6)	8 (+4)	8 (+11)	8 (+10)	8 (+11)	8 (+10)	8 (+10)	8 (+10)	7 (+10)
Health	6 (-8)	7 (-11)	3 (+2)	3 (+2)	2 (-9)	2 (-22)	3 (-31)	5 (-16)	7 (-2)	8 (+3)	8 (+7)	8 (+10)	8 (+10)	7 (+9)	7 (+9)
Environment	9 (+27)	6 (+32)	3 (+50)	3 (+32)	4 (+45)	5 (+58)	7 (+92)	8 (+92)	8 (+72)	8 (+47)	8 (+31)	8 (+17)	8 (+13)	8 (+12)	7 (+11)

Table 1: Article coverage of broad categories at each level of breadth first traversal of category hierarchy starting from the broad category. Each cell entry is of the form x (+/-y), where x is the relative percentage of articles covered in by a given broad category in Wiki-14 and y is the percentage increase (or decrease) in the number of articles covered by the broad category from Wiki-12.

3.4 Disambiguation Page Evolution

Wikipedia a has a concept of disambiguation pages[1]. Disambiguation pages on Wikipedia are used to resolve conflicts in article titles that occur when a title is naturally associated with multiple articles on distinct topics. Each disambiguation page organizes articles into several groups, where the articles in each group pertain only to a specific topic. Disambiguations may be seen as paths in a hierarchy leading to different articles that arguably could have the same title. For example, the title *Apple*[2] can refer to a plant, a company, a film, a television show, a place, a technology, an album, a record label, and a news paper daily.

We observed that in Wiki-14, there has been an increase of 13% in disambiguation pages. Wiki-12 has 219404 and Wiki-14 has 248322 number of disambiguation pages. We found that 61235 number of disambiguation pages from Wiki-12 are updated in Wiki-14, which covers additional 120479 articles. This shows that Wikipedia editors are consciously improving the disambiguation pages.

3.5 Link Evolution

Links in Wikipedia are used to reference association between two categories (say parent-child) or between two articles (one article referring other article) or between a category and a page (page belongs to a category). The Table 2 summarizes the link statistics for each type of link. Most of the category-category links get dropped because of splitting of a category into more categories as shown in the Figure 4. B and R are new categories. The links A-P and A-Q are dropped, because of re-grouping of P,Q under B.

3.6 Evolution Summary

In Table 2 we summarize all these statistics.

[1] http://en.wikipedia.org/wiki/Wikipedia:Disambiguation

[2] http://en.wikipedia.org/wiki/Apple_(disambiguation)

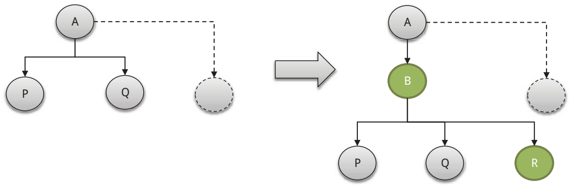


Figure 4: Regrouping of articles resulting in link deletion and addition

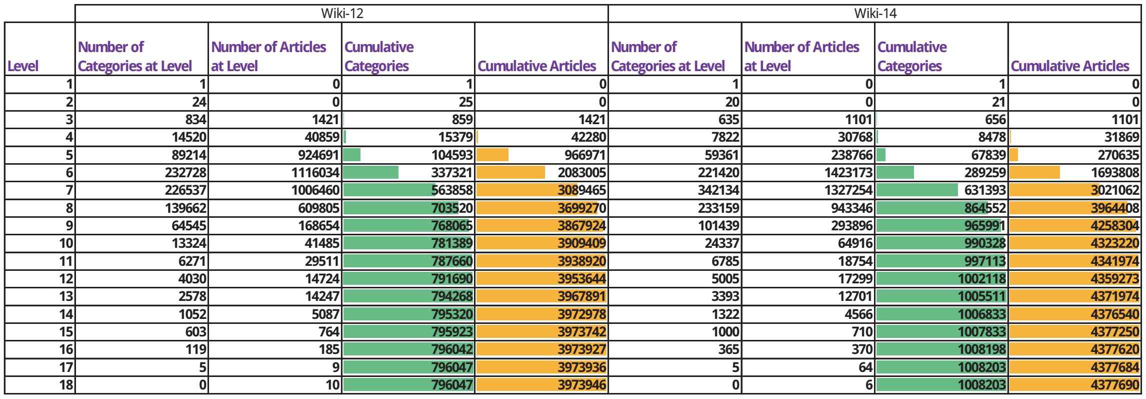
Head	Wiki-12	Wiki-14
Total Categories	906775	1132932 (+25%)
Total Articles	4119421	4627167 (+12%)
Total Category-Category Links	2243031	3137644 (+40%)
Total Category-Article Links	24486557	30505543 (+24%)
Total Disambiguation Pages	219404	248322 (+13%)
Total Administrative Categories	85195	93388 (+10%)
New Categories	-	258483
Deleted Categories	-	32328
New Articles	-	569456
Deleted Articles	-	61710

Table 2: Summary of evolution statistics

4 Category hierarchy

4.1 Category hierarchy statistics

Starting June 2004 Wikipedia added the concept of categories to organize the articles. Each category can have multiple parent and child categories. Wikipedia editors are free to create or link existing categories as a parent or a child. This has evolved the category hierarchy into a massive cyclic graph. However, we found that, most of the documents are covered (by breadth first traversal) within a depth of 9 or 10 in Wiki-12 and Wiki-14. Figure 5 shows the distribution of this coverage with the depth.



	Wiki-12				Wiki-14			
Level	Number of Categories at Level	Number of Articles at Level	Cumulative Categories	Cumulative Articles	Number of Categories at Level	Number of Articles at Level	Cumulative Categories	Cumulative Articles
1	1	0	1	0	1	0	1	0
2	24	0	25	0	20	0	21	0
3	834	1421	859	1421	635	1101	656	1101
4	14520	40859	15379	42280	7822	30768	8478	31869
5	89214	924691	104593	966971	59361	238766	67839	270635
6	232728	1116034	337321	2083005	221420	1423173	289259	1693808
7	226537	1006460	563858	3089465	342134	1327254	631393	3021062
8	139662	609805	703520	3699270	233159	943346	864552	3964408
9	64545	168654	768065	3867924	101439	293896	965991	4258304
10	13324	41485	781389	3909409	24337	64916	990328	4323220
11	6271	29511	787660	3938920	6785	18754	997113	4341974
12	4030	14724	791690	3953644	5005	17299	1002118	4359273
13	2578	14247	794268	3967891	3393	12701	1005511	4371974
14	1052	5087	795320	3972978	1322	4566	1006833	4376540
15	603	764	795923	3973742	1000	710	1007833	4377250
16	119	185	796042	3973927	365	370	1008198	4377620
17	5	9	796047	3973936	5	64	1008203	4377684
18	0	10	796047	3973946	0	6	1008203	4377690

Figure 5: Coverage of articles as a function of depth.

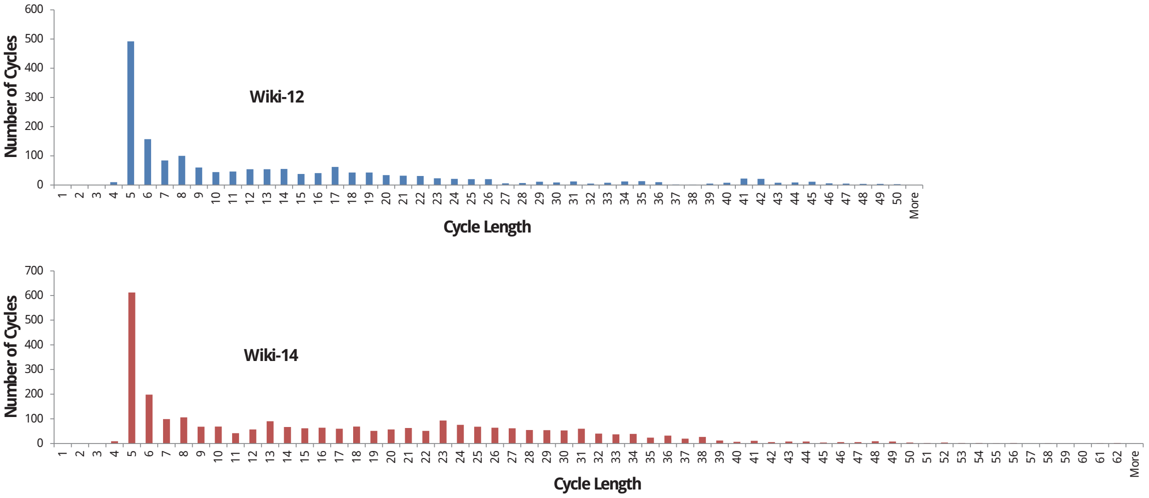


Figure 6: Cycle Length Histogram

4.2 Cycles in category hierarchy

Cycles are formed in the Wikipedia category hierarchy when a category A is assigned as category B's ancestor and the category B is assigned as A's ancestor. We have found 1766 number of cycles in Wiki-12 and 2804 in Wiki-14. These cycles vary in length from 4 to 50 in Wiki-12 and 4 to 62 in Wiki-14. The Figure 6 shows the distribution of these cycles over the cycle length.

4.3 Category hierarchy as isa graph

In many of the knowledge discovery tasks, Wikipedia category hierarchy is treated as an *isa* graph. Though it makes sense (in most of the cases) to treat the concept represented by a child category as a specific type of a broader concept represented by a parent category (e.g., Computer_science is a type of Applied_sciences), it is often the case that long distance *isa* relationship in category hierarchy does not make sense. For example, we can trace a path from the category "Optical_storage" to "Biology" as follows:

Optical_computer_storage is a descendant of *Biology* as per the following relation hierarchy: Optical_computer_storage → Computer_storage_devices → Recorders → Equipment → Technology → Intellectual_works → Creativity → Intelligence → Neuroscience → Biology

Similarly, from "Automotive_testing_agencies" to "Algebra" as follows:

Automotive_testing_agencies is a descendant of *Algebra* as per the following relation hierarchy: Automotive_testing_agencies → Automobiles → Private_transport → Transport → Industries → Economic_systems → Economics → Society → Structure → Dimension → Manifolds → Geometric_topology → Structures_on_manifolds → Algebraic_geometry → Abstract_algebra → Algebra

As explained in Section 3.1, our manual inspection of a few (39) randomly sampled documents show that, is-a relation does not make sense beyond 7 levels. We found that, between 3 to 5 levels, we can get a reasonably good *isa* relation.

5 Conclusion

We reported various statistics related to the categories and pages evolution between English Wikipedia 2012 and 2014. We found that evolution of categories at the rate of 25% to be higher than the evolution rate of articles which is 12%. The category structure can be treated as a cyclic graph covering 99% of the articles within a depth of 10 from the root category *Main_topic_classification*s. Although Wikipedia category graph can be treated as an *isa* hierarchy, our observations show that beyond certain levels, *isa* relation does not hold. We also analyzed the evolution of Administrative categories and Disambiguation pages in Wikipedia. Our observations show that, evolution is restructuring existing categories, articles and disambiguation pages to keep them coherent.

References

Kittur, A.; Chi, E. H.; and Suh, B. 2009. What's in wikipedia?: mapping topics and conflict using socially annotated category structure. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, 1509–1512. New York, NY, USA: ACM.

Ponzetto, S. P., and Strube, M. 2011. Taxonomy induction based on a collaboratively built knowledge repository. *Artif. Intell.* 175(9-10):1737–1756.

Suchecki, K.; Salah, A. A. A.; Gao, C.; and Scharnhorst, A. 2012. Evolution of wikipedia's category structure. *CoRR* abs/1203.0788.