First publ. in: Proceedings of International Workshop Bridging the Gap Between Semantic Web and Web 2.0, 4th European Semantic Web Conference (ESWC ´07), Innsbruck, Austria, June 7, 2007

Revision and Co-revision in Wikipedia*

Detecting Clusters of Interest

Ulrik Brandes and Jürgen Lerner**

Department of Computer & Information Science, University of Konstanz

Abstract. The online encyclopedia Wikipedia gives rise to a multitude of network structures such as the citation network of its pages or the co-authorship network of users. In this paper we analyze another network that arises from the fact that Wikipedia articles undergo perpetual editing. It can be observed that the edit volume of Wikipedia pages varies strongly over time, often triggered by news events related to their content. Furthermore, some pages show remarkably parallel behavior in their edit variance in which case we add a co-revision link connecting them. The goal of this paper is to assess the meaningfulness of the co-revision network. Specific tasks are to understand the influence of normalization (e.g., correlation vs. covariance) and to determine differences between the co-revision network and other relations on Wikipedia pages, such as similarity by author-overlap.

1 Introduction

Wikipedia[1] is a Web-based collaborative authoring environment, where anyone on the Internet can create and modify pages about encyclopedic topics. Since its creation in 2001, Wikipedia enjoys increasing popularity. At the end of 2006, Wikipedia has more than five million articles—about 1.5 million alone in the English Wikipedia—and grows by several thousand articles per day.[2]

There are several fundamental differences between Wikipedia pages and traditional articles (e. g., articles written for scientific journals or conference proceedings or entries in printed encyclopedias). Firstly, Wikipedia articles are written without centralized supervision, i. e., there are no editors deciding over which topics are treated and how much space is reserved for a certain entry. Furthermore, articles can be included and edited without a prior review process. Secondly, Wikipedia pages are written by up to thousands of authors, potentially having different education, knowledge, interests, and opinions on the topic. Thirdly, Wikipedia pages are never finished but undergo perpetual and frequent editing.

In this paper we focus on the latter two properties. Thereby, we have two goals in mind: The first is to better understand the content-generation process

* Research supported by DFG under grant Br 2158/2-3

** Corresponding author, lerner@inf.uni-konstanz.de

[1] http://www.wikipedia.org/

[2] http://stats.wikimedia.org/

of Wikipedia by tackling questions such as what is the typical edit volume of a Wikipedia page and how does it evolve over time, which pages are frequently revised during the same time periods, and which pages have many common authors. The second goal is to exploit these two properties to define similarity between pages: a *co-author similarity* measuring how much the author sets of two pages overlap and a *co-revision similarity* measuring to what extend two pages are edited during the same time intervals. Here we want to tackle how these measures have do be defined (e. g., which normalization is appropriate) such that meaningful and non-trivial similarity is obtained.

1.1 Related Work

Wikipedia has been established in 2001 to collectively create an encyclopedia. Maybe due to its size, popularity, and relevance for understanding new forms of collective knowledge creation, Wikipedia receives increasing interest in research. A study carried out by *Nature* in 2005 suggests that the accuracy of Wikipedia articles about scientific topics comes close to the accuracy of their counterparts in the *Encyclopaedia Britannica* [3]. Viégas *et al.* [7, 8] proposed a *history flow* approach for the visual analysis of the page history. A difference to our paper is that [7, 8] focus on the text of the page and we on the revision behavior. Work in [6] analyzes the information quality of Wikipedia articles by defining and measuring attributes such as authority, completeness, and volatility. The growth of Wikipedia is described in, e. g., [9, 1], whereas [4] analyze category-membership of articles. Other papers (e. g., [2, 5]) use the collection of Wikipedia articles to improve machine learning techniques for text categorization and detection of semantic relatedness of terms.

1.2 Input Data

Wikipedia makes its complete database (containing all versions of every article since its initial creation) available in XML-format.[3] The files containing the complete history of all pages can be extremely large. For instance, the complete dump for the English Wikipedia unpacks to more than 600 gigabytes (GB).[4] Wikipedia makes also available so-called stub-files. These files contain meta-data about every revision but not the text and are still quite large. For the present study we used the stub-file for the English Wikipedia (which is the largest one) from the 2006-11-30 dump with a size of 23 GB. (Note that this dump includes some revisions from December 2006, since it takes several days to create it.) More precisely, we used only the information "who performed when a revision to which page." Parsing the XML-document has been done with a Java implementation of the event-based SAX interfaces[5] which proved to be very efficient for parsing such huge files. Constructing the whole document tree, as

[3] `http://download.wikimedia.org/`
[4] `http://meta.wikimedia.org/wiki/Data_dumps`
[5] `http://www.saxproject.org/`

this is normally done by DOM parsers,[6] would simply be impossible (at least very inefficient and/or requiring uncommonly huge memory), given the file sizes. In the whole paper we consider only pages from the main namespace (i. e., we do not consider, discussion pages, user pages, user-talk pages, etc.). Some computations (especially in Sects. 3 and 4) are performed only for those pages that have more than 2000 edits. There are 1,241 pages in the 2006-11-30 dump satisfying this criteria (compare the remarks at the beginning of Sect. 3).

2 Statistics on Single Pages

The time-stamp of a revision denotes the exact second when this edit has been inserted in Wikipedia. When comparing the edit volume of Wikipedia pages over time, however, we adopt a much coarser point of view and consider their *weekly* number of edits. The decision "one week" is in a certain sense arbitrary and exchangeable by longer or shorter intervals of time. Furthermore, this decision certainly has an influence on the co-revision network defined in Sect. 3. However, we have chosen a week as this marks how people normally organize their work. Thus, a page that undergoes every week the same number of revisions but that is edited more often on week-ends than during the week is not considered to have a varying edit volume.

A second difficulty arises from the fact that Wikipedia pages are not all created at the same time. For instance, the page `2006 Israel-Lebanon conflict` does not even have the possibility to exist before 2006 (assuming that no author tries to predict the future). While this does not matter when we consider single pages, the problem has to be solved how to compare the edit volume of two pages that have different lifetimes. A first convention is to ignore the time when only one page existed, a second is to consider the longer time interval and take the point of view that pages received zero edits during the time when they did not yet exist. We will adhere to the second convention (more precisely we always consider the time from January 2001 until December 2006) for two reasons. Firstly, we do not want to ignore the fact that some pages are created earlier than others, as this already marks a difference between them. Secondly, measures like the covariance of the edit volume of two pages (used in Sect. 3) are hard to compare if we take them over different numbers of intervals (considering only the lifetime of the youngest Wikipedia page is obviously not an option, as this is simply too short).

Let p be a Wikipedia page and let $r_i(p)$ denote the number of revisions on page p in week i, where the weeks are assumed to be indexed with $i = 1, \ldots, K$. The value $R(p) = \sum_{i=1}^{K} r_i(p)$ is the total number of revisions on page p, $r_{\max}(p) = \max_{i=1,\ldots,K} r_i(p)$ the maximum number of weekly edits on page p, and $\mu_r(p) = R(p)/K$ the *mean* value (average number of edits per week). Furthermore, $\sigma_r^2(p) = \sum_{i=1}^{K}(r_i(p) - \mu_r(p))^2/K$ is the *variance* of p's edit volume (denoting the expected squared difference to its mean value) and $\sigma_r(p) = \sqrt{\sigma_r^2(p)}$ the *standard deviation*.

[6] `http://www.w3.org/DOM/`

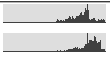
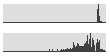
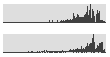
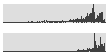
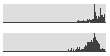
For a page p, let $A(p)$ denote the set of authors (logged-in or anonymous) that performed at least one edit to p and let $a(p) = |A(p)|$ denote the size of p's author set. Authors that are logged-in are identified by their username. The anonymous authors are identified by the IP-address of the computer from which they made the contribution. A problem arising from the inclusion of anonymous authors is that the same person might be logged-in using different IP-addresses, in which case we would count him/her several times. We have chosen to include anonymous authors since we observed that some of them make valuable and frequent contributions. Nevertheless, interpretation of the numbers of authors should take it into account that they probably contain many duplicates.

It is straightforward to aggregate values over a set of pages. For instance, if P is the set of all Wikipedia pages (from the main namespace), then $r_i = \sum_{p \in P} r_i(p)$ is the edit volume of Wikipedia in week i (for $i = 1, \ldots, K$).

2.1 Most-edited Pages

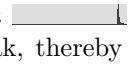
Table 1 lists the ten pages with the maximal average number of edits per week. Since we took the average over the same number of weeks for all pages (see above), these are also the Wikipedia pages having the highest number of edits in total. The last row in Table 1 denotes the values obtained by summing up the weekly edit number over all pages.

Table 1. The ten pages with the maximal average number of edits per week (real numbers are rounded to integer). The diagrams in the second column show the number of edits per week. The horizontal time-axis in these diagrams is the same for all pages (i. e., it goes over six years). In contrast, the vertical axis is scaled to unit height, so that the same height means a different number for different pages (maximum number of edits per week, corresponding to unit height, is denoted in the third column).

title(p)	$r_i(p)_{i=1,\ldots,K}$	$r_{\max}(p)$	$\mu_r(p)$	$\sigma_r(p)$	$a(p)$
`George W. Bush`		992	105	164	10,164
`Wikipedia`		630	70	115	9,275
`United States`		635	54	91	5,926
`Jesus`		735	50	87	4,302
`2006 Israel-Lebanon conflict`		3,679	48	**319**	2,755
`Adolf Hitler`		415	46	70	5,218
`World War II`		507	45	77	5,260
`Wii`		998	44	114	4,585
`RuneScape`		505	43	85	4,650
`Hurricane Katrina`		3,946	41	**246**	4,527
all pages		942,206	198,179	281,802	5.2 million

The topics of the most-edited pages span a broad range from people over countries and historic events to online games and a game console. However, the focus of this paper is on the differences and similarities in the revision characteristics of pages rather than their topics.

The numbers counting edits and authors appear to correlate quite well. A slight deviation from this rule is the page `2006 Israel-Lebanon conflict` having a smaller number of authors than other pages with so many edits (this page has only rank 68 in the list of pages having the most authors). The correlation (see the definition of correlation in Sect. 3.2) between the number of authors and the number of revisions (computed over all pages having at least 2,000 revisions, compare Sect. 3) is 0.88. Thus, pages with many authors indeed tend to have many revisions and vice versa.

Much more significant are the differences in the standard deviation (and thus also in the variance) of the edit volumes. For instance, the high variance of the pages `2006 Israel-Lebanon conflict` and `Hurricane Katrina` (printed in bold in Table 1) is probably due to the fact that interest in these pages is triggered by the events they describe. While theses two pages did not exist before the respective event, it turns out later that some pages that existed much earlier received also a high increase in interest at the respective time points. The edit plots of `2006 Israel-Lebanon conflict`  and `Hurricane Katrina`  show both a very narrow peak, thereby making the high variance visible.

The edit plots also reveal characteristics of other pages that are not so extremely reflected in the variance. For instance, the edit volume of `George W. Bush`  first increases. Then it suddenly drops and remains rather constant at a certain level. Probably the reason is that this page is a frequent target of vandalism and was the first page that became protected (compare [8]) resulting in a decrease in the number of edits.

The aggregated number of weekly edits for all Wikipedia pages is generally increasing, so that Wikipedia as a whole is more and more edited. Not surprisingly, the aggregated plot  is much smoother than those of single pages.

3 The Co-revision Network

Some Wikipedia pages appear to have quite parallel edit volumes, so that interest in these pages is raised at the same time. In this section we analyze the network arising if we consider two such pages as similar. In doing so we have two goals in mind: firstly to understand better which kind of pages are frequently *co-revised* and secondly to assess whether co-revision helps us to establish meaningful and non-trivial similarity of pages. Special emphasis is given in developing appropriate normalization for the edit plots and similarity values. In Sect. 4 we compare the co-revision network with the network derived from author-overlap.

In this section we are confronted with the problem that computing the co-revision network on all Wikipedia pages leeds to unacceptably high running times and memory consumption, since the matrix encoding the co-revision for all pairs of pages is obviously quadratic in their number (1.5 million in the main namespace). One possibility to make the computation fast enough (and still to analyze hopefully all interesting pages) is to reduce the number of pages

by considering only those that received a minimum number of edits. (A second advantage when doing so is that several normalizations become more stable since we do not divide by numbers that are too small.) In Sects. 3 and 4 we considered only those pages from the main name space that have at least 2000 edits (1,241 pages satisfy this criterion).

3.1 Covariance

The (*weekly*) *edit covariance* or (*weekly*) *revision covariance* of two pages p and q is defined to be

$$cov_r(p,q) = \sum_{i=1}^{K} [r_i(p) - \mu_r(p)] \cdot [r_i(q) - \mu_r(q)]/K \ .$$

The covariance is symmetric, i. e., $cov_r(p,q) = cov_r(q,p)$.

To get an overview of the covariance network we applied a quite simple method, described in the following, which will also be applied to the (much more useful) correlation network (see Sect. 3.2) and author-overlap network (see Sect. 4). Let $Cov = (cov(p,q))_{p,q \in P}$ denote the matrix containing the covariance values for all pairs of pages and let k be an integer. The graph G^k_{cov} is the graph whose edges correspond to the k entries in Cov having the highest values and whose vertices are the pages incident to at least one edge. The graph resulting from the 50 strongest edges is shown in Fig. 1. This network contains only the pages with the highest variances. In conclusion, covariance does not seem to yield insightful similarity of pages.

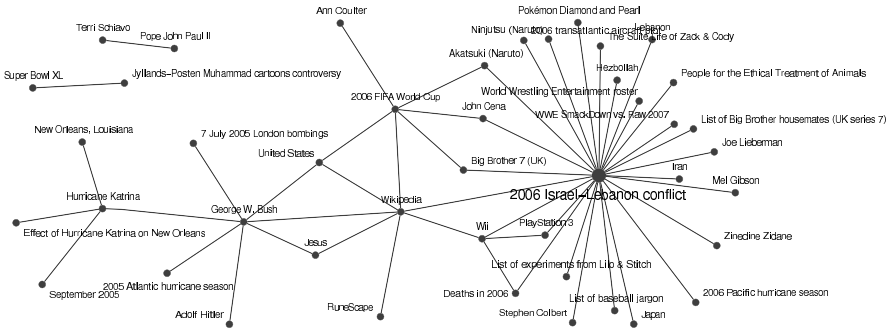


Fig. 1. Image of the graph G^{50}_{cov}. The central vertex corresponding to `2006 Israel-Lebanon conflict` is shown larger. This page has the highest variance of all Wikipedia pages and dominates the covariance network.

3.2 Correlation

Since edit covariance is highly influenced by the variance of the pages' edit volume it is reasonable to normalize these values. The (*weekly*) *edit correlation* or (*weekly*) *revision correlation* of two pages p and q is defined to be

$$corr_r(p,q) = \frac{cov_r(p,q)}{\sigma_r(p)\sigma_r(q)} \quad .$$

The correlation is symmetric (i. e., $corr_r(p,q) = corr_r(q,p)$) and in $[-1,1]$.

The two pages with the highest correlation are `Hurricane Katrina` and `Effect of Hurricane Katrina on New Orleans` which reach a correlation of 0.97 (i. e., close to the maximum). As for the covariance network we construct the graph G^{50}_{corr} (see Fig. 2) from the 50 edges corresponding to the largest correlation values.

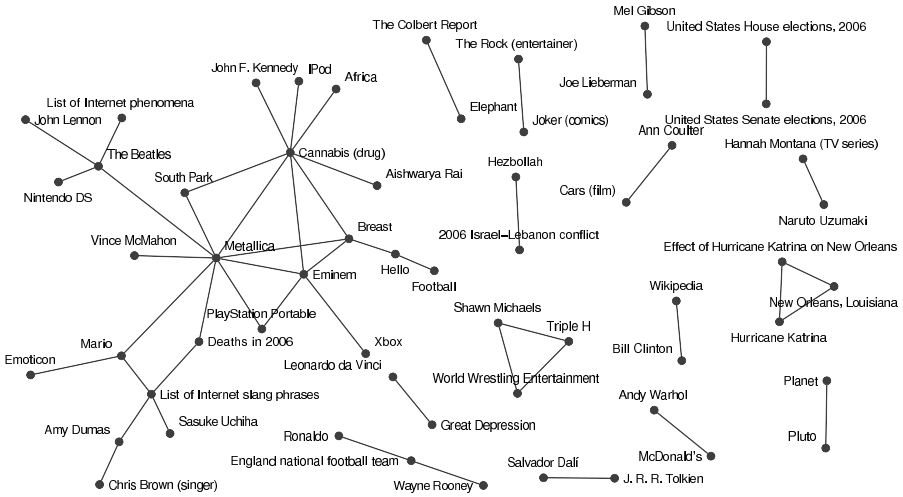
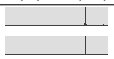
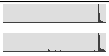
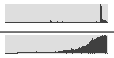


Fig. 2. Graph constructed from the 50 edges with highest correlation values.

Some of these correlations appear to be meaningful, others not. For instance it is reasonable that the three pages related to "Hurricane Katrina", the two pages related to the "2006 Israel-Lebanon conflict", and also the two pages `Pluto` and `Planet` are frequently revised at the same time (Pluto lost its status as a planet in 2006). Indeed, as Table 2 shows, some of the associated edit plots look remarkably similar, although they reach very different maximal values.

On the other hand, some correlations seem to be quite arbitrary. To understand why these pages are nevertheless so highly correlated we look at prominent members of the largest connected component in the left part of Fig. 2 and show

Table 2. Edit plots of selected pages showing a high correlation (compare Fig. 2).

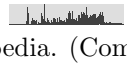
title(p)	$r_i(p)_{i=1,\ldots,K}$	$r_{\max}(p)$
`Hurricane Katrina`		3,946
`Effect of Hurricane Katrina on New Orleans`		1,099
`New Orleans, Louisiana`		533
`2006 Israel-Lebanon conflict`		3,679
`Hezbollah`		681
all pages		942,206
`Metallica`		138
`Cannabis (drug)`		221
`South Park`		212
`Eminem`		313

their edit plots in Table 2. What can be observed is that the edit plots of these pages do not look very special with respect to the aggregated edits of all pages. Especially the plots of `Cannabis (drug)` and `Metallica`, which are the most connected pages in Fig. 2, are very similar to the aggregated plot. So our current hypothesis is that some pages are just similar with respect to edit correlation because they are edited like the average Wikipedia page.

In conclusion, the similarity values derived by correlation of the weekly number of edits are much better than those derived from covariance. However, while a high correlation might point to a meaningful connection between the pages it is not necessarily so. The major drawback of correlation seems to be that pages that are edited as the average Wikipedia are assigned high similarity values, independent on whether they treat related topics. In the next subsection we attempt to filter this out.

3.3 Relative Edit Volume

Considering the strongly skewed aggregated edit volume of Wikipedia and having in mind the remarks at the end of the previous subsection, it may be worthwhile to consider the *relative edit volume* of individual pages, i. e., the percentage that a specific page receives from the weekly edits done in the entire Wikipedia. So, let $r_i(p)$ denote the number of edits of page p in week i and r_i denote the total number of edits on all Wikipedia pages in week i. Then $r_i(p)/r_i$ is called the *relative number of edits* of page p in week i. This yields the measures *relative edit covariance* and *relative edit correlation*, compare Sects. 3.1 and 3.2.

The plots showing the relative edit volume reveal some interesting characteristics of the pages. For instance the page `George W. Bush`  receives high (relative) interest already in the early days of Wikipedia. (Compare the plot showing the absolute number of edits  which begins to rise later.) Even more extreme is the difference between the relative edit plot of `Rammstein`  showing a single peak at the beginning of Wikipedia and its absolute edit plot  which indicates more interest in later years.

The comparison between the relative and the absolute edits also provides a distinction between pages that are solely edited during a certain event and pages that only show a strong increase in interest during events. For instance, the absolute edit plots of `2006 Israel-Lebanon conflict`  and `Hezbollah`  are very similar. On the other hand, their relative plots reveal that the page `2006 Israel-Lebanon conflict`  is relative to the whole Wikipedia still focused on that event, whereas the page `Hezbollah`  receives the most edits (relative to the whole Wikipedia) much earlier.

Motivated by such examples we thought that *relative covariance* and *relative correlation* would yield similarity values which are more reliable than their counterparts derived from the absolute edit volume. However, it turned out that this is not the case. Instead, both the relative covariance and the relative correlation are dominated by a few edits in the early days of Wikipedia when the aggregated number of edits was by orders of magnitude smaller than in later years.

In conclusion, normalizing the edit volumes by the aggregated number of edits seems to be a natural way to prevent that pages become similar just because they behave like the average page (compare Sect. 3.2). However, since the aggregated edit volume  is highly skewed this involves division by very small numbers (compared to the largest ones) and thus yields a highly unstable method. It is an issue for future work to develop a more appropriate normalization.

4 The Co-author Network on Pages

Some Wikipedia pages have thousands of authors. In this section we consider similarity of pages derived from overlapping author sets. As in Sect. 3 we have two goals in mind: firstly to understand better which kind of pages are frequently *co-authored* and secondly to assess whether co-authoring helps us to establish meaningful and non-trivial similarity of pages. In addition we want to compare the co-revision and co-author network. The term "co-author network" often denotes networks of authors (in contrast, we construct a network of pages) connected by commonly written articles. However, in this section we consider only the network of Wikipedia pages resulting from overlapping author sets.

A first possibility is to define similarity of pages by simply counting the number of common authors, i. e., taking the values $a(p, q) = A(p) \cap A(q)$ as a measure of author overlap between two pages p and q. (We remind that $A(p)$ denotes the set of authors (logged-in or anonymous) of a page p and $a(p) = |A(p)|$ denotes the number of its authors.)

The two pages with the highest number of common authors (namely $1,074$) are `George W. Bush` and `Wikipedia` which are also the two pages having the largest number of authors (both roughly $10,000$). As for the co-revision network (compare Figs. 1 and 2) we construct the graph arising from the 50 strongest values in $a(p, q)$, see Fig. 3. As it could be expected, the un-normalized co-authoring similarity $a(p, q)$ is highly biased towards pages with large author sets

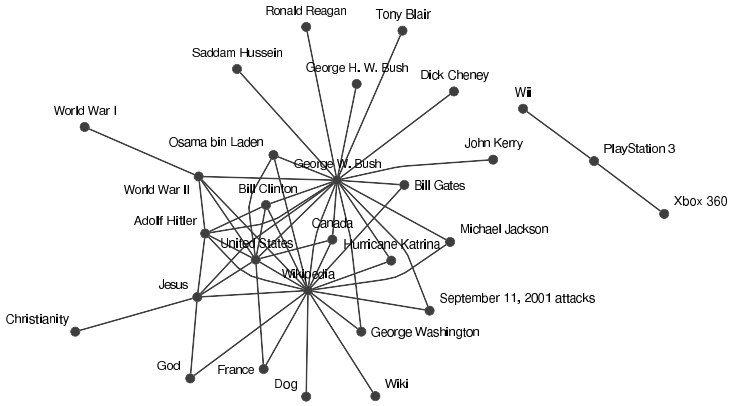


Fig. 3. Graph whose edges are the 50 pairs having the highest numbers of common authors. This graph is dominated by pages having the largest numbers of authors.

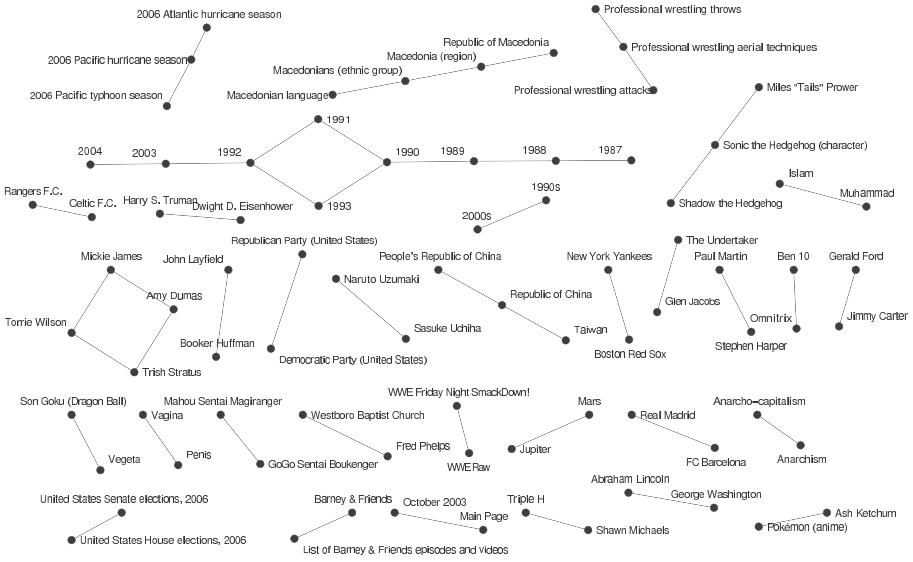


Fig. 4. Graph whose edges are the 50 pairs having the highest normalized numbers of common authors. The connected components correspond in most cases well to pages with similar topics.

although a reasonable cluster containing three pages about game consoles is identified.

Similar to the normalization of covariance to correlation we normalize the number of common authors by dividing with the geometric mean of the numbers of authors:

$$a_{\cos}(p,q) = \frac{a(p,q)}{\sqrt{a(p)a(q)}} \ .$$

(The notation $a_{\cos}$ has been chosen since this measure is the cosine of the angle between the characteristic vectors of the two author sets.) The normalized number of common authors $a_{\cos}(p,q)$ ranges between zero and one. The highest value is between the two pages `2006 Atlantic hurricane season` and `2006 Pacific hurricane season` (reaching a value of 0.27). The 50 strongest values give rise to the graph shown in Fig. 4. The connected components of this graph appear to be quite reasonable as they normally consist of pages treating strongly related topics.

It is remarkable that the graphs stemming from correlation in the number of weekly edits (Fig. 2) and from normalized author overlap (Fig. 4) are almost disjoint (an exception are the United States elections 2006). Indeed, from a qualitative point of view the former seems to connect pages that are related to the same events and the latter to connect pages having related topics. Co-revision and co-authoring also are quite independent from a quantitative point of view: the correlation between the values $a_{\cos}(p,q)$ and $corr_r(p,q)$ (see Sect. 3.2) is 0.18 and the correlation between the values $a(p,q)$ and $cov_r(p,q)$ (see Sect. 3.1) is 0.39. (Note that we performed this computation only for all pairs of pages that have more than 2000 edits, so that the results might not generalize to the set of all Wikipedia pages.) These low correlation values indicate a rather weak dependence between co-revision and co-authoring similarity. The somewhat higher correlation between the un-normalized versions is probably due to the fact that pages with higher covariance normally have a higher total number of edits and, thus, more authors (compare Table 1). In conclusion, co-revision and co-authoring seems to be quite different—at least for the pages with many edits.

5 Discussion and Future Work

In contrast to traditional articles, Wikipedia pages have huge author sets and are permanently edited. This paper analyzes these two properties and achieves some initial findings.

The plots of the edit volume of pages over time reveal some interesting characteristics: For instance, some pages show an overall increase in interest; others are mostly edited during certain events. Furthermore, the plots of two pages shown simultaneously reveal whether these pages are edited in parallel.

When analyzing co-revision similarity, it became evident that correlation performs much better than covariance, since the latter is biased towards pages of very high variance. The similarity derived from correlation seems to be quite

meaningful for some pages and rather arbitrary for others. Our current hypothesis is that pages that are edited as the average Wikipedia page receive quite large correlation values, independent on whether they are somehow related or not. Attempts to filter this out by considering the relative edit volumes failed due to the highly skewed distribution of the aggregated edit volume. It is an open problem to develop a better normalization strategy.

The normalized version of the co-authoring similarity seems to yield quite meaningful associations. Co-authoring similarity and co-revision similarity appear to be rather unrelated, so that co-revision might point to complementary relatedness of pages. Furthermore, co-revision can be applied to relate pages written in different languages whose author sets are normally non-overlapping.

Further issues for future work include developing appropriate clustering algorithms for the co-revision or co-authoring networks, analyzing how the content of a page changes after it received a peak of interest (pages such as `Hezbollah` or `New Orleans, Louisiana`, compare Table 2), and comparing co-revision and co-authoring to other network structures such as hyperlinks pointing from one page to another or common membership in categories.

References

1. L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the wikigraph. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*, pages 45–51, 2006.
2. E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21'st National Conference on Artificial Intelligence*, 2006.
3. J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.
4. T. Holloway, M. Božičević, and K. Börner. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. arXiv:cs/0512085.
5. M. Strube and S. P. Ponzetto. WikiRelate! computing semantic relatedness using wikipedia. In *Proceedings of AAAI'06*, 2006.
6. B. Stvilia, M. B. Twindale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality*, 2005.
7. F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582, 2004.
8. F. B. Viégas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in Wikipedia. In *Proceedings of HICSS 40*, 2007.
9. J. Voss. Measuring Wikipedia. In *Proceedings of the International Conference of the International Society for Scientometrics and Informetrics*, 2005.