On the Evolution of Wikipedia

Rodrigo B. Almeida
UCLA Computer Science Department
Los Angeles - USA
barra@cs.ucla.edu

Barzan Mozafari
UCLA Computer Science Department
Los Angeles - USA
barzan@cs.ucla.edu

Junghoo Cho
UCLA Computer Science Department
Los Angeles - USA
cho@cs.ucla.edu

Abstract

A recent phenomenon on the Web is the emergence and proliferation of new social media systems allowing social interaction between people. One of the most popular of these systems is Wikipedia that allows users to create content in a collaborative way. Despite its current popularity, not much is known about how users interact with Wikipedia and how it has evolved over time.

In this paper we aim to provide a first, extensive study of the user behavior on Wikipedia and its evolution. Compared to prior studies, our work differs in several ways. First, previous studies on the analysis of the user workloads (for systems such as peer-to-peer systems [10] and Web servers [2]) have mainly focused on understanding the users who are *accessing* information. In contrast, Wikipedia's provides us with the opportunity to understand how users *create* and *maintain* information since it provides the complete evolution history of its content. Second, the main focus of prior studies is evaluating the implication of the user workloads on the system performance, while our study is trying to understand the evolution of the data corpus and the user behavior themselves.

Our main findings include that (1) the evolution and updates of Wikipedia is governed by a *self-similar process*, not by the Poisson process that has been observed for the general Web [4, 6] and (2) the exponential growth of Wikipedia is mainly driven by its rapidly increasing user base, indicating the importance of its open editorial policy for its current success. We also find that (3) the number of updates made to the Wikipedia articles exhibit a power-law distribution, but the distribution is less skewed than those obtained from other studies.

Keywords

Wikipedia, user behavior, social systems

1. Introduction

The Web has been, from its very beginning, much different from other content creation media. While the previous media were mainly governed by centralized organizations (e.g. publishers in the case of books or TV networks for television shows), the Web allowed anyone to freely create and publish content without any need of third-party approval. Because of this uncoordinated and decentralized nature there was no easy way of predicting beforehand how it would evolve over time. So far, several studies have focused on understanding and characterizing the evolution of this huge repository of data [5, 11].

Recently, a new phenomenon, called social systems, has emerged from the Web. Generally speaking, such systems allow people not only to create content, but also to easily interact and collaborate with each other. Examples of such systems are: (1) Social network systems such as MySpace or Orkut that allow users to participate in a social network by creating their profiles and indicating their acquaintances; (2) Collaborative bookmarking systems such as Del.icio.us or Yahoo's MyWeb in which users are allowed to share their bookmarks; and (3) Wiki systems that allow collaborative management of Web sites.

The focus of this paper is on the third kind, more specifically, on Wikipedia which is the largest publicly available Wiki [17]. The idea of Wikis was introduced in 1995 by Ward Cunnungham [12, 18, 17] as a programming language pattern enabling people not only to access the data of a Web site but also to change it. Every editable page in a Wiki has an option that allows (registered or anonymous) users to change it according to their own interests. Although some pages in a Wiki site may have restricted access, e.g. some pages can be modified only by administrators, the majority of them can be freely edited.

Wikipedia was launched in January 2001 by Jimbo Wales and Larry Sanger [18] with the goal of becoming a collaborative, free-content, encyclopedia using the Wiki approach. As of October 2006, the Wikipedia is comprised of more than 5 million articles on a wide variety of subjects written in more than 200 different languages. A recent, much discussed, article from Nature [8] compares Wikipedia with The Britannica Encyclopedia and argues that, despite its anarchical functioning, the former comes close to the latter in terms of the accuracy of its science entries.

This paper tries to model the behavior of users contributing to Wikipedia (hereafter called contributors) as a way of understanding its evolution over time. It presents what we believe to be the first extensive effort in that direction. This understanding will allow us, in the future, to create a model for Wikipedia evolution that will be able to show its trends and possible effects of changes in the way it is managed. Several studies have characterized the user behavior of different services, such as peer-to-peer systems, streaming media, etc., but our work goes in a different direction since Wikipedia allows us to model the behavior of *information producers* instead of the behavior of information consumers and to track the effects of such behavior.

Metric type	Metric	Value
Article	# of article entries	2.5 million
	# of actual articles	1.3 million
	# of redirected articles	1.2 million
Graph structure	# of links connecting articles	58.9 million
	# of broken links	6.5 million
	# of links to redirected articles	6.8 million
Article history	# of revisions	48.2 million
	# of identified revisions	33.4 million
	# of distinct registered contributors	495.7 thousand
	# of distinct IP addresses	3.8 million

Table 1: *General statistics of the Wikipedia data used in this study*

Our hope is that the characterization presented will be useful both in the design of more effective social systems and in increasing our understanding of the underlying driving forces governing content creation by humans. The main findings of this study are summarized below:

- Wikipedia evolution follows a *self-similar process* both in terms of revisions to existing articles and creation of new ones. This result is in contrast to the findings from prior studies [4, 6], where researchers found that the changes to the Web pages follow a Poisson Process.
- The number of articles on Wikipedia has been growing exponentially since its creation in 2001. This growth is mainly driven by the exponential increase in the number of users contributing new articles, indicating the importance of the Wikipedia's open editorial policy in the current success. Interestingly, we find that the number of articles contributed by each user has decreased over time as the Wikipedia grows larger.
- We observe a clear separation of Wikipedia contributors into two distinct groups when we look at the total number of articles contributed by each user. We also observe that most users tend to revise existing articles rather than creating new ones.
- Article popularity, in terms of the frequency of updates to it, follows a power-low distribution, but the distribution is less skewed than the ones reported in the study of other read-dominated workload.
- Although contributors tend to have a wide range of interests with respect to the articles they contribute to, in a single interaction with Wikipedia, they tend to center their updates around a single main article.

The remainder of this paper is organized as follows. Section 2 discusses related work. The data acquisition and cleansing is discussed in Section 3. The results of our characterization are presented in Section 4. Finally, Section 5 presents conclusions and possible future work directions.

2. Related work

Since its inception, Wikipedia has been gaining popularity and nowadays it is consistently ranked in the top 20 most popular sites according to Alexa (`http://www.alexa.com`). In spite of its popularity only an small number of studies have focused on it. Some effort has been dedicated to improving Wikipedia's functionality [18] and to use it as a source of information for Information Retrieval tasks [1]. Our work goes in a different direction: we aim at understanding the behavior of Wikipedia contributors as a way to understand its general evolution. The work in [17] follows a similar goal but is different from ours in the sense that the authors do not try to model contributor interaction in Wikipedia as a whole. Instead, they focus on proposing a visualization framework that can be used to evaluate cooperation and conflict patterns on individual articles, in addition to providing some general statistics regarding the whole Wikipedia.

A number of characterizations of different workloads types have been conducted, such as peer-to-peer systems [10], chat rooms [7], e-mail [3], streaming media [16], e-business [15], and Web servers [2]. We borrow from these studies the techniques they have used, but there is a fundamental difference between their analysis and ours. Their focus was primarily performance evaluation its influences on system design and we use the tools to understand contributor behavior and its impact on Wikipedia evolution.

3. Data acquisition and cleansing

In this section we discuss the data acquisition and cleansing steps of our work. Section 3.1 discusses data acquisition and Section 3.2 discusses data cleansing.

3.1 Data acquisition

Wikipedia is currently maintained by a non-profit organization called Wikimedia Foundation which has the goal of encouraging the growth, development and distribution of free, multilingual content. Moreover Wikimedia also aims at providing the full content of these Wiki-based projects to the public free of charge. Wikipedia is only one of many other Wikis supported by Wikimedia such as Wiktionary, Wikibooks, Wikiquote, Wikinews, Wikisource, Wikimedia Commons, and Wikispecies [19].

Because of its goal, not only does Wikimedia make the content of its Wikis available through specific Websites but also through dumps of their whole data that can be downloaded by end users (see `http://download.wikimedia.org/`). Although the idea was to have a weekly dump, because of the size of the data and computational resource limitations, the Wikimedia Foundation has not been able to do so very effectively. The dumps, specially the ones for the largest Wikis (e.g. the English version of the Wikipedia), often do not contain all of the data they should because of some uncorrected errors that occurred during their generations.

Our study is based on a dump of the English version of the Wikipedia generated on August 16^{th} of 2006. This was

the latest full dump available when we were gathering data for our study and contained the whole history of Wikipedia. The dump is comprised of 15 Gigabytes of compressed data that, when uncompressed, takes more than 700 Gigabytes of storage.

3.2 Data cleansing

The dumps provided by Wikimedia Foundation contain several files in different formats (e.g. SQL statements, XML files, etc.) which can be parsed and loaded on a regular relational DBMS using MediaWiki [14], the software used by Wikimedia Foundation to manage its Wiki sites. This method would not be of practical use to us because of computational resource limitations and also the fact that we do not need to access the whole data for our analysis.

Our approach was to create a custom-made parser for Wikipedia's data that extracts only the information that was useful to us. After we parsed the data we were left with the list of Wikipedia's articles and their links; the list of updates that have been made to each article and the contributor/IP address responsible for such update. We use the term *update* both for revisions of existing articles and for the creation of a new one. Whenever a differentiation between these two types of updates needs to be made, it is noted in the text.

When an article is updated by an unregistered contributor, the contributor information is not available and the IP of the machine is recorded instead. Wikimedia also has some automated scripts that are used every now and then to update articles from Wikipedia. These updates are also logged as anonymous contributions. Therefore, it is hard identify which anonymous modifications were done by real humans and which were batch scripts run by the Wikimedia Foundation. Due to this difficulty, in our study, we only considered the updates made by registered contributors.

Another important observation is that the data available in the dump contains not only the actual articles from Wikipedia, but also other articles such as help, disambiguation or administrative articles. These different types of articles are organized in different namespaces. We ignored these "meta" data and considered only the namespaces that are used to store the actual Wikipedia articles.

(a) Original hyperlink structure (b) Post-processed structure

Fig. 1: *The redirection framework for Wikipedia*

Table 1 summarizes the information obtained by our parser. These results are further discussed below. The number of article entries found was 2.5 million. These entries were then, subdivided into two categories: actual articles and redirect articles. The creation of redirect articles is the way Wikipedia handles the fact that one concept may have different names. Instead of creating one specific article for each of the different names a concept has, one main article is created and all of the articles for the other different names just mention that the user should be redirected to the main one. Figure 1(a) exemplifies this mechanism. In this example we have three articles named R, A and B. A points to B and B points to R which in turn is a redirection to A, i.e. whenever a contributor requests the article named R, she is automatically redirected to the article named A and is informed that this redirection was taken. The number of actual articles and the number of redirection articles is approximately 1.2 million each.

Figure 1(b) shows our general approach for handling redirection which is to aggregate all of the information related to R into the information of A. As a result of this procedure: all of the links directed to R became directed to A; all of updates for R became updates for A and so on. It is important to point out that our analysis is based on the current version of Wikipedia and as a consequence of that, we consider only the current redirections; therefore, it may be the case that in the past R and A could have been different articles about the same thing and at some point, some contributor decided to merge this two articles by redirecting R to A. Similarly, the redirection from R to A may be removed in the future and R to A may become completely distinct articles.

With respect to the graph structure formed by the Wikipedia's, we found that the total number of links in this graph is 58.9 million (an average of 45 links per article). The number of links that had to be rewritten as a result of redirection is 6.8 million. Also, among 58.9 million links, 6.5 links were broken.

As a final remark of the general characteristics of the data, the total number of revisions to the articles on Wikipedia is 48.2 million. Among this, 33.4 million revisions have been made by 495.7 thousand registered contributors while the other revisions have been made by 3.8 million distinct IP addresses. The greater number of IP addresses found compared to the number of registered contributors can be accounted to by many factors such as the use of DHCP or the fact that the number of different people willing to register and to effectively help improve Wikipedia is smaller than the number of people that have eventually played with it and decided not to become active contributors.

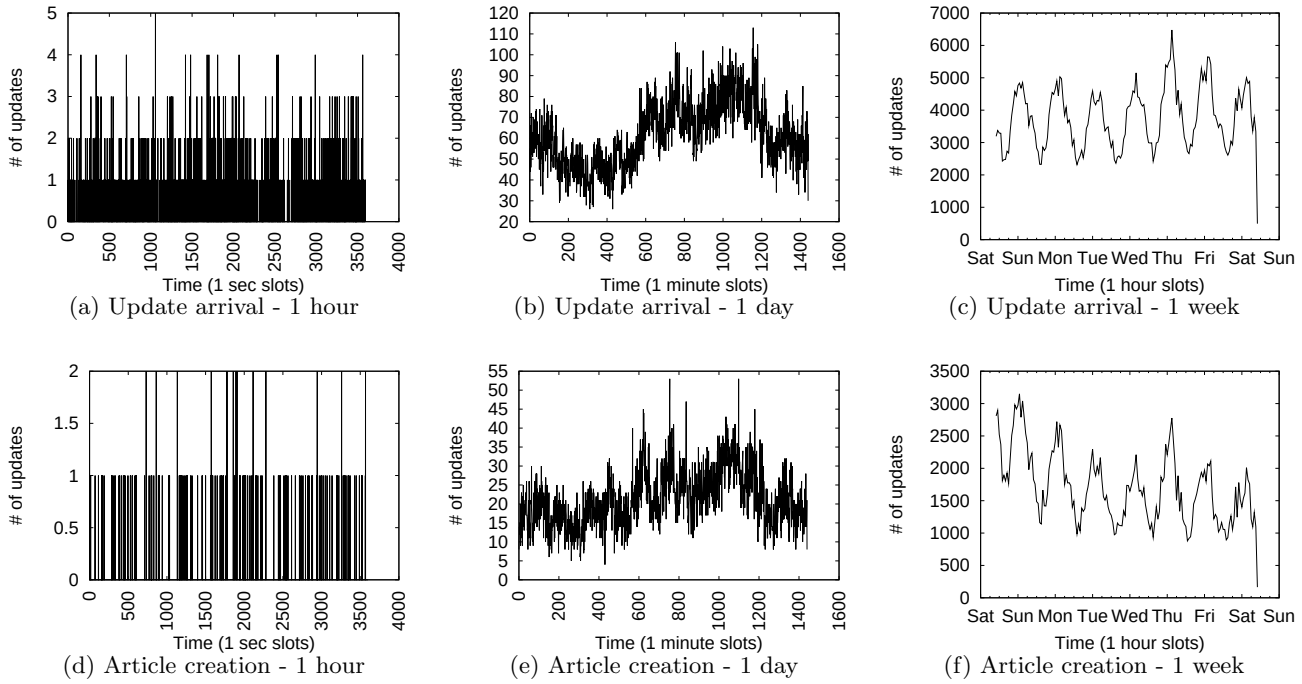
4. Characterization results

In order to understand the interaction of contributors with Wikipedia, we looked at their interactions at three levels — the individual update level, the session level and the contributor level — similarly to work done for e-commerce [2], streaming media [16] and static Web servers [15].

At the individual update level, we do not differentiate the contributor of each update and analyze the aggregate behavior of the updates made by all contributors. At the second level, we group the updates into a set of sessions and analyze their properties. A session is composed of a set of updates from a specific contributor that were issued at a similar point in time and are used to model individual interactions of the contributors with the service. In particular, we have set the session threshold as being 3 hours, i.e. any period of inactivity from any contributor that is larger than 3 hours triggers the creation of a new session.[1] At the third, contributor level, we aggregate all updates made by each contributor and study the individual properties of each contributor.

In the next few sections, we report the main results of the characterization we have conducted and discuss the effects

[1] We also used different threshold values for the session timeout, and the results were roughly equivalent to what we report here.



(a) Update arrival - 1 hour (b) Update arrival - 1 day (c) Update arrival - 1 week

(d) Article creation - 1 hour (e) Article creation - 1 day (f) Article creation - 1 week

Fig. 2: *Time series for the arrival processes for updates and article creation*

of the results on the evolution of Wikipedia. Section 4.1 discusses the general processes governing contributor interaction with Wikipedia in terms of updates. The growth of Wikipedia is discussed in more detail in Section 4.2. Section 4.3 shows some results based on the analysis of contributor-centered metrics. Finally, Section 4.4 shows the analysis in terms of an article centered approach.

4.1 Self-similarities in Wikipedia evolution

One of the key characteristics of stochastic processes is their arrival process.

In our case, we wanted to examine two processes: (1) the arrival process of updates in general and (2) the arrival process of the creation of new articles. A usual way of analyzing such data is through the the time series plots for the events generated by the process. Let X_t be a random variable that captures the number of events of a certain process that happened in the time interval $[t-1, t)$. For example, if the unit time is one second, then X_1 represents the number of events that occurred during the first second, X_2 represents the number of events that occurred during the second second and so on. Figure 2 shows the time-series plots the updates received ((a), (b) and (c)) and for the creation of new articles ((d), (e) and (f)) at three different time scales, with the unit time varying from 1 second to 1 hour.

Recent studies have shown that the evolution of the Web as a whole follows a Poisson process [4, 6]. In the Poisson process the time between successive events follows an i.i.d. exponential distribution. As a result of this characteristic the Poisson process is said to aggregate well. In other words, as the unit time goes from a small timescale, e.g. 1 second, to a larger one, e.g. 1 hour, X_t tends to get smoother approaching a straight line. Our plots show that the X_t time-series for the Wikipedia data still preserves some of its variability and burstiness, differently from the prediction of the Poisson model.

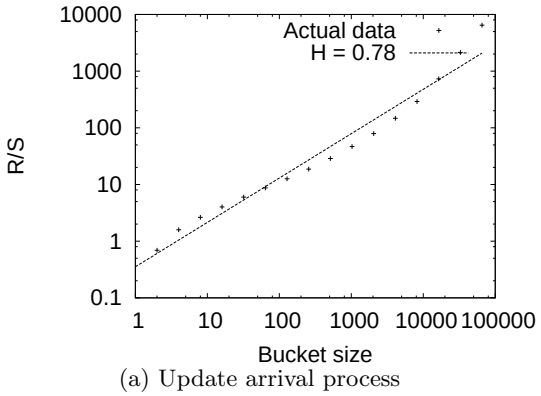
Another class of processes, the so called self-similar processes, have been extensively used to model user access and network traffic patterns. These kinds of processes typically exhibit what is called a "fractal-like" behavior, i.e. no matter what time scale you use to examine the data, you see similar patterns. The implication of such a behavior is that: (1) X_t is bursty across several time scales, (2) there is no natural or expected length of a burst, and (3) X_t does not aggregate as well as a Poisson process, which is close to what is observed in Figure 2.

To investigate the appropriate model for this process more formally, we analyzed the data using a statistical method. There exists several methods to evaluate the self-similarity nature of a time-series. Each of this methods explores one or more of the statistical properties of self-similar time-series as a way to estimate the self similarity of a sample data. These properties include the slowly decaying variance, long range dependency, non-degenerate autocorrelations and the Hurst effect.

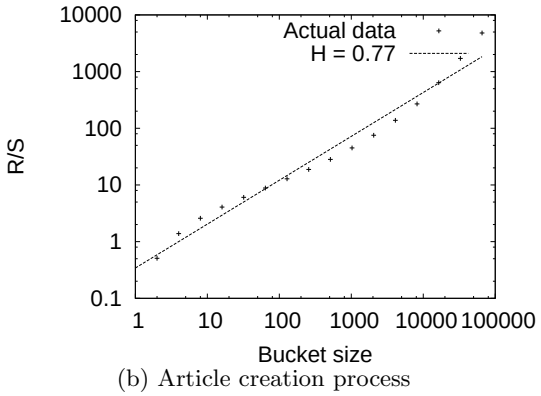
In this study we have used the rescaled adjusted range statistics in order to estimate the Hurst effect for the processes being analyzed [9]. Let X_t be the time series being analyzed, the Hurst effect is based on the fact that for almost all of the naturally occurring time series, the rescaled adjusted range statistic (R/S statistic) for a sequential sample of X_t with size n obeys the following relation:

$$E[R_n/S_n] = Cn^H \quad (1)$$

The parameter H of Equation 1 is called the Hurst parameter and lies between 0 and 1. For Poisson processes $H = 0.5$. For self similar processes $0.5 < H < 1$, so this parameter is easy-to-use single parameter that is able to characterize whether



(a) Update arrival process



(b) Article creation process

Fig. 3: *Hurst parameter estimation*

or not a process is self-similar.

Therefore, in order to find the Hurst parameter for the process we are analyzing, we took sequential samples from X_t with different sizes and compute the value of R_n/S_n for these samples. Let $\bar{X_n}$ be the average for a sample and σ_n be its standard deviation, then, the R_n/S_n value for each sample can be computed as follows:

$$R_n = max(W_1, W_2, ..., W_n) - min(W_1, W_2, ..., W_n) \quad (2)$$

where:

$$W_k = (\sum_{1 \le i \le k} X_i) - k\bar{X_n} \quad (3)$$

and

$$S_n = \sigma_n \quad (4)$$

Figure 3 shows the R/S plots for the two processes studied and the different samples sizes. The graph show the empirical values for the R/S statistics that have been found from our time series and the value for H that was found by the best fit for Equation 1 over the actual data. As can be seen from the figure, both processes are self-similar since $H > 0.5$ for all of the time series. These results show that the change in Wikipedia articles is different from the change in the whole Web which is known to follow a Poisson process [4, 6]. The fact that changes to a data-source follows a Poisson process have been used to devise techniques for keeping copies of a certain data-source up-to-date in an effective way and given our result, it might have been the case that the techniques need to be reevaluated to see whether or not they are still effective in a context in which the evolution of a data-source does not follow a Poisson process.

4.2 Wikipedia growth

Another point regarding the page creation process is how it determines the number of articles available in Wikipedia over time. Let A_t be the cumulative number of articles that have been created until time t. Figure 4(a) shows a plot for A_t as a function of time. In the figure, the vertical axis is logarithmic. From the figure, we can see that the number of articles on Wikipedia had been growing more than exponentially in the beginning, but it has stabilized into an exponential growth since year 2003. We curve fitted this portion of the graph to the following equation

$$A(t) = C \times e^{a \times t} \quad (5)$$

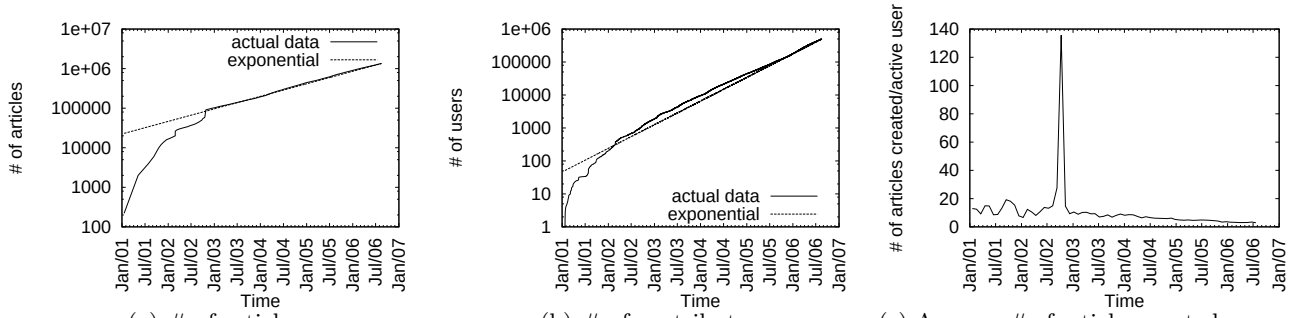
and found that $a = 2.31 \times 10^{-8}$.

A related question that arises from this fact is what is the main cause of such an exponential growth. Potentially, there are two sources for the growth: (1) the increase in the productivity of the Wikipedia contributors and/or (2) the increase in the number of new contributors to Wikipedia. In order to see which of these two factors is the primary source of the Wikipedia growth, we have analyzed the cumulative number of unique contributors over time and also the average number of articles created per contributor. These results can be seen in Figures 4(b) and (c) respectively. From Figure 4(b), we can clearly see that the number of unique contributors have also increased exponentially with the a parameter of the exponential curve being 5.23×10^{-8}. The fact that the a parameter of this curve is significantly larger than that of Figure 4(a) strongly indicates that the growth of the Wikipedia articles are mainly due to the rapidly expanding contributor base, not by the increase of the contributor productivity. To further verify this fact, Figure 4(c) shows the average number of articles created by each contributor that was active monthly, i.e. the number of articles created over the number of contributors that have created at least one article in that month. The analysis of this plot corroborates the fact that the average productivity of each contributors instead of getting higher is, in fact, decreasing. The sudden peak in the graph around October of 2002 was caused by a massive bot-creation of pages for US towns based on census data.

By a more careful investigation of average productivity of our users, we observed that different generations of users exhibit different behaviors. To see this we categorized our users into different groups according to the time they have registered in Wikipedia, and interestingly we can observe that as our old users get tired of creating new articles and lose their enthusiasm, users that recently joined Wikipedia are getting more passionate about creating new articles. As an example of this difference we have shown the average number of article creation per user, for two groups: those who have registered during 2001 and those who have registered during 2005, Figure 5. We have also examined different periods and different granularities but the general behaviour was almost the same. The most important point here is that when looking at the whole group of our users together, we can conclude that their average productivity is decreasing overall, Figure 4(c).

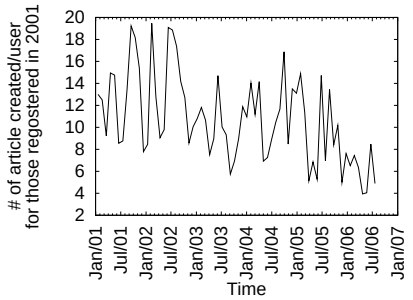
4.3 Contributor centered analysis

Previous studies have shown that when considering scientific paper authoring, the number of authors that made n con-

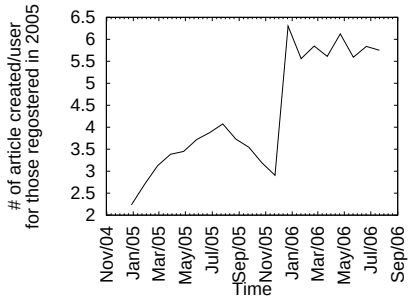


(a) # of articles (b) # of contributors (c) Average # of articles created per contributor per month

Fig. 4: *Wikipedia growth*



(a) Users registered from Jan 2001 to Dec 2001



(b) Users registered from Jan 2005 to Dec 2005

Fig. 5: *Average contribution for user registered at different times*

tributions is about $\frac{1}{n^a}$, which is known as Lotka's law [13]. Moreover, people have found that a is often nearly 2. Another way of seeing this is to order authors by the number of updates they have made and to plot for each author its ranking and number of updates. Figure 6 show the result of this analysis. Let $P(r)$ be the number of updates made by the contributor in position r of the rank. Then the Zipf's law [20], which is another way of seeing Lotka's law, states that the $P(r) \propto \frac{1}{r^k}$. Interestingly, the graph in Figure 6 seems to follow *two* Zipf's law curves with parameters 0.65 and 1.63.

These two Zipf's law curves seem to indicate that there exist two distinct groups of Wikipedia contributors; a small number of contributors (roughly 5000 of them) who are very productive and contribute a large number of articles (more than 1000) and the vast majority of contributors who contribute mostly well below 1000 articles. The exact reason for this clear separation is not clear, but it may be because most individuals have a limited number of updates that they can make and keep track of.

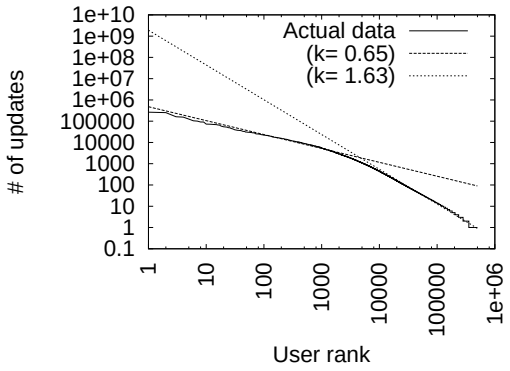


Fig. 6: *Contributor rank in terms of updates made.*

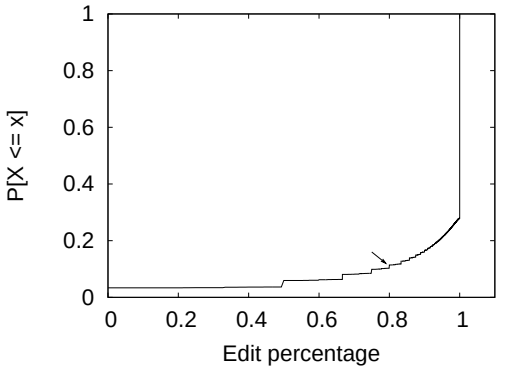


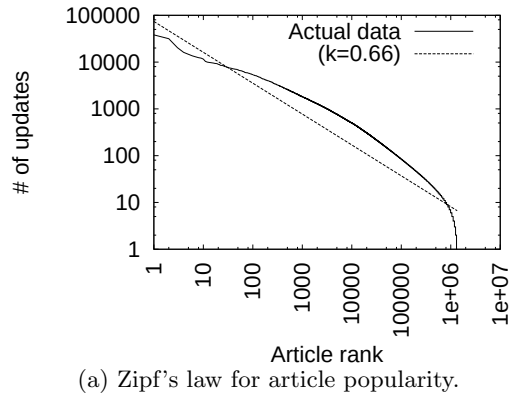
Fig. 7: *Distribution of percentage of revisions made by contributor.*

Another question concerning contributors is whether they have a tendency to only create new articles, to only edit existing ones or if they are interested in both. Figure 7 shows the cumulative distribution of the percentage of updates that were revisions to previously created articles. This plot can be interpreted in the following way. Consider the point (0.80, 0.11) which is highlighted by the arrow in the plot. This point

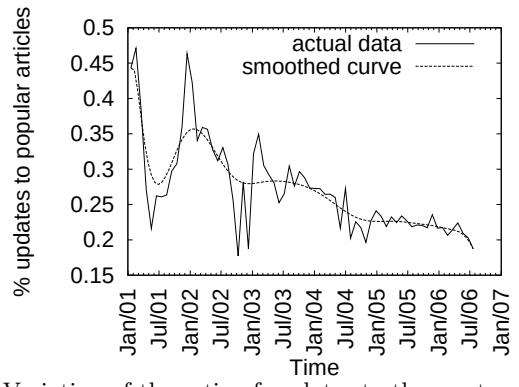
indicates that if a contributor is randomly selected then the probability that 80% of her updates are revisions is 11%. This plot, therefore, shows that approximately 70% of Wikipedia contributors did not create any articles at all and were only interested in revising articles that already existed (notice the sudden increase on the plot for $P[X <= 1]$) and that the burden of creating new articles is concentrated on only 30% of the contributors. This value remains almost constant throughout the whole period analyzed.

4.4 Article centered analysis

So far we have been trying to model the processes in which contributors interact with Wikipedia without paying much attention to which specific articles are being modified or created. But this is an important characteristic that may help us understand the reasoning behind contributor actions and may help to explain the results we have found so far.



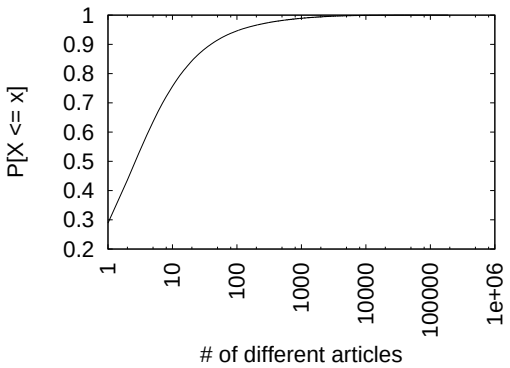
(a) Zipf's law for article popularity.



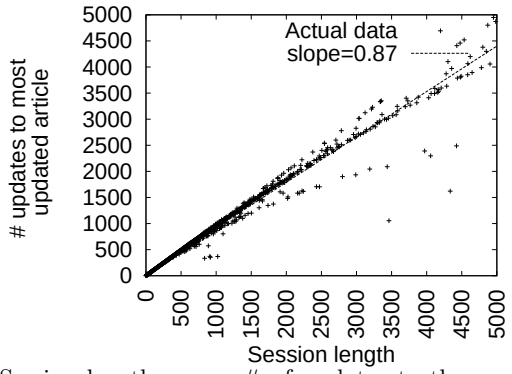
(b) Variation of the ratio of updates to the most popular 1% of the articles to the total updates per month.

Fig. 8: *Article popularity results.*

Figure 8(a) shows the ranking of the articles when they are ordered by number of updates they received. As expected, this plot roughly follows a Zipf's law. The difference in our workload to other workloads that have been previously studied is that the curve is less skewed than the Zipf's law with parameter $2/3 \leq k \leq 1$, which has been found before. Compared to the result from other workload analysis, unpopular articles get relatively more updates and popular articles get less updates. We conjecture that this behavior is caused by the fact that as the number of articles on Wikipedia increases, competition for the contributors' attention takes place and people tend to devote more time to edit the newly created articles than the old popular ones; the old ones may have already had several revisions and are supposed to somewhat reflect the general opinion of people. To verify this conjecture, Figure 8(b) shows a plot of the relative ratio of updates of the most popular 1% articles to the total number of updates through time. As can be seen by this plot, it is in fact the case that the chance of updating a popular article decreases through time.



(a) Distribution of the number of different articles per contributor



(b) Session length versus # of updates to the most updated article in each session

Fig. 9: *Contributor focus while interacting with Wikipedia.*

Another interesting experiment related to the articles being updated by each contributor, is whether or not contributors have a diverse set of interests and how do these interests are represented in individual sessions of contributors. Figure 9(a) shows the cumulative distribution of the number different articles updated per contributor. This plot shows that 40% of the contributors have made updates to at least 4 different articles. It also shows that there are certain users with a very large number of articles updated, i.e. 1% of the contributors have contributed to at least 1000 different articles. The general conclusion is that users seem to have a varied interest set in terms of the number of articles they update.

Figure 9(b) shows a plot of the session length, i.e. number of updates in each session versus the number of updates that were made to the most updated article in each session. Let's say that during a single session, a user updated article a_1 twice, and article a_2 once. Then, the session length for this session would be 3 and the number of updates to the most updated article would be 2. Only sessions of size less than or equal to 5000 were considered in this analysis. The reason for

limiting the number of sessions considered is that we believe that the larger sessions may represent behavior of automatic software robots that would not correspond to the behavior of actual users. The sessions considered comprise more than 99% of the of total sessions found. This plot also shows how the actual data could be fitted by a straight line. As can be seen, although contributors may have a varied set of interests (Figure 9(a)) when we consider the interactions in a single session, the updates tend to be centered around an specific article (Figure 9(b)), i.e. on average 87% of the updates in a single session go to the same article.

5. Conclusion and future work

This paper presents a characterization of the behavior of contributors creating content on the Wikipedia focusing on understanding the evolution of Wikipedia through the understanding of this behavior. Based on this characterization we were able to find that Wikipedia evolution is a self-similar process growing exponentially mostly because of its increasing number of contributors. Moreover, we show that Wikipedia contributors are naturally split into distinct groups based on their behavior and that although the contributors have a broad range of interests in most of their visits they only focus on a single article. On the article side we were able to see that the number of changes to an article follows a power law that is less skewed than one would expect based on other workload studies.

To the best of our knowledge, this is the first extensive effort in this direction. Research directions we intend to pursue in the future include the analysis of other Wiki sites as a way to validate which of the characteristics found are invariants and which are specific to the English version of the Wikipedia, the study of the anonymous workload and the comparison of the behavior of anonymous contributors to registered ones, and the study of the evolution of other characteristics from Wikipedia such as its graph structure.

Acknowledgments

This work is partially supported by NSF grants, IIS-0534784 and IIS-0347993, and a CAPES/Fulbright grant, BEX: 2408058. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding institutions.

References

[1] D. Ahn, V. Jijkoun, G. Mishne, K. Muller, M. de Rijke, and S. Schlobach. Using wikipedia in the trec qa track. In *Proceedings of the Thirteenth Text Retrieval Conference*, November 2004.

[2] M. F. Arlitt and C. L. Williamson. Web server workload characterization: the search for invariants. In *Proceedings of the 1996 ACM SIGMETRICS Conference*, May 1996.

[3] L. Bertolotti and M. C. Calzarossa. Models of mail server workloads. *Performance Evaluation*, 46(2-3), 2001.

[4] B. E. Brewington and G. Cybenko. How dynamic is the web. In *Proceedings of the International Conference on World Wide Web*, 2000.

[5] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of the 9th international World Wide Web conference*, May 2000.

[6] J. Cho and H. Garcia-Molina. Synchronizing a database to improve freshness. In *Proceeding of the 2000 SIGMOD Conference*, 2000.

[7] C. Dewes, A. Wichmann, and A. Feldmann. An analysis of internet chat systems. In *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, 2003.

[8] J. Giles. Internet encyclopaedias go head to head. `http://www.nature.com/news/2005/051212/full/438900a.html`, December 2005.

[9] M. Gospodinov and E. Gospodinova. The graphical methods for estimating hurst parameter of self-similar network traffic. In *Proceedings of the 2005 International Conference on Computer Systems and Technologies*, 2005.

[10] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proceedings of the 2003 SOSP*, 2003.

[11] R. Kumar, P. Raghavan, S. Rajapolan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)*, 2000.

[12] B. Leuf and W. Cunningham. *The Wiki Way: Quick Collaboration on the Web*. Addison-Wesley Professional, April 2001.

[13] A. J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12), 1926.

[14] Mediawiki website. http://www.mediawiki.org/.

[15] D. Menasce, V. Almeida, R. Riedi, F. Pelligrinelli, R. Fonseca, and W. Meira. In search of invariants for e-business workloads. In *Proceedings of the 2000 ACM Conference on Electronic Commerce*, October 2000.

[16] E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin. A hierarchical characterization of a live streaming media workload. In *IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, 2002.

[17] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with *history flow* visualizations. In *Proceedings of the 2004 Conference on Human Factors in Computing Systems*, April 2004.

[18] M. Vlkel, M. Krtzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic wikipedia. In *Proceedings of the 15th International Conference on World Wide Web*, May 2006.

[19] Wikimedia website. http://meta.wikimedia.org/wiki/Wikimedia.

[20] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949.