Serial Sharers: Detecting Split Identities of Web Authors

Einat Amitay, Sivan Yogev, Elad Yom-Tov
IBM Research, Haifa, Israel
{einat;sivany;yomtov}@il.ibm.com

ABSTRACT

There are currently hundreds of millions of people contributing content to the Web. They do so by rating items, sharing links, photos, music and video, creating their own webpage or writing them for friends, family, or employer, socializing in social networking sites, and blogging their daily life and thoughts. Of those who author Web content there is a group of people who contribute to more than a single Web entity, be it on a different host, on a different application or under a different username. We name this group *Serial Sharers*. In this paper we analyze patterns in the contributions of Serial Sharers. We examine the overlap between their individual contributions and propose a method for detecting their pages in large and diverse collections of pages.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering, Information filtering.

General Terms

Algorithms, Measurement, Experimentation, Human Factors, Standardization.

Keywords

Web authorship, profiling Web authors, publicly shared spaces.

1. INTRODUCTION

The idea for this paper stemmed from reading an interesting visualization paper about authorship in Wikipedia [12] in which the authors, Holloway et al., describe the contribution patterns of the top 10 most zealous Wikipedians. The thought that such productive contributors can actually change or influence a domain like "law" or "science" to an extent that they dictate the structure of the whole domain was intriguing.

Taking this thought even further, how many people dedicate their writing on the Web to advocate "open source" and what is their influence on current trends by merely expressing their stand in online forums, in blogs, and in virtual communities like Wikipedia? For example, Figure 1 demonstrates that there are nearly 1000 single authors who contributed over 1000 edits (contribution to a single Wikipedia entry in a given time) to the English portion of Wikipedia. Some people annotated the collection with over 100,000 text edits. This small group of people who contribute so much content to a single collection like Wikipedia may create either intentionally or maliciously a distortion in the way information is interpreted.

ACM SIGIR 2007 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

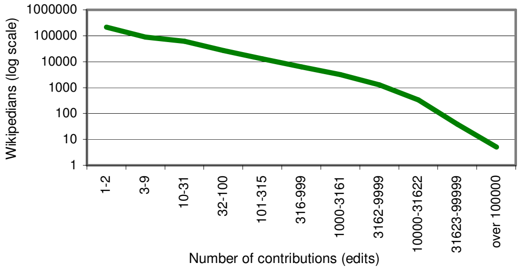


Figure 1 – a histogram of the number of contributions per single author to the English portion of Wikipedia, until September 2006.

Another anecdotal example is the size of the entry for each country in the English Wikipedia plotted alongside the population size of the country, as shown in Figure 2. The trend line traces the decrease in entry size in kb with the decrease in country size. Assuming that there are certain facts that should be common to the description of all countries, like size, population, government, etc., this decrease may be explained by the fact that there are many more social and cultural aspects to describe, but it may also be explained by the number of authors who contribute to each entry. This assumption is supported by the nearly equal size of the entries in the CIA Factbook online (around 100 kb for each country). This authorship "voting" system is a democracy in which the one with the loudest voice wins. Being loud on the Web simply means producing a lot of content on many different pages.

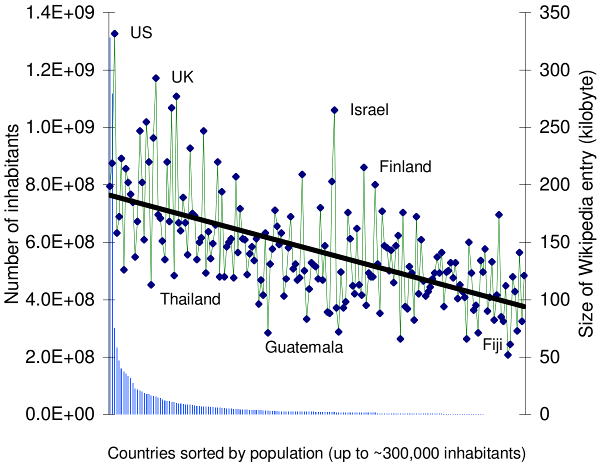


Figure 2 – a comparison of country population size and Wikipedia entry size in kilobytes.

According to recently published studies, about 35% of American Web users contribute some form of content to the Web [12].

Similarly, 31% of urban Chinese Web users create or update Web pages [18], while 20% of British users publish content on the Web [11]. Although the ratio between those who contribute content to those who do not contribute seems quite low, the real numbers translate to tens of millions of Web authors who constantly create and publish new content.

Table 1 lists the various forms in which people contribute content to the Web. They rate products, share links, photos, music and video, create their own webpage or write them for friends, family, or employer, socialize in social networking sites, and blog their daily life and thoughts. The younger the users the more zealous this diverse activity becomes. A recent study shows that 61% of 13 to 17 year-olds in the US have a personal profile on sites such as MySpace, Friendster, or Xanga. Half have posted pictures of themselves online and 37% of those teens maintain a blog [9].

Table 1 – Web authorship – what % of Web users actually contribute content and essentially writes the Web. Source: Pew Internet & American Life Project Surveys [19].

% of Web users who have done this	Activity	Survey Date
35%	Posted content to the internet	December 2005
30%	Rated a product, service or person using an online rating system	September 2005
14%	Created or worked on own webpage	December 2005
11%	Used online social or professional networking sites like Friendster or LinkedIn	September 2005
8%	Created or worked on own online journal or blog	February-April 2006

Among those who hide behind the numbers in Table 1 there are people who produce several types of content. Good examples for these are university professors and students who maintain their own personal Web page on a different host and also a page on their faculty site. This paper is about those authors who shout the loudest on the Web. They not only contribute content to the Web but do so on several different hosts and in various different forms, be it by tagging public material, through their homepage, by blogging, by contributing portions to Wikipedia, and the likes. These authors are not spammers in the trivial sense. Most have no intention of manipulating search results, or influencing world-wide information. They simply enjoy utilizing everything the virtual world offers. We call them *Serial Sharers*.

1.1 Serial Sharers

In a recently published study [21] it was found that 37% of American bloggers had a personal website before they started blogging and that 43% of all bloggers maintain at least two blogs. The actual numbers show that several millions of people in the US alone have authored more than a single page of content and published it online. The portions of content produced by such prolific authors may be considered as a distribution of their online identity. Overall if we took the sum of all the content contributed by a single author we may better describe the interests and thus better profile such a user. The example shown in Figure 3 is a real collection of eight different pages authored by the same person. The pages have some features in common such as the name of the author, some links, some images, some sentences or words, but the overall layout is different, the amount of information provided varies from page to page, the purpose and audience of the pages are different, and so are the hosts where those pages reside.

1.2 Possible Applications

Knowing that the same person authored a collection of not trivially-related pages may be used to enhance and create new applications where knowledge about users is essential. Analyzing and using information about a single author which is extracted from different sources may add new dimensions to user information, such that is not easily available today.

1.2.1 User profiling

Analyzing the identified set of pages written by the same author may help in tailoring user profiles for personalization or for expertise location. Such user profiles may be derived from information the author chose to include in some or all of the pages. For ***personalization*** the profile may be modeled according to the choice of publication media and the information presented in each media; by the shared structure of the documents; by color choice; by syntactic and lexical choice; by layout decisions, by the inclusion of images, etc.

Such information may be used to create user-driven defaults of color and layout choices tailored for each individual user. It may also be used to display advertisements that match the profile of who the user's readership is across all sites, which is the readership most likely to visit the documents in the set. Looking at profiling the audience of a whole site, such collections of authorship-driven profiles spread over several media types and may help to better understand use patterns. For example, what information people choose to share in blogs versus what information they choose to publish on their homepages. It may also help determine the influence of individuals on a collection, to better track a community and those who shape its shared content.

For ***expertise location*** profiling the whole set may reveal and strengthen evidence for knowledge repeating itself in several documents. Also, by using link analysis techniques it may be possible to better reflect the interest the author attracts by looking at all the incoming links to the whole set of documents rather than to a single document. Analyzing social networks based on the whole set of pages written by the same author reveal different patterns than those networks found in homogenous collections consisting only of blogs or of online forum messages. Such information may serve businesses like recruiting services, online dating services, people search indices, and so on.

1.2.2 Noise reduction

Serial sharers may also affect search engine ranking since a single author may produce the same idea in identical or similar forms on some or all of the published pages. This may introduce quite considerable noise to the index of a search engine that relies on any kind of host counting, link counting, term counting or even clickthrough counting.

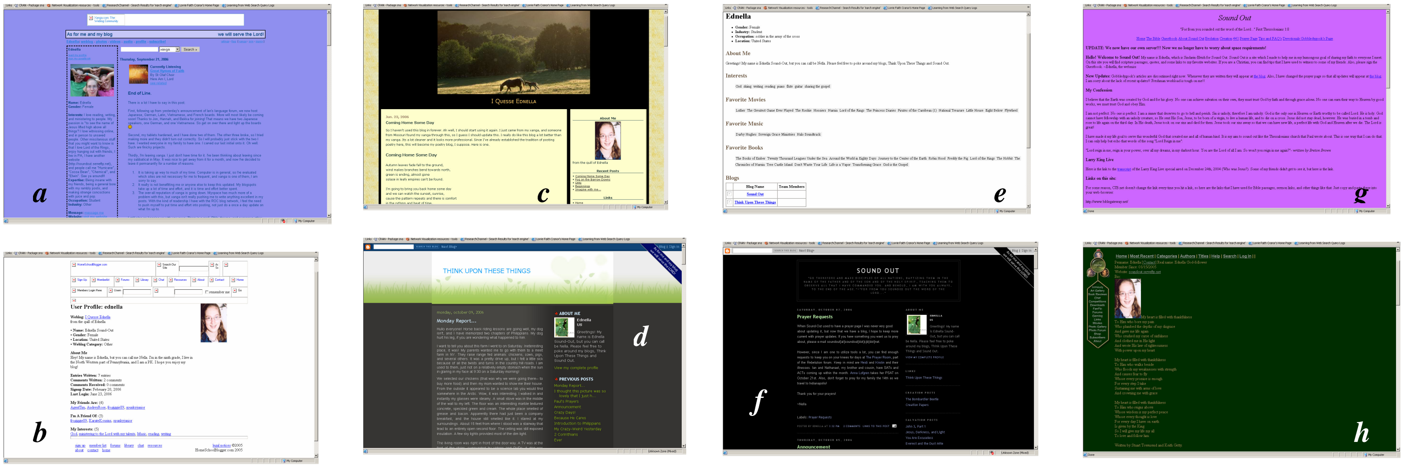


Figure 3 – Eight pages written by the same author and hosted on different sites (*a, c, d, f* blogs; *b, e, h* profiles; *g* unknown type)

On narrow scale or esoteric topics the phenomenon may even influence content per subject. So, assume that there is a band with only 50 content references created by online fans. One specific fan has authored several of them, describing a specific favorite song on two blogs, a homepage, a social networking profile and also on the same fan's YouTube page along with the appropriate link to the mp3 file of that song. Thus, a tenth of the content about the band was produced by a single author. Even if all the other fans disagree with the author on which is the favorite song, the prolific author's voice is loud enough to make a difference. The content contributed is definitely not spam and should not be considered spam.

Serial sharers do not produce spam. They simply use the media in the way it was intended to be used. As demonstrated earlier, today's youth have a higher percentage of users contributing blogs and general content to the Web's collection. When those teens grow up, being a serial sharers will most probably be the norm. This will eventually lead to the Web being a collection of many voices associated with many echoes. The echoes introduce noise into search engine indices. The noise may skew results retrieved for certain topics like "open source" where few people write a lot of content distributed on different hosts.

There are some solutions that come to mind for using author detection to reduce noise in search engine indices. The first is similar to the idea of site collapse where results coming from the same author may be displayed in a cluster or appear after a "show more results by this author" button is pressed.

Another option, which is harder to implement, is to reduce the set to a single file, sort of a summary file that will represent the whole set written by the same author as a single entity in the collection. Creating a single file or a connected set of files may also help aggregate clickthrough data received for a set of same-author pages to better reflect the interest in the whole set rather than in portions of it.

1.2.3 Sizing Web site's unique user community

A different usage for collecting the whole set of pages written by the same author is size estimation of user communities publishing on Blogger, YouTube or Facebook. This will allow for more realistic calculation of the number of unique users who contribute content to the site compared to a different site. Such a comparison may provide stronger evidence about the adoption of certain applications and the rejection of others. For example, if a smaller hosting site is able to prove that its audience consists solely of artists who usually do not publish in any other space this makes the site unique and marketable for advertisement to art supplies companies. On the other hand, a site that has most of its authors publish similar content elsewhere has less value in terms of uniqueness and targeted marketing offerings.

Owners of Web sites may be able to produce a seed of documents labeled with their respective authors taken from the collection and compare those samples with those of other sites. This will help create a benchmark against which user community sizing may be performed.

2. RELATED WORK

In a search system, the problem of author detection resembles, in a sense, the problems of Duplicate Page Detection [6] and Mirror Site Detection [5], both of which use multi-dimensional aspects of the page to describe duplication in features such as size, structure, content, similar naming of the URL, etc. Duplication and mirroring are artifacts of hosting similar information on different machines or hosts in order to facilitate access to those pages in a desired context (e.g. hosting a mirror of a software library on a public university server). Author Detection is somewhat similar in the sense that information written by the same author, such as a user profile or a homepage, is sometimes partially duplicated by mentioning similar topics, expressing similar opinions, repeating the same links or usernames, etc.

However, sometimes each page written by the same author comprises of exclusively unique segments. In the collection we describe in section 4.1.1 there are authors who make a clear distinction between pages about their hobbies such as mountain biking, and their professional pages where they write about academic research or their family.

Many studies explore the field of author detection or author attribution in restricted domains. For instance, Argamon et al. [2], Li et al. [16] and Zheng et al. [25] employ machine learning and shallow parsing methods to detect authors in various collections of newsgroups. Using similar methods, Novak et al. [20] cluster short messages on online message boards for detecting users who mask their identity. Abbasi and Chen [1] analyze online forums in Arabic and English, employing machine learning techniques to learn a distinctive and large set of linguistic features for each user. Others have studied author detection using similar methods in blogs [14] and in emails [10].

However, there have been very few papers published about author detection across several different collections and domains. Rao & Rohatgi [22] tried to align authors from both mailinglists and newsgroups. They report that the stylistic conventions practiced by users of the different media resulted in very poor detection rates with learning and shallow parsing methods.

In this paper we intend to show the feasibility of performing author detection over several media types such as blogs, user profiles, personal tagging spaces, professional and personal homepages and any other identifiable personal information that can be attributed to a single author. Figure 3 is an example for the kind of variety we seek to explore. The set of eight different pages all written by the same author and published on different hosts consists of several traits that are visually similar, like images and layout, and several traits that are different like title, length, and intended readership.

3. DETECTION BY COMPRESSION

The studies described in section 2 all look at very controlled and contained domains. However, to solve the problem of author detection on the Web it is very costly to employ methods of shallow parsing and machine learning for several reasons. First, feature extraction is a costly process which requires analyzing many aspects of the page and then producing large data structures for storing such information. Secondly, feature extraction in such an uncontrolled environment cannot scale up, as observed by Keogh et al. [14]. The authors follow the work of Benedetto et al. [4] who applied off-the-shelf compression software to extract the compression distance for each pair of pages. Benedetto et al. managed to cluster the world languages by using this feature alone. They have also tried to detect similar authors in a small pool (90 documents) of academic papers. Their reported success rate on this restricted domain is over 95% for pairing texts by the same author. Kukushkina et al. [16] explain the linguistic motivation behind using compression to represent author specific repetition frequencies. Recently, Cilibrasi & Vitanyi [8] explained the theoretical rational behind using compression to represent and then compare entities with complex features.

Using compression instead of textual and structural feature extraction is advantageous for our task since there are so many ways in which two pages written by the same author can be similar. They may share themes, content terms, relative URL path, linking patterns, page layout, color scheme, image filenames, etc. Encoding such a feature set for a collection of pages is a very subjective task. If the feature set is large enough to describe all possible aspects its usage will not scale to large collections such as the Web. Compression captures all of the features that repeat themselves in a single page and treats them as information redundancy. So it may capture HTML structure redundancies as well as stylistic redundancies. The final size of the compressed page is determined by the repeating patterns detected in the compression. By using compression for author detection we hypothesize that every author has a unique compression signature that is similar across all the pages of the same author.

3.1 Compression Distance

The *Normalized Compressor Distance* (NCD) was suggested in [4] (with formal justification in [8]) as a tool for detecting document similarity. Given a compressor *C* and two documents *x*, *y*, we define:

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

Where *C*(*x*), *C*(*y*) and *C*(*xy*), are the bit-wise sizes of the result sequences when using *C* to compress *x*, *y* and the concatenation of *x* and *y*, respectively[1]. NCD assesses the similarity between a pair of documents by measuring the improvement achieved by compressing an information-rich document using the information found in the other document.

In this paper we use a variation of NCD which we term 2-sided NCD (2NCD), with the following definition:

$$2NCD(x, y) = \frac{[C(xy) - C(x)] \cdot [C(xy) - C(y)]}{C(x) \cdot C(y)}$$

2NCD measures separately how much the compression of **each** of the documents is improved by using the information included in the other document. The compression distance assigned to the document pair is the product of these two measurements.

4. EXPERIMENT

We designed an experiment to test whether authors can be detected using only their compression signature, even across different types of writing styles and Web publication types. We collected nearly 10,000 pages including blogs, user profiles, del.icio.us spaces, Flickr photo collections, Wiki style pages, personal homepages, etc., written by 2201 different authors. We then conducted several experiments based on this collection.

4.1.1 Data Collection

In order to collect data for such a large scale experiment it is necessary to ask people to provide a list of Web pages that they have authored. It is not possible to simply crawl the Web for such information without prior knowledge as pointed out by Bar Ilan [3], since it may be that a person is described by others such as in the case of corporate executives, and famous movie actors. Obviously, people also write under pseudonym and it will be difficult to detect them without prior knowledge. We first tried to ask people to send us their collection by email, however we received only several dozens of replies which is not enough for our task. One of the replies stated that the full list of his authored pages can be found on ClaimID.com. ClaimID is an experimental site set up by two students from the University of North Carolina. The site is described by Stutzman & Russell [23] as a system for managing online personal identities. ClaimID allows its users to list URLs that were authored by them and/or about them. The site is a list of user profiles with detailed lists of what information was produced by the author and what was not. We crawled the site,

[1] Assuming that C is a normal compressor (see [8]), and therefore *C*(*xy*) = *C*(*yx*).

which is publicly available to search engine crawlers, and collected over 8000 unique user information. We then filtered this list and stored only authors who had at least two pages authored by them hosted on two different hosts. We also removed those who had simply duplicated the content of one site and put it up as a mirror on another host (assuming this will be revealed by simple duplicate- or mirror- site detection).

We ended up with 2201 users who authored 9834 different pages. Figure 4 describes the distribution of page types in our collection. This is a very crude division, based on the occurrence of terms in the URL, the anchor or the short description appearing in the ClaimID profile. For example, we labeled a page with the term "blog" if any if the fields contained, even partially, any of the terms *blog*, *livejournal*, *typepad*, *wordpress*, and *fotolog*. "Community-Share" label was assigned to social-space pages marked with *del.icio.us*, *simpy.com*, *blinklist.com*, *ma.gnolia.com*, *connotea.org*, *scuttle.org*, *wists.com*, *shadows.com*, *digg.com*, *slashdot.org*, *myspace.com*, *deviantart.com*, *youtube.com*, etc. "Unknown type" means that there was no trivial way to automatically detect the type of the page from its host name or from the description provided by its author on ClaimID. Manually inspecting some "unknown type" pages revealed that many came from sources such as professional or work-related sites, newspaper articles, contributions to school projects, etc.

We left the files intact, including all HTML and scripts. This was done in order to achieve realistic results that could potentially be applied to any collection of Web pages without any pre-processing. Also, removing HTML markup may have affected the detection of structure and layout characteristics unique to individual authors.

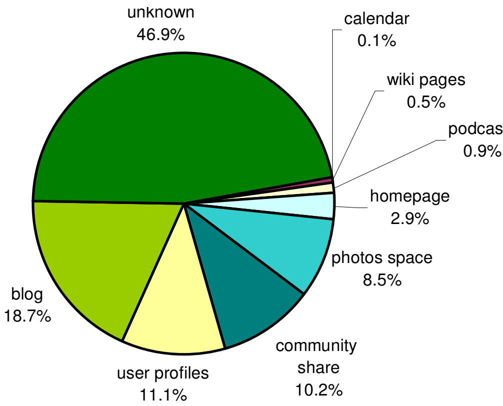


Figure 4 – The percentage of each detected page type in our collection of 9834 pages coming from 2201 different authors.

4.1.2 Common links as baseline comparison

Following Calado et al. [7] who recently tested linkage similarity measures and found link co-citation to yield the best results for topic similarity between documents, we decided that our baseline comparison should be link co-occurrence between each pair of documents.

As a first step to test the existence of link co-occurrence between sets of documents known to be produced by the same author we calculated the amount of shared links for each set. It turned out that about 60% had common links while 40% had no common links between the different pages they have written. The most prolific author had 1283 links appearing repeatedly in the set of the pages he authored. We did not compare against shared textual content since it was not a measure that could scale up to our collection. We also considered using duplicate detection methods, however, after inspecting the documents it seemed that this approach will not yield better results than simply comparing common links.

4.1.3 Detection by Compression Experiment

Motivated by efficiency considerations, we sampled our collection and extracted two smaller sets comprising 1043 documents for the first set and 1109 documents for the second set. The sampling was arbitrary and was designed to sample authors rather than pages. All the pages written by the same author were grouped together and the two samples did not include the same author twice. We worked with these samples to compare each possible pair of documents using link co-occurrence and compression distances.

For each document we computed its shared links with every other document in the sample. For each such pair we also calculated their compression distance by first compressing each document on its own and then compressing the pair together.

For the compression task we used 7za.exe[2], an open source free compressor, which has a relatively large buffer. We found the large buffer to be advantageous for Web pages. The large buffer size also supports our assumption that the compressor is symmetric. We also tried MATLAB's built-in ZIP compressor but found it to be less effective.

4.1.4 Detection by Compression Results

The results of the compression distances computed for each document pair (using 2NCD) are shown in Figure 5 and Figure 6. The figures are the histogram of the values received for each comparison. The green bars represent pairs that actually belong to the same author, while the red bars indicate pairs that were written by different authors. For both samples it is obvious that the green bars accumulate on the left-most side of the chart. This accumulation clearly demonstrates the strength of the compression distance as a method for representing authorship encoded information.

In Figure 5 and Figure 6, the green bars display a bimodal distribution, which is typical to cluster-containing data [23]. In our studied domain we contend that there are two types of relations between documents written by the same author. The first type consists of the cases where a person writes several Web pages with a similar motivation, such as a professional blog and a professional homepage. Since the underlying function of these documents is the same, and they reflect the same purpose, the resulting documents are very similar and therefore the compression distance is very low. This may explain the green slopes on the left end of Figure 5 and Figure 6.

The other type of relation consists of documents which were written by the same author but serve different purposes, such as a a personal calendar and a dig.com entry. These pages will have many dissimilar features. However, since the author is the same the resemblance between these documents will remain. Those documents probably comprise the green hills which spread from compression distance 0.01 to 0.035 in the above figures. Between

[2] http://en.wikipedia.org/wiki/7-Zip

the two types of pages lays a continuum of similarity values, some overlapping with those of unrelated authors.

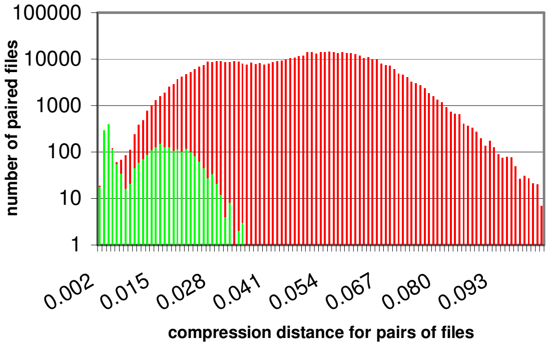


Figure 5 – A histogram of the compression distances computed for each pair of documents in the first sample consisting of 1043 documents. The green bars represent true document pairs. The red bars represent false document pairs.

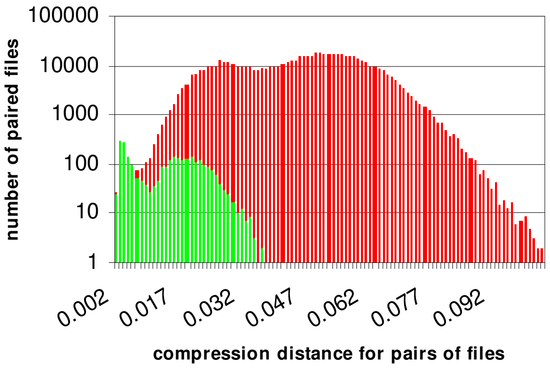


Figure 6 – A histogram of the compression distances computed for each pair of documents in the second sample consisting of 1109 documents. The green bars represent true document pairs. The red bars represent false document pairs.

In order to better visualize the results of the compression-based similarity, we generated a graph known as the Receiver Operating Characteristic (ROC) curve. This curve plots the sensitivity versus the specificity of a system. In our case, each point on the curve plotted in an ROC is a threshold similarity. The horizontal axis of the ROC curve represents the probability that two pages that have a compression similarity index smaller than the threshold will not be from the same author. The vertical axis shows the probability that two pages which have a compression index smaller than the threshold will indeed be from the same author. The ideal curve would touch the upper left corner of the graph, while a random decision will result in a curve from the bottom left corner to the upper right-hand corner. An ROC is usually parameterized by the area under the curve, where 0.5 represents random decision and 1.0 an ideal system.

Figure 7 and Figure 8 show the results of compression-based similarity compared to using the number of co-occurring links as a method for detecting authorship. The area obtained by the latter method is 0.6, only slightly better than chance. Compression-based similarity achieves an area of greater than 0.97, which is close to the ideal detector. Thus, the compression-based similarity offers a superb method for identifying authorship.

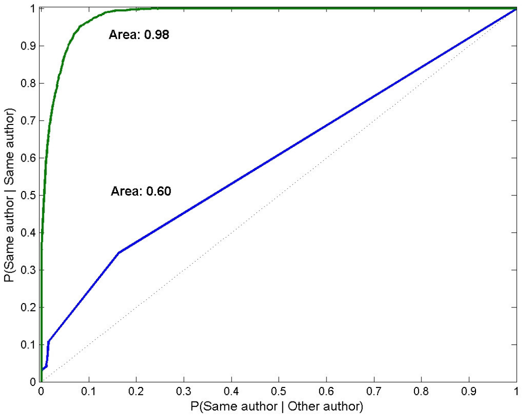


Figure 7 - Receiver Operating Characteristic (ROC) curve plotted for the first experiment. The grey line represents equal chance, blue line represents probability of being correct using common links, and green line represents the probability of being correct using compression sizes.

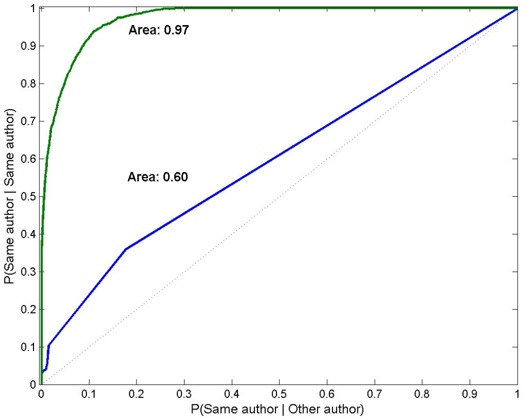


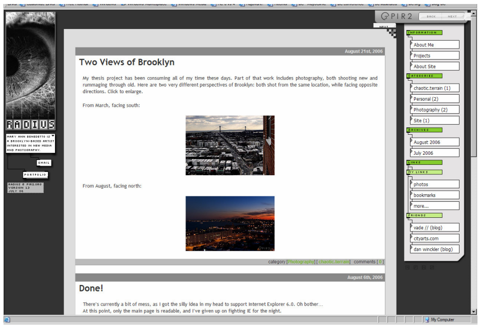
Figure 8 - Receiver Operating Characteristic (ROC) curve plotted for the second experiment. The grey line represents equal chance, blue line represents probability of being correct using common links, and green line represents the probability of being correct using compression sizes.

Table 2 is a color-coded matrix of compression distances calculated for the eight document examples displayed in Figure 3. All the document pairs were assigned low compression distance values which means they were considered similar.

There were no falsely paired documents of that same-author set until the compression distance value doubled from the last true pair. The falsely paired document, appearing in Figure 9, was matched to documents g (0.016), d (0.018), f (0.018), and h (0.018). This brings us to the problem of chaining or clustering together all the scored pairs to create the original set of pages produced by the same author. The next section describes a naïve attempt to cluster the paired documents using only the information provided by the compression distance.

Table 2 - color-coded matrix of compression distances calculated for the pages presented in Figure 3

	a	b	c	d	e	f	g	h
a		0.0051	0.0048	0.0051	0.0065	0.0051	0.0028	0.0028
b	0.0051		0.0033	0.0036	0.0041	0.0036	0.0036	0.0036
c	0.0048	0.0033		0.0038	0.0044	0.0038	0.0033	0.0033
d	0.0051	0.0036	0.0038		0.0041	0.0036	0.0036	0.0036
e	0.0065	0.0041	0.0044	0.0041		0.0041	0.0041	0.0041
f	0.0051	0.0036	0.0038	0.0036	0.0041		0.0036	0.0036
g	0.0028	0.0036	0.0033	0.0036	0.0041	0.0036		0.0036
h	0.0028	0.0036	0.0033	0.0036	0.0041	0.0036	0.0036	



August 21st, 2006

Two Views of Brooklyn

My thesis project has been consuming all of my time these days. Part of that work includes photography, both shooting new and rummaging through old. Here are two very different perspectives of Brooklyn: both shot from the same location, while facing opposite directions. Click to enlarge.

From March, facing south:

From August, facing north:

category [Photography] [chaotic.terrain] comments [0]

August 6th, 2006

Done!

There's currently a bit of mess, as I got the silly idea in my head to support Internet Explorer 6.0. Oh bother...
At this point, only the main page is readable, and I've given up on fighting IE for the night.

Figure 9 – a page which was the first to be falsely correlated with several of the pages in Figure 3 (with g: 0.016, with d: 0.018, with f: 0.018, and with h: 0.018)

4.1.5 *Document clustering*

In order to cluster the paired documents we used a naïve clustering algorithm as follows: Given a distance function *D* and a threshold *t*, let G = (V, E) be a graph whose vertices are all of the documents in a collection, with an edge connecting every pair of documents (*x*, *y*) such that *D*(*x*, *y*) ≤ *t*. A cluster of single-author documents is a connected component in G.

The results of applying this algorithm using 2NCD with different thresholds on the two sample sets are given in Figure 10. It should be noted that the data was not manually verified and therefore it may include some noise (for instance a person who registered on ClaimID under two different usernames). The number of same-author pairs is presented along with the error rates produced by using different thresholds. The lines show the number of detected same-author pairs while the bars show the error rate for each threshold. We labeled the document pairs whose compression distance is below the threshold "Original", and the pairs resulting from running the clustering script "Clusters". The total number of true same-author pairs is 2705 and 2745 in sample sets 1 and 2, respectively.

An important observation from this figure is that up to a threshold of 0.008, both error rate and the number of pairs added by the clustering algorithm are relatively small (approximately 10% or lower). This means that given a set of very similar documents, the compression distance identifies almost every pair in the set as related, with relatively few errors. At threshold 0.008, the number of clustered pairs is approximately 3/8 (37.5%) from the total number of truly related pairs.

Estimating the number of those authors who have more than a single Web page to be half of those who maintain blogs yields about 6 million users with at least 12 million pages in the US alone. Detecting nearly 40% of the pages authored by such serial sharers reveals a newly detected community which calls for new methods of exploration and research.

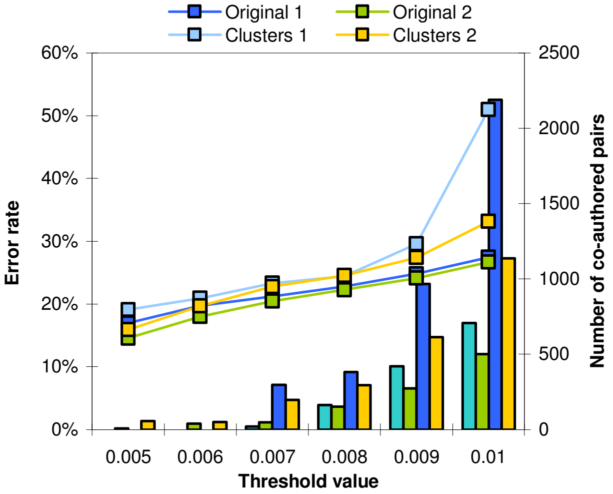


Figure 10 - The number of detected same-author pairs according to compression distance (Original) and clustering algorithm (Clusters), along with the error rates using different thresholds. The lines show the number of detected same-author pairs (out of approx. 2700 real co-authored pairs in each sample), while the bars show the error rate.

5. CONCLUSION & FUTURE WORK

We presented the problem of author detection over a collection of pages originating from different sources and written to serve different online functions. We applied a detection-by-compression algorithm to compute the compression distance for each pair of documents in a collection of pages with a known author. We then showed that it is possible to correctly determine authorship for a considerably large portion of the Web pages based on such a distance, and went on to chain the pairs into document clusters.

It is evident from the studies presented earlier that the youth of today is much more likely to have authored multiple Web pages. When those teens become adults they will probably share much more content on the Web than today's adults. If this prediction is correct then the title "serial sharer" will apply to many more people around the world. Hundreds of millions of people will have their contributions stored all over the Web, managing their personal archiving and memoirs online. Search engines need to prepare for that day with a mechanism for automatically detecting and labeling such individual productivity.

The good news is that search engines already use compression in storing cached versions of documents. The only caveat is the fact that in order to calculate the compression distance for each pair, both files need to be compressed together. This challenge may give rise to new solutions for candidate file pairing that will allow search engines to reduce the number of paired files to be compressed. Such solutions may take usernames found in the URL as a first "rule of thumb" comparison candidacy. Similarly, solutions may be found in computing the probabilities of people co-publishing in certain places, for instance, if a person publishes in del.icio.us they are likely to also have a page on blogger.com, etc.

Such solutions will lead to finding patterns in cross domain adoption of Web applications. It will be easier then to decide which application attracts a larger number of unique users by aligning sites like del.icio.us with blogger and myspace to find common authors. This alignment may also provide insight about what content people choose to publish on one site and not on the other, and why people decide to split their identity and write in several different places.

Incorporating author identification into search engines will advance features such as profiling, expertise location, finer granularity in trend analysis, and may help generating better insights about the sources and motivation for the publication of the retrieved results.

6. REFERENCES

[1] Abbasi, A. and Chen, H. 2005. Applying Authorship Analysis to Extremist-Group Web Forum Messages. IEEE Intelligent Systems 20(5):67-75.

[2] Argamon, S., Šarić, M., and Stein, S. S. (2003). Style mining of electronic messages for multiple authorship discrimination: first results. In Proceedings of ACM KDD'03, pp. 475-480.

[3] Bar–Ilan J. (2006). False Web memories: A case study on finding information about Andrei Broder. First Monday, volume 11, number 9 (September 2006), URL: http://firstmonday.org/issues/issue11_9/barilan/index.html

[4] Benedetto D., Caglioti E., Loreto V. (2002). Language trees and zipping. Physical Review Letters, 88(4):048702.

[5] Bharat, K. and Broder, A. 1999. Mirror, mirror on the Web: a study of host pairs with replicated content. WWW8, appeared in Computer Networks & ISDN Systems, 31(11-16):1579-1590.

[6] Broder A. Z., Glassman S. C., Manasse M. S., Zweig G. (1997). Syntactic clustering of the Web. WWW6, appeared in Computer Networks & ISDN Systems, 29(8-13):1157-1166.

[7] Calado P., Cristo M., Moura E.S., Gonçalves M.A., Ziviani N., Ribeiro-Neto B. (2006). Link-based Similarity Measures for the Classification of Web Documents. Journal of the American Society for Information Science and Technology (JASIST) 57(2):208-221.

[8] Cilibrasi R., Vitanyi P.M.B. (2005). Clustering by compression. IEEE Transactions on Information Theory, 51(4):1523-1545.

[9] Cox Communications, Press Release, March 2006. Available online: http://www.cox.com/takecharge/survey_results.asp

[10] de Vel, O., Anderson, A., Corney, M., and Mohay, G. (2001). Mining e-mail content for author identification forensics. SIGMOD Record, 30(4):55-64.

[11] Dutton W. H., di Gennaro C., Hargrave A. M. (2005). Oxford Internet Survey 2005 Report: The Internet in Britain. Available online: http://www.worldinternetproject.net/publishedarchive/oxis2005_report.pdf

[12] Holloway T., Bozicevic M., Börner K. (2005). Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. under review. Available online: http://arxiv.org/ftp/cs/papers/0512/0512085.pdf

[13] Horrigan J.B. (2006). Broadband Adoption 2006. Pew Internet & American Life Project. Available online: http://www.pewinternet.org/pdfs/PIP_Broadband_trends2006.pdf

[14] Keogh E., Lonardi S., Ratanamahatana C. A. (2004). Towards parameter-free data mining. In Proceedings of ACM KDD 2004, pp. 206-215.

[15] Koppel M., Schler J., Argamon S., Messeri E. (2006). Authorship Attribution with Thousands of Candidate Authors. ACM SIGIR 2006, pp. 659 - 660.

[16] Kukushkina O.V., Polikarpov A.A., Khmelev D.V. (2000). Using Literal and Grammatical Statistics for Authorship Attribution. Problemy Peredachi Informatsii, 37(2):96-108. Also translated in “Problems of Information Transmission”, pp. 172-184. Available online: http://www.math.toronto.edu/dkhmelev/PAPERS/published/gramcodes/gramcodeseng.pdf

[17] Li J., Zheng R., Chen H. (2006). From fingerprint to writeprint. Commun. ACM 49(4):76-82.

[18] Liang G. (2005). Surveying Internet Usage and Impact in Five Chinese Cities. Report of the Research Center for Social Development, Chinese Academy of Social Sciences (November 2005). Available online: http://news.bbc.co.uk/1/shared/bsp/hi/pdfs/10_02_06_china.pdf

[19] Madden M., Fox S. (2006). Riding the Waves of "Web 2.0". Pew Internet & American Life Project's Report, October 2006. Available online: http://www.pewinternet.org/pdfs/PIP_Web_2.0.pdf

[20] Novak, J., Raghavan, P., and Tomkins, A. (2004). Anti-aliasing on the web. In Proceedings of WWW '04. pp. 30-39.

[21] Princeton Survey Research Associates International for the Pew Internet & American Life Project. (2006). Blogger Callback Survey Final Revised Topline 7/6/06, Data for July 5, 2005 – February 17, 2006. Available online: http://www.pewinternet.org/pdfs/PIP_Bloggers_Topline_2006.pdf

[22] Rao J.R., Rohatgi P. (2000). Can pseudonymity really guarantee privacy? In Proceedings of the 9th USENIX Security Symposium, pages 85–96.

[23] Steinbach M., Ertoz L., Kumar V. (2004). Challenges of clustering high dimensional data. In New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition. Springer, 2004.

[24] Stutzman F., Russell T. (2006). ClaimID: a system for personal identity management. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL), p. 367.

[25] Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 57(3):378-393.