The Dangers of Webcrawled Datasets

by Graeme B Bell

E-mail: datasetspaper@graemebell.net

Abstract:

This article highlights legal, ethical and scientific problems arising from the use of large experimental datasets gathered from the Internet - in particular, image datasets. Such datasets are currently used within research into topics such as information forensics and image-processing. This paper strongly recommends against webcrawling as a means for generating experimental datasets, and proposes safer alternatives.

Contents

0. Introduction

"Imagine the scene: you have woken up to discover four police officers at your door. They tell you that your university's systems administrators have found many highly illegal images stored on your research server. Furthermore, the files were downloaded and later directly accessed by your user account. While protesting your innocence, you suddenly realise that you have no idea what was among the two million images your webcrawler downloaded over the previous month..."

Over the last 15 years, research within fields such as image-processing and information forensics has been conducted on increasingly large datasets. In steganalysis, for example, the growing use of statistical classification techniques has lead to extremely large 'clear image' datasets being gathered. These are used either to train classifiers or measure the false positive rate during classification (Provos and Honeyman, 2003; Medhi et al, 2005). The justification for gathering these large datasets is that they are intended to represent a very broad sample of the many different types of image found across the Internet. Consequently, the datasets can sometimes include millions of image samples which have been gathered blindly and automatically from the Internet by webcrawler software (Provos and Honeyman, 2003).

Webcrawler programs explore the World Wide Web automatically, storing the data encountered as they traverse within and between websites. These programs are used because of the effort and expense involved in manually gathering large quantities of material from the Internet. Unfortunately, as the opening paragraph of this introduction shows, this method for gathering large sets of experimental data may in fact be dangerous and inappropriate as a research practice.

This paper begins by briefly introducing webcrawling approaches (Section 1). The paper then discusses the potential for very serious negative consequences, when viewed from legal, moral and scientific perspectives (Section 2). A number of alternative approaches to dataset construction are proposed for use in future research (Section 3). Finally, brief conclusions are offered, and are followed by notes and references.

1. Webcrawling

Webcrawling is the practice of recursively traversing the World Wide Web as though it were a graph, where pages are vertices of the graph, and links are edges. Despite the dynamic and error-prone nature of Web links, even naive approaches are seen as being effective in several areas of computing. For example: cataloguing page contents for search engines (Eichmann, 1994; Brin and Page, 1998); estimating the size and structure of the Web (Albert et al, 1999); or gathering resources automatically, such as images or sounds (Provos and Honeyman, 2003). This paper looks at the practice of using webcrawling to build large media file datasets - in particular, image datasets - with the intention of subsequently processing them as part of empirical research.

Generally, a webcrawler's automated traversal of the Internet can be described as 'blind', because it takes place fully automatically and without human interaction. The researcher controlling the webcrawler is usually only required to provide a starting location, some parameters to control the speed, breadth and depth of search, and a list of media types to be gathered. It is also possible to provide a blacklist of sites to be avoided. For example, a steganalysis researcher might deliberately exclude specific websites that they know will contain steganography.

There are five main forms of interaction with the material that is gathered by the crawler. When initially gathering the experimental dataset, the images must first be accessed and transmitted over a network from the Web server to the research computer running the crawler. Generally, this leaves a record of the access on the remote system and sometimes also on intervening systems. Secondly, the retrieved data must be stored on the research computer until the experiment is carried out, and possibly shared between collaborating researchers. Thirdly, the files must be accessed. In general, this will take place automatically, and a program will process the images without 'viewing' them as a human might. Fourthly, to allow scientific reproducibility, a researcher is likely to continue to store the data long-term. Finally, the original research group is expected to share, publish or redistribute the dataset, to allow external researchers to repeat the experiment.

Unfortunately, these research activities may result in unexpected consequences. There are many significant problems with the webcrawling approach for dataset generation that have not been previously discussed in published literature. These problems include (but are not limited to) the transmission, storage, use and redistribution of inappropriate images that have been gathered unknowingly. This paper now describes a variety of problems that may be caused by webcrawling.

2. Problems with automatically gathered Internet image datasets

Experiments based upon automatically gathered image datasets require remote access, transmission, storage, local access and redistribution of data. Two main types of problem can arise as a result of these actions. These are 'legal and social problems' and 'scientific problems'. Although the examples given here involve image datasets, many of the following issues can also affect text, audio and video datasets.

2a. Legal and Social Problems

The essential problem with webcrawling is that the software has no way to know the semantics of the data it is gathering. As search proceeds more deeply or broadly into the Internet from the starting page, almost any part of the Internet may be accessed in principle. Within two steps, one may start at Google, traverse to a major user-driven news or social-media site [1], and from there traverse to almost any site on the Internet. Some of the accessed content may be illegal and/or socially unacceptable in a variety of ways, which will now be considered.

Copyright: In some countries, it may be illegal to download, store, process or distribute images that specifically prohibit users from doing so via copyright warnings on a webpage or embedded in the image itself. While personal use of copyrighted material for research purposes may be permitted in some countries, it may not be permitted universally. Further, the actions that may be taken (transmission, storage, access, redistribution) may be restricted. It is almost impossible for a researcher to manually audit a million image dataset to be sure that the image data, metadata and originating site has not been marked with a copyright notice prohibiting or limiting use outside a Web browser. The problem is even worse with other data-types: what percentage of one million blindly webcrawled audio or video files would be free of copyright concerns?

Abusive/Distasteful: Images downloaded blindly from the Internet may portray racial/ethnic/sexist slurs, extreme violence, murder, rape, child abuse, or a wide array of other sexually explicit or distasteful activities. It is reasonable to assume that in many cases, these will not be legal within your country, or may not be acceptable under your institution's ethics policies. Unfortunately, there is no way to know for certain that you did not encounter such material when blindly gathering a one million image dataset.

Context-driven: Storage and redistribution can alter context. Person A distributes an image labelled with some self-deprecatory remark or ethnic/racial epithet. Person B unknowingly downloads a copy of the image during a webcrawl, and later stores or redistributes it within their dataset to allow reproduction of their experiment. Person B may now appear to libel Person A, (e.g. perhaps appearing as a racist), if their dataset is explored by an outsider. A discussion of the complexities relating to the context of abusive terms can be found in (Hom, 2008).

Religion: It can be both illegal and unwise to transmit, possess, use, or redistribute images that relate to religion in your datasets. A surprising number of countries specifically prohibit blasphemy, and historically, laws relating to religious blasphemy have been enforced against media re-distributors as well as original authors (Wikipedia, 2009). Some religions have also actively used copyright law to prevent the reproduction of religious materials (Dibbell, 2009).

Still, the issue of legality pales into insignificance when one considers the worldwide reaction to the 12 graphical cartoons of the Prophet Mohammed, published in Denmark in 2005 as a test of Islamic tolerance (Curry, 2008). The consequences included worldwide riots and protests calling for the 'butchering of those who mock Islam' (Bowcott, 2006). Ultimately, reactions to these images lead to the death of over 100 people, an international boycott of Danish products, and the destruction and arson of European and Danish embassy buildings in several countries. Death threats were issued against the cartoonists and the editor of the newspaper. The Prime Minister of Denmark described the consequences of distributing these 12 images as being the worst international crisis for the country since World War 2 - no small claim for any European country.

Threats of further violence relating to these images (and others) are taken seriously by institutions even today. National newspapers, television stations and universities alike have declined to republish these cartoons even when directly discussing them (Cohen, 2009). Consider that these 12 images have been published worldwide on many Web servers since 2005. They have also been repeatedly republished online in international media and in counter-protests supporting free speech, and have been considered high-profile image data for several years. Consequently, the images are almost certain to feature in any sufficiently large image-gathering webcrawl. Can any researcher in digital imaging be absolutely sure that their datasets do not contain these images?

These examples of legal and social difficulties are far from exhaustive. Consider the consequences of accidentally transmitting, storing or distributing materials that promote nazi-ism, deny the Jewish holocaust (Council of Europe, 2002), promote homophobia or promote extreme political views [2]. Despite provisions already in place to address the issue of intentionality in publication and distribution, there is still the potential for difficulties.

In general, whether images are blindly or intentionally interacted with, there may be significant consequences. Furthermore, research is often international, yet images which are permitted in one country or culture may not be permitted in another. In 2003, a Peruvian woman living in America lost access to her own children as the direct result of possessing a single photograph of herself breastfeeding her own 1-year-old child. This photograph would have been quite acceptable and normal within Peruvian society (Korosec, 2003). As researchers move between different countries with their digital research materials, they may find themselves in breach of unexpected laws that carry very serious short-term and long-term consequences.

Both the law and common sense suggest that even a single unacceptable image accessed, transmitted, stored, used, or redistributed could be one too many. Yet what is the chance of collecting millions of images from the Internet and not finding even one such image among them? The facts show that even a single image can be a surprisingly powerful thing.

2b. Frequency of inappropriate data on the Internet

It is difficult to have confidence in the reliability of published material describing the frequency of inappropriate data on the Web. Firstly, there are few sources to rely upon. Secondly, many of the groups providing such data have a vested interest in promoting a particular view of the Internet and of Internet users. For example, companies who present statistics as part of their marketing for 'copyright infringement detection' software, or religious groups promoting a particular viewpoint about the Internet. Thirdly, in order to calculate certain statistics, one must presumably break the law by finding and accessing illegal data. This strongly discourages investigation and data-gathering. However, some indicative examples are suggested below.

Copyrighted images: While numerous studies into the scale of copyright infringement of music and video exist, it is difficult to find any that have studied infringement of copyrighted images. However, Vivozoom, a company whose software tracks the use of stock photography by commercial websites, has produced a PR release (De la Vina, 2009) stating that "some 85 percent of the rights-managed images detected on commercial websites are being misused". The CEO comments that "people may not realize that there is a cost associated with the use of some images." (De la Vina, 2009).

*Religious imagery****:*** A Google Images search conducted by the author on 16 October 2009 for "jyllands posten muhammad OR mohammed OR muhammed" brought up 314,000 image results (Google, 2009). Many of the images found (especially among the top results) are known to be extremely offensive to many Muslims.

*Abusive imagery****:*** (Stanley, 2001) summarises results from several research papers, suggesting that even in 2001, there were an estimated 14 million pornographic websites, carrying approximately 1 million pornographic images of children between them. In 2003, the Independent reported data presented by the UK's National Society for the Prevention of Cruelty to Children, stating that around 20,000 images of child abuse were being added to the Internet each week at that time [3] (Frith, 2003; Renold and Creighton, 2003). The researchers noted that "little systematic and reliable data is available" on this topic (Renold and Creighton, 2003).

2c. How would unacceptable data be noticed, if it is only a few images within millions?

- The most likely situation is perhaps that police gain access to an Internet server that contained illegal data for download (or are monitoring network traffic to it). They might then find a research computer listed in the Web access logs (along with a record of the data that was downloaded).

- Local network administrators may discover apparent staff access to 'unusual' websites when they are monitoring, auditing or casually inspecting network traffic records or web-proxy logs.

- Local systems administrators may audit individual computers and storage devices using software designed to find particular types of data (e.g. 'are there many pink pixels?') or meta-data (originating site, author, filename) believed to be associated with illegal imagery.

- A colleague or external research group might manually discover data by chance while examining images, if the dataset becomes commonly used; 'Many eyes make all unacceptable imagery shallow'.

- A hostile individual (e.g. a disgruntled undergraduate student) might search through data manually, seeking incriminating material to use against colleagues/supervisors.

- Re-publication. If the experimental dataset is put online for download, it may enter search engine image databases with a high priority, and thus be shown frequently to the general public. This problem is made worse by search engine page-ranking algorithms that assign high priority to academic sites. The offensive data may become clearly and directly associated with your institution, in full view of the public.

There are presently no publicised cases of image-processing or information forensics researchers being arrested for accidental download of illegal material. However, there are reported cases of other well-intentioned computer-literate users running into serious problems with the police after their computer automatically downloaded image files without their direct knowledge (Oates, 2006).

2d. Plausible deniability and broader risks

Cases A and B, below, illustrate how failure to address this problem could allow criminals to use plausible deniability as a defence against prosecution, at the expense of the prestige of the research field.

Case A: A researcher or hobbyist uses a webcrawler to gather a large dataset from home to test out some well-known algorithms, and puts results online. However, their dataset *accidentally* contains some illegal images. This is discovered at the customs department of an airport. The researcher protests their innocence.

Case B: A researcher or hobbyist uses a webcrawler to gather a large dataset from home to test out some well-known algorithms, and puts results online. However, their dataset *deliberately* contains some illegal images. This is discovered at the customs department of an airport. The researcher protests their innocence.

As an outsider, it would be very difficult to distinguish between these two situations. The underlying problem is that it is very difficult to know post-hoc whether a webcrawler program collected a small amount of illegal material deliberately, or accidentally. It is also impossible to be certain of how the data was used in the past, intentionally or otherwise. Even traversal logs are of little use in a world where Internet links can change continually and dynamically on a second-by-second basis.

Furthermore, in areas such as steganography and CAPTCHA development, there is a considerable volume of genuine research and study that is undertaken outside of research institutions. Any member of the public could therefore plausibly claim a purely academic motive in gathering material. It is hard to imagine that a court could allow such an obvious 'get out of jail free' opportunity for criminals who wish to store illegal data. Indeed, reports indicate that police are very unlikely to accept ignorance as an excuse in such matters (Soghoian, 2007).

Ultimately, even if a researcher believes that the risk to themselves is acceptably small, there still remains the secondary danger of harm to the prestige of research fields as a whole. How many national stories portraying information forensics researchers as gatherers of unacceptable imagery would be necessary to bring that field into disrepute? Even a single highly publicised case could be a publicity disaster for researchers in a particular field, particularly for those seeking tenure or research grant support. The current practice of using and encouraging webcrawling for the purpose of gathering large datasets carries significant potential dangers, both for individuals and for entire research fields.

2e. Science-related problems

While the use of large datasets in science generally brings with it a measure of confidence about the results found, the means by which such datasets are gathered can be as important as the scale on which gathering takes place. This paper now raises concerns about data quality, expense, and reproducibility, in the context of blind webcrawling for millions of images.

Accidental positives: Generally, when a researcher gathers images on a large scale from the Internet, they are attempting to collect many 'normal' images. It is well known that in practice, if material is gathered blindly from across the Internet, it may be less 'normal' than expected. For example, images may contain steganography or watermarks, or may produce equivalent results to that of marked media (Provos and Honeyman, 2003). Each of these situations can have an effect upon the behaviour of an algorithm, especially if it is meant to be training upon a set of clear images. Similarly, unexpected true positives gathered by accident among a set of true negatives may seem to be unavoidable false positive results.

Unpredictable malformed data: Large image sets that are automatically downloaded and unchecked by humans may contain files that are partial or incomplete, contain corrupted data, or have an incorrect data-type or metadata. They may have been produced by unreliable software. Unusual and unanticipated inputs may cause experimental algorithms to behave in unpredictable ways that have not been observed during earlier testing. This may substantially reduce the overall reliability of experimentation. This problem was noted by the founders of Google (Brin and Page, 1998): "Invariably, there are hundreds of obscure problems which may only occur on one page out of the whole Web and cause the crawler to crash, or worse, cause unpredictable or incorrect behavior. Systems which access large parts of the Internet need to be designed to be very robust and carefully tested". This is as true for image data as it is for hypertext data.

It is typically impossible for a small research group to manually check millions of samples of data and millions of algorithm executions over unknown data. It is therefore impossible to be certain that the overall results have not been skewed by erroneous program behaviour. In this sense, the use of large webcrawled datasets raises concerns akin to those involved with large automatically generated mathematical proofs (see (Swart, 1980; Wilson, 2002) for discussion). In mathematics, such proofs are looked upon as 'acceptable, but undesirable'. This is because there is often no possibility of direct human verification of the complete behaviour of the program or complete detail of the result.

Why should it be any different in image-processing research? Surely, results based primarily on blindly gathered and blindly processed datasets should be viewed with greater skepticism than those from a dataset that is known to be suitable and correctly-processed, through direct verification by a human?

Expense: A researcher may feel pressure to conduct an experiment on millions of images for many reasons. Perhaps they feel that since other papers have conducted experiments on such a large scale, their work will be unwelcome if they are not seen to be operating 'at the same level'. Alternatively, a reviewer may assert that conducting a large-scale experiment using millions of blindly webcrawled data samples is a necessary part of publication (in fact, such an incident prompted this paper). Alternatively, it may be the reviewer themselves, seeking to duplicate results produced from a large dataset, who incurs the expense.

Gathering millions of examples of any media file is an expensive and time-consuming business for any academic. However, it presents a particularly significant problem for academics who are unfunded or who are operating outside of a well-equipped academic environment.

Firstly, there is the cost of equipment suitable for storing and processing large datasets, and the electricity to run the equipment throughout the period of collection. Secondly, the cost of human time in maintaining and administrating the equipment and data. Thirdly, the cost of bandwidth for all webcrawls carried out, which is far from trivial for an academic working from home or working within a developing country. For example, broadband charges in some African states average more than US$1300 per month, with relatively slow connection speeds (UNCTAD, 2009). Fourthly, delay: researchers operating with connections of limited speed must incur a time cost that may be measured in months. This time cost delays authorship, review and publication. Finally, there may be costs in terms of human administration and correspondence. Brin and Page note that "since large complex systems such as crawlers will invariably cause problems, there needs to be significant resources devoted to reading the email and solving these problems as they come up." (Brin and Page, 1998)

The costs in terms of equipment, human time, financial cost and delay due to data transfer, may make experimentation (or duplication) at this scale impractical and inaccessible to all but the most well-provisioned academics. Consequently, reviewers who expect such datasets to be used - without highly compelling justification - may effectively prohibit publication by authors who lack sufficient time, funding and equipment.

Reproducibility: Assuming that a dataset containing millions of media files has been gathered, how can the experiments be reproduced by others? It is not reasonable to assume that image datasets measured in terms of hundreds of gigabytes (or MP3 or video datasets measured in terabytes) can be quickly or conveniently shared between researchers. It is convenient to pass a 1000-image dataset by e-mail or FTP to an editor, to share with several anonymous reviewers. In contrast, it is far less convenient to buy and mail several large hard drives, particularly if they must be forwarded back and forth via an editor.

Similarly, it is not reasonable to assume that all researchers will have the resources necessary to process such a large volume of data, in terms of computational power, time, and human effort. If anything, multi-million-sample datasets strongly discourage reproduction of results - and this is bad for science.

Imagine the case of a reviewer who seeks to demonstrate that a result based upon a million-image dataset is incorrect. If they demonstrate their case on a self-gathered 1000-image dataset, the original author may imply the new result is merely a consequence of the smaller dataset size, and insist that the reviewer should develop their own million-sample dataset (at their own cost!) in order to make their case. Similarly, once a technique has been tested on a very large webcrawled dataset, the belief may arise that any reproduction or validation of the experiment should take place on the same scale - a practice shown here to be both expensive and dangerous. Surely it is better if research experiments take place on a scale that can be privately and independently reproduced in a convenient, fast, affordable, and legally safe manner?

Furthermore, the Internet is a fast-changing environment, and even a 500GB slice of the Internet is now a relatively small fraction of the whole. If the very large volumes of data that have been gathered through webcrawling are not carefully kept and backed up in perpetuity, it may be impossible to reproduce the original result. Indeed, the rate of change of the Internet makes it impossible to be certain that the entire contents of any large dataset were truly gathered in full from the Internet to begin with.

Ethics/legality and reproducibility: There are concerns that relate simultaneously to experimental reproducibility and also to the earlier issues of copyright and media content. While an individual researcher may decide that they are willing to take risks with such issues, is it reasonable to expect that any future researchers seeking to verify the result using the original dataset must take the same risks?

Furthermore, while researchers may be willing to access, download and process samples blindly gathered from the Internet, they may be less willing to freely and fully redistribute their dataset if it is known or suspected to contain unacceptable material. This may hamper or completely prevent subsequent reproduction of experiments.

Alternatively, researchers may have a policy of removing objectionable or illegal data from their datasets when discovered, so that the dataset can continue to be distributed freely. However, this means that the results of a paper can no longer be completely reproduced without all of the original data, which is undesirable. This problem could be particularly important among very large datasets used for benchmarking purposes among many different research groups.

Summary: It appears that the mild statistical benefits provided by using a very large dataset for experiments (as opposed to a dataset containing perhaps only thousands of samples) are offset by many serious scientific concerns relating to reliability & human verification, expense, and reproducibility. Consequently it seems difficult to justify the use of blindly gathered datasets that are too large for any single researcher to manually audit.

Perhaps the use of 'unauditably large' datasets should be discouraged, in order to promote and encourage external validation of results, manual validation of results, and participation from poorly-resourced scientists. Considering the Pareto Principle (the 80/20 rule), perhaps research communities in fields such as image-processing and information forensics should collectively determine standardised sizes for sample sets. These sizes could be chosen so as to optimise both for utility of information provided and also for convenience (safety, expense, reproducibility...). It might well be the case that multiple independent sample sets containing hundreds of 'safe' images are more than sufficient for many purposes.

3. Alternative approaches

A number of possible solutions to some of the problems highlighted in this paper are now introduced. Several of these solutions could be quickly achieved, and might also result in significant spin-off benefits for researchers.

- Ignore the problem and/or use a disclaimer. I strongly recommend against this strategy. A disclaimer asserting intentional ignorance of a possible crime is unlikely to provide strong legal protection, especially when alternative solutions exist that do not involve breaking the law. Researchers opting for this strategy should perhaps be ready to justify it to university presidents and local police forces.

- Maintain a list of 'official' researchers in fields such as steganalysis. This would exclude researchers from the field unnecessarily and would probably not be recognised internationally by law enforcement agencies. It also apparently provides an excellent excuse for any criminal capable of writing a steganalysis paper. It is not an effective solution.

- Gain access to publicly audited datasets. It is possible that companies such as Google and Youtube might be willing to share a large dataset of images that have been vetted as 'safe' by the public. Unfortunately, 'safety' (as decided by the public) is more likely to refer to issues such as overt pornography, rather than issues such as copyright. Further, even with cross-checking of results, it is still possible that some unacceptable images will slip through surveillance. *Consequently, the researcher still bears a risk for every image that they have not manually inspected themselves*. This is only a partial solution at best.

- Use synthetic data? It may be possible to generate suitable datasets automatically, in a variety of ways. Perhaps by using fractals or procedural texture generation techniques; or by modifying and combining a smaller set of known-safe images together combinatorially, using transformations and mappings found to be suitable for research purposes (i.e. that do not introduce unnatural artefacts). Alternatively, researchers could set up 3-D models and animations, and select thousands of frames from animations. Such methods have the advantages of being legal, ethical, easily reproducible and quick to transmit. They are also more likely to produce valid data structures that do not cause unexpected program behaviour during experimentation. Synthetic methods based on 3-D animation or procedural generation might use only a few kilobytes of program code to generate thousands or millions of unique and distinct images. The automated generation of large-scale synthetic datasets that fit the needs of fields such as information forensics and image-processing represents an interesting problem for future research.

- Use appropriate quantities of data. In medicine, genuine life-or-death decisions are sometimes made on the basis of empirical studies with only tens or hundreds of sample points. In that light, can 'millions of samples' truly be justified as a requirement in areas such as information forensics and image-processing? I suggest that it is far more important to discover whether a technique is approximately 0, 10, 50, 90 or 100 percent effective - in a safe and easily reproducible way - than to try to place four or five significant figures of accuracy on the initial experimental results. A broad picture can often be usefully established with relatively few data points. A more precise model of accuracy is most effectively built through numerous independent studies rather than a single large study. Datasets of around 100-1000 images should be sufficient for many practical purposes, and can be audited by hand for quality and legality. The use of small, safe, effective, affordable and easily-reproduced datasets should therefore be encouraged.

- Instead of pushing for individual researchers to carry out large experiments, perhaps it would be more valuable to encourage researchers to offer to validate each other's techniques with smaller-scale experiments? Three independent research groups producing identical results with unique implementations and independently constructed 500-image datasets, may provide a stronger degree of confidence in a technique than could be achieved by a single researcher with one million blindly selected samples.

It might even be possible to automate the process of experimentation in fields such as steganography. I envisage a 'steganography clearing centre' website, to which researchers would submit their novel techniques as source code or virtual machine code. This centre would contain many known-legal samples that have been gathered manually and audited by hand. The submitted algorithm could then be automatically tested against standardised sample-sets on a controlled playing field.

This would provide a stronger degree of confidence in the reliability of experimental results, as the data and experimental process would be out of the control of individual researchers. It would also provide increased flexibility during experimentation (perhaps offering a variety of alternative experimental datasets). External reproducibility would become straightforward; simply a matter of logging in and clicking the correct buttons to re-run the experiments. It would remove all of the legal and ethical risks outlined in this paper from individual researchers, and would standardise and strengthen the quality of research in the field.

A useful model for such a clearing centre is the Redcode 'King of the Hill server' (Metcalf, 2009), to which any Internet user may submit source code by e-mail. Submitted code is automatically compiled and experimentally tested against a variety of competing techniques, with results returned within minutes by e-mail. The datasets used in such a clearing centre could be easily audited for both legality and quality, and could be updated annually to stay in line with real world Internet data. Essentially, I strongly advocate establishing a centralised, standardised, Internet-based, quality controlled, ethical benchmarking system for empirical research in fields such as image processing and information forensics.

4. Conclusions

Many of the problems that have been highlighted in this paper apply to text, audio and video datasets that are gathered online as much as they apply to research into computer imagery. The use of large, webcrawled datasets presently seems like a disaster waiting to happen, that could jeopardise the reputations of individual researchers and research fields alike. In conclusion, I strongly encourage computing researchers to devise and use new mechanisms for collecting, auditing, using and redistributing Internet-based materials for empirical studies in computing.

About the author

Graeme Bell holds a Ph. D. in Computer Science from the University of St Andrews, UK. He was the top Science graduate from the University of St Andrews in 2001 and also the winner of the 2001 Scottish Young Software Engineer of the Year award. His research interests include artificial intelligence, bioinformatics, robotics, image-processing, steganography, and the Internet.

E-mail: datasetspaper /at/ graemebell /./ net

Notes

[1] For example: Slashdot.org, Digg.com, Reddit.com, Facebook.com, Myspace.com, 4chan.org.

[2] A complete taxonomy of objectionable material is left as an exercise for the reader.

[3] The author was unable to find more recent publications addressing this difficult topic.

References

Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi, 1999. "The diameter of the World-Wide Web," Nature, volume 401, pp. 130-131.

O. Bowcott, 2006. "Arrest extremist marchers, police told," The Guardian, (6 February), at http://www.guardian.co.uk/uk/2006/feb/06/raceandreligion.muhammadcartoons, accessed 23 October 2009.

S. Brin and L. Page, 1998. "The anatomy of a large-scale hypertextual Web search engine," Proc. Seventh International World-Wide Web Conference, Australia. Elsevier Science Publishers B. V., pp. 107-117.

P. Cohen, 2009. "Yale press bans images of Muhammad in new book by Jytte Klausen," New York Times (12 August), at http://www.nytimes.com/2009/08/13/books/13book.html, accessed 23 October 2009.

Council of Europe, 2002. "Additional protocol to the convention on cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems," November 2002. Council of Europe, Strasbourg, and at http://conventions.coe.int/Treaty/en/Treaties/Html/189.htm, accessed 23rd October 2009.

A. Curry, 2008. "Is cartoon controversy history? Danish library wants to preserve inflammatory drawings," Spiegel Online, (30th January) at http://www.spiegel.de/international/europe/0,1518,532057,00.html, accessed 23rd October 2009.

M. De la Vina, 2009. "Online theft costs the photo industry up to $10 billion," Press Release, Vocus/PRWEB (18 September) at http://www.prweb.com/releases/Vivozoom/microstock_images/prweb2898414.htm, accessed 23 October 2009.

J. Dibbell, 2009. "Scientology: The Web's first copyright-wielding nemesis," Wired Magazine, (21 September) at http://www.wired.com/culture/culturereviews/magazine/17-10/mf_chanology_sidebar, accessed 23rd October 2009.

D. Eichmann, 1994. "The RBSE spider: balancing effective search against Web load," Proc. of the First World Wide Web Conference, Geneva, Switzerland. Elsevier Science BV, pp. 113–120.

M. Frith, 2003. "20,000 child porn images a week put on Internet, says NSPCC," The Independent (8 October) at http://www.independent.co.uk/news/business/news/20000-child-porn-images-a-week-put-on-Internet-says-nspcc-582609.html, accessed 23 October 2009.

Google at http://www.google.com, accessed 16 October 2009.

Christopher Hom, 2008. "The semantics of racial epithets," The Journal of Philosophy, volume 105, number 8, pp. 416-440.

M. Kharrazi, H.T. Sencar, and N. Memon, 2005. "Benchmarking steganographic and steganalysis techniques," Proc. SPIE, Security, Steganography, and Watermarking of Multimedia Contents VII, SPIE volume 5681, pp. 252-263.

T. Korosec, 2003. "1-Hour arrest," Dallas Observer, (April 17) at http://www.dallasobserver.com/content/printVersion/281491, accessed 23 October 2009.

J. Metcalf, 2009. "Corewar: King of the hill," at http://corewar.co.uk/hills.htm, accessed 23 October 2009.

J. Oates, 2006. "German police seize TOR servers," The Register (September 11) at http://www.theregister.co.uk/2006/09/11/anon_servers_seized/, accessed 23 October 2009.

Neils Provos and Peter Honeyman, 2003. "Hide and seek: An introduction to steganography, " IEEE Security and Privacy, volume 1, number 3, pp. 32-44.

E. Renold and S.J. Creighton, 2003. *Images of abuse: a review of the evidence on child pornography*. NSPCC Publications, London.

C. Soghoian, 2007. "Tor anonymity server admin arrested," CNET News (16 September) at http://news.cnet.com/8301-13739_3-9779225-46.html, accessed 23 October 2009.

J. Stanley, 2001. "Child abuse and the Internet," Child Abuse Prevention Issues, volume 15, pp. 1-20, and at http://www.aifs.gov.au/nch/pubs/issues/issues15/issues15.pdf, accessed 23 October 2009.

E.R. Swart, 1980. "The Philosophical Implications of the Four-Color Problem," American Mathematical Monthly, volume 87, number 9, pp. 697-702.

UNCTAD, 2009. "Africa catches up in mobile phones but is falling behind in broadband access," United Nations Conference on Trade and Development, UNCTAD/PRESS/PR/2009/057 (October 22), and at http://www.unctad.org/Templates/webflyer.asp?docid=12273&intItemID=1528&lang=1, accessed 24 October 2009.

Wikipedia, 2009. "Blasphemy law," at http://en.wikipedia.org/wiki/Blasphemy_law, accessed 23 October 2009.

R Wilson, 2002. *Four Colours Suffice: How the Map Problem Was Solved*. The Penguin Press.

License