Communications of the Association for Information Systems

Volume 18 Article 29

11-29-2006

Anything You Search Can Be Used Against You in a Court Of Law: Data Mining in Search Archives

Blake Ives
University of Houston

Vlad Krotov
University of Houston, vokrotov@uh.edu

Communications of the Association for Information Systems

ANYTHING YOU SEARCH CAN BE USED AGAINST YOU IN A COURT OF LAW: DATA MINING IN SEARCH ARCHIVES

Blake Ives
Vlad Krotov
Information Systems Research Center
University of Houston
blake.ives@uh.edu

ABSTRACT

AOL's recent public release of user search information resulted in a heated privacy debate. This case study is a detailed account of this incident. The case is designed as an in-class teaching aid covering managerial, legal, and ethical issues related to privacy. It consists of four sections (A, B, C, and D). Each section is fairly short and is designed to be read in class, separated by discussion of the previous section. Alternatively, the first section might be distributed in advance; though this runs the risk of students identifying the case and jumping ahead in the discussion (AOL's identity is concealed from students until the end of section B). A set of potential discussion questions for each section appears in the appendix. While there are too many questions to be covered in a single class, instructors can choose questions based on their particular teaching objective. A teaching note is also available from the authors.

Keywords: AOL, search data, privacy, teaching case study, ethics

Editor's Note: A teaching note is available from the first author (blake.ives@uh.edu) to faculty requiring it who are listed in the MISRC-ISWorld Faculty Directory.

I. DATA MINING IN SEARCH ARCHIVES (A)

On May 30, 2006, three computer scientists presented a scholarly paper on information search behavior at a conference in Hong Kong. The paper, "A Picture of Search", described the trio's investigation of factors related to the efficiency and effectiveness of large-scale search services. Among these factors were (Pass et al., 2006):

1. The nature of the users' search queries and how queries change over time.
2. How users form search queries.
3. How users interact with a search service.
4. The runtime efficiency of a search service.
5. The nature of search results produced.
6. Who uses search engines?

The findings reported that entertainment, shopping, and pornography were the most common queries, that a small percentage of users perform the majority of queries, and that a high proportion of the queries come from the east and west coasts of the United States (Pass et al., 2006). Among the conclusions was the prediction that further research could produce invaluable improvements in large-scale search services.

Hoping to stimulate such research, the team made their archive of search data available on the Internet in late July (or early August) of 2006. The database contained approximately 20 million web queries performed by over 650,000 users for three months during the spring of 2006. Each record in the database contained the following fields (see Exhibit A1 for a detailed description of the dataset) (Sadetsky, 2006a):

- **AnonID** –anonymous user ID; in order to protect user privacy, user logins were replaced with a unique random number
- **Query** - the query submitted by the user
- **QueryTime** - the time and date at which the query was submitted
- **ItemRank -** the order number of the search result on which the user clicked on the search results page (if none was clicked – the field was marked as nil)
- **ClickURL** – the web address of the page which the user chose to visit (if none was clicked – the field was marked as nil)

Exhibit A2 contains illustrative search data for one user.

This information was reportedly made available for the purpose of providing real query log data of real users which can be used for "for personalization, query reformulation or other types of search research" (Sadetsky, 2006b).

EXHIBIT A1

Dataset Description

(Reprinted from Sadetsky, 2006b)

This collection is distributed for NON-COMMERCIAL RESEARCH USE ONLY. Any application of this collection for commercial purposes is STRICTLY PROHIBITED.

Brief description:

This collection consists of ~20M web queries collected from ~650k users over three months. The data is sorted by anonymous user ID and sequentially arranged. The goal of this collection is to provide real query log data that is based on real users. It could be used for personalization, query reformulation or other types of search research.

The data set includes {AnonID, Query, QueryTime, ItemRank, ClickURL}.

- AnonID - an anonymous user ID number.
- Query - the query issued by the user, case shifted with most punctuation removed.
- QueryTime - the time at which the query was submitted for search.
- ItemRank - if the user clicked on a search result, the rank of the item on which they clicked is listed.
- ClickURL - if the user clicked on a search result, the domain portion of the URL in the clicked result is listed.

Each line in the data represents one of two types of events:

1. A query that was NOT followed by the user clicking on a result item.
2. A click through on an item in the result list returned from a query.

In the first case (query only) there is data in only the first three columns/fields -- namely AnonID, Query, and QueryTime (see above).

In the second case (click through), there is data in all five columns. For click through events, the query that preceded the click through is included. Note that if a user clicked on more than one result in the list returned from a single query, there will be TWO lines in the data to represent the two events. Also note that if the user requested the next "page" or results for some query, this appears as a subsequent identical query with a later time stamp.

CAVEAT EMPTOR -- SEXUALLY EXPLICIT DATA! Please be aware that these queries are not filtered to remove any content. Pornography is prevalent on the Web and unfiltered search engine logs contain queries by users who are looking for pornographic material. There are queries in this collection that use

SEXUALLY EXPLICIT LANGUAGE. This collection of data is intended for use by mature adults who are not easily offended by the use of pornographic search terms. If you are offended by sexually explicit language you should not read through this data. Also be aware that in some states it may be illegal to expose a minor to this data. Please understand that the data represents REAL WORLD USERS, un-edited and randomly sampled, and that AOL is not the author of this data.

Basic Collection Statistics

Dates:
01 March, 2006 - 31 May, 2006
Normalized queries:
36,389,567 lines of data
21,011,340 instances of new queries (w/ or w/o click-through)
7,887,022 requests for "next page" of results
19,442,629 user click-through events
16,946,938 queries w/o user click-through
10,154,742 unique (normalized) queries
657,426 unique user ID's

EXHIBIT A2

Selected Queries for User No. 1247763

User ID	Search Keywords	Date	Website
1247763	teeth pain does not go away	3/4/06 23:38	http://www.yourdentist.com
1247763	Prozac side effects	3/9/06 23:21	http://www.emtionalhealth.com
1247763	bill collectors Missouri	3/13/06 13:43	http://www.howardpeopleslawyer.net
1247763	fidelity	3/13/06 16:39	http://www.fidelity .com
1247763	antique rifle	3/13/06 19:48	
1247763	how to stop collection calls	3/16/06 21:54	http://www.oag.us
1247763	Missouri debt collection act	3/16/06 22:08	http://www.debtcollectionlawyers.net
1247763	brushing sensitive teeth	3/17/06 0:00	http://www.oralcare.com
1247763	oral-b	3/17/06 12:34	http://www.oral-b.com
1247763	stop assignee calls	3/18/06 13:35	
1247763	stop assignee calls collection	3/18/06 13:36	http://p32218.ezboard.com
1247763	marble bust Roman	3/18/06 18:51	http://www.tradeantiques.com
1247763	marble bust Greek	3/18/06 18:54	http://cgi.ebay.com
1247763	Victorian bust woman	3/18/06 19:09	http://www.victorianera.org
1247763	sore jaw after going to dentist	3/22/06 10:48	
1247763	Myspace	3/24/06 17:12	http://www.myspace.com
1247763	buy stamps Angola	3/25/06 12:20	http://www.stampcollection.com
1247763	acoustic guitar	3/25/06 12:34	http://www.musicstore.com
1247763	acoustic guitar sale	3/25/06 12:42	http://www.smithguitars.com
1247763	Hotel California lyrics	3/25/06 13:42	http://www.allsongs.com
1247763	Yankee parade	3/25/06 14:42	http://www.ushistory.org
1247763	Missouri college gpa teacher Adair district	3/26/06 12:33	http://www.educationboard.org
1247763	teachers gpas Adair district	3/26/06 13:18	http://www.teachingjobs.org
1247763	Missouri teacher's average gpa	3/26/06 13:27	
1247763	state board education Missouri	3/26/06 13:28	http://www.mbec.state.mi.us
1247763	pick a professor	3/27/06 20:05	http://www.pickaprof.com
1247763	Missouri Department of Corrections	4/7/06 18:45	
1247763	inmate search	4/7/06 19:05	http://www.correction.state.tx.gov
1247763	Missouri inmates	4/7/06 19:08	http://www.ancestorhunt.org
1247763	peanuts jaw pain	4/14/06 23:31	http://www.healthtips.com
1247763	Civil War reenactment clothing	4/15/06 0:13	http://www.texas-wars.org
1247763	sword copy for sale	4/15/06 22:48	http://www.maricopa.edu/~hashim
1247763	multiple sclerosis	4/20/06 20:41	http://www.nationalhealthsociety.org
1247763	trunk refinishing supplies	4/20/06 21:06	

User ID	Search Keywords	Date	Website
1247763	trunk supplies	4/20/06 21:15	http://www.antique-hardware.com
1247763	record conversation Missouri	4/20/06 21:31	http://www.spyemporium.com
1247763	jaw pain after exercise	4/20/06 23:57	
1247763	Nazi flag authentic	4/22/06 20:53	http://www.epier.com
1247763	show antique New York	5/2/06 21:03	http://www.maineantiquedigest.com
1247763	Wade Stein antique show New York	5/2/06 21:05	http://www.maineantiquedigest.com
1247763	silly hats	5/2/06 22:28	
1247763	median gpa Missouri	5/6/06 17:42	http://www.4lawschool.com
1247763	Missouri State University	5/6/06 17:54	
1247763	Roman coins	5/6/06 20:29	http://www.romancoins.com
1247763	rate professor	5/10/06 20:10	http://www.ratemyprofessor.com
1247763	Sony video camera	5/11/06 23:08	http://www.videoworks.com
1247763	Canon Power Shoot Zoom	5/11/06 23:25	
1247763	Ted Kaczynski	5/12/06 10:58	http://www.postmodernism.com
1247763	wagon wheel west	5/12/06 22:27	http://www.farmequipmentsupply.com
1247763	median gpa Missouri State University	5/14/06 19:37	http://www.businessweek.com
1247763	median gpa sociology graduate school Texas	5/14/06 19:43	
1247763	General Lee	5/15/06 13:43	http://www.historyproject.org
1247763	Ancestors Civil War	5/15/06 13:54	http://www.americanheroes.com
1247763	single women marriage	5/15/06 16:45	http://www.plentyoffish.com
1247763	Yahoo personals	5/15/06 16:55	http://www.personals.yahoo.com
1247763	tobacco jaw pain	5/17/06 13:43	http://www.quit-assist.com
1247763	Metallica tour Texas	5/18/06 21:11	http://www.metallica.org
1247763	best guitarists	5/18/06 21:23	http://www.rockmusic.org
1247763	Metallica tickets	5/18/06 21:46	http://www.ticketmaster.com
1247763	bipolar symptoms	5/19/06 19:15	http://www.psychology.umich.edu
1247763	schizophrenia signs	5/19/06 20:23	http://www.schizopreniahelp.org
1247763	Cliff notes das Kapital	5/20/06 10:37	http://www.cliffnotes.com
1247763	Jessica Ashford Kennedy High	5/22/06 22:38	http://www.classmates.com
1247763	refinancing credit card debt visa	5/23/06 9:38	http://www.refinance-now.com
1247763	Swimming lessons Alexandria Missouri	5/23/06 10:12	http://www.personaltrainers.com
1247763	swimming pulse rate	5/24/06 22:13	http://www.lifefitness.com
1247763	mold kitchen	5/25/06 11:40	http://www.homecare.com
1247763	find roommate	5/25/06 15:20	http://www.roommates.com
1247763	Mexican restaurants Alexandria	5/25/06 18:38	http://www.citysearch.com

User ID	Search Keywords	Date	Website
	Missouri		
1247763	stake house Alexandria	5/26/06 17:35	http://www.southernstakehouse.com
1247763	anxiety	5/26/06 23:59	http://www.anxietydisorder.org

II. DATA MINING IN SEARCH ARCHIVES (B)

Starting on August 6th, 2006 Internet bloggers became aware of the database and were concerned about its potential uses. Mike Arrington (2006), a former high-tech lawyer, wrote in his *TechCrunch* blog:

> *The most serious problem is the fact that many people often search on their own name, or those of their friends and family, to see what information is available about them on the Net. Combine these ego searches with porn queries and you have a serious embarrassment. Combine them with 'buy ecstasy' and you have evidence of a crime. Combine it with an address, Social Security number, etc., and you have an identity theft waiting to happen. The possibilities are endless.*

Two days later, *The Wall Street Journal* picked up the story (Richmond, 2006):

> *The incident recalled a controversy sparked earlier this year by Google, Inc. when it refused to turn over to the Justice Department a sample of anonymous search queries conducted by its users, citing competitive concerns and consumers' privacy fears.*

The Wall Street Journal article was referring to January 19, 2006, when the Department of Justice subpoenaed Google, Microsoft, Yahoo, and AOL for data on user queries submitted to their online search engines and a random sample of one million indexed web addresses (Mohammed, 2006). The Department of Justice sought to use this data to combat child pornography.

Microsoft, AOL, and Yahoo announced that they complied with the request on a limited basis (Mohammed, 2006). For instance, AOL spokesperson Andrew Weinstein commented (Mohammed, 2006):

> *We did provide the DOJ with some information that we thought would be of use to them, but it was not the information requested in the subpoena and there were no privacy implications for our users... [we provided a] generic list of aggregate and anonymous search terms*.

Google refused to provide any data to the Department of Justice, but a U.S. District Court judge subsequently ruled that Google had to provide a random sample of 50,000 URLs in its index database, but not the search query data (Perez, 2006).

On February 8, 2006, in response to the controversy surrounding Google's refusal to release sensitive search data to the DOJ, Massachusetts Democratic Representative to the U.S. Congress, Ed Markey introduced the Eliminate Warehousing of Consumer Internet Data Act (EWOCID) (see Exhibit B1) (Markey, 2006). The bill would require "owners of Internet websites to destroy obsolete data containing personal information" (Markey, 2006). Markey proposed that the violators of this requirement be punished by the FTC. The bill was subsequently "bottled up" in Congress (McCullagh, 2006).

Some technology lobbying groups and free market advocates expressed skepticism towards EWOCID (McCullagh, 2006). Sonia Arrison, the director of the Pacific Research Institute in San Francisco, a free market advocacy group, commented (McCullagh, 2006):

> *Rep. Markey's bill seeks to micromanage technology firms, which would be an enormous step in the wrong direction… Why on Earth would anyone think the FTC would do a better job at managing data than Google or Yahoo?*

Apparently, Google's concern that the release of search query data might compromise consumers' privacy was well placed. On August 9, 2006 reporters for *The New York Times* studied the newly released database of search queries and were able to identify user No. 4417749 (Barbaro et al., 2006):

> *No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything." And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."*

Number 4417749, now identified as Thelma Arnold, a 62-year-old widow from Lilburn, Georgia, reacted to the public release of her searches (Barbaro et al., 2006):

> *My goodness, it's my whole personal life, I had no idea somebody was looking over my shoulder… We all have a right to privacy. Nobody should have found this all out.*

While Ms. Arnold's queries were innocent enough, the search queries of other AOL members were disturbing. Another user, for instance, performed the following searches (Frind, 2006):

> *how to kill your wife*
> *wife killer*
> *how to kill a wife*
> *dead people*
> *pictures of dead people*
> *killed people*
> *dead pictures*
> *murder photo*
> *steak and cheese*
> *photo of death*
> *death*
> *dead people photos*
> *photo of dead people*
> *www.murderdpeople.com*
> *decapitated photos*
> *car crashes*
> *car crash photo*

A blogger highlighted another disturbing search pattern (Frind, 2006):

> *A particular user, 8581027, searches for preteen sex for awhile. Then he searches for a specific person in a specific physical location, and later the same person with the birth date of 1999. [...] Then searches for "grants for english teachers in mississippi".*

Other bloggers pointed out that over one hundred users in database had been searching for child pornography (Frind, 2006).

On August 6, 2006 Executives at AOL responded to the unwelcome press attention.

EXHIBIT B1
Eliminate Warehousing of Consumer Internet Data Act of 2006

(Reprinted from Markey, 2006)
HR 4731 IH
109th CONGRESS
2d Session
H. R. 4731

To require owners of Internet websites to destroy obsolete data containing personal information.

IN THE HOUSE OF REPRESENTATIVES

February 8, 2006

Mr. MARKEY introduced the following bill which was referred to the Committee on Energy and Commerce.

A BILL

To require owners of Internet websites to destroy obsolete data containing personal information.

> *Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled*,

SECTION 1. SHORT TITLE.

> This Act may be cited as the `Eliminate Warehousing of Consumer Internet Data Act of 2006'.

SECTION 2. FINDINGS.

> Congress finds the following:
>
> > (1) As the Nation's communications networks continue to grow and become ever more sophisticated, more individuals and industries will be using such networks to communicate and conduct commercial transactions.
> >
> > (2) The ease of gathering and compiling personal information during such communications, both overtly and surreptitiously, is becoming increasingly efficient and almost effortless due to advances in digital telecommunications technology and the widespread use of the Internet.
> >
> > (3) Consumers have an ownership interest in their personal information.
> >
> > (4) Information gathered about consumers over the Internet can provide detail about some of the most intimate aspects of an individual's life, including their Internet interests, communications with other citizens, purchases, information inquiries, and political or religious interests, affiliations, or speech.
> >
> > (5) Certain information about Internet searches or website visits conducted from a particular computer can be obtained and stored by websites or search engines, and can be traced back to individual computer users.
> >
> > (6) Fair information practices include providing consumers with knowledge of any data collection, conspicuous consumer notice of an entity's data practices, consumer choice to provide consent or deny authorization for such practices, access to data collected, safeguards to ensure data integrity, and contact information.
> >
> > (7) In order to safeguard consumer privacy interests, companies that gather personal information that can identify individual consumers should cease to store such information after it is no longer necessary to render service to such consumers or to conduct any legitimate business practice.
> >
> > (8) Cable operators, who can gather personal information about a subscriber's use of the cable system and obtain information about a consumer's video programming choices and use of their cable modem are currently required under section 631 of the Communications Act of 1934 (47 U.S.C. 551) to destroy any personal information gathered from a subscriber after it is no longer necessary

for the purpose for which it was gathered and if there are no other pending legal requests for such information.

(9) A similar obligation should govern information gathered about consumers by Internet websites, which often possess information about computer users which is more detailed, and arguably more personalized, than information cable operators typically gather.

SECTION 3. DESTRUCTION OF DATA WITH PERSONAL INFORMATION BY INTERNET WEBSITES.

An owner of an Internet website shall destroy, within a reasonable period of time, any data containing personal information if the information is no longer necessary for the purpose for which it was collected or any other legitimate business purpose, or there are no pending requests or orders for access to such information pursuant to a court order.

SECTION 4. ENFORCEMENT BY THE FEDERAL TRADE COMMISSION.

A violation of Section 3 shall be treated as a violation of a rule defining an unfair or deceptive act or practice prescribed under section 18(a)(1)(B) of the Federal Trade Commission Act (15 U.S.C. 57a(a)(1)(B)). The Federal Trade Commission shall enforce this Act in the same manner, by the same means, and with the same jurisdiction as though all applicable terms and provisions of the Federal Trade Commission Act were incorporated into and made a part of this Act.

SECTION 5. DEFINITIONS.

As used in this Act the following definitions apply:

(1) The term `Internet' means collectively the myriad of computer and telecommunications facilities, including equipment and operating software, which comprise the interconnected world-wide network of networks that employ the Transmission Control Protocol/Internet Protocol, or any predecessor or successor protocols to such protocol, to communicate information of all kinds by wire or radio.

(2) The term `personal information':

(A) means information that allows a living person to be identified individually, including the following: the first and last name of an individual, a home or physical address of an individual, date or place of birth, an email address, a telephone number, a Social Security number, a tax identification number, birth certificate number, passport number, driver's license number, credit card number, bank card number, or any government-issued identification number; and

(B) does not include any record of aggregate data that does not permit the identification of particular persons.

(3) The term `web page' means a location that has a single Uniform Resource Locator or another single location with respect to the Internet, as the Federal Trade Commission may prescribe.

(4) The term `Internet website' means a collection of web pages that are presented and made available by means of the Internet as a single website (or a single web page so presented and made available), which web pages have any of the following characteristics:

(A) A common domain name.

(B) Common ownership, management, or registration.

III. DATA MINING IN SEARCH ARCHIVES (C)

On August 6, 2006, approximately two weeks after its release, AOL took down the search archive (Arrington, 2006).

By then the site had been repeatedly mirrored on the Internet with the records still easily retrievable weeks later. Reportedly, hundreds of Internet users downloaded the approximately 500MB data files and began circulating it on the Web. Some Internet users created simple online interfaces allowing other users to mine the database.

On August 7th, 2006 Andrew Weinstein, an AOL spokesperson, made an official statement regarding the incident (Arrington, 2006):

> *This was a screw up, and we're angry and upset about it. It was an innocent enough attempt to reach out to the academic community with new research tools, but it was obviously not appropriately vetted, and if it had been, it would have been stopped in an instant.*
>
> *Although there was no personally-identifiable data linked to these accounts, we're absolutely not defending this. It was a mistake, and we apologize. We've launched an internal investigation into what happened, and we are taking steps to ensure that this type of thing never happens again.*

The statement was posted on various blogs and emailed to others.

By August 9, 2006 several privacy experts had commented on the issue.

Parry Aftab, executive director of an Internet safety and help group wiredsafety.org, said that if AOL violated its own privacy policy (Jones, 2006):

> *There could be really serious consequences. The lawyers and regulators will be all over this. The FTC has given fines in the millions of dollars for breaching privacy, but the real cost is going to be the brand.*

According to Aftab (Jones, 2006), if AOL violated its privacy policy, AOL could be sued under the consumer fraud laws. It is also possible that the company had violated the Electronic Consumer Privacy Act of 1986 (Jones, 2006).

Among the possibly relevant sections from the AOL privacy policy are the following (AOL, 2006):

> *"We do not keep track of where you go on the World Wide Web".*
>
> *"We do not use any information about where you personally go on AOL or the Web, and we do not give it out to others".*
>
> *"Each and every AOL employee must abide by AOL's privacy policy. Only authorized AOL employees are permitted to have access to your personal information and such access is limited by need".*
>
> *"Employees who violate our privacy policies are subject to disciplinary action, up to and including termination".*

Meanwhile the Electronic Frontier Foundation threatened legal action against AOL (Jones, 2006).

Jonathan Zittrain, an Oxford University professor of Internet governance and regulations and co-founder of Harvard Law School's Berkman Center of Internet and Society, stated that he did not believe AOL had violated its privacy policy (Jones, 2006). Additionally, he believes search data to be a valuable resource for scientists trying to glimpse into the realm of the human mind (Jones, 2006). Zittrain also said that it would be challenging for the users affected by the incident to prove that any harm was done to them (Jones, 2006).

In an interview on August 9, 2006 Andrew Weinstein stated that AOL's research team violated internal policies by making the search log data public (Jones, 2006). He said that AOL had launched an internal investigation in order to understand why this happened and to prevent an incident like that from ever happening again. He also announced that the company would no longer retain linked search terms for over 30 days (Jones, 2006).

IV. DATA MINING SEARCH ARCHIVES (D)

On August 14, 2006 The Electronic Frontier Foundation (EFF) filed a complaint against AOL with the FTC (EFF, 2006):

> *The Electronic Frontier Foundation ("EFF"), having reason to believe that AOL LLC ("AOL") has violated the Federal Trade Commission Act, and that investigation and injunctive relief is in the public interest, alleges that AOL committed unfair and deceptive trade practices by intentionally and publicly disclosing Internet search histories of more than half a million AOL users.*

On August 16, 2006 the World Privacy Forum (WPF) filed a similar complaint.

On August 21, 2006 Jon Miller, chairman and CEO of AOL, announced the resignation of Maureen Govern, AOL's Chief Technology Officer (Karnitschnig and Mangalindan, 2006). The division responsible for the release of search data was reportedly under her supervision (Karnitschnig and Mangalindan, 2006). Jon Miller wrote in an internal memo (SiliconValley.com, 2006):

> *I wanted to let you know that Maureen Govern, our Chief Technology Officer, has decided to leave AOL effective immediately. I want to thank Maureen for her hard work during her time with AOL, and we wish her all the best as she pursues new opportunities.*

In another memo dated that same day, Jon Miller wrote (SiliconValley.com, 2006):

> *…After the great lengths we've taken to build our members' trust and be an industry leader on privacy, it was disheartening to see so much good work destroyed by a single act….This incident took place because some employees did not exercise good judgment or review their proposal with our privacy team. We are taking appropriate action with the employees who were responsible.*

Also that day, AOL fired its Chief Architect for Research, Dr. Abdur Chowdhury. Reportedly, he started the incident by posting the search data online and, shortly after (on August 4, 2006), notifying the research community about the availability of data by posting a message to the SIGIR list (see Exhibit D1) (Hurst, 2006).

An employee who was responsible for supervising Chowdhury's research, and who reported to Maureen Govern, was also fired (Mills and Broache, 2006).

Dr. Abdur Chowdhury was affiliated with the Information Retrieval Laboratory at the Illinois Institute of Technology which receives funding from AOL (Zeller, 2006). Eric Jensen, a researcher at the same laboratory, commented on the firing of Dr. Chowdhury (Zeller, 2006):

> *… A lot of questions out there in academia […] can't get addressed without this kind of data… [AOL's] response has served to inflame the situation rather than address the problem. I think that rather than making scapegoats out of people, you could have explained what this data was for and say, "Yes, there's a privacy discussion to be had there, so let's try to figure it out".*

Dr. Jensen also said that the data was made public after being approved by all of the appropriate executives at AOL, including Maureen Govern (Zeller, 2006). Dr. Chowdhury's lawyer said that Mr. Chowdhury declined to comment on the issue (Zeller, 2006).

While researchers from other universities generally agreed that real world search data is invaluable for research, they hesitated to use the released data (Hafner, 2006). For example, William W. Cohen, an Associate Professor in the Machine Learning Department at Carnegie Mellon University, said in an interview with *The New York Times* (Hafner, 2006):

> *I would feel personally uncomfortable looking too closely at searches showing things like marriages breaking up. I don't want to do research in order to see if my algorithms are working correctly, while delving into the details of people's lives.*

Jon Miller also wrote that in response to the incident AOL was implementing a plan designed to achieve the following objectives (SiliconValley.com, 2006):

1. Creation of a cross-departmental task force for developing best practices for handling web search information and other sensitive data and improving AOL Privacy Policy.

2. Imposition of additional access restrictions on databases containing sensitive data.

3. Evaluation and creation of new tools for removing sensitive information from databases used in research.

4. Privacy awareness programs for employees at al levels.

Jon Miller concluded that from now on AOL would be committed to earning the trust of its customers back "each and every day and with each and every action we take" (SiliconValley.com, 2006).

EXHIBIT D1
Abdur Chowdhury Message Posted to the SIGIR List

(Reprinted from Hurst, 2006)

Sender: Abdur Chowdhury <xxx@xxx.org>
Subject: research.aol.com

AOL is embarking on a new direction for its business - making its content and products freely available to all consumers. To support those goals, AOL is also embracing the vision of an open research community, which is creating opportunities for researchers in academia and industry alike.

We are introducing AOL Research to everyone, with the goal of facilitating closer collaboration between AOL and anyone with a desire to work on interesting problems. To get started, we invite you to visit us at http://research.aol.com, where you will find:

- 20,000 hand labeled, classified queries
- 3.5 million web question/answer queries (who, what, where, when, etc.)
- Query streams for 500,000 users over 3 months (20 million queries)
- Query arrival rates for queuing analysis
- 2 million queries against US Government domains

Also, please feel free to provide feedback on the site, datasets you'd like to see in the future, and any other comments about our vision.

Thanks,
Abdur Chowdhury

ACKNOWLEDGEMENTS

We would like to thank the numerous bloggers (both identified and anonymous) and news reporters for detecting and documenting the incident.

REFERENCES

> *Editor's Note:* The following reference list contains hyperlinks to World Wide Web pages. Readers who have the ability to access the Web directly from their word processor or are reading the paper on the Web, can gain direct access to these linked references. Readers are warned, however, that
> 1. these links existed as of the date of publication but are not guaranteed to be working thereafter.
> 2. the contents of Web pages may change over time. Where version information is provided in the References, different versions may not contain the information or the conclusions referenced.
> 3. the author(s) of the Web pages, not AIS, is (are) responsible for the accuracy of their content.
> 4. the author(s) of this article, not AIS, is (are) responsible for the accuracy of the URL and version information.

AOL (2006) "AOL Privacy Policy," http://legal.web.aol.com/policy/aolpol/privpol.html (Current September 1, 2006)

Arrington, M. (2006) "AOL Proudly Releases Massive Amounts of Private Data," TechCrunch, http://www.techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/ (Current September 1, 2006)

Barbaro, M., T. Zeller, and S. Hansell (2006) "A Face Is Exposed for AOL Searcher No. 4417749," New York Times, http://select.nytimes.com/gst/abstract.html?res=F10612FC345B0C7A8CDDA10894DE404482 (Current September 1, 2006)

EFF (2006), EFF Complaint to FTC, http://www.eff.org/Privacy/AOL /aol_ftc_complaint_final.pdf (Current September 1, 2006)

Frind, M (2006) "AOL Search Data Shows Users Planning to Commit Murder," The Paradigm Shift, http://plentyoffish.wordpress.com/2006/08/07/aol-search-data-shows-users-planning-to-commit-murder/ (Current September 1, 2006)

Jones, K.C (2006b) "AOL's Unprecedented Release," Yahoo News, http://news.yahoo.com/s/cmp/20060811/tc_cmp/191901270 (Current September 1, 2006)

Hafner, K (2006) "Researchers Yearn to Use AOL Logs, but They Hesitate," The New York Times, http://www.nytimes.com/2006/08/23/technology/23search.html?ref=technology (Current September 1, 2006)

Hurst, M (August 2006) Personal Communication

Karnitschnig, M. and M. Mangalindan (2006) "AOL Fires Technology Chief After Web-Search Data Scandal," The Wall Street Journal Online1

[1] The article was subsequently taken off the Wall Street Journal web site, since it stated that AOL's CTO was fired. It became later known that the CTO was not fired; she resigned.

Markey, E. (2006) “Eliminate Warehousing of Consumer Internet Data Act of 2006,” HR 4731 IH, The Library of Congress, http://thomas.loc.gov/cgi-bin/query/z?c109:H.R.4731: (Current September 1, 2006)

McCullagh, D. (2006) “AOL Gaffe Draws Capitol Hill Rebuke,” ZDNet, http://news.zdnet.com/2102-9588_22-6104040.html (Current September 1, 2006).

Mills E. and A. Broache (2006) “Three Workers Depart AOL after Privacy Uproar,” ZDNet News, http://news.zdnet.com/2100-9588_22-6107830.html?tag=nl.e589 (Current September 1, 2006)

Mohammed, A. (2006) “Google Refuses Demand for Search Information,” The Washington Post, http://www.washingtonpost.com/wp-dyn/content/article/2006/01/19/AR2006011903331.html (Current September 1, 2006)

Pass, G., A. Chowdhury, and C. Torgeson (2006) “A Picture of Search,” The First International Conference on Scalable Information Systems, Hong Kong, June, 2006

Perez, J. C. (2006) “Judge: Google Must Hand Over Index Data,” PC World, http://www.pcworld.com/article/id,125133-page,1-c,privacylegislation/article.html (Current September 1, 2006)

Richmond, R (2006) “AOL Releases Web-Search Data Of 650,000 Users,” The Wall Street Journal Online, http://online.wsj.com/article/SB115498329554829021.html?mod=bolcrnews (Current September 1, 2006)

Sadetsky, G. (2006a) “AOL Search Data Mirrors,” http://www.gregsadetsky.com/aol-data/ (Current September 1, 2006)

Sadetsky, G. (2006b) “Original Description of the Dataset,” http://www.gregsadetsky.com/aoldata/U500k_README.txt (Current September 1, 2006)

SiliconValley.com (2006) “Reason for Leaving Last Job: Violated the Privacy of 600,000 Company Customers,” http://blogs.siliconvalley.com/gmsv/2006/08/the_heads_have_.html (Current September 1, 2006)

WPF (2006), WPF Complaints to FTC, http://www.worldprivacyforum.org/pdf/WPF_FTCcomplaint8162006fswp.pdf (Current September 1, 2006)

Zeller, T. (2006) “AOL Moves to Increase Privacy on Search Queries,” New York Times, http://www.nytimes.com/2006/08/22/technology/22aol.html (Current September 1, 2006)

APPENDIX I. POSSIBLE DISCUSSION QUESTIONS

Part A

1. Using a few sentences, describe the person who conducted the search queries listed in Exhibit A2.
2. How might a search company use this kind of data for adding value to advertising services provided to, for instance, a seller of consumer goods or electronics?
3. What do you as a consumer see as the potential advantages to you that could arise from effective use of such data?
4. What cautions would you recommend to anyone attempting to generalize from the search terms provided by a particular user?
5. What value might the U.S. department of Homeland Security see in such data?
6. What concerns do you as a user of this search service have?

Part B

1. Was it an ethical responsibility for the search engines to comply fully with the DOJ subpoena?
2. Was "limited compliance" ethical?
3. Is it ethical for search engines to report only those users who search for "illegal" keywords?
4. Given that search engines agree to report users searching for “illegal” keywords, what actions should the DOJ take against these users?
5. Should the Congress pass EWOCID bill?
6. What actions should the DOJ or local law enforcement take against those users who search for "illegal" key words in the sample revealed by AOL?
7. What action should AOL take on August 6th?

Part C

1. How can the incident potentially impact AOL?
2. How can the incident potentially impact AOL subscribers and Internet users in general?
3. Was AOL's "damage control" effective? What else could AOL do to minimize the damage?
4. What changes should be made to:
    a. AOL's privacy policy?

b. AOL's internal policies?

Part D

1. Were EFF and WPF complaints to the FTC justified?
2. To what extent do you think Dr. Abdur Chowdhury and Maureen Govern were responsible for the incident?
3. Do you think it was the right decision on behalf of AOL to fire the employees responsible for the incident?
4. Should researchers use the released data set for their research?
5. Should AOL continue making the search data available to the public? If so, what should be done to protect consumer privacy while allowing researchers to use the data?
6. What do you think AOL should do to win back the trust of its customers?

LIST OF ACRONYMS

AOL	America Online
CEO	Chief Executive Officer
CTO	Chief Technology Officer
DOJ	Department of Justice
EFF	Electronic Frontier Foundation
EWOCID	Eliminate Warehousing of Consumer Internet Data Act
FTC	Federal Trade Commission
LLC	Limited Liability Company
URL	Uniform Resource Locator
WPF	World Privacy Forum

ABOUT THE AUTHORS

Blake Ives is Past President and Fellow of the Association for Information Systems and Director of the University of Houston’s Information Systems Research Center. He is a past Editor-in-Chief of the Management Information Systems Quarterly and has twice served as Conference Chair for the International Conference on Information Systems. Dr. Ives is also a founder of ISWorld. A frequent contributor to CAIS, he has also published in MIS Quarterly, Information Systems Research, Sloan Management Review, Journal of Management Information Systems, Decision Sciences, Management Science, Communications of the ACM, IBM Systems Journal, and variety of other journals. He has been a visiting Fellow at both Oxford and Harvard. He holds the Charles T. Bauer Chair of Business Leadership at the C.T. Bauer College of Business at the University of Houston.

Vlad Krotov is a doctoral student at the Department of Decision and Information Sciences, Bauer College of Business, University of Houston.

.

Communications of the Association for Information Systems

ISSN: 1529-3181