Digital Footprints and Data Mining

The goals of these activities are developing an understanding of what a digital footprint is and discovering more about your individual digital footprint.

Prerequisite concepts: Internet, Web, IP/DNS, Search Engines, Privacy, Security

Associated reading: *The Tao of Computing, 2nd Ed.* (chapters 10–14)
Nine Algorithms that Changed the Future (chapters 2, 4, 6, 9)

Digital Breadcrumbs — Surfing the Web

Respond in your own words to each of the following.

1. Use your web browser to visit the following site:
   http://www.cs.grinnell.edu/~walker/fluency-book-2/web-info.php

   i. Summarize what this web server knows about your IP address, your browser, your computer, and your browsing history.

   ii. The web page at grinnell.edu uses cookies to track your visits to the site. Describe what the "History" section of that page reports. Then "reload" the page several times in your browser and describe what happens. Look at the HTTP_COOKIE information listed under "More Browser Information" and explain how the values of the two cookies LAST_VISIT and VISIT_NUMBER change when you "reload" the page.

1. Use your web browser to visit the following site:
   http://browserspy.dk

   i. Try out several of the available tests found on the left-hand side of the BrowserSPY home page. Describe the specific tests you invoked and the observed results.

   ii. Summarize what this web server knows about your system, software, and connection.

2. Identify 3 practices that an individual might follow that would help make that person's information more secure on a computer. For each practice, (a) describe the practice, (b) explain how it helps, and (c) identify behaviors or personal traits that might undermine the practice.

3. Identify 3 practices that an individual might follow that would help make that person's internet footprint less informative. For each practice, (a) describe the practice, (b) explain how it helps, and (c) identify behaviors or personal traits that might undermine the practice.

4. Include a brief reflection that addresses your experience with this assignment.

Digital Footprint — Searching the Web

1. Search the web for the phrase "digital footprint"[1] and discuss your findings with other class participants.

2. Try to determine as much as you can about yourself and one additional person (a classmate, someone you know, a well-known individual, etc.) and try to determine as much as you can. Among the information you seek, include at least the following:

- A parent's name
- The person's address
- A photograph of the person's home
- The person's hobbies

2a. Summarize your results

2b. Summarize how they were obtained (describe the techniques used)

2c. Reflect on the amount of information, difficulty of obtaining information, and what you would do to prevent someone from discovering information about yourself.

.....

Related concept: Boolean logic (AND, OR, NOT) applied to “advanced search”

[1] Some URLs with information about digital footprints:
- http://www.kidsmart.org.uk/digitalfootprints/
- http://www.pewinternet.org/Reports/2007/Digital-Footprints.aspx
- http://www.commonsensemedia.org/videos/digital-footprint
- http://book.mydigitalfootprint.com/footprint-cms/MY_DIGITAL_FOOTPRINT.html

Information Sharing — Technologies & Risks

1. Consider a group of people who wish to collaborate on the development of a project and its associated confidential report.

1a. Identify a set of risks associated with sharing project and report files using FTP.

1b. Identify a set of risks associated with sharing project and report files using e-mail.

1c. Identify a set of risks associated with sharing project and report files use a shared file server.

2a. How do you know whether or not the word-processing program you use is making another copy of your document that could be accessed by someone other than yourself and without your permission?

2b. How do you know whether or not new software you purchase contains a virus or other malware? (Note that commercially available products have been contaminated in the past.)

2c. When your computer is connected to the Internet, how do you know whether or not your computer is automatically transmitting information about your files without your explicit request or permission?

2d. When you receive a message over the Internet, how do you know whether or not the message was actually written by the specified sender? When you send a message over the Internet, how do you know if someone other than the specified recipient are accessing the information in the message?

3. Include a brief reflection that addresses your experience with this assignment.

Data Mining

Given that you now have an understanding of digital footprints, the next goal is to understand how a digital footprint can be used to obtain and use additional information.

I. Data Mining

Review the following articles about data mining. Consider how seemingly disconnected pieces of information can be brought together to form a more extensive model of an individual or a population.

Pappalardo, J. "NSA Data Mining: How It Works." *Popular Mechanics*, September 11, 2013.[2]

Arbesman, S. "Five Myths About Big Data." *Washington Post*, August 16, 2013.[3]

Duhigg, C. "How Companies Learn Your Secrets." *New York Times*, February 16, 2012.[4]

Briefly summarize (less than 250 words) the most important ideas or information from these articles.

II. Data and Analysis

Working as a group, reach consensus about the meaning of the following terms and phrases:

- Big Data
- Analytics
- Meta-Data
- Patterns in Noise
- Data Mining
- False Positive

For each of these terms, write a clear definition that is easily understood by someone who is not already familiar with the term and that are capable of standing alone (that is, they do not require someone such as yourself to provide additional explanation or clarification).

[2] www.popularmechanics.com/technology/military/news/nsa-data-mining-how-it-works-15910146

[3] articles.washingtonpost.com/2013-08-16/opinions/41416362_1_big-data-data-crunching-marketing-analytics

[4] www.nytimes.com/2012/02/19/magazine/shopping-habits.html

III. Marketing

Working as a group, consider the following scenario and questions.

You are working on a project for a supermarket and have been asked to develop marketing strategies based on the collection of data from their customers. The store offers a "shopper reward card" which provides discounts on some items when presented by the customer at check-out.

1. What information can you collect from supermarket customers?
2. How will you collect that information?
3. Will the customers know the data is being collected? If so, how will they be informed? If not, why not?
4. What other information can be added with the collected information so that together they provide additional and useful insights?
5. How can you use the data to develop general marketing strategies?
6. How can you use the data to develop a strategy to market the store's brand of dairy products?

As a group, prepare a presentation (no longer than 9 minutes duration) that addresses items 1, 2, 3, 4, and 6.

IV. What is popular?

Many companies are starting to mine data contained in real-time social media communications such as Twitter microblogs. The purpose of this activity is to investigate the information potential of these resources and again to consider what could be learned about people based on their use of social media.

An overview and introduction to these ideas can be found in the *Google Flu Trends* project which provides the latest flu activity based on aggregated Google searches. It also contains articles that provide background and describe how the system works.

In this activity you will work collaborative in a small group to mine a database of social media for information. Using basic search-query tools, the objective is to gain insight into how a large collection of information from a variety of different sources can be used to obtain specific useful information not apparent from any single source.

While there are a variety of publicly available databases that you can use, I recommend using the Twitter Inc. archives[5] or the Topsy Labs, Inc. resource[6].

Address the following questions, providing quantitative validation for your results:

1. What movies are most popular right now?
2. What songs were most popular last month?
3. What TV show was most popular just before the one that is most popular now?

Assume that you do not have any first-hand information or familiarity with current cultural and social phenomena and thus would not know any particular title; thus, for example, searches for “Hunger Games”, “Timber”, “Bieber”, or “NCIS” are not appropriate.

Your group will present the results and a reflection on the search process itself. You should include insight, observations, and advice about coming up with search queries that retrieve the information for which you were looking.

Your group must prepare and deliver a presentation that addresses the results of your research and reflections on the process.

.....

Related concept: Boolean logic (AND, OR, NOT) applied to “advanced search”

---

[5] twitter.com/search-home/

[6] www.topsy.com/

REFERENCES

The Tao of Computing, Second Edition by Henry M. Walker; Chapman and Hall/CRC (2012) ISBN-10: 1439892512 [http://amzn.to/19XwIYO]

From the publisher's website:

Describing both the practical details of interest to students and the high-level concepts and abstractions highlighted by faculty, *The Tao of Computing, Second Edition* presents a comprehensive introduction to computers and computer technology.

It uses a question-and-answer format to provide thoughtful answers to the many practical questions that students have about computing. Among the questions answered, the book explains:

- What capabilities computers have in helping people solve problems and what limitations need to be considered
- Why machines act the way they do
- What is involved in getting computers to interact with networks

The book offers a down-to-earth overview of fundamental computer fluency topics, from the basics of how a computer is organized and an overview of operating systems to a description of how the Internet works. The second edition describes new technological advances including social media applications.

I. Underlying Building-Block Questions
   1. How Are Computers Organized?
   2. How Are Numbers and Characters Represented in a Computer (and Who Cares)?
   3. How Are Images Represented (and Does It Matter)?
   4. Where are Programs and Data Stored?
   5. What is an Operating System and What Does It Do?

II. Software/Problem-Solving Questions
   6. What Can Computers Do for Me?
   7. What Should I Know about the Sizes and Speeds of Computers?
   8. How Are Software Packages Developed?

III. Networking/Distributed System Questions
   9. How Are Computers Connected?
   10. How Do Computers Share Information, So That I Can Exchange Materials with Others Using a Computer Network?
   11. When Can I Consider My Personal Data Secure?

IV. Web/Internet Questions
   12. How Does the Internet Work?
   13. How Do Web Applications Work?
   14. How Private (or Public) Are Web Interactions?

V. Social and Ethical Questions
   15. How Universal Is Access to Computers and the Web?
   16. Can I Use Web-Based Materials in the Same Way as I Use Printed Sources?
   17. Can Computers Think (Now or In the Future)?

Nine Algorithms That Changed the Future: The Ingenious Ideas That Drive Today's Computers by John MacCormick; Princeton University Press (2013) ISBN-10: 0691158193 [http://amzn.to/1bLo6AQ]

From the publisher's website:

Every day, we use our computers to perform remarkable feats. A simple web search picks out a handful of relevant needles from the world's biggest haystack: the billions of pages on the World Wide Web. Uploading a photo to Facebook transmits millions of pieces of information over numerous error-prone network links, yet somehow a perfect copy of the photo arrives intact. Without even knowing it, we use public-key cryptography to transmit secret information like credit card numbers; and we use digital signatures to verify the identity of the websites we visit. How do our computers perform these tasks with such ease?

This is the first book to answer that question in language anyone can understand, revealing the extraordinary ideas that power our PCs, laptops, and smartphones. Using vivid examples, John MacCormick explains the fundamental "tricks" behind nine types of computer algorithms, including artificial intelligence (where we learn about the "nearest neighbor trick" and "twenty questions trick"), Google's famous PageRank algorithm (which uses the "random surfer trick"), data compression, error correction, and much more.

These revolutionary algorithms have changed our world: this book unlocks their secrets, and lays bare the incredible ideas that our computers use every day.

ACKNOWLEDGEMENTS

These activities are adapted from previous works of Henry M. Walker (Grinnell College) and Paul T. Tymann (Rochester Institute of Technology) whose contributions to computer science education are much appreciated and gratefully acknowledged.