Discussion Report: The Database of Intentions

Shivnath Babu
Duke University
shivnath@cs.duke.edu

This report summarizes the class discussion on "The Database of Intentions". The assigned reading for this discussion was the first chapter of the book "The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture", written by John Battelle.

1 The Database of Intentions

In the past few years, search has become a common way in which we interact with the Internet. Every day, many millions of people search the Internet for information about (i) topics they may want to know more about, (ii) items they may want to buy, (iii) competing services they want to choose from, and many other things. We could consider keeping a *database* [2] of all these searches. For example, we could store the data as a conventional row-and-column *table* of records. Let us call the table "SearchData".

Each record in the SearchData table will have three columns, or *attributes*:

1. *ID*: The ID attribute will store some information about who did the search. For example, the ID attribute may store the *cookie* [1] that comes with the search request. Another option may be to store the *IP address* [4] of the computer from which the search request originated.

2. *Search keywords:* This attribute stores the keywords used in the search. For example, a search for "Duke University" contains two keywords: "Duke" and "University".

3. *Timestamp:* This attribute stores the day and time when the search was done.

There is a gold mine of information in this data. For example:

- This data reflects the current thoughts, desires, fears, and intentions of people. For example, we can aggregate the data in the SearchData table to find how many people searched last month for "Iraq War".

- This data can be used to determine trends over time. For example, we can compare the number of people who searched for "Iraq War" in December 2006 Vs. the number of people who searched for "Iraq War" in January 2007. If the number for December 2006 is significantly lower than the number for January 2007, then it could be that general interest in the Iraq War is increasing.

- This data can be used to determine geographical trends provided we can find the location (e.g., the city) from which the search was done; see [5]. Such an analysis can enable us to find trends like: "People in Durham, NC, are searching more for basketball-related topics than people in Palo-Alto, CA."

The search data and the mine of information that it encapsulates is collectively termed the "Database of Intentions". Note that Google is not the only organization that has access to such a database of intentions. Other search engines—e.g., A9, MSN, Yahoo!—can collect search data as well. However, the popularity of the search engine will determine how representative the search data is about the interests and intentions of all people. For example, a search engine that is used by geeks only will not have data that represents the interests and intentions of non-geek communities.

2 Using and Abusing the Database of Intentions

This section outlines some of the many uses of the Database of Intentions. Unfortunately, it is hard sometimes to draw the line between use and abuse in this setting.

The Database of Intentions can be used to find find trends about shopping. Businesses may be able to use this information to stock stores with the items that customers are most interested in; which improves business profits as well as increases customer satisfaction. We give a hypothetical example. Suppose the data collected by Google shows that many people are searching for "Battelle" and "Book" together. Bookstores like Borders and Barnes and Noble will be very interested in this information. (They may have to buy this information from Google.) Since an increased interest in Battelle's book will likely lead to more purchases of this book, the bookstores would want to ensure that enough copies of the book are available on their shelves.

The above example shows that businesses can find effective ways to profit from the Database of Intentions. As described in the assigned reading: "Search drives clickstreams and clickstreams drive profits". While we can argue that the use of the Database of Intentions in the above example benefits the customers—since they can get the book from the store immediately, instead of having to wait while the store orders the book—this may not always be the case as illustrated in the next example.

Consider a student studying at an U.S. university who happens to be a citizen of some country in the Middle-East. As part of an essay for his class project, the student may need to find information about chemicals that are used to make explosives. Consequently, he may search on Google to locate such information on the Internet. If Google's search logs are made available to U.S. government agencies, it can very well happen that authorities get suspicious about the student. (The assigned reading mentions the USA PATRIOT Act in this setting.)

Note that the previous example is not really an example about the Database of Intentions. This example assumes that the U.S. government is able to get personally-identifiable information about the person doing the search. However, we can come up with other examples where the behavior of a community of users (e.g., people in Durham, NC) can be extracted from the Database of Intentions, and used to discriminate against them. Such instances raise concerns about using the Database of Intentions to attain ill-gotten ends, and raises questions about privacy, security, rights, and ownership. Currently, customers have no option other than trusting the entity that collects the data (e.g., Google) to refrain from unlawful usage of the data; particularly, to never sell personally-identifiable information to others. (In this setting, the assigned reading mentions the Electronic Communications Privacy Act.)

To summarize this section: Battelle raises some interesting questions about (i) how the Database of Intentions can be used/abused and (ii) who/what safeguards the interests of web searchers in this setting. These topics will be discussed in more detail later during this semester.

3 Technological and Social Advances

Recent technological advances have contributed to making the Database of Intentions a reality:

- The cost of hard disks [3] for storing large amounts of data has dropped significantly over the last few years. This drop in price makes it possible to store the massive and rapidly-growing data repository of web-search data.
- Hard disks have become faster at writing (or saving) data.
- It has become possible to use a large cluster of computers to do tasks in parallel. Such parallelism enables us to store and process data faster than if we could devote only one computer to the task.

In addition to technological advances, social changes have also contributed to the use and impact of the Database of Intentions. A larger and more diverse community of people use web search today than, say, 10 years ago. As we discussed earlier, web-search data may not be useful (or reliable) if the data represents the interests and intentions of few users only.

4 Limitations of Web-Search Today

While search engines have made significant progress in terms of performance and usability over the last decade, they are still far from perfect. (Udi Manber, ex-CEO of the A9 search engine, claims that the search problem is only 5% solved.) The primary complaint about search engines is that they provide a very restricted interface that a user can use to express her queries. (We will discuss this topic later this semester, where we will look at other query interfaces, e.g., natural-language-based interfaces where users can enter queries in English.)

Furthermore, human-level intelligence may be needed to answer some queries. Given the huge amount of data on the Internet today and the large number of search queries asked per day, it is not possible to build a search engine where humans answer all queries. The Computer-Science research community has made significant progress towards creating artificial intelligence, which could be applied to make search engines more user-friendly. (The assigned reading mentions two interesting entities that are relevant here: the Turing test and the Cyc database of commonsense rules.)

References

[1] *Wikipedia entry for HTTP Cookie*. http://en.wikipedia.org/wiki/HTTP_cookie.

[2] *Wikipedia entry for Database*. http://en.wikipedia.org/wiki/Database.

[3] *Wikipedia entry for Hard Disk*. http://en.wikipedia.org/wiki/Hard_disk.

[4] *Wikipedia entry for IP address*. http://en.wikipedia.org/wiki/IP_address.

[5] *Finding the approximate geographic location of an IP address.* Use http://whatismyipaddress.com to get your IP address. Then, enter this IP address into the search box at http://www.ip2location.com/free.asp to find the geographic location of the address.