Dan Cohen’s Digital Humanities Blog » Blog Archive » The Wikipedia Story That’s Being Missed

With all of the hoopla over Wikipedia in the recent weeks (covered in two prior posts), most of the mainstream as well as tech media coverage has focused on the openness of the democratic online encyclopedia. Depending on where you stand, this openness creates either a Wild West of publishing, where anything goes and facts are always changeable, or an innovative mode of mostly anonymous collaboration that has managed to construct in just a few years an information resource that is enormous, often surprisingly good, and frequently referenced. But I believe there is another story about Wikipedia that is being missed, a story unrelated to its (perhaps dubious) openness. This story is about Wikipedia being free, in the sense of the open source movement—the fact that anyone can download the entirety of Wikipedia and use it and manipulate it as they wish. And this more hidden story begins when you ask, Why would Google and Yahoo be so interested in supporting Wikipedia?

This year Google and Yahoo pledged to give various freebies to Wikipedia, such as server space and bandwidth (the latter can be the most crippling expense for large, highly trafficked sites with few sources of income). To be sure, both of these behemoth tech companies are filled with geeks who appreciate the anti-authoritarian nature of the Wikipedia project, and probably a significant portion of the urge to support Wikipedia comes from these common sentiments. Of course, it doesn’t hurt that Google and Yahoo buy their bandwidth in bulk and probably have some extra lying around, so to speak.

But Google and Yahoo, as companies at the forefront of search and data-mining technologies and business models, undoubtedly get an enormous benefit from an information resource that is not only open and editable but also free—not just free as in beer but free as in speech. First of all, affiliate companies that Yahoo and Google use to respond to queries, such as Answers.com, primarily use Wikipedia as their main source, benefiting greatly from being able to repackage Wikipedia content (free speech) and from using it without paying (free beer). And Google has recently introduced an automated question-answering service that I suspect will use Wikipedia as one of its resources (if it doesn't already).

But in the longer term, I think that Google and Yahoo have additional reasons for supporting Wikipedia that have more to do with the methodologies behind complex search and data-mining algorithms, algorithms that need full, free access to fairly reliable (though not necessarily perfect) encyclopedia entries.

Let me provide a brief example that I hope will show the value of having such a free resource when you are trying to scan, sort, and mine enormous corpora of text. Let's say you have a billion unstructured, untagged, unsorted documents related to the American presidency in the last twenty years. How would you differentiate between documents that were about George H. W. Bush (Sr.) and George W. Bush (Jr.)? This is a tough information retrieval problem because both presidents are often referred to as just "George Bush" or "Bush." Using data-mining algorithms such as Yahoo's remarkable Term Extraction service[1], you could pull out of the Wikipedia entries for the two Bushes the most common words and phrases that were likely to show up in documents about each (e.g., "Berlin Wall" and "Barbara" vs. "September 11″ and "Laura"). You would still run into some disambiguation problems ("Saddam Hussein," "Iraq," "Dick Cheney" would show up a lot for both), but this method is actually quite a powerful start to document categorization.

I'm sure Google and Yahoo are doing much more complex processes with the tens of gigabtyes of text on Wikipedia than this, but it's clear from my own work on H-Bot[2] (which uses its own cache of Wikipedia) that having a constantly updated, easily manipulated encyclopedia-like resource is of tremendous value, not just to the millions of people who access Wikipedia every day, but to the search companies that often send traffic in Wikipedia’s direction.

Update [31 Jan 2006]: I’ve run some tests on the data mining example given here in a new post. See Wikipedia vs. Encyclopaedia Britannica for Digital Research[3].

This entry was posted on Tuesday, December 20th, 2005 at 2:08 pm and is filed under Google[4], Search[5], Text Mining[6], Wikis[7], Yahoo[8]. You can follow any responses to this entry through the RSS 2.0[9] feed. You can leave a response[10], or trackback[11] from your own site.

References

1. ^ Term Extraction service (developer.yahoo.net)
2. ^ H-Bot (chnm.gmu.edu)
3. ^ Wikipedia vs. Encyclopaedia Britannica for Digital Research (www.dancohen.org)
4. ^ View all posts in Google (www.dancohen.org)
5. ^ View all posts in Search (www.dancohen.org)
6. ^ View all posts in Text Mining (www.dancohen.org)
7. ^ View all posts in Wikis (www.dancohen.org)
8. ^ View all posts in Yahoo (www.dancohen.org)
9. ^ RSS 2.0 (www.dancohen.org)
10. ^ leave a response (www.dancohen.org)
11. ^ trackback (www.dancohen.org)

Excerpted from *Dan Cohen’s Digital Humanities Blog » Blog Archive » The Wikipedia Story That’s Being Missed*

http://www.dancohen.org/2005/12/20/the-wikipedia-story-thats-being-missed/