Chris Krause

Assignment 3 – Wikipedia

Introduction

Wikipedia is a collaborative web-based encyclopedia which markets itself as being free to edit, reproduce and view. All content, ranging from the controlled vocabulary which founds it to the plain text and images of the articles themselves are all submitted and engineering by volunteer contributors. Wikipedia is the computer age's most daring and bold project: it is an attempt at making all knowledge free and all claims corroborated. While claims in print are static and unchanging, only revised by subsequent printings or not at all, Wikipedia offers dynamic articles that are continually updated to improve the breadth of knowledge, understanding and the impacts and significance of current events. In this way Wikipedia is leading to the day in which books will become primary sources rather than sources of reliable knowledge, as historical artifacts rather than as sources of firm reference. As Wikipedia is made freely available, free of corporate patrons to appease, it has become a neutral ground to disseminate information to all users of the internet and beyond. 'Beyond' in the sense that Wikibooks and other subprojects are aiming to provide quality open source textbooks to those who are not privy to the costly antecedent alternatives.

Wikipedia poses a significant threat to traditional librarianship: in the world of Wiki all users are potential librarians, and while many lack the technical expertise of a master of library and information science, Wikipedia contributors have become expert

archivists, catalogers and information retrieval specialists. The fundamental difference between the old guard and Wiki is that the latter does not have a top down hierarchy but a horizontal hierarchy: collaboration and deliberation is used to produce and maintain content, to make executive decisions, in contrast to the professionalism of traditional librarianship. Herein is the pinnacle of web 2.0 genius making a play of classical technology: Wikipedia is the next generation library, putting everything we know about the discipline into question, including the necessity and utility of our jobs. If the majority of users are skilled at information retrieval then of what use is our kind? One could make the argument that we will always serve as wise guardians and gateways to knowledge in the same sense Vergil was to Dante, informing the tools needed to safely and efficiently access and retrieve information. But what will conclude when the day comes in which the human race no longer needs our aid in learning of information systems, when these complex rhythms become basic facets of the rearing process and information retrieval becomes a rudimentary function of human experience? The digital divide is slowly eroding and so surely this latter time will come. Wikipedia is the cause of the appropriate consternation the reader may be feeling.

Yet Wikipedia is not without flaw. A casual user has no means of determining if an article has been vandalized, if critical information has been removed or if claims the article is making are corroborated by sound evidence.[1] Vandalism is possible, although usually reverted by vigilant administrators and automated bots within minutes.[2] As edits are potentially anonymous, Wikipedia can be an unreliable tool for information retrieval for politically charged or controversial topics, as slander and trolling is a concern.[3] Other

critics argue that Wikipedia's claims are substantiated by consensus rather than by the rigor of evidence.[4] Yet the majority of these concerns and criticisms are fallacious. Articles which are repeatedly vandalized or covering topics of a charged political nature are locked and carefully scrutinized by dozens of experts.[5] All modifications to Wikipedia articles are listed in a central location which is patrolled by innumerable volunteer administrators on a perpetual basis. Creating nonsense articles, false claims and simply vandalizing content is increasingly difficult as Wikipedia increases in popularity and appeal. Claims without citations are frequently marked up with a "citation needed" tag, and citations must meet an extensive and objective notability and credibility criteria to be included in articles.[6] In 2007 "Wikiscanner" was established to detect corporate infiltration of Wiki content. Policies and procedures were accordingly established to prevent skewed edits and corporate agendas from mangling the neutrality of the encyclopedia.[7]

While it is difficult to separate user content generation from user access in analyzing Wikipedia, for purposes of this paper we will focus on the latter.

User Model

Wikipedia is a web-based hypertext portal which can be accessed with a basic internet connection. The website is slickly designed to web 2.0 design standards, utilizing cascading style sheets and XHTML. The design is as follows: an initial portal (Wikipedia.org) presents the user with a list of languages, with dozens of hypertext links to different language versions of the site. These different parts of Wikipedia are

not translations or mirrors of the English Wikipedia, but feature unique content created in foreign languages. The English, German, French, Italian, Portuguese, Japanese, Spanish, Polish, Russian and Dutch languages represent the biggest presence on Wikipedia (500,000+ articles each), but dozens of other languages also have their own section of the website. After the user selects a language he or she is presented with a welcome splash page including various sections designed to draw the user into exploratory and "edutainment" browsing: featured article, "did you know," "in the news," "on this day," featured picture and many links in the lower section of the page which link to Wikipedia subprojects, services and affiliates.

A search bar is featured on the left margin. This is one means by which the user can access the massive store of content not listed on the front page. Another means is through the listing of content portals, a sort of extensive table of contents. This list is lengthy and displays the most significant dedicated categories of Wikipedia's content, which accordingly are linked to lead to dependant sub-portals. These latter portals are typically imitations of the front page, with their own "did you know" sections catered to content relevant to the subject matter. Yet this is not uniformly the case: subject sub-portals are similar to academic department websites: they are personalized to fit the spirit of the subject, and have a more personal feel than the sterile initial pages one must navigate through to get to the heart of the content. Many such pages are detailed with artwork, original designs, quotes and creative, amusing sections.

One innovative and original feature of Wikipedia is inline hypertext linking. Every word, person, place or topic which an editor believes links to a unique article or piece of

knowledge is linked within the text itself. This is a tremendous aid to findability, usability and exploratory research, as it allows seamless transitioning to relevant articles on demand, which are in turn also formatted to this standard.

Wikipedia's record structure is complex and difficult to describe. The most basic cataloging unit is called a "category." Categories are simply collections of articles organized by topic, which can then be further broken down into subcategories. Both categories and subcategories can be formatted for presentation in a myriad of ways: alphabetically, chronologically, numerically, randomly, or by custom lists tailored to the subject matter. A very anarchic system, Wikipedia's categories are determined by the editors, and record structure can change seamlessly and dynamically based upon summary edits or through discussion. The category structure is listed at the bottom of any Wikipedia article, and various subcategories can be accessed via hypertext links in order to list all the sorted articles under that particular subcategory alone. How complex the record structure is for a given group of articles is determined by the editors: some articles are sorted at the end of ten or twenty subcategories, while other articles are only vested under one or two subcategories. It is possible to directly retrieve an article through search or by clicking on a hyperlink rather than navigating the various categories and portals. Academic precedents are typically utilized to form as a grammar of cataloging, especially in regard to scientific and medical articles. In this sense records are not cataloged randomly on Wikipedia, even though the system technically allows for disorder.

Wikipedia's audience is general population adults, although a notable minority of content, especially some more obscure medical, mathematical and science articles tend to be written in academic or highly technical language. Wikipedia can be used as a serious research tool for college or post-graduate work if used very carefully: special attention must be paid to the citations to ensure the article was informed by rigorous evidence. Often the sources the Wikipedia editors used can be located in Google Scholar or in Wikisource (a repository of public domain primary and secondary documents) and then considered for further independent research. Aside from the general population and researchers Wikipedia also has a sizable user population of experts who visit the website to contribute to it. The only population Wikipedia would be really inappropriate for is young children, as the language of most articles is written at the high school level or above. That being said, Wikijunior, a subdivision of Wikibooks, is intended for providing free textbooks to children of 8-12 years.

Criteria

Several criteria will be selected for evaluating the efficiency of Wikipedia's access interface and searching capability. These capabilities are chosen on the basis of ideal qualities of an information retrieval system and are not specifically suited for evaluating Wikipedia. An effective information system must be fast and efficient, must return the content the user is searching for (we will refer to this as "findability"), must be able to interpret the intent behind a poor query in order to redirect the user to relevant content (heuristics) and must offer some aspect of personalization and feedback. These criteria speak to the heart of web 2.0 navigation, experience and accessibility and are what we

have come to expect of the digital services we now take for granted ranging from Google to Twitter. Without those capabilities, an information retrieval system has very limited usefulness and remains a tool to accomplish a focused task. A system designed around the enlightened principles of web 2.0 technologies is not only capable of finishing a task, but also serves as an invaluable creative boon to achieve excellence in research and education.

While some of those capabilities can be evaluated using empirical means of measurement, i.e. speed and efficiency can be calculated in seconds; the other categories will require a thorough observation and reasoned judgment. In this sense they will either pass or fail the quality standard. Anecdotes and observations are applicable evidence in this regard, as the evaluation portion of this paper will function as a review.

Before we proceed to that avenue I would like first to comment aside on the subject and importance of heuristics. I believe that the ability of search software to interpret the verbose and often scattered reasoning of a searcher and direct him or her to quality content is one of the principle qualities which make it stand apart from alternatives. During a thick bout of research and information filtering, the expedience of the search may very well influence the quality of the work being done. A frustrating battle with an ineffectual system to come up with any relevant content defeats the research process itself. Ideal systems do not require the complex search patterns which dominate traditional software and academic databases, but instead, function as naturally as conversation with a reference librarian in asking for help, being directed to

what we are looking for. When the user conducts original research, consisting of thousands of searches, he or she is averse to battle through every search to find what they are looking for. The search software must have the heuristic ability to interpret the researcher's intent and to return content which approximates that interest. From there the researcher, guided by the creative element of the work, has the ability to further refine the search; the user should not be forced to refine a search before it even begins.

Search Examples

1. Objective: Find an article on blood diseases, without knowing what this area of medical study is called.

Action: Search query “blood diseases” returned Hematology

2. Objective: Find a listing of the various divisions participating in the German expeditionary force to North Africa (‘Afrika Korps’) during World War II

Action: Search query “Afrika Korps” leads me to a search results page instead of bringing me to the Afrika Korps article of the same title, strange. Section on “composition and terminology” has what I was looking for.

3. Objective: Find out if it’s true that imbibing pop rocks and carbonated beverages can cause children to explode.

Action: Search query “pop rocks and coke” returned search results with articles of carbonation, myth busters and punk rock in the top 3. “Pop rocks” is mentioned in both

carbonation and myth busters, I choose carbonation. I realize that there is nothing on the page about this phenomenon so I check the page thoroughly and notice a hypertext link to “pop rocks” in the “see also” section. Therein is a section on “urban legend” which puts to rest my ignorance.

4. Objective: Find information on a band I heard about called “Buhrzum”

Action: Search query “Buhrzum” returns a “did you mean” for “Burzum.” I click and am re-directed to the band in question.

5. Objective: Find out the weight of various roman coins.

Action: Search query “weight roman coins” returns various articles featuring the words “roman” and “coin.” None of the articles succinctly or comprehensively answer my question.

6. Objective: Learn about astrophysics.

Action: Clicked the science portal link. An initial overview of the portal yields little results, and then I notice a tab called “categories and main topics” which yields multiple hypertext links to astronomy topics.

Comment: Additional searches of this sort could be performed, but they are typically slower than searching the database directly with a query. Trouble might occur when attempting to find very specific information if browsing through the portal table of contents; this latter medium is effective for exploratory browsing rather than focused information retrieval.

Evaluation

Wikipedia's speed and efficiency cannot be questioned; it is truly a marvel of web 2.0 engineering. For most articles the plaintext content is returned in less than a second, and even the largest articles have all of their images loaded and cached in less than 10. Of course these access times are relative and in no way authoritative: speed is influenced not only by the current load of the Wikipedia servers but also by the functioning of name servers, local ISPs and infrastructure, all the way down to computer issues and spyware. We can however conclude generally that Wikipedia's speed and efficiency is remarkable considering the site traffic: according to Alexa as of November 28th 2009 the average load time for Wikipedia.org is 2.079 seconds, faster than 57% of the other websites on the internet. At the same time Wikipedia is the #6 most viewed website on the internet, occasionally shooting to the top 3 when major current events are developing. The website draws between .5 and .6% of all traffic on the internet. This translates to trillions of pageviews a day, although exacting figures on traffic does not appear to be available. Wikipedia has also made a shift in the past few years to SVG file format for images. These images are scalable vector graphics which are rendered from lines of XML via the web client and use up drastically less memory than raw data heavy bitmap and JPEG images. This is one way, among many, that Wikipedia has streamlined the efficiency and speed of their website while simultaneously adopting avant-garde web 2.0 technologies, reducing overall bandwidth consumption.[8]

In the early days of Wikipedia content was different to locate and access and when querying the database using the search function, would only return articles with identical spelling and grammar. Over the years the search has optimized, but it is still less effective than other search engines at locating what the user is attempting to find. While you are able to enter in complex and verbose questions and sentences into Google and it is fairly effective at bringing you to the most relevant content, no matter how fuzzy the logic of the query, Wikipedia's search uses basic inline text searching and keyword saturation analysis to return records. If you type in a search query and it does not match the title of an article, Wikipedia will quickly return highlighted, truncated instances of those words being used in other articles. This is a surprisingly primitive plaintext search, and is perhaps only advantageous in its almost immediate speed.

Over the past year Wikipedia has taken serious measures to improve the search, taking several cues from the model search engine: Google. It is now possible to search within different forms of data: content pages, multimedia and help articles. Wikipedia has also instituted a "did you mean" functionality identical to Google. This function allows Wikipedia to heuristically estimate the intent behind a poor query and to offer suggestions to further refine the search. This recent addition will guarantee the long term viability and usability of Wikipedia, as the previous incarnation of the search function was one of the website's most distinct flaws. While this ability is not quite as sophisticated or accurate as Google's equivalent, it does perform basic heuristic calculations which should enable users with computer literacy issues to better access and utilize the website's resources.

One simple way in which Wikipedia has attempted to improve findability is by implementing editor created redirection vocabularies. Blank pages are created that automatically redirect a user to a more relevant article when a query is entered into the search that is misspelled or otherwise requires disambiguation. Example: if you type “blood disease” into the search bar it will redirect you to Hematology, an article on the medicine of diseases of the blood. This is what a good heuristic engine should do, although in this case it is only the illusion of true heuristics as the redirections were manually plotted by concerned editors. Regardless, most major articles have dozens of redirects, greatly expanding their visibility and steering users away from flimflam and less enlightening alternatives.

Finally we will consider the capability of Wikipedia to offer personalized feedback and a sense of greater interactivity to the user. Wikipedia was founded on content collaboration: every article is created automatically with an attached discussion page, in which modifications to the article can be reasoned and considered. In recent years discussion pages have also come to include relevant discussion of interest rather than strictly deliberations on proposed article changes. This trend is in tune with the introduction of a user account system. Originally created as a means of identifying editors and reducing vandalism, the system is still in beta and offers some basic customizations to the access experience. It is possible to “subscribe” to articles in order to monitor updates, track your own editorial history, modify the appearance of the website interface to utilize several skins and format text to preference, message other users and create a basic profile.

While these measures are positive steps toward web 2.0 standards of interactivity, the experience still falls short of the personalized feedback scripts of Amazon or the social experience of Twitter. While Wikipedia may be hesitant from adopting similar systems so as to preserve its appearance as a respectable encyclopedia and not a social hub, those features are productive in enhancing the information retrieval experience. Amazon's recommendation systems are effective at leading users toward other relevant products, and in the same way Wikipedia could use similar means for steering users toward other relevant articles and topics of research. The trick is to balance the presentation so that it does not become invasive or forced, but if it was carefully implemented, it would act as a service rather than a disservice to the users of Wikipedia.

---

[1] Robert McHenry, Encyclopedia Britannica editor-in-chief referred to it as a "faith-based encyclopedia." Caslon Analytics. http://www.caslon.com.au/wikiprofile1.htm

[2] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. "Studying Cooperation and Conflict between Authors with History Flow Visualizations". Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI) (Vienna, Austria: ACM SIGCHI). 2007. 575–582.

[3] Torsten Kleinz.. "World of Knowledge." The Wikipedia Project (Linux Magazine). 2005.

[4] Simon Waldman. "Who knows?". Guardian.co.uk. 2004. http://www.guardian.co.uk/technology/2004/oct/26/g2.onlinesupplement.

[5] Wikipedia's semi-protection policy: http://en.wikipedia.org/wiki/Protection_policy#Semi-protection

[6] Reliable sources/citations requirements: http://en.wikipedia.org/wiki/Wikipedia:Reliable_sources

[7] Katie Hafner. "Seeing Corporate Fingerprints From the Editing of Wikipedia". New York Times. 2007.

[8] Chris Lilley. *Scalable Vector Graphics.* http://www.w3.org/Graphics/SVG/