Visualizing the Overlap between the 100 Most Visited Pages on Wikipedia for September 2006 to January 2007

Anselm Spoerri
SCILS, Rutgers, The State University of New Jersey,
aspoerri@scils.rutgers.edu

Abstract

This paper compares the monthly lists of the 100 most visited Wikipedia pages for the period of September 2006 to January 2007. searchCrystal is used to visualize the overlap between the five monthly Top 100 lists to show which pages are highly visited in all five months; which pages in four of the five months and so on. It is shown that almost 40 percent of a month’s top 100 pages are visited in all five months, whereas 25 percent are highly visited only in a single month. The presented visualizations make it possible to gain quick insights into the overlap and topical relationships between the monthly lists.

Introduction

Wikipedia is one of the most visited web sites and a major online information resource used by a diverse set of users. According to comScore’s web traffic rankings for January 2007, Wikipedia has become the ninth most visited site in the United States with 43 million unique visitors (comScore, 2007). Many popular web sites, such as YouTube\(^1\), Digg\(^2\) or Del.icio.us\(^3\), offer a “most popular” feature that lets you browse the most viewed videos, the news stories with the most votes or the web page bookmarked by the largest number of users. This paper
investigates how the 100 most viewed and thus "most popular" pages on the English version of Wikipedia stay the same or change over the five-month period of September 2006 to January 2007. This paper aims to show how information visualization can be used to gain quick insights into the overlap structure and topical relationships between the monthly most visited Wikipedia pages. For example, the visualizations will help you see that a much smaller percentage of the popular Wikipedia pages is related to typical encyclopedic topics, such as geography, history or politics, than you would expect.

**Method**

The tool “WikiCharts” is used to identify the 100 most visited pages on the English version of Wikipedia in a given month (Weber, 2006). The visualization tool “searchCrystal” is used to compare and visualize the overlap between the “Top 100” lists for the months of September 2006 to January 2007 (Spoerri, 2004a).

**WikiCharts**

Leon Weber developed WikiCharts by writing a javascript program that relays the name of the viewed Wikipedia page to the Wikipedia toolserver. The script is executed with a probability of less than one to avoid overloading the toolserver. The data is collected in a log file without recording the IP address of the computer requesting the page. The logged requests are added into a MySQL database every 15 minutes. This database can be accessed via a PHP script, where you can specific the month and the number of most visited pages to display. WikiCharts became available and started collecting data in August 2006. In this paper, the monthly lists of the 100 most visited pages for the months September 2006 to January 2007 are used and analyzed.
searchCrystal

The five monthly “Top 100” lists are compared in searchCrystal to visualize the degree of overlap between them. Similar to a bullseye display, searchCrystal shows the pages contained in all the lists in the center of the display. The pages listed in four out of the five lists are shown further away from the center so that the number of lists that contain the same page decreases toward the periphery of the display.

searchCrystal is an information visualization toolset that consists of several complementary views: the Category, Cluster, Spiral and List View (Spoerri, 2004a; 2004b; 2004c). Each view helps you explore specific aspects of the overlap structure between the lists or sets of data being compared.

The Category View groups all the pages that are contained in the same combination of lists. It shows how many pages are included in which specific combinations of lists (see Figure 1). At its periphery, star-shaped icons with a single color represent the specific lists being compared. Each list is assigned a unique color and the number inside a star-shaped icon indicates the number pages in the list. The interior of the Category View consists of circular icons whose colored sectors indicate which specific lists contain the same page. The size of a circular icon indicates how many pages are contained in a specific combination of lists. At the edge of a circular icon, the two pages with the highest list positions are also shown. Small icons represent these pages; the shape reflects the number of lists that contain the page and the colors indicate which lists. The page title is displayed, but it can be truncated to prevent titles from overlapping.
The Cluster View shows how the individual pages are related to the lists being compared. In this view, the star-shaped icons at the periphery act like “magnets” that pull a page icon toward them based on the page’s list positions (Spoerri, 2004a). Thus, the position of a page icon reflects the relative difference between the page’s positions in the lists that include it. Further, pages are mapped into the same circular ring if they are contained in the same number of lists. The closer a page is placed toward the display center, the higher the average of its list positions (see Figure 2).

A page icon has multiple visual properties to help you determine how many and which specific lists contain the page and the page’s average position in the lists. The shape of a page icon indicates the number of lists that contain the page and the colors indicate which lists. The size of a page icon reflects the average position of the page in the lists that contain it. The greater the size and the stronger the color saturation of a page icon, the higher up it is placed in the lists. Thus, both the position of a page icon inside its designated ring and its size and color saturation indicate whether a page is highly placed in the lists that include it. The orientation of a page icon is such that its colored sides face the lists it is related to. The page title is displayed next to a page icon, but it can be truncated to prevent titles from overlapping.

The Spiral View places all pages sequentially along an expanding spiral (see Figure 3). As for the Cluster View, pages that are included in all five lists are located in the center ring. The icons for pages that are contained in the same number of lists are placed consecutively along the spiral and in the same concentric ring. Title fragments are displayed in the radial direction to make effective use of the white
space in the spiral layout (Spoerri, 2004b). The Spiral View, which can be rotated, makes it possible for you to rapidly scan a large number of pages and their titles.

**Results**

Using the Category View, Figure 1 provides a compact overview of the overlap between the monthly lists of the 100 most visited English pages on Wikipedia for the months September 2006 to January 2007. The number and size of the circular icon in the very display center indicate that 39 out “Top 100” pages are contained in all five lists. The two pages with the highest positions in all five lists are the “Main_Page” and the page about “Wikipedia”. The former page is the most visited page in each of the five months studied.

<INSERT Figure 1>

Examining the circular icons, which are located farthest away from the center, you can observe that on average 25 of the “Top 100” pages in a month are highly visited only in a single month (see Figure 1). The titles of the two small page icons, which are attached to the circular icons with a single color, provide a quick insight into the dominant news or social events that are specific to a given month. In September 2006, the death of Steve Irwin by a stingray captured the public’s attention. Halloween and North Korea are the top pages that are highly visited only in October 2006. Sacha Baron Cohen, the creator of the film “Borat”, which was released in November 2006, and Thanksgiving are two top pages that are highly visited only in November. In December 2006, Google honored Edvard Munch with a Scream Logo and the anniversary of the attack on Pearl Harbor occurred. In January 2007, former President Gerald Ford was buried and the “Gerald Ford” and
"Deaths in 2007" pages were the two top pages that were highly visited only in that month.

The circular icons that represent pages that were highly visited in several months make explicit the events, personalities or topics that held the public’s attention over an extended period. For example, the film “Casino Royale” was released in November 2006 and generated great interest in “James Bond” in the months of November, December and January. Further, the film “Borat” and Albert Einstein - what a contrast - are the two top pages that are highly visited in the fourth quarter of 2006.

<INSERT Figure 2>

Using the Cluster View, Figure 2 shows how the individual Wikipedia pages are related to the five monthly “Top 100” lists. The center ring of the Cluster View has been magnified so that only the pages that are included in all five lists are visible. The other rings have been reduced in size and moved toward the periphery of the display. You can select and drag the circular rings in the Cluster and Spiral Views to dynamically increase specific parts of the display to be able to examine specific groups of pages in more detail.

The ten most visited Wikipedia pages in all five months cluster toward to center of the Cluster View, which implies that they are located toward the very top of all five lists. They are: 1) “Main page”, 2) “Wikipedia”, 3) “Wiki”, 4) “United States”, 5) “WII”, 6) “World War II”, 7) “Sex”, 8) “Naruto”, 9) “List of sex positions”, 10) “PlayStation 3”.

The page about the “September 11, 2001 attacks” is also contained in all five lists. This page is located close to the center of the Cluster View, almost equidistant from the star-shaped icons that represent
"09/2006", "10/2006" and "11/2006" and further away from "12/2006" and "01/2007" (see Figure 2). This suggests that the “September 11, 2001 attacks” page has higher list positions in September, October and November than in December and January. If you want to know the specific list positions, then you place the cursor over a page icon and a “Details-on-Demand” display appears. In Figure 2, the “Details-on-Demand” display for the “September 11, 2001 attacks” page shows that the position of this page steadily decreases as time progresses.

<INSERT Figure 3>

Using the Spiral view, Figure 3 shows in a single display all the 230 unique pages that are contained in the five “Top 100” lists that are being compared. As in the Cluster View, you can click & drag the circular rings to magnify specific areas of the Spiral View. In Figure 3, the different rings have the same width. Due to the large number of pages contained in all five lists, the page icons are tightly packed in the center ring and only the titles of pages further away from the center can be fully displayed. The titles for all the page icons in the subsequent rings can be shown. This makes it possible for you to rapidly scan a large number of pages and their titles, especially since the spiral can be rotated. A standard list display can only show a limited number of pages and you have to scroll extensively to explore all the pages. The Spiral View enables you to get a quick insight into the major topics that are covered by the most popular Wikipedia pages, such as pages related to entertainment, history, politics, geography, or sexuality.
Discussion

Using the Cluster View, it becomes apparent that topics related to “Sexuality” represent a large percentage of the pages that are contained in all five lists (see Figure 2). Using the Spiral View, pages related to “Entertainment”, such as music, films or video games, are becoming increasingly frequent as you move away from the display center (see Figure 3). There are also many pages related “Geography”, such specific countries or places. Further, pages that are related to “Politics”, such as political figures, or to “History”, such as wars and specific events, represent a major group of popular Wikipedia pages. Specifically, pages about “World War I”, “World War II” and the “Vietnam War” are highly visited in Wikipedia. However, the current war in Iraq is not represented by a set of pages that make it into the Top 100 in any of the months studied. searchCrystal can be used to compare lists that contain more than 100 pages and WikiCharts can return at most the thousand most visited Wikipedia pages. If the Top 150 or 200 pages are compared, then only one page is related to the Iraq war and it shows up only in the December 2006 list.

The fact that pages related to geography, history, or politics are highly visited in Wikipedia is what you would expect, since they represent the prototypical topics to be found in an encyclopedia. However, an informal analysis of the most popular Wikipedia pages, which is made easy by the Spiral View, suggests that pages related to “Entertainment” and “Sexuality” make up a much larger share of the highly visited Wikipedia pages than would be expected from an online encyclopedic resource. This fact (and surprise for the author) warrants further analysis. Future research will categorize the popular Wikipedia pages to identify the major topics of interest and their exact
percentages as function of the number of monthly "Top 100" lists that contain them. This will enable us to determine if some these major topics of interest are “timeless” since their related pages tend to be contained in all monthly lists.

A related research project will investigate why many of the most visited Wikipedia pages are related to the most popular search topics. If you examine the Google and Yahoo most popular search queries, then you will find that many queries are related the same media celebrities, films or TV shows that have Wikipedia pages that are highly visited.

**Conclusion**

The monthly lists of the 100 most visited in September 2006 to January 2007 were compared and their overlap was visualized using searchCrystal. It was shown that almost 40 percent of a month’s top 100 pages are visited in all five months, whereas 25 percent are highly visited only in a single month. Further, it was illustrated how searchCrystal can be used to gain quick insights into the overlap structure and topical relationships between the monthly "Top 100” lists. In particular, the Spiral View made it easy to identify the major topics of interest and see that pages related to entertainment and sexuality represent a large percentage of the most popular Wikipedia pages.

**About the author**

Anselm Spoerri is an Assistant Professor in School of Communication, Information and Library Studies (SCILS) at Rutgers, The State University of New Jersey. He was a researcher at AT&T Bell Labs after completing his Ph.D. research at MIT, where he developed InfoCrystal, which is a precursor of searchCrystal.
Acknowledgments

The author would like to thank Sacha Stanton for developing a mechanism for reading the WikiCharts data output.

Notes

1. YouTube: Popular Videos at http://youtube.com/browse?s=mp&t=m&c=0&l=
3. Del.icio.us: Popular Bookmarks at http://del.icio.us/popular/

References


Submitted to online journal FirstMonday on 21 February 2007.


Figure 1: Category View displays the overlap between the five monthly “Top 100” Wikipedia pages for September 2006 to January 2007.
Figure 2: Cluster View shows the pages contained in all five “Top 100” lists, where the center ring has been magnified. (Bottom) shows the Details-on-Demand display for the “September 11, 2001 attacks” page.
Figure 3: Spiral View displays all the 230 unique Wikipedia pages that are contained in the monthly "Top 100" lists for September 2006 to January 2007.