Exploring Wikipedia with Höpara

David Milne       Ian H. Witten
Department of Computer Science, University of Waikato
Private Bag 3105, Hamilton, New Zealand
+64 7 838 4246
{dmilne, ihw}@cs.waikato.ac.nz

Categories and Subject Descriptors
H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia – navigation, user issues.

General Terms
Experimentation, Human Factors.

Keywords
Exploratory Search, Information Visualization, Wikipedia.

1. INTRODUCTION

Anyone who has browsed Wikipedia has likely experienced the feeling of being happily lost, browsing from one interesting topic to the next and encountering information that they would never have searched for explicitly. With some 3M articles and 70M links, Wikipedia represents an extreme example of large-scale hypertext. We consider it to be a rich and challenging platform for investigating navigation and disorientation in large interconnected information spaces.

This demonstration showcases Höpara, a new search engine that aims to make Wikipedia and its underlying link structure easier to explore. It works on top of the encyclopedia’s existing link structure, abstracting away from document content and allowing users to navigate the resource at a higher level. It utilizes semantic relatedness measures to emphasize articles and connections that are most likely to be of interest, visualization to expose the structure of how the available information is organized, and lightweight information extraction to explain itself.

2. HÖPARA

Höpara’s interface is shown in Figure 1. The upper area is a classic search box where the user has entered Tacoma Narrows Bridge. Below and to the left of the search box is a visualization of the related topics: the engineers involved, other similar bridges, and some of the engineering concepts connected to its demise. To the right is an extract of the relevant article.

The query is ambiguous: Wikipedia contains an article about the bridge that famously collapsed in 1940, and another about its replacement. When queries have multiple interpretations, the system selects one sense automatically but makes the remaining senses available with the link entitled or did you mean. The system also allows for synonymy: e.g. the query galloping gertie (the bridge’s nickname) takes the user to the same result.

The visualization on the left of the interface displays the user’s query in the center, and four categories or groupings of suggestions surrounding it. Within each grouping is a small graph whose nodes are topics and edges are semantic relations between them. Larger topics, such as Leon Moisseif (the project’s lead engineer) are more strongly related to the query. An edge between two topics indicates that they are semantically related to each other, and thicker edges indicate stronger relations.

Only four groupings of topics can be shown at a time. Moving clockwise from the top left corner of Figure 1, the first three represent the categories containing the strongest, most relevant suggestions. Their size represents their expected value to the searcher. The fourth grouping, indicated with a dotted outline, shows the best topics that did not belong to the top categories. If this is clicked, the visualization smoothly rotates to reveal the remaining, less relevant categories. Thus the system can scale to show many categories of topics, without panning or zooming.

Similarly, only the four best topics are shown within each grouping. If more are available, then scalability is achieved by allowing categories to be expanded. If the user clicks Suspension Bridges in Figure 1, for example, then system smoothly transitions to the layout shown in Figure 2. Animation is used carefully to minimize disorientation: the desired category slickly expands and the others shrink and move out of the way. More nodes are added to the graph, which is smoothly but quickly rearranged using a force-directed layout that encourages related topics to be clustered together spatially, and others (such as the general topic Suspension Bridge, or Millennium Bridge—the only one not located in the U.S.) to be separated.

Mousing over any topic link reveals a tooltip containing the first sentence of the article in question, as shown for Suspension Bridge in Figure 2. The user can click on any topic to open a box on the right side of the interface, as shown for Tacoma Narrows Bridge (1940) in Figure 1 and Bronx-Whitestone Bridge in Figure 2. The box contains the first paragraph and an image extracted from the article, and a link to Wikipedia. It also contains sentence snippets to explain how the topic relates to the original query. In Figure 6 this reveals that the Bronx-Whitestone Bridge and the Deer Isle Bridge suffered similar design problems as Galloping Gertie, which explains why Höpara emphasizes them over the other bridges.

On the top right corner of the Bronx-Whitestone Bridge topic box in Figure 2 is a set of three buttons that control how it can be investigated further. The first explores it as a new query. The second adds it to the current query, to explore things that relate to or connect both bridges. The third button removes it from the query—it is disabled in Figure 2 because Bronx-Whitestone Bridge is not part of the current search.

Multi-topic queries can also be built directly in the search box. For example, the query bridge failure is automatically recognized...
as two distinct topics: Bridge and Structural failure. The interaction is much the same as for the mono-topic query in Figures 1 and 2, except that the visualization on the left is narrowed to contain only suggestions that relate to (or bridge between) both query topics (e.g. Catastrophic failure, Structural design), and multiple connection snippets are shown within each topic box (because there are multiple query topics to connect to).

The backend of Hōpara is built upon the Wikipedia Miner toolkit, a Java-based API and suite of web services for navigating and making use of the structure and content of Wikipedia. This provides much of the functionality that Hōpara depends on, including matching search terms to articles, gathering related topics, grouping these suggestions into categories, and explaining connections between topics.

Try Hōpara out for yourself at http://www.nzdl.org/hopara

---