Information Systems 35 (2010) 260–269

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Interactive visualization for opportunistic exploration of large document collections

Simon Lehmann *, Ulrich Schwanecke, Ralf Dörner

Department of Design, Computer Science and Media, RheinMain University of Applied Sciences, Kurt-Schumacher-Ring 18, 65197 Wiesbaden, Germany

ARTICLE INFO

Article history:
Received 28 August 2009
Received in revised form
4 October 2009
Accepted 5 October 2009
Recommended by: D. Shasha

Keywords:
Information visualization
Opportunistic exploration
Browsing
Searching
Wikis

ABSTRACT

Finding relevant information in a large and comprehensive collection of cross-referenced documents like Wikipedia usually requires a quite accurate idea where to look for the pieces of data being sought. A user might not yet have enough domain-specific knowledge to form a precise search query to get the desired result on the first try. Another problem arises from the usually highly cross-referenced structure of such document collections. When researching a subject, users usually follow some references to get additional information not covered by a single document. With each document, more opportunities to navigate are added and the structure and relations of the visited documents gets harder to understand.

This paper describes the interactive visualization *Wivi* which enables users to intuitively navigate Wikipedia by visualizing the structure of visited articles and emphasizing relevant other topics. Combining this visualization with a view of the current article results in a custom browser specially adapted for exploring large information networks. By visualizing the potential paths that could be taken, users are invited to read up on subjects relevant to the current point of focus and thus opportunistically finding relevant information. Results from a user study indicate that this visual navigation can be easily used and understood. A majority of the participants of the study stated that this method of exploration supports them finding information in Wikipedia.

1. Introduction

A common approach to collect information about a specific subject in large, cross-referenced document collections like the Wikipedia is to start with an already known term and open the associated article. Within the article, links to other terms probably relevant to the research are found. If more than one article has to be read to get the desired information, a user has to follow some of the links to other articles. They usually have to be read one at a time and again contain links to further articles probably relevant to the subject. While navigating between several articles, a user has to keep track of what they already have read and how the piece of information they are currently reading relates to everything they already have read. Additionally, a user might want to know what other links they encountered in all the previously read articles and which therefore are probably worth following, especially if many of them lead to a single article.

The problem arises from the complex structure of highly cross-referenced articles. They form a directed graph, which consists of hundreds of thousands of articles

* Corresponding author. Tel.: +49 611 9495 1294; fax: +49 611 9495 1210.
E-mail address: simon.lehmann@hs-rm.de (S. Lehmann).

0306-4379/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2009.10.004

and usually significantly more links. The English language version of Wikipedia currently comprises 3 million articles and over 70 million links between them [28].

Common web-browsers used for exploring web-based information resources are not providing any means to help users with the specific tasks presented when researching a subject in such a large network of articles. They only provide a history of visited pages which can be navigated forwards and backwards. Besides this simple linear history of pages, they do not tell the users what other links they might consider following and how two different pages are related to each other.

In this paper, we present a navigation concept for the exploration of such large information resources, which visualizes the structure of articles a user has already seen and where a user might find further information related to the already researched subject. While there have been similar approaches to on-line, interactive visualizations of hypertext document structures in general, our work focuses on the additional visualization of all related links a user might want to follow, and how to help with choosing potentially interesting articles. By applying a weighting function to the yet unread articles, we can highlight more important articles and help a user to opportunistically explore the vast amount of information.

This paper is organized as follows. In Section 2 we will review other approaches taken to interactively visualize large document structures. Section 3 will briefly explain the concept of opportunistic exploration and how we are applying it to the task of searching for information. The visualization and interaction concept we developed is then described in Section 4. Section 5 is outlining the implementation of our concept in the web-based application *Wivi*.[1] In Section 6, the set-up and results of the user study we conducted are described and analyzed respectively. In Section 7, we finally give our conclusion and present potential future work.

2. Related work

There have been several approaches to interactively visualize large document collections in order to make exploring and finding relevant information easier.

The InfoSky system developed by Andrews et al. [3] is a visual explorer for news articles which contain no cross-references themselves, but are classified into a hierarchy many levels deep. It visualizes all articles at once in a *galaxy of stars*, where each article is visualized as a star. These stars are clustered by the hierarchical structure of the articles they represent. Like earlier work in this field—such as the landscape metaphor [7,29] or the hyperbolic browser [20]—this approach visualizes all documents at once and makes the documents explorable by using a semantic zoom technique which reveals individual documents. This works well with a fully known set of documents which can be hierarchically grouped (either manually or automatically). Even though visualizing large numbers of documents in the range of ten to hundred thousands is possible, visualizing millions of documents at once in an interactive environment is still a problem [14].

Very large document collections like the Wikipedia are too large to be visualized interactively at once and in the case of the world wide web, the exact amount of documents and their structure is unknown. One approach to deal with such large and only partially known document collections is to exclusively visualize the immediate surroundings of the document space already explored by a user. The NESTOR tool implemented by Eklund et al. [11] simply visualizes the history of a user's browsing of the world wide web. The WebOFDAV system implemented by Huang et al. [17] (and similar, more recent systems [18,19]) provide the immediate neighbors of visited pages as navigational elements to find new pages to visit. Most of those systems use various force-directed layout algorithms for displaying the graph of pages and some also use clustering methods to improve the generated layouts. Those systems simply display what a user has seen and what can be immediately reached from there, but provide no means to help the user deciding which pages to visit next.

Besides the interactive browsers for general hypertext documents there have been attempts to make more specialized visual browsers for knowledge spaces like the Wikipedia. Hirsch et al. [16] developed two interactive visualizations of Freebase (a "Semantic Wiki") and Wikipedia named Thinkbase and Thinkpedia respectively. Their approach uses similar techniques for visualization as those used for general hypertext documents, but utilizes semantic web data for nodes and links. This leads to a different visualization concept, because some nodes represent actual articles while other nodes represent semantic concepts belonging to the articles. The Thinkpedia visualization also introduces weighting of the semantic concepts based on the relevance value computed by the semantic web service used for retrieving all concepts relevant to an article. This helps a user in finding more relevant data, but since navigating to another article results in a complete regeneration of the graph which does not contain the previously visited articles anymore, it is only of limited use for extended exploration of the Wikipedia.

Another accentuation technique for expeditiously finding relevant terms in text documents is the textarc visualization [22] which displays each line of a single text on a large circle with a very tiny font size. Inside the circle, the words of the text are displayed with different sizes and positions according to their frequency and distribution in the text. Words of higher frequency are made larger, and thus more visible. The position of a word is determined by the centroid of the points where it occurs in the text on the circle. This allows to visually spot words which are most important, and where the words are used predominantly in the text. Words which are used throughout the text are positioned near the center of the circle, while those that appear only in a certain section are positioned close to that section on the circle. This visualization provides a good overview of the important terms of a longer text like a book, and makes it easy to

[1] http://wivi.slashslash.de/