Interactive Visualization and Exploration of Information on Philosophers (and Artists, Scholars, & Scientists) in an e-Learning Portal for Digital Humanities

Sofia J. Athenikos
College of Information Science and Technology
Drexel University
Philadelphia, PA, USA
1-215-895-2474
sofia.j.athenikos@acm.org

ABSTRACT
Since its introduction in 2001, Wikipedia has grown rapidly to become one of the most prominent information sources on the Web. Nevertheless, the manner of information search and retrieval on Wikipedia remains that of conventional keyword-based search and retrieval of a list of articles instead of more focused results. It would be highly desirable to have tools that could maximize the effective and innovative use of a vast amount of information content in Wikipedia. The objective of the PanAnthropon project is to create a large-scale e-learning/e-knowledge portal for digital humanities based on the information extracted from Wikipedia (as well as other complementary information sources). The portal is intended to enable semantics-based and visually-enhanced search and exploration concerning major philosophers, scholars, artists, and scientists, in particular, their explicit/implicit intellectual and cultural connections. The aforementioned PanAnthropon project is conceived as a large-scale extension of a related pilot project, entitled WikiPhiloSofia, which concerned extraction, analysis, and visualization of semantic and hyperlink information on major philosophers using the Wikipedia data. By exploiting the synergy between the Social Web and the Semantic Web and by employing interactive visualization as an effective mode of information presentation, these projects contribute to a paradigm shift toward the next generation of online information services/systems that incorporate information aesthetics and edutainment. In this paper I present the results from the WikiPhiloSofia project.

Categories and Subject Descriptors
H 3.3 [Information Storage and Retrieval]: Information Search and Retrieval – search process; H 3.5 [Information Storage and Retrieval]: Online Information Services – web-based services; H 3.7 [Information Storage and Retrieval]: Digital Libraries – collection; H 5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – hypertext navigation and maps; H 5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia – navigation, user issues.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
IVICA 2009, June 19, 2009, Austin, TX, USA.
Copyright 2009 ACM …$5.00.

General Terms

Keywords

1. INTRODUCTION
Since its inception in 2001 as a collaborative Web encyclopedia project, Wikipedia (http://www.wikipedia.org) has grown rapidly to become one of the most sought-after resources on the Web. What makes Wikipedia a particularly valuable knowledge source for various applications consists in its unique combination of structural/semantic features, i.e., a dense hyperlink structure, a hierarchical category structure, infoboxes, wikitables, etc., and a vast amount of user-generated content. Nevertheless, the mode of information search and retrieval on Wikipedia remains that of conventional keyword-based search and retrieval.

The objective of my proposed project, entitled PanAnthropon [4], is to create a large-scale Web-based learning/knowledge portal by extracting, presenting, and visualizing meaningful and interesting facts, relations, and networks, through the automatic processing of the structural features and semantic (and multimedia) content of Wikipedia (as well as other complementary information sources). While the methodology can be applied to various domains, the immediate intended domain of application is digital humanities. The aim is to provide a useful and user-friendly Web interface for humanities scholars/students to conduct data-driven studies, by enabling semantics-based and visually/aesthetically-enhanced search and exploration on major philosophers, scholars, artists, and scientists, in particular, their explicit and implicit intellectual and cultural connections. The PanAnthropon project constitutes a large-scale extension of my pilot project, called WikiPhiloSofia (aka The WikiPhil Portal) [2-3], which concerned extraction and visualization of facts, relations, and networks involving major philosophers using Wikipedia.

In this paper I present the results from the WikiPhiloSofia project, with a special focus on interactive visualization. The remainder of the paper is organized as follows: Section 2 describes the data collection process. Section 3 describes the Web portal interface. Section 4 illustrates presenting the information via interactive visualization. Section 5 briefly discusses related work. Finally, Section 6 concludes the paper.
2. WIKIPEDIA DATA COLLECTION

The WikiPhiloSofia project used the English version of Wikipedia as the sole data source. The last download of Wikipedia pages was done on 23 December 2008.

Data extraction and information collection proceeded through a three-step process as follows: (i) Step 1: Extract a chronological list of 300 major philosophers from the “Timeline of Western Philosophers” page. (ii) Step 2: Extract the hyperlink connections and academic/biographical facts on the philosophers from their Wikipedia article pages (including infoboxes and wikitables), and store the data in a MySQL database in the form of semantic triples (Subject—Predicate—Object). (iii) Step 3: Retrieve information needed for visualization by querying the database, and store the results as XML files marked up with GraphML and TreeML.

The academic/biographical information extracted during Step 2 was augmented by automatically inferred information, in cases where the relational information involving a pair philosophers (e.g., influenced-by/influenced) appears in only one philosopher’s page. Table 1 summarizes the types of information extracted.

Table 1. Types of information extracted

<table>
<thead>
<tr>
<th>Basic Bio Info</th>
<th>Academic/Biographical Info</th>
<th>Info on Hyperlink Connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>Occupations</td>
<td>(Outgoing) Hyperlinks</td>
</tr>
<tr>
<td>Timeline</td>
<td>Fields/Main Interests</td>
<td>Categories</td>
</tr>
<tr>
<td>Life</td>
<td>Schools/Traditions</td>
<td>Philosophers Linked via Out-Links</td>
</tr>
<tr>
<td>Birth</td>
<td>Notable Ideas (Known For)</td>
<td>Philosophers Linked via In-Links</td>
</tr>
<tr>
<td>Death</td>
<td>Notable Works</td>
<td>Philosophers Linked via Bi-Links</td>
</tr>
<tr>
<td>Names</td>
<td>Notable Awards</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Religions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noted</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Venerated In</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Influenced</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Influenced By</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Notable Teachers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Notable Students</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 shows the basic statistics concerning the dataset obtained through the procedure in Step 2.

Table 2. Basic statistics on the Wikipedia dataset

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of (non-administrative) hyperlinks in 300 philosopher pages</td>
<td>68,301</td>
</tr>
<tr>
<td>Total # of hyperlink connections among 300 philosopher pages</td>
<td>6,298</td>
</tr>
<tr>
<td>Total # of overlapping categories for 300 philosopher pages</td>
<td>3,484</td>
</tr>
<tr>
<td>Total # of triplets extracted on academic/biographical facts</td>
<td>5,698</td>
</tr>
<tr>
<td># of philosophers not linked to other philosophers via out-links</td>
<td>9</td>
</tr>
<tr>
<td># of philosophers not linked to other philosophers via bi-links</td>
<td>22</td>
</tr>
<tr>
<td># of philosophers not linked to or from any other philosopher</td>
<td>3</td>
</tr>
<tr>
<td># of philosophers with no bi-directional link connections</td>
<td>45</td>
</tr>
<tr>
<td># of philosophers who have no “influenced” relations</td>
<td>139</td>
</tr>
<tr>
<td># of philosophers who have no “influenced-by” relations</td>
<td>120</td>
</tr>
<tr>
<td>Max # of philosophers connected via out-links from 1 philosopher</td>
<td>60</td>
</tr>
<tr>
<td>Max # of philosophers connected via in-links to 1 philosopher</td>
<td>105</td>
</tr>
<tr>
<td>Avg # of philosophers connected via out-in-links per philosopher</td>
<td>14</td>
</tr>
<tr>
<td>Max # of philosophers connected via bi-links with 1 philosopher</td>
<td>49</td>
</tr>
<tr>
<td>Avg # of philosophers connected via bi-links per philosopher</td>
<td>9</td>
</tr>
</tbody>
</table>

The number of data files created for visualization during Step 3, concerning the facts, relations, and networks involving one philosopher, two philosophers, all 300 philosophers, is 632,410.

3. WEB PORTAL INTERFACE

A prototype Web portal interface was created, whereby users can issue queries on the facts, relations, and networks involving 300 philosophers and explore the results using diverse modalities of interactive information visualization as will be illustrated in the next section. Figure 1 shows the homepage of the WikiPhiloSofia portal (http://research.cis.drexel.edu:8080/sofia/WPS/).

Upon entering the portal, the user is presented with a main menu with options for search and exploration, as shown in Figure 2.

Figure 1. Homepage of the WikiPhiloSofia portal.

Figure 2. Main menu of user options.

In case a user chooses to focus on one philosopher, for example, the user is then directed to the query menu shown in Figure 3.

Figure 3. Menu for query on one philosopher.

Table 3 summarizes various query options and result display modalities.

Table 3. Options for semantics-based search and exploration
4. INTERACTIVE INFO VISUALIZATION

This section illustrates presenting the results of user queries on the philosophers via interactive visualization implemented by using Prefuse information visualization toolkit (http://prefuse.org/).

4.1 One Philosopher Query

4.1.1 Academic/Biographical Facts

Suppose the user wishes to view all academic/biographical facts on a particular philosopher, e.g., Anaximander. Such information can be represented, e.g., in a radial graph view, as in Figure 4. Even though some labels are illegible, when the user moves the cursor over any node, its label is highlighted and also displayed at the bottom of the window. The user can drag a node to a viewable position, search for a certain label, or transform the shape of graph by clicking on a node thereby moving it to the center.

![Figure 4. Academic/biographical facts on Anaximander.](image)

4.1.2 Direct Links/Influences

Suppose the user wishes to see philosophers that have hyperlink/influence connections with a given philosopher. Figure 5 shows a radial graph representing philosophers Berkeley is connected with via in-links. As shown, each node is labeled with a philosopher name followed by the link count. Figure 6 shows a colored graph representing philosophers Cicero is connected with via out-links. As shown, node colors are differentiated according to the link connection strengths (measured by link counts).

![Figure 5. Philosophers connected with Berkeley via in-links.](image)

4.1.3 Extended Links/Influences

Suppose the user wishes to explore link/influence connections beyond that of direct (1st-degree) connection. Such information can be represented in a fisheye tree view, as shown in Figure 7, which expands/contracts a node according to the user choice so that the user can easily explore many alternative paths. Note that the tree contains all extended influence connection paths within 6 degrees of separation and that, as a result, a node representing the same philosopher may appear multiple times.

![Figure 7. Extended (6-degree) influences from Descartes.](image)
4.2 Two Philosopher Query

4.2.1 Direct Relations

For this type of query, the user can select two philosophers and see the relations that the first (focus) philosopher has with respect to the second (target) philosopher. Figure 8 shows an example with Engels set as the focus and Feuerbach set as the target.

Figure 8. Direct relations between Engels and Feuerbach.

4.2.2 Commonalities

For this type of query, the user can select two philosophers and view all commonalities that exist between the two. Figure 9 shows an example with Hegel and Kierkegaard.

Figure 9. Commonalities between Hegel and Kierkegaard.

4.2.3 Direct (Common) Links/Influences

For this type of query, the user can select two philosophers and the type of link/influence connection and see all philosophers connected via the selected link/influence connection with the two chosen philosophers, including those that are connected with both. Figure 10 shows an example concerning bi-link connections with Locke and Marx. Figure 11 shows an example concerning influenced-by connections with Nietzsche and Parmenides.

Figure 10. Philosophers bi-linked with Locke and/or Marx.

Figure 11. Philosophers influenced by Nietzsche and/or Parmenides.

4.3 All Philosopher Query

4.3.1 Strongest Link/Influence Networks

The method that is used here to effectively visualize the networks of philosophers, highlighting the most significant nodes and their interconnections, is a graph simplification method that has been developed in the project, called the strongest link paths (SLP).

As the name suggests, the method selects, for each node in a given network graph, only the strongest link (in terms of the edge weight representing the link count or other connection strength/significance measure), taking a greedy algorithmic approach. The resultant graph contains a single link per source node, thereby substantially simplifying the graph topology.

Even though SLP is rather simpler than other well-known graph scaling methods such as pathfinder network [14] or main path analysis [9], it has been found that the application of SLP results in not only substantial data reduction but also a fairly meaningful representation of the dominant figures and their interconnections within the networks of philosophers. (In addition, the utility of SLP has been confirmed in a different context, with respect to the visualization of Web forum social networks.)
In the case of strongest link/influence networks, a straightforward version of SLP is used, which selects, for each node, only the link with the highest hyperlink/influence count. The resultant graph consists of distinct clusters clearly separated from one another. Figure 12 shows a close-up of one of the largest clusters in the strongest bi-link network, which centers on Russell. Figure 13 shows a close-up of (part of) the largest cluster in the strongest influenced-by connection network, which centers on Kant.

4.3.2 Non-Overlapping Extended Links/Influences

In the case of networks emerging from extended link/influence connections involving a given philosopher, a variation of SLP is used, which adds all 1st-degree links from the origin node, and then, for each subsequent degree, only adds links to the nodes that are not already covered, thereby eliminating overlapping edges.

The graph that results from applying SLP by eliminating edge crossing consists of one large cluster which in turn consists of subclusters. Figure 14 shows the non-overlapping extended out-link network originating from Quine. Figure 15 shows the non-overlapping extended influenced connection network originating from Thales, which reveals that Thales, the first philosopher on the chronological list of 300 philosophers, can reach Foucault, the last one on the list, within merely 3 degrees of separation (via Anaximander and Heidegger).

4.3.3 Statistical Rankings

Figure 16 shows a tag cloud representation of philosophers’ relative rankings (solely) based on the number of philosophers connected via in-links.
5. RELATED WORK
Corresponding to the rapid, exponential growth of the size of Wikipedia, recent years have witnessed an increasing number of researchers working on various projects concerning Wikipedia. Various studies concerning Wikipedia as a knowledge source can in general be divided into those that aim at extracting lexical/semantic/encyclopedic knowledge from Wikipedia and those that aim at enhancing Wikipedia (and wikis in general) by embedding explicit semantic features (e.g., [10]). The former can further be divided into those concerned with structural mining and those concerned with content mining (cf. [18]). Examples of the former include those concerned with extracting domain/corpus-specific thesauri by exploiting the hyperlinks, categories, and redirects in Wikipedia [12], deriving a taxonomy using Wikipedia’s category structure [13], constructing a large-scale ontology by exploiting the structural features of Wikipedia [15], and deriving a large number of RDF statements from Wikipedia templates [1]. The summary distinctions of the WikiPhiloSofia project (as well as the PanAnthropon project) from these studies, however, consist in its incorporation of interactive visualization as an effective mode of knowledge representation and in its focus on extracting/deriving intellectual/conceptual connections among the entities of interest. The latter point of distinction also separates the current work from most applications of various forms of information visualization involving Wikipedia, which have been mainly concerned with visualizing the activity pattern of content contributors (e.g., [5],[7],[16],[17]) or with visualizing the conceptual structure of Wikipedia itself (e.g., [8],[11]).

6. CONCLUSION
In this paper I have demonstrated visualizing various facts, relations, and networks involving major philosophers using the hyperlink/semantic data extracted from Wikipedia. The results from the WikiPhiloSofia project suggest that fairly meaningful and interesting connections among intellectual figures can be explored effectively via interactive information visualization. The proposed PanAnthropon project will greatly extend the scope of coverage (both in terms of domains and data sources) and also incorporate more effective data extraction/storage methods and data analysis/visualization tools. The Web-based learning portal thereby created will serve as a useful and user-friendly interface for humanities students, as well as the general public, to conduct semantics-based and visually-enhanced exploration of the information concerning influential philosophers, scholars, artists, and scientists.

7. REFERENCES