# WikiPhiloSofia and PanAnthropon: Extraction and Visualization of Facts, Relations, and Networks for a Digital Humanities Knowledge Portal

Sofia J. Athenikos[1]
College of Information Science and Technology
Drexel University
Philadelphia, PA, USA
1-215-299-1299
sofia.j.athenikos@acm.org

## ABSTRACT
Wikipedia, with its unique structural features and a vast amount of user-generated content, is being increasingly recognized as a valuable knowledge source for various applications. Nevertheless, the mode of information search and retrieval on Wikipedia remains that of conventional keyword-based search and retrieval. The objective of my (soon-to-be-proposed) thesis project, entitled *PanAnthropon*, is to create a large-scale Web-based knowledge portal for digital humanities, which enables semantics-based and visually-enhanced search and exploration concerning influential philosophers, scholars, artists, and scientists, in particular, explicit and implicit intellectual and cultural connections among them, based on the data extracted from Wikipedia (and complementary information sources). The *PanAnthropon* project extends my pilot project, *WikiPhiloSofia*, which concerned extraction, analysis, and visualization of semantic and hyperlink information concerning major philosophers using Wikipedia. By exploiting the synergy between the Social Web and the Semantic Web and by employing visualization as an effective mode of information presentation, the aforementioned projects contribute to a paradigm shift toward the next generation of Web-based information service/system that explicitly incorporates information aesthetics and edutainment. In this poster I present the promising results I have obtained so far.

## Categories and Subject Descriptors
H 3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation*, *search process*; H 3.5 [**Information Storage and Retrieval**]: Online Information Services – *web-based services*; 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*.

[1] The author is a doctoral student at the College of Information Science and Technology at Drexel University, Philadelphia, PA, USA. Her advisor is Dr. Xia Lin (xlin@drexel.edu).

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.
*Hypertext 2009*, June 29 – July 1, 2009, Torino, Italy.
Copyright 2009 ACM …$5.00.

## General Terms
Algorithms, Design, Human Factors.

## Keywords
Wikipedia, Social Semantic Web, Digital Humanities, e-Learning, Information Visualization, Social Network Analysis, HCI.

## 1. EXTENDED ABSTRACT
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 (soon-to-be-proposed) thesis project, entitled *PanAnthropon*, is to create a large-scale Web-based knowledge portal for e-learning, by extracting, (re)presenting, and visualizing meaningful and interesting facts, relations, and networks, through the automatic processing of the structural features and semantic/ 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 produce a useful, user-friendly Web interface for humanities scholars/students to conduct data-driven studies, by enabling semantics-based, reasoning-supported, 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 aforementioned *PanAnthropon* project constitutes a large-scale extension of my pilot project, called *WikiPhiloSofia* (aka *The WikiPhil Portal*), which concerned extraction/visualization of facts, relations, and networks involving major philosophers.

In this poster I present some of the promising results I have obtained from the *WikiPhiloSofia* project [1-3] (and possibly also the early results from the *PanAnthropon* project [4]). The rest of this extended abstract provides a summary overview of the main content of the poster.

(1) Wikipedia Data Extraction

The data extraction process proceeded as follows: (i) Step 1: Extract a (chronological) list of 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, and store the data in a MySQL database. (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 types of information extracted are summarized in Table 1:

**Table 1. Types of information extracted**

| | | |
|---|---|---|
| Period | Occupations | (Outgoing) Hyperlinks |
| Timeline | Fields/Main Interests | Categories |
| Lifetime | Schools/Traditions | Philosophers Linked via Out-Links |
| Birth | Notable Ideas (Known For) | Philosophers Linked via In-Links |
| Death | Notable Works | Philosophers Linked via Bi-Links |
| Names | Notable Awards | |
| | Religions | |
| | Venerated In | |
| | Influenced By | |
| | Influenced | |
| | Notable Teachers | |
| | Notable Students | |

(2) Semantic Query and Exploration

The above information can be explored using various foci, facets, and visualization modalities, as shown in Table 2:

**Table 2. Options for semantics-based search and exploration**

| Foci | Facets | Visualization Modalities |
|---|---|---|
| One Philosopher | Academic/Biographical Facts | Radial Graph View, Graph View, Tree View |
| | Direct Links/Influences | Radial Graph View, (Colored) Graph View |
| | (All 6-Degree) Extended Links/Influences | Tree View |
| Two Philosophers | Direct Relations | Radial Graph View |
| | Commonalities | Radial Graph View |
| | Direct (Common) Links/Influences | Radial Graph View |
| All Philosophers | Strongest Link/Influence Networks | Graph View |
| | (Non-Overlapping) Extended Link/Influence Networks | Radial Graph View, Graph View |
| | (Purely Statistical) Rankings | Tag Cloud View |

(3) Web Portal Interface

Figure 1 shows the homepage of the *WikiPhiloSofia* portal site (http://research.cis.drexel.edu:8080/sofia/WPS/).


**Figure 1. Homepage of the *WikiPhiloSofia* portal.**

(4) Interactive Information Visualization

The interactive information visualization is implemented by using the Prefuse information visualization toolkit (http://prefuse.org/). Here I present some examples. (The poster contains more.)

Figure 2 shows extended influences originating from Plato. Figure 3 shows commonalities between Descartes and Leibniz.

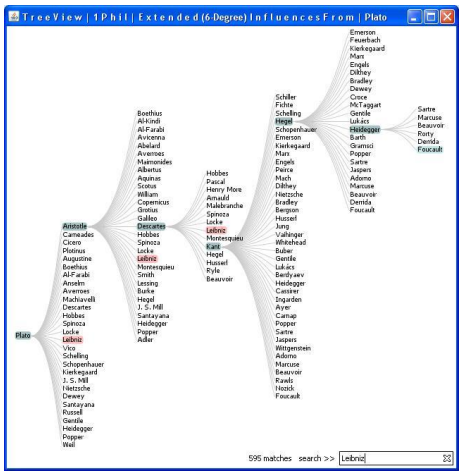


**Figure 2. Extended (6-degree) influences from Plato.**

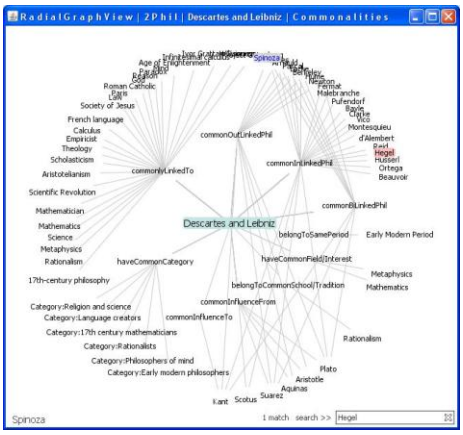


**Figure 3. Commonalities between Descartes and Leibniz.**

The method that I use to effectively visualize a network, in particular, in order to highlight the most significant nodes and their interconnections, is a graph simplification method I have developed in the project, called the *strongest link paths* (SLP) [2]. 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.

Here I use two variations of SLP. In the case of the strongest hyperlink/influence networks, I use the straightforward version of SLP by selecting, for each node, only the link with the highest hyperlink/influence count. In the case of networks emerging from extended hyperlink/influence relations, I add all 1st-degree links and then, for each subsequent degree, I only add links to nodes that are not yet covered, thereby eliminating overlapping edges.

The graph that results from applying SLP by selecting only the links with highest hyperlink/influence counts consists of distinct clusters clearly separated from one another. Figure 4 shows a close-up of the largest cluster in the strongest out-link network, which centers on Plato and Aristotle.

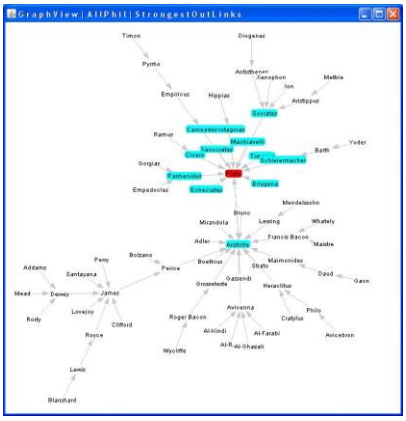

**Figure 4. Largest cluster in the strongest out-link network.**

The graph that results from applying SLP by eliminating edge crossing consists of one large cluster which in turn consists of subclusters. Figures 5-6 show non-overlapping extended in-link and influenced relation networks involving Heidegger and Thales.

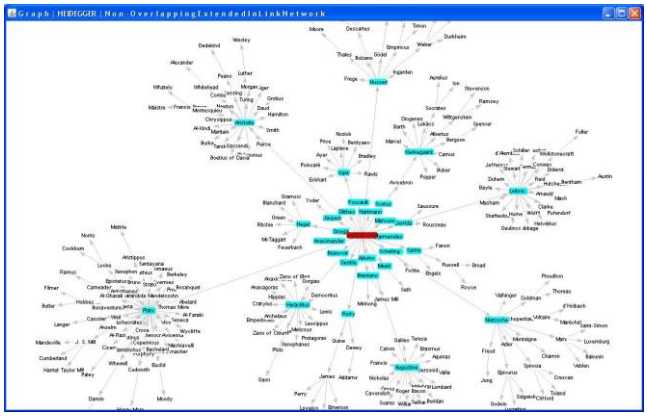

**Figure 5. Non-overlapping extended in-links to Heidegger.**

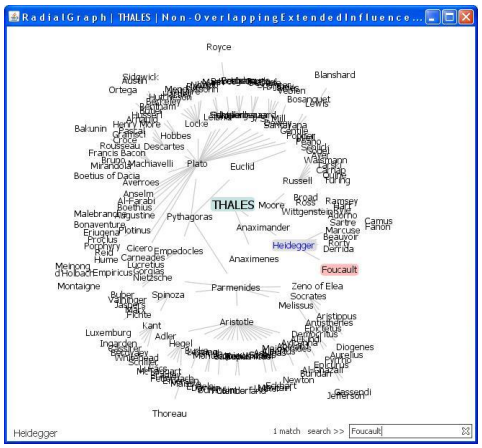


**Figure 6. Non-overlapping extended influences from Thales.**

Building on the results of the *WikiPhiloSofia* project, as briefly described above, the *PanAnthropon* project will extend the scope of the pilot project by incorporating data sources other than Wikipedia and domains other than philosophy, by incorporating more efficient and effective data extraction methods, data storage/ representation mechanisms/formalisms, and data analysis and visualization tools, and by incorporating non-textual multimedia information resources. The technical significance of the project consists in its expected contributions to the fields of information extraction, information retrieval, information visualization, digital libraries, digital humanities, etc. The intellectual significance of the project consists in the generation/visualization of intellectual landscapes representing explicit/implicit connections among influential philosophers, scholars, artists, and scientists. The practical significance of the project consists in the utility of the Web portal as an interface for humanities scholars/students to conduct data-driven scholarship.

## 2. ACKNOWLEDGMENTS

I would like to thank the members of my thesis committee – Profs. Xia Lin, Howard White, and Il-Yeol Song at Drexel University, Prof. John Hopcroft at Cornell University, and Prof. Stéfan Sinclair at McMaster University, Canada – for their encouraging recognition of the merits of *WikiPhiloSofia* and *PanAnthropon*.

## 3. REFERENCES

[1] Athenikos, S.J., and Lin, X. The WikiPhil Portal: extraction, analysis, and visualization of philosophical connections using Wikipedia. Poster. Won student poster award at the Fall 2008 North East DB/IR Day (University of Pennsylvania, Philadelphia, PA, USA, 14 October 2008).

[2] Athenikos, S.J., and Lin, X. The WikiPhil Portal: visualizing meaningful philosophical connections. Presented at the 2008 Chicago Colloquium on Digital Humanities and Computer Science (DHCS 2008) (University of Chicago, Chicago, IL, USA, 1-3 November 2008). Forthcoming in the proceedings.

[3] Athenikos, S.J., and Lin, X. WikiPhiloSofia: extraction and visualization of facts, relations, and networks concerning philosophers using Wikipedia. To be presented at the 2009 Digital Humanities Conference (DH 2009) (University of Maryland, College Park, MD, USA, 22-25 June 2009).

[4] Athenikos, S.J. PanAnthropon: e-knowledge portal for digital humanities toward semantic exploration and visualization of intellectual, cultural, & scientific connections. To be presented at the Doctoral Consortium of the 9th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2009) (Austin, TX, USA, 15-19 June 2009).