Vispedia: On-demand Data Integration for Interactive Visualization and Exploration

Bryan Chan, Justin Talbot, Leslie Wu, Nathan Sakunkoo, Mike Cammarano, Pat Hanrahan*
Stanford University, California, USA

ABSTRACT

Wikipedia is an example of the large, collaborative, semi-structured data sets emerging on the Web. Typically, before these data sets can be used, they must transformed into structured tables via data integration. We present Vispedia, a Web-based visualization system which incorporates data integration into an iterative, interactive data exploration and analysis process. This reduces the up-front cost of using heterogeneous data sets like Wikipedia. Vispedia is driven by a keyword-query-based integration interface implemented using a fast graph search. The search occurs interactively over DBpedia's semantic graph of Wikipedia, without depending on the existence of a structured ontology. This combination of data integration and visualization enables a broad class of non-expert users to more effectively use the semi-structured data available on the Web.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*interactive data exploration/discovery, data and knowledge visualization*

General Terms

Design, Human Factors

Keywords

Information visualization, Data integration, Semantic web

1. INTRODUCTION

Within the database community there has been considerable interest in how to handle the considerable growth of heterogeneous data sets online. The size of these data sets and the inconsistency of their schemata imply that visual tools will be necessary to support the exploration and analysis of the data [2]. Within the information visualization community, recent research has focused on making visualization techniques available to a broader audience of non-expert users. However, a major barrier to this effort is the necessary up-front cost of finding, integrating, and normalizing data. If the costs of obtaining data are too high, many users will give up, even before attempting to create a visualization. We present Vispedia, a system that incorporates data integration within an interactive visual data exploration loop (see Figure 1). Though this work was done primarily in the information visualization community, we believe that this system will be of particular interest to SIGMOD researchers.

Our data integration approach starts from one of the many tables available on Wikipedia. Since any given table will seldom contain all attributes of interest, we allow users to integrate additional attributes on demand. Guided by the user's keyword queries, Vispedia follows links within Wikipedia from the source table to find potential query responses contained in other Wikipedia articles. This approach leverages Wikipedia's wealth of of semi-structured, hyperlinked data, in the form of infoboxes, lists, and categories.

In addition to permitting users to rapidly create and refine standalone visualizations, Vispedia also supports new browsing and data integration scenarios. For example, users who create a visualization of data from a Wikipedia article can then follow links in the visualization back to related articles to fix incorrect or missing data.

In 2007, we presented the keyword-search-based integration and visualization formalism and provided a non-interactive algorithm for ranking potential integration paths against keywords [5]. In 2008, we described Vispedia, an interactive system based on the formalism and evaluated its usability [6]. In a first-use formative

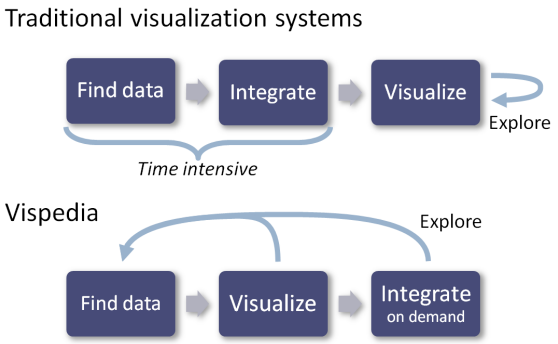

Figure 1: (top) In traditional visualization systems, obtaining data is a slow preprocess, so exploration is limited to already structured data. (bottom) Instead, Vispedia uses on-demand data integration as part of the exploratory loop, making visual exploration over large semi-structured data spaces possible.

*{bryanc, jtalbot, lwu2, sakunkoo, mcammara}@stanford.edu, hanrahan@cs.stanford.edu

SIGMOD'09, June 29–July 2, 2009, Providence, Rhode Island, USA.

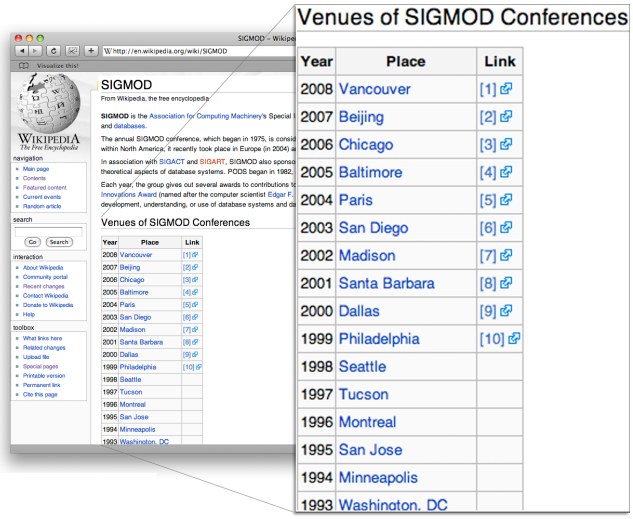

Figure 2: Wikipedia page for the SIGMOD conference.

evaluation study of the system (n=7), participants successfully created compelling visualizations from Wikipedia data.

2. DEMONSTRATION SCENARIO

The demonstration audience will be able to interact with the complete Vispedia system running in a browser. In case of limited connectivity, a subset of the Wikipedia data set will be hosted on the demonstration machine. Readers are invited to try the live system at http://vispedia.stanford.edu.

First, the user browses Wikipedia for articles of interest. Any Wikipedia table, such as this list of SIGMOD locations (Figure 2), can be selected as the "seed" table for integration / visualization.

After clicking on the Vispedia-provided browser bookmarklet ("Visualize this!"), the user can select tables and visualization types directly inside of the Wikipedia page (Figure 3). In this case, we choose to visualize SIGMOD locations on a map.

The user is redirected to the Vispedia site and is shown a *visualization template* (Figure 4). This specifies the necessary attributes for creating a visualization and allows the user to enter keyword queries which the Vispedia system will use to locate data within Wikipedia. In this case, Vispedia will begin a search from the data and links contained within the SIGMOD table looking for link

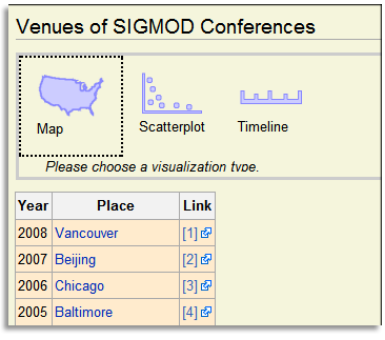

Figure 3: Using the Vispedia bookmarklet allows the user to select the SIGMOD table and an appropriate visualization.

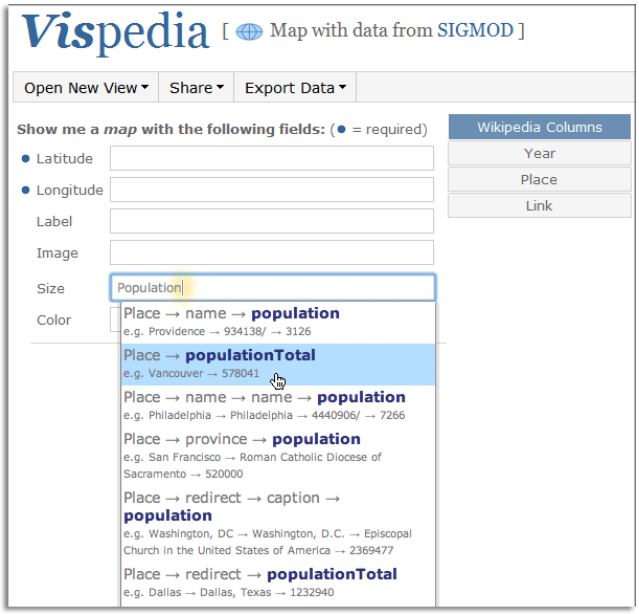

Figure 4: On the Vispedia site, the user can use keyword queries to integrate data from the Wikipedia semantic graph for use in a visualization (in this case, a map).

paths that terminate at a numeric value within a Wikipedia infobox and which contain keywords that match those provided by the user (e.g. "Population"). To help the user locate existing link paths within Wikipedia, the query boxes provide a drop-down list of path suggestions, much like a standard search interface.

Due to the heterogeneity of data and links on Wikipedia, multiple paths may be used to gather information for all the rows in the original data table. An additional drop down allows more advanced users to explicitly control which paths are used (not shown).

When the user changes keywords in the visualization template, the visualization immediately updates, maintaining an interactive experience and supporting the exploration process.

While creating the visualization, the user finds that this year's conference is not yet included in the Wikipedia article. The user adds the information to the Wikipedia article. Vispedia re-extracts the table and a revised visualization is generated which now includes Providence.

After filling in a number of the fields, the visualization shows the locations of SIGMOD conferences (Figure 5). The year of the conference is shown by the color of the circle; this data is pulled directly from the original table. Size of the circle shows the relative population of the host cities. Vispedia automatically finds this information by traversing the Wikipedia link graph.

In addition, Vispedia generates a data table showing the integrated results. Clicking on a column reveals the data provenance, the link path followed through Wikipedia to locate the used data (Figure 6). The user can click through to Wikipedia to see the source of the underlying data.

In a few minutes, the user has created a useful visualization from Wikipedia. During this process, he or she interactively performed time-consuming data integration tasks using a simple, easy to understand interface.

3. TECHNICAL DETAILS

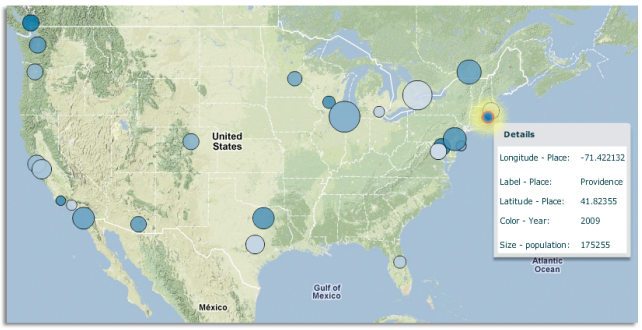

Figure 5: A map created by the Vispedia user showing the location of SIGMOD host cities in the United States and Canada and their populations. Darker circles are more recent venues.

The underlying semantic graph used by Vispedia is part of the January 2008 release of DBpedia 3.0 [3] that includes information extracted from the infoboxes, redirections, images, and geocoordinates in Wikipedia. This subset has 13 million nodes representing articles and literal values, and 36 million labelled edges corresponding to attributes and hyperlinks. We intend to incorporate other graph-structured data in the future.

We opted to store this graph in 4GB of RAM using our own memory-mapped property list implementation backed by a Berkeley DB for the strings in node and edge labels. Existing semantic graph stores [10, 1] are not designed to handle the best-first graph enumeration algorithms that we use.

As described in the scenario, users begin by selecting a table from a Wikipedia article. Vispedia transforms the current version of this table into a graph representation, using a node for each row, and it connects this graph with the DBpedia graph.

Next, Vispedia performs data integration through an interactive process that resembles *schema matching* without a predefined ontology. Each attribute in the visualization, like the size of a circular marker, is described by an expected XML datatype and a set of keywords provided by the user. Vispedia matches this description to data found in the semantic graph by following link paths.

The system starts from each table row, enumerates paths to data that meets the datatype constraint, and returns a list of the most relevant paths by computing a similarity metric between the keywords and path edge labels. This ranked list of paths summarizes the heterogenous graph structure for the user and helps them iteratively refine their query. To make this approach work interactively, we use a time-limited A* search, a best-first search that explores available paths in order of decreasing relevance until 1 second has elapsed.

More details on ranking method and and graph search algorithm are contained in our previous papers [5, 6].

ide ▸]	Label [Place]	Size [population]	Color [Year]
.42	Providence	**Place**: Providence → **redirect**: Providence, Rhode Island → **populationTotal**: 175255	2009
3.1	Vancouver	**Place**: Vancouver → **populationTotal**: 578041	2008
.39	Beijing	**Place**: Beijing → **populationTotal**: 17430000	2007
.62	Chicago	**Place**: Chicago → **populationTotal**: 2833321	2006
.62	Baltimore	**Place**: Baltimore → **redirect**: Baltimore, Maryland → **populationTotal**: 640961	2005
.33	Paris	**Place**: Paris → **name**: 2988507/ → **population**: 2138551	2004
.15	San Diego	**Place**: San Diego → **redirect**: San Diego, California → **populationTotal**: 1256951	2003
.38	Madison	**Place**: Madison, Wisconsin → **populationTotal**: 223389	2002

Figure 6: The resulting integrated table. Expanded columns reveal the data provenance and links back to the Wikipedia source.

4. RELATED WORK

The challenge of performing data integration at web-scale has been explored by others. The Cimple [7] project created a web platform for supporting the social aspects of community-driven data integration. Madhavan et al. promote a formal "Pay as You Go" [11] model where web-scale data integration is done on demand and implicit or explicit user feedback is incorporated to improve the results. Like our project, they use queries and ranking to enable integration. In concurrent work, Cafarella et al. [4] describe how to extract and integrate tables from web sources. Google Base [9] and Freebase [8] are two commercial ventures whose goal is to create large collaboratively-authored semantic databases.

In contrast to these systems, Vispedia emphasizes visualization as the major focus, and as part of a sensemaking loop rather than as the end point of a data integration pipeline.

5. CONCLUSION

We have described a novel system, Vispedia, that combines visualization and data integration into a single system supporting interactive data exploration.

6. REFERENCES

[1] Virtuoso. http://www.openlinksw.com/.

[2] R. Agrawal, A. Ailamaki, P. A. Bernstein, E. A. Brewer, M. J. Carey, S. Chaudhuri, A. Doan, D. Florescu, M. J. Franklin, H. Garcia-Molina, J. Gehrke, L. Gruenwald, L. M. Haas, A. Y. Halevy, J. M. Hellerstein, Y. E. Ioannidis, H. F. Korth, D. Kossmann, S. Madden, R. Magoulas, B. C. Ooi, T. O'Reilly, R. Ramakrishnan, S. Sarawagi, M. Stonebraker, A. S. Szalay, and G. Weikum. The Claremont report on database research. *SIGMOD Rec.*, 37(3):9–19, 2008.

[3] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proc. ISWC 2007*.

[4] M. J. Cafarella, A. Y. Halevy, D. Z. Wang, E. W. 0002, and Y. Zhang. Webtables: exploring the power of tables on the web. *PVLDB*, 1(1):538–549, 2008.

[5] M. Cammarano, X. L. Dong, B. Chan, J. Klingner, J. Talbot, A. Halevey, and P. Hanrahan. Visualization of heterogeneous data. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1200–1207, 2007.

[6] B. Chan, L. Wu, J. Talbot, M. Cammarano, and P. Hanrahan. Vispedia: Interactive visual exploration of wikipedia data via search-based integration. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1213–1220, 2008.

[7] A. Doan, R. Ramakrishnan, F. Chen, P. DeRose, Y. Lee, R. McCann, M. Sayyadian, and W. Shen. Community information management. *IEEE Data Eng. Bull.*, 29(1):64–72, 2006.

[8] freebase. `http://www.freebase.com`.

[9] Google Base. `http://base.google.com/`.

[10] Jena. http://jena.sourceforge.net/.

[11] J. Madhavan, S. Cohen, X. L. Dong, A. Y. Halevy, S. R. Jeffery, D. Ko, and C. Yu. Web-scale data integration: You can afford to pay as you go. In *CIDR*, pages 342–350, 2007.