WiGiPedia: Visual Editing of Semantic Data in Wikipedia

Svetlin Bostandjiev* **John O'Donovan** **Brynjar Gretarsson** **Christopher Hall**
Tobias Höllerer†
Department of Computer Science, University of California Santa Barbara

INTRODUCTION

Wikipedia is emerging as the dominant global knowledge repository. Recently, large numbers of Wikipedia users have collaborated to produce more structured information in the online encyclopedia. For example, the information found in tables, categories and infoboxes. Infoboxes contain key-value pairs, manually appended to articles based on the unstructured text therein. The wiki contains some structured information which can be crawled by DBpedia [2], which attempts to organize wiki data into into a database of subject-predicate-object triples. By leveraging this data we generate an interface, which we call *WiGipedia*, embedded on every Wiki article as an interactive graph visualization where entities represent articles, categories and relational entities, with typed edges between them. This intelligent web interface is designed to simplify the elicitation of semantically structured information in Wikipedia (Figure 1). Our motivation is to both inform the user of interesting contextual information pertaining to the current article, and to provide a simple way to introduce and/or repair semantic relations between wiki articles. User actions result in improved accuracy and consistency of structured data spread across multiple articles. *WiGipedia* provides users with an intuitive interface that allows single-click edits without knowledge of the Wiki markup language, templates, etc. An online demo of the interface can be found at *www.wigipedia-online.com*.

RELATED WORK

WiGipedia combines facets from semantic web research, focusing on the gathering of rich semantic data in a collaborative manner, and information visualization within a wiki. While there are many forms of visualizations available that will serve our purpose, in this case we focus on a node-link graph because it highlights relations between entities in an intuitive way [3]. *WiGipedia* harnesses semantically rich data from DBpedia that has been originally extracted from wiki infoboxes and categories. DBpedia is queryable in SPARQL and supports complex queries such as "Find similarities between Led Zeppelin, Pink Floyd, and Deep Purple" (i.e. visualization in Figure 1). RelFinder [5] is another tool that supports relational queries over a wiki, however, in contrast to our approach, it cannot be used to modify the data. The edits made in *WiGipedia* modify the wiki markup, which in turn is remined by DBpedia. In this sense, our system enables the user to 'close the loop' between Wikipedia and DBpedia. Additionally, *WiGipedia*'s simple single-click edits address users' difficulties in editing Wikipedia shown in studies such as the Wikimedia Usability Initiative [1]. *WiGiPedia* builds on the *WiGis* toolkit [4] to provide interactive browser-based visualizations (*www.wigis.net*).

USAGE SCENARIO

Figure 1 highlights the steps in the process flow of *WiGipedia*. For Bob, a casual Wikipedia user, the interaction experience occurs as follows: Bob searches Wikipedia for an article of interest, for example, details on his favorite band, Pink Floyd. The article page shows the standard wiki page and an infobox containing a picture with some facts about the band. In addition to the infobox, Bob notices a graph with nodes and edges embedded in the wiki page. The graph contains nodes representing Pink Floyd and a range of other contextually relevant information such as music genres, places of origin, years of performance, and a selection of similar bands, e.g. Led Zeppelin and Deep Purple. Bob highlights a few nodes and notices that he can move them around to reconfigure the entire graph layout into meaningful arrangements which highlight important information. When he is satisfied with his layout, Bob then notices that all three bands on the periphery of the graph are linked to the node "English rock music groups", but only two of them are linked to "England". Bob decides to create a link between the nodes "Pink Floyd" and "England" and a suggestion box appears above the graph. The box contains a drop-down list of recommendations for the edge label. Bob clicks on "Origin" and a labeled edge appears on the graph. Bob also notices that the Wikipedia article infobox has changed, and that the text "England" has appeared and is highlighted in yellow, alongside a green check mark and a red X. Bob clicks on the green check mark and confirms his Wikipedia update. In a similar fashion, Bob can remove or update existing edges.

DESIGN AND IMPLEMENTATION

The goal of the visualization is to explain how data in the current Wikipedia article relates to similar data spread across related articles, and allow for discovery and correction of inconsistencies in the data: passively as exploration tool, and actively as contribution interface. The graph is comprised of two tiers of nodes. The "principal set" of nodes, representing the initial set of wiki articles, are arranged around a fixed circular perimeter and rendered in dark blue. The sub-nodes can

*alex@cs.ucsb.edu
†holl@cs.ucsb.edu

Workshop on Visual Interfaces to the Social and Semantic Web (VISSW2011),
Co-located with ACM IUI 2011,
Feb 13, 2011, Palo Alto, US.

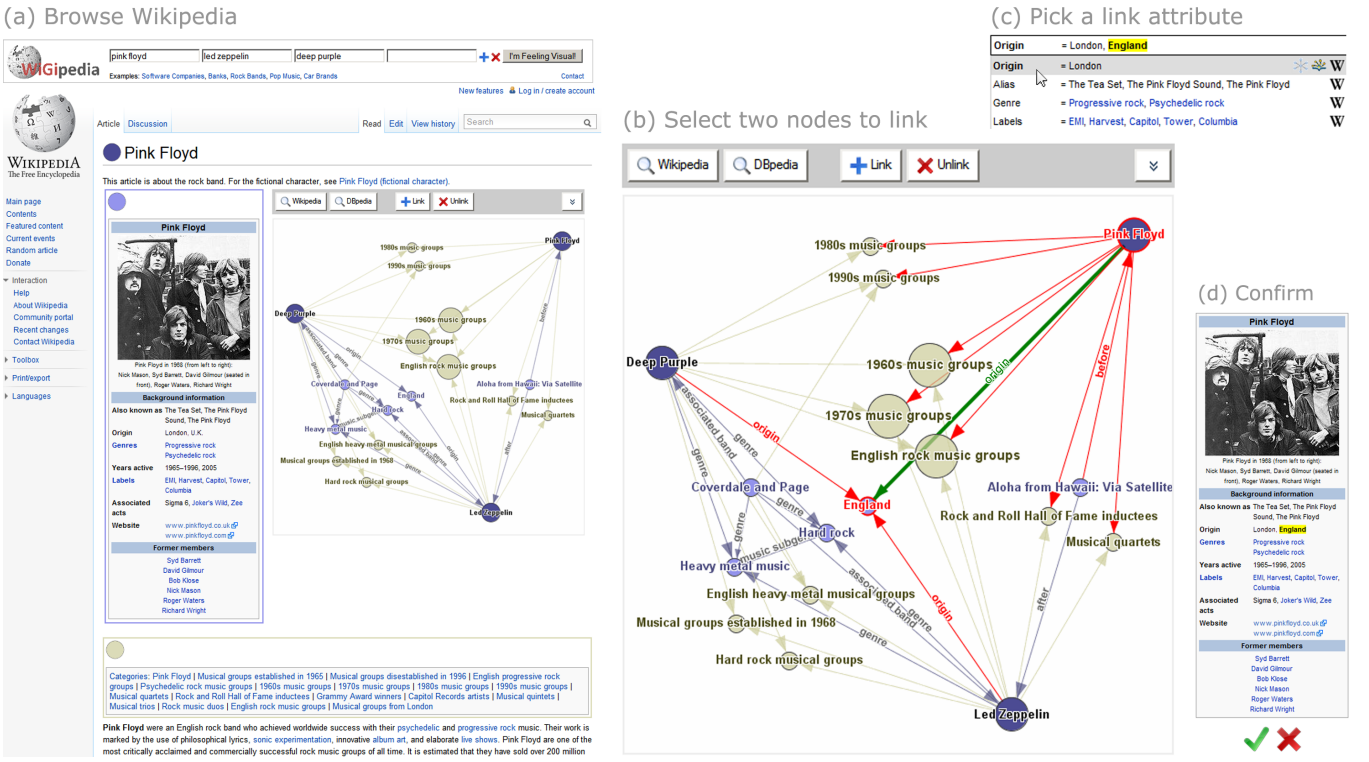

Figure 1. WiGiPedia Components: (a) shows a wiki article for "Pink Floyd" with an embedded contextual graph of semantically linked articles. In (b) dark blue nodes represent the starting point for the graph generation. In this case three English bands: "Pink Floyd", "Led Zeppelin", and "Deep Purple". The remainder of the graph represents commonalities between seed articles: light blue nodes are other wiki articles (i.e. "England") and tan nodes are Wikipedia categories (i.e. "1980s music groups"). A user created edge is highlighted in green. (c) shows the drop-down menu with a list of suggested labels for the new edge. In this example "Pink Floyd" and "England" are linked by "Origin". (d) shows a confirmation step that occurs before the new update is propagated to Wikipedia.

be either categories (rendered in tan), or article pages themselves (rendered in light blue), and represent an item held in common between at least two nodes in the principal set. Every node-edge-node triple corresponds to a subject-predicate-object RDF triple extracted from DBpedia. To help the user make the connection between the types of graph nodes and their corresponding article elements, a circle of matching color is placed next to the wiki article title, infobox, and category list. To generate the graph we begin with the current Wikipedia article the user has navigated to, along with a selection of related articles picked from the most highly linked wiki articles in the current article. These articles form the principal set of nodes. Next, a DBpedia SPARQL endpoint is queried for the relations held in common between each pair of nodes in the principal set. To lay out the graph the principal nodes are pinned around an outer circle and then a Fruchterman-Reingold force directed algorithm is applied to position all other nodes. When the user creates a new link between two nodes a set of candidate edge label recommendations is generated. This further facilitates simple updates, but also upholds consistency and coherence within the existing Wikipedia corpus. We rank edge label suggestions from three different but interdependent sources, each providing a unique angle: the displayed graph, a semantic DBpedia query, and the source article's infobox template.

CONCLUSION

While recent collaborative efforts have provided more structured data such as infoboxes to Wikipedia, large amounts of data remains inconsistent, incorrect or missing. This paper introduced *WiGipedia*, a novel interface that facilitates the input of semantic information from regular Wikipedia users. By querying DBpedia for semantic relations between a selected set of articles and generating a graph visualization, the system provides context to the article being read. However, the main contribution of *WiGipedia* is as an input modality, supporting single-click semantic updates to Wikipedia, based on a users comprehension of relations in the graph. Looking forward, we believe that contributing to the consistency and structure of Wikipedia is a step towards the creation of a rich Wikipedia ontology, capable of supporting complex analytical queries over this huge knowledge repository.

ACKNOWLEDGEMENT

This research was partially sponsored by an ARO MURI award for proposal #56142-CS-MUR and by funding from the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

1. Wikimedia usability initiative, 2010. http://usability.wikimedia.org.
2. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*.
3. C. Chen. *Information Visualization*. Springer, July 2004.
4. B. Gretarsson, S. Bostandjiev, J. O'Donovan, and T. Höllerer. Wigis: A scalable framework for web-based interactive graph visualizations. In *GD'09: Proceedings of the International Symposium on Graph Drawing*, 2009.
5. P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, and T. Stegemann. Relfinder: Revealing relationships in rdf knowledge bases. In *SAMT*, volume 5887 of *Lecture Notes in Computer Science*, pages 182–187. Springer, 2009.