Integration of Wikipedia and a Geography Digital Library*

Ee-Peng Lim, Z. Wang, D. Sadeli, Y. Li, Chew-Hung Chang, Kalyani Chatterjea, Dion Hoe-Lian Goh, Yin-Leng Theng, Jun Zhang, and Aixin Sun

Nanyang Technological University, Singapore

Abstract. In this paper, we address the problem of integrating Wikipedia, an online encyclopedia, and G-Portal, a web-based digital library, in the geography domain. The integration facilitates the sharing of data and services between the two web applications that are of great value in learning. We first present an overall system architecture for supporting such an integration and address the metadata extraction problem associated with it. In metadata extraction, we focus on extracting and constructing metadata for geo-political regions namely cities and countries. Some empirical performance results will be presented. The paper will also describe the adaptations of G-Portal and Wikipedia to meet the integration requirements.

Keywords: Web-based encyclopedia, geography digital libraries, integration.

1 Introduction

G-Portal is a digital library that supports learning of geography based on publicly available web information[1, 4, 5]. In G-Portal, we treat geography related web pages and other types of web objects as content resources. Metadata records of these web objects are created and stored in a database server, and can be shared by different learning projects. G-Portal also provides the map-based and classification-based interfaces to view and query metadata records.

Wikipedia represents a very appealing source of information to G-Portal as it contains many entries related to geography in addition to its many interesting features mentioned earlier. Wikipedia has 44 sub-categories under the *geography* category, and at least another 600 sub-categories at the next level. Each geography entry or article is jointly authored and edited by a group of Wikipedians from different parts of the world, and hence is of quality generally much better than the other non-reviewed web pages. With so much knowledge embedded in Wikipedia, it is interesting to explore how the Wikipedia resources can be utilized in research and education.

Since Wikipedia is not a digital library, it lacks several important features that are essential to digital library users. At present, metadata of Wikipedia entries

* The project is funded by Centre for Research in Pedagogy and Practice, National Institute of Education with Project ID "CRP25/04 LEP".

S. Sugimoto et al. (Eds.): ICADL 2006, LNCS 4312, pp. 449–458, 2006.

are very basic and it does not provide sufficient information for searching and browsing at metadata level. Wikipedia has strong system features for users to edit entries, while digital libraries have better support for tracking information of interest to each user or a group of users. For example, G-Portal allows metadata records to be selected into different groups known as *projects*, and each project is designed for the learning needs of one or more user.

There are clearly many benefits accrued from integrating Wikipedia and G-Portal. Firstly, the addition of Wikipedia metadata to G-Portal will not only enlarge the latter's metadata collection, but also create a special web content resources that are of high quality and are constantly updated with the latest information. Secondly, users can now exploit the use of G-Portal features to access the metadata of Wikipedia entries and to further organize them for different learning purposes. For example, one can create a G-Portal project consisting of metadata of both Wikipedia and non-Wikipedia content for some geography learning activity.

In this research, we have therefore investigated different approaches to integrate Wikipedia and G-Portal. On one extreme is a tightly coupled integrated system with a single user interface supporting a combined set of functions based on a single database. On the other extreme is a loosely coupled system with Wikipedia and G-Portal retaining their original user interfaces, functions and databases. For the practical reason of keeping each system autonomous, we have chosen the latter integration approach. This approach minimizes changes to Wikipedia and G-Portal so that both can continue to service their own existing users. Nevertheless, the loosely coupled integration still poses some technical challenges, namely identifying relevant Wikipedia entries, and extracting metadata attribute information from the entries.

Our work focuses on constructing the metadata for geo-political regions such as countries and cities. Geo-political regions form a basic but significant and important set of entries in Wikipedia (and also important to geography in general). The descriptions of these regions are usually very comprehensive in Wikipedia and they often serve as the background context for discussing other geography concepts.

We aims to develop new extensions to G-Portal to make it easy to invoke G-Portal from Wikipedia with the requested metadata record shown prominently in its user interface. At the same time, G-Portal's metadata embeds links to access Wikipedia entries.

2 System Architecture for the Integrated System

The essential system modules of Wikipedia and G-Portal, together with those modules for integrating them are shown in Figure 1.

Wikipedia itself is implemented using a server engine known as MediaWiki that maintains a large collection of entries in a MySQL database[7]. To display a Wikipedia page in a web browser, MediaWiki retrieves the user requested entry from the database, formats it in HTML and returns the formatted HTML page[3].

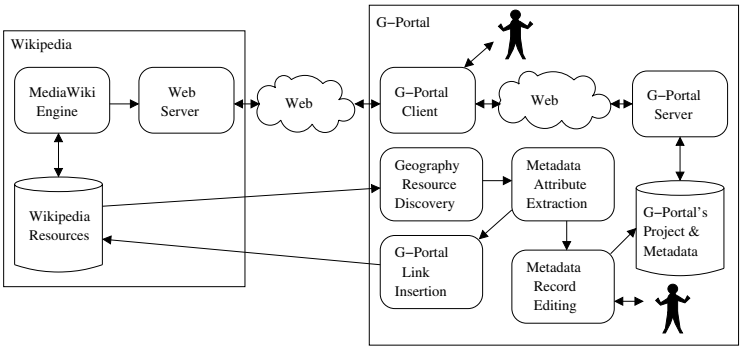


Fig. 1. System Modules for the Integration of Wikipedia and G-Portal

G-Portal, on the other hand, maintains its metadata and project information in its own database implemented on an XML database system and another relational database system. G-Portal has a server module that handles retrieval and updates of its database. Upon user request, G-Portal client obtains metadata records under the selected project from the server and displays them.

To support a two-way integration framework, a user viewing a Wikipedia entry using a web browser should be able to invoke G-Portal client whenever required to display the metadata record of the Wikipedia entry. On the other hand, a user of G-Portal should be able to view the metadata records of geography-related Wikipedia entries using the G-Portal client.

The metadata records of geography-related Wikipedia entries are constructed by two modules, namely the geography resource discovery and metadata element extraction module as shown in Figure 1. Geography resource discovery involves scanning through the Wikipedia entries in the database and finding those that are related to the various geography topics.

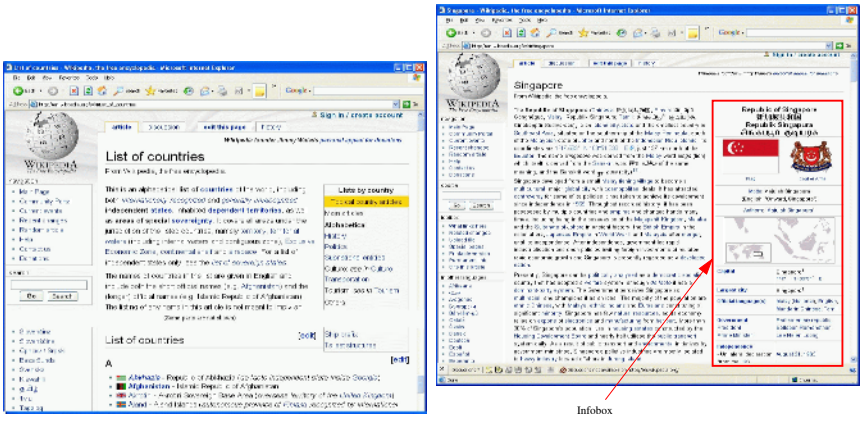
Once the relevant Wikipedia entries are found, their metadata records have to be constructed by extracting metadata attributes and their values from the Wikipedia page content. This metadata creation process can be a painstaking task for large collection of entries if it is done manually. In many cases, even with automated metadata attribute extraction, the metadata records still have to be further manually edited before they can be imported into G-Portal.

As new metadata records are constructed and added to G-Portal, the Wikipedia pages need to be modified to have the links to G-Portal Client inserted. These links will allow Wikipedia users to invoke the G-Portal Client which displays the corresponding metadata record.

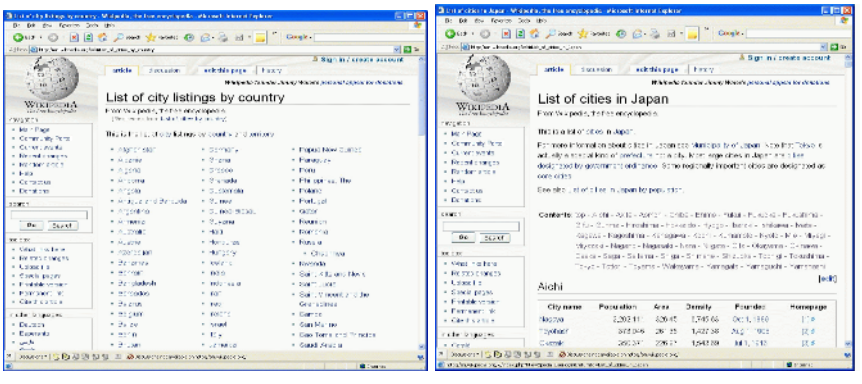
3 Identification of Relevant Wikipedia Articles

3.1 Organization of Wikipedia Entries for Geo-political Regions

Every Wikipedia entry is associated with an article that contains the entry description and this also applies to geo-political regions entries. In addition to



(a) List of Countries" page (b) "Singapore" page



(c) "List of Cities by Country" page (d) "List of Cities in Japan" page

Fig. 2. Wikipedia Pages

having an article to describe each geo-political region (e.g., country, state, city, etc.), Wikipedia also supports some *List_of_⟨region⟩* entry that enumerate all articles of some ⟨region⟩ type. We call these the *list-type* pages. For example, a list of countries covered by Wikipedia can be found at: "http://en.wikipedia.org/wiki/List_of_countries" as shown in Figure 2(a). The link of each country leads to a Wikipedia article about the country as shown in Figure 2(b).

A list-type page also exists for cities grouped by countries (see Figure 2(c)). This page provides the link to a list of cities page for each country. For example, the list-type page for the cities of Japan is shown in Figure 2(d). Every city of Japan can be found by navigating the links from this page.

The existence of such list-type pages simplifies the task of identifying geo-political regions covered by Wikipedia and extracting the corresponding entries. Nevertheless, there are still some complications caused by the nature of Wikipedia:

- The links to country entries in the "List_of_countries" article are formatted manually and there are other links in the article not linked to country entries.
- The city entries in the "List_of_cities_in⟨country ⟩" also demonstrate a wide variety of formats across different countries complicating the way they can be identified. For example, the way Japan cities are formatted in the "List_of _cities _in_Japan" entry is different from those formatted in the "List_of_cities _in_China" entry.
- Some regions may not be covered by Wikipedia but their links may still exist in the list-type entry. A city name may appear in the "List_of_cities_in_China" entry but there is no corresponding Wikipedia entry.

While the above observation of list-type entries were made on geo-political entries, we also found similar style of organizing entries in other subject domains, e.g., list of mathematics topics, list of songs. This suggests that our methods developed for geo-political entries can be adopted for entries of other domains.

3.2 Entry Link Extraction Algorithm for List-Type Pages

In this section, we proposed an algorithm to extract the links to entries for geo-political regions. We focus on the extraction of links to country and city entries since they are the ones with list-type pages.

Our algorithm makes the following assumptions:

- For each geo-political region type (say, city), there is a single root list-type page that provides the links to the correspoding entries either directly or indirectly via a set of other intermediate list-type pages (e.g., List_of_cities_by_country page). This root page serves as the input to our algorithm.
- A set of region names are available for training purpose. Very often, names of some well known regions are already known in the public domain. For example, a set of countries and cities are available at ESRI's website. These training sets may not be able to cover all countries and cities in Wikipedia[1], they serve as useful information for training our extraction patterns.

To avoid developing different extraction algorithms for list-type pages in Wiki-Text and HTML, our algorithm first represents a WikiText or HTML page as a DOM tree. The DOM tree consists of leaf nodes each representing a basic text segment and internal nodes each representing a tagged document element in the WikiText or HTML page. The children of an internal node represents a

[1] The coverage of countries is usually much better than that of cities in the training set.

set of text segments or document elements nested within the parent document element.

As shown in Algorithms 1 and 2, our algorithm first collects a set of tag paths (*PathSet*) for the training regions found in the DOM tree. The tag paths are then used to extract all the region names by traversing the DOM tree as shown in Algorithm 2. Due to space constraint, we will leave out the experimental results of these algorithms. In general, they performed reasonably well in finding the country and city regions from the list-type pages.

Algorithm 1 Extract_Entry_Links(T, C)

Input: DOM Tree T,
A set of training region names C
Output: extracted entry links L

1: initialize $L := \phi$; initialize $PathSet := \phi$;
2: Construct_Path($T.root$, C, (), $PathSet$)
3: Extract_Leaf_by_Path($T.root$, $PathSet$, (), L)
4: **Return**

Algorithm 2 Construct_Path(n, C, $Path$, $PathSet$)

Input: Tree node n,
A set of training region names C
Input/Output: Tag path $Path$ (initially ())
A set of tag paths $PathSet$ (initially ϕ)

1: **if** n is leaf node **then**
2: **if** $n.text$ contains some $c \in C$ **then**
3: Add $Path$ to $PathSet$
4: **end if**
5: **else**
6: Push $tag(n)$ to $Path$
7: **for** each child n' of n **do**
8: Construct_Path(n',C,$Path$,$PathSet$)
9: **end for**
10: Pop a tag from $Path$
11: **end if**
12: **Return**

Algorithm 3 Extract_Leaf_by_Path(n, $PathSet$, $Path$, R)

Input: Tree node n (initially $T.root$)
A set of paths $PathSet$
Current tag path $Path$ (initially ())
Input/Output: A set of region names R (initially ϕ)

1: **if** n is leaf node **then**
2: **if** $Path \in PathSet$ **then**
3: Add n to R
4: **end if**
5: **else**
6: Push $tag(n)$ to $Path$
7: **for** each child n' of n **do**
8: Extract_Leaf_by_Path(n',$PathSet$,$Path$,R)
9: **end for**
10: Pop a tag from $Path$
11: **end if**
12: **Return**

Fig. 3. Algorithms for Identifying Geo-political Entries

4 Extraction of Metadata Attributes

Instead of taking a text extraction approach looking for possible metadata attributes and values from Wikipedia pages, we examine the *infoboxes* commonly found in the Wikipedia entries. An infobox is essentially a table included in articles with a common subject, e.g. country, city, etc.. Infobox templates are also available for the tables to have a common look in different Wikipedia articles. An example of a country infobox is shown in Figure 2(b) where the infobox contains information about the capital, largest city, official language, government, and others. It turns out that many country and city articles in Wikipedia

contain good quality metadata attribute information in infoboxes. Infoboxes are also widely used for representing metadata attributes of Wikipedia entries from other domains (e.g., science, technology, arts and entertainment, etc.).

The task of extracting metadata attribute information from infoboxes consists of two parts: (i) identifying the existence of infoboxes (as not all infoboxes contain the metadata information, and some pages may not even contain a relevant infobox); and (ii) extracting the metadata attributes and their corresponding attribute values. To tackle the task, we adopted a set of training metadata attributes for recognizing the relevant infoboxes, as well as for recognizing their formats. The set of training metadata attributes for country (denoted by $A_{country}$) was first obtained by manual inspection from the relevant infobox of USA Wikipedia page while the training metadata attributes (A_{city}) was obtained from the Beijing Wikipedia page in a similar way. We chose them because of their rich sets of metadata attributes in the infoboxes. There were 15 attributes in $A_{country}$ (namely, Motto, Anthem, Capital, Largest city, Official languages, Government, Independence, Constitution, Area, Population, GDP, Currency, Time Zone, Internet TLD, and Calling Code), and 17 attributes in A_{city} (namely, Origin of name, Administration type, CPC Beijing Committee Secretary, Mayor, Area, Population, GDP, Major nationalities, City trees, City flowers, Country-level divisions, Township-level divisions, Postal code, Area code, Licence plate prefixes, ISO 3166-2, and Official website).

To identify the relevant infobox from a Wikipedia page, the infoboxes in the page are ranked by the numbers of metadata attributes from $A_{\langle region \rangle}$ found in them. The highest ranked infobox is chosen for the second sub-task of metadata extraction where columns in the infobox are further ranked by the number of metadata attributes. The column with the highest rank will be assumed to contain the metadata attribute names while the other columns are treated as attribute values. Each metadata attribute value pair is then extracted from the row corresponding to the metadata attribute.

5 Modification of G-Portal Client and Wikipedia Entries

The access to G-Portal metadata from Wikipedia allows Wikipedia users to easily switch from the web browser's page-based view of a resource to the G-Portal's map-based and category-based views of the corresponding metadata record together with the metadata records of other Wikipedia resources. This switch is extremely useful when one wishes to visualize different Wikipedia resources on a map or in a hierarchy of categories.

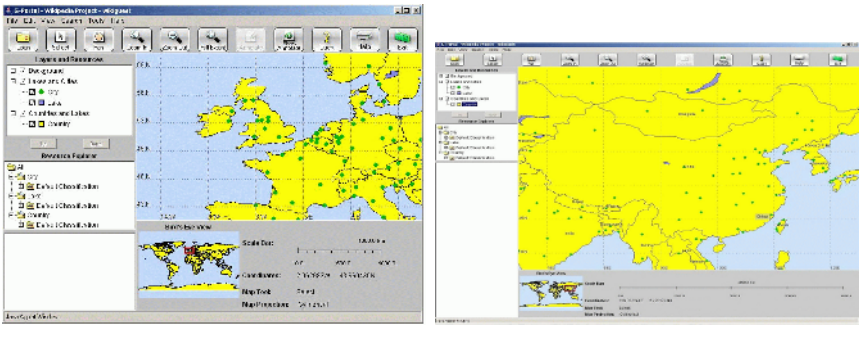
To provide access to G-Portal metadata from Wikipedia, there are some modifications to be made to both G-Portal and Wikipedia, namely: (a) Automatic login for Wikipedia project; (b) Parameterized record centric display for G-Portal's map-based interface; and (c) Insertion of G-Portal client links into Wikipedia resources. The first two affects the design of G-Portal while the last involves some changes to Wikipedia articles.

5.1 Automatic Login

G-Portal users are usually required to manually log in with their user names and passwords before they are allowed to view the metadata of selected projects. In contrast, Wikipedia supports public access and passwords are not required. Furthermore, G-Portal users are also required to select the specific project for viewing metadata. For Wikipedia users to directly access G-Portal metadata without having to worry about logging in and selecting projects, G-Portal has to be modified to receive login and project parameters upon invocation. This is done by extending the G-Portal client applet to take the above input parameters. A PHP script (known as "gportal.php") is also generated for Wikipedia users to invoke G-Portal client applet and passing to the latter the user ID, password and name of the corresponding project. In our case, there is so far only one project created for Wikipedia resources and it is known as "Wiki Project". Note that the above mechanism can be easily extended for other G-Portal projects whenever automatic login and project selection are required.

5.2 Parameterized Metadata Centric Display

G-Portal, in its original user interface design, displays the full extent of the map of a project before the user starts browsing the metadata within the project. This design assumes that the user has no specific metadata record to view at the beginning. This assumption however does not hold when we integrate G-Portal with Wikipedia since Wikipedia users access G-Portal from some Wikipedia resources and it is necessary to show clearly where the metadata of these resources are located in the map interface. A simple solution to this is to highlight a selected metadata record in the map interface but this method suffers from some shortcomings. Firstly, there may be many other metadata records located nearby. A full map extent display thus will not be able to display the highlighted



(a) London at the center (b) China in the map interface

Fig. 4. G-Portal's map interface with computed bounding rectangle region

one clearly. Secondly, metadata records on the map are of different shape and sizes, it is visually unpleasant to show a highlighted metadata record using the same zoom level.

We address the above issues by making G-Portal client display the map region with the selected metadata record at the center and at the appropriate zoom level. We have extended G-Portal client to accept a geographical region defined by a bounding rectangle as another input parameter. To provide this input to the G-Portal client, the G-Portal script "gportal.php" (for invoking the G-Portal client) has been further modified to (1) take resource id of the metadata record; (2) retrieve the spatial shape of the record from the G-Portal's database; and (3) compute the bounding rectangle containing the record at the center with the appropriate zoom level. For point represented metadata, the point location and a pre-defined zoom level is used to determine the bounding rectangle (see Figure 4(a)). For polygon or line represented metadata, the script computes the bounding box containing the shape (see Figure 4(b)).

5.3 Insertion of G-Portal Client Links into Wikipedia Resources Using Reverse Proxy

For Wikipedia users to directly call on G-Portal, it is necessary to insert links to G-Portal client (or more accurately, the links to the G-Portal PERL script which invokes the client) within the Wikipedia resource web pages. The insertion can be automated by having a program editing the content of Wikipedia pages. Nevertheless, this insertion will inevitably modify the operational Wikipedia entries and affect the use of Wikipedia. To minimize changes to Wikipedia, we have chosen to implement a reverse proxy[9] for Wikipedia. Our reverse proxy is a virtual mirror of Wikipedia which takes a URL request for Wikipedia page and forwards it to real Wikipedia server so as to receive the corresponding HTML page. As part of the proxy, we developed a filter to add appropriate G-Portal links to the received HTML page if the request concerns a country or city entry before the page is further returned to the client of reverse proxy. Furthermore, the filter modifies all links in the HTML page by their proxy links so as to ensure that further browsing from this page will also involve the reverse proxy.

6 Conclusion

Interoperability among digital libraries has been an very active area of research. There are standardization efforts for metadata representations such as Dublin Core[2], and protocols for sharing and querying metadata records, e.g., OAI[6], Z39.50[8]. The work however focuses more on the integration among different digital libraries as opposed to between digital libraries and non-digital libraries. We have also observed that the ongoing development of Wikipedia has very much on content construction and content review activities instead of integrating with digital libraries. The research issues related to integrating Wikipedia with digital libraries are therefore very much unexplored.

In this paper, we therefore studied the task of finding and extracting metadata from Wikipedia articles. We specifically focused on the metadata of Wikipedia articles of geo-political regions. Our experiments have shown that our proposed extraction algorithms performed quite well on country and city articles. This paper also reports how G-Portal and Wikipedia have to be modified to support the proposed integration. Interested users can try out G-Portal's Wiki Project at http://gportal.cais.ntu.edu.sg/GPortal/index.htm.

References

1. Chew Hung Chang, John Hedberg, Yin-Leng Theng, Ee-Peng Lim, Tiong-Sa Teh, and Dion Hoe-Lian Goh. Evaluating G-Portal for Geography Learning and Teaching. In *ACM/IEEE Joint Conf. on Digital Libraries*, Denver, USA, June 2005.
2. Makx Dekkers and Stuart Weibel. Dublin Core Metadata Initiative Progress Report and Workplan for 2002. *D-Lib Magazine*, 8(2), 2002.
3. Wikimedia Foundation. http://www.mediawiki.org/wiki/mediawiki.
4. Dion Hoe-Lian Goh, Aixin Sun, Wenbo Zong, Dan Wu, Ee-Peng Lim, Yin-Leng Theng, John Hedberg, and Chew Hung Chang. Managing Geography Learning Objects Using Personalized Project Spaces in G-Portal. In *European Conf. on Research & Advanced Technology for Digital Libraries*, Vienna, September 2005.
5. Ee-Peng Lim, Dion Hoe-Lian Goh, Zehua Liu, Wee-Keong Ng, Christopher Soo-Guan Khoo, and Susan Ellen Higgins. G-Portal: A Map-based Digital Library for Distributed Geospatial and Georeferenced Resources. In *ACM/IEEE Joint Conference on Digital Libraries*, Portland, July 2002.
6. Xiaoming Liu, Kurt Maly, Mohammad Zubair, and Michael L. Nelson. Repository Synchronization in the OAI Framework. In *ACM/IEEE Joint Conf. on Digital Libraries*, Houston, May 2003.
7. MySQL. www.mysql.com/.
8. Michalis Sfakakis and Sarantos Kapidakis. Expression of Z39.50 Supported Search Capabilities by Applying Formal Descriptions. In *European Conf. on Research and Advanced Technology for Digital Libraries*, Vienna, September 2005.
9. Wikipedia. Reverse proxy — wikipedia, the free encyclopedia, 2006. [Online; accessed 19-June-2006].