Facetedpedia: Enabling Query-Dependent Faceted Search for Wikipedia

Ning Yan, Chengkai Li, Senjuti B. Roy, Rakesh Ramegowda, Gautam Das
Department of Computer Science and Engineering, University of Texas at Arlington
Arlington, TX, USA
ning.yan@mavs.uta.edu,cli@uta.edu,
{senjuti.basuroy,rakesh.ramegowda}@mavs.uta.edu, gdas@uta.edu

ABSTRACT

Facetedpedia is a faceted search system that dynamically discovers query-dependent faceted interfaces for Wikipedia search result articles. In this paper, we give an overview of Facetedpedia, present the system architecture and implementation techniques, and elaborate on a demonstration scenario.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Design, Experimentation

Keywords: faceted search, data exploration, Wikipedia

1. INTRODUCTION

Faceted search [2] is a useful technique for information exploration, especially when a user needs to browse through a long list of articles or objects, which, without any auxiliary facility, could be time consuming and painstaking. A faceted interface for a set of objects is a set of category hierarchies, where each hierarchy corresponds to an individual facet (dimension, attribute, property) of the objects. In a facet, the user can navigate through the hierarchy of categories and ultimately a specific "property" value if necessary, thus reaching those objects associated with the categories and the value. The user navigates multiple facets and the intersection of the chosen objects on individual facets are brought to the user's attention.

In [3] we developed Facetedpedia (`http://idir.uta.edu/facetedpedia`), a faceted search system for Wikipedia. It focuses on the *dynamic* discovery of *query-dependent* faceted interfaces. Given a set of top-s ranked Wikipedia articles as the result of a keyword search query, Facetedpedia produces an interface of multiple facets for exploring the result articles. The facets cannot be pre-computed due to the query-dependent nature of the system. In applications where faceted interfaces are deployed for relational tuples or schema-available objects, the tuples/objects are captured by prescribed schemata with clearly defined dimensions, therefore a query-independent static faceted interface, either manually or automatically generated, may suffice. By contrast, Wikipedia articles are lacking such pre-determined dimensions that could fit all possible dynamic query results, therefore efforts on static facets would be futile.

Facetedpedia is a challenging undertaking. The concept of faceted interface is built upon two pillars: facets and the category hierarchy associated with each facet. Web pages, unlike relational tables, are lack of predefined schemata and controlled vocabulary that would otherwise readily provide the facet dimensions and category hierarchies. Therefore we must answer two questions: (1) *facet identification*– What are the facets of an article?; and (2) *hierarchy construction*– Where does the category hierarchy of a facet come from?

The gist of our approach is to exploit the *collaborative vocabulary* in Wikipedia, including the hyperlinks in articles, the categories of articles, and the hierarchical relationships between different categories. With regard to the concept of facet, the Wikipedia articles highly related to (e.g., hyperlinked from) a search result article are exploited as its attributes. With regard to the concept of category hierarchy, the "grassroots" category system in Wikipedia provides the category-subcategory relationships for the category hierarchy on a facet dimension.

Given the sheer size and complexity of Wikipedia, the space of possible faceted interfaces for a dynamic query is prohibitively large. Therefore in solving the faceted interface discovery problem, we addressed the following two issues:

Facet Ranking Metrics: A faceted interface is for a user to navigate through the associated category hierarchies and to finally reaching the target articles. Therefore we proposed metrics for ranking individual facet hierarchies by user's navigational cost. Moreover, the utilities of multiple facets do not necessarily build up linearly: Since the facets in an interface should ideally describe diverse aspects of the result articles, a set of individually "good" facets may not be "good" collectively. We thus further designed metrics for ranking interfaces (each with k facets) by both their average pairwise similarities and average navigational costs.

Facet Search Algorithms: It is infeasible to directly apply the above ranking metrics exhaustively on all possible choices, due to the prohibitively large search space. Furthermore, the interactions between the facets in a faceted interface make the computation of its exact "goodness" score intractable. We thus developed greedy algorithms and top-k algorithms that optimize the ranking metrics, for effective and efficient search of the "good" faceted interfaces.

To the best of our knowledge, Facetedpedia is the first query-dependent faceted search system. Existing research prototypes or commercial systems mostly cannot be applied to meet our goals, because they either are based on manual or static facet construction, or are for structured records or text collections with prescribed metadata. None of the existing systems is fully dynamic in both facet identification and hierarchy construction. Recently, another faceted inter-

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0099-5/10/10.

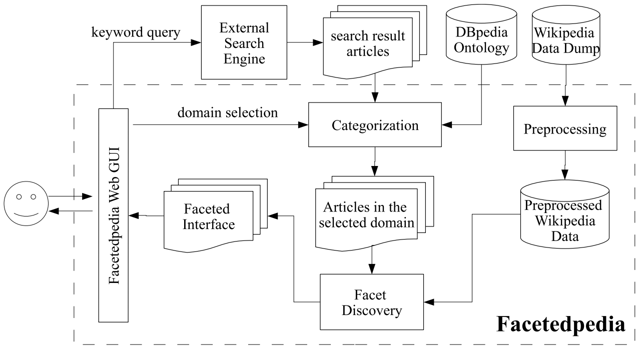


Figure 1: The architecture of Facetedpedia.

face for Wikipedia has been developed as part of the DBpedia project [1]. However, they use the same facets for articles in the same domain, therefore the faceted interfaces are query independent. The facets are generated based on structured data extracted from infoboxes of Wikipedia articles instead of the textual content of the articles. Detailed review of related systems can be found in [3].

2. SYSTEM IMPLEMENTATION

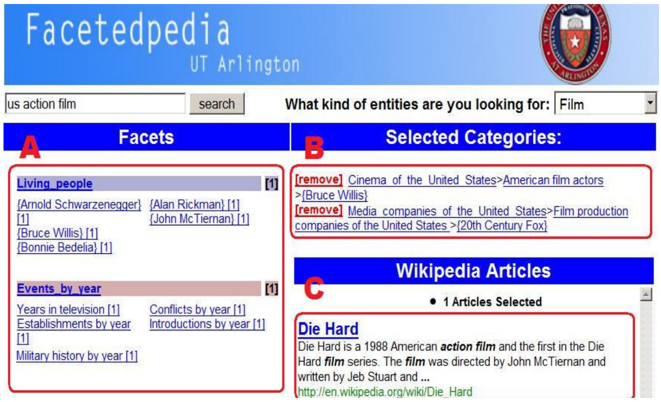
Facetedpedia consists of four major components, as shown in Figure 1. Below we introduce them in more detail.

Preprocessing Wikipedia Data Dump: We used the Wikimedia MySQL data dump generated at July 24th 2008. We imported four tables (Page, Pagelinks, Categorylinks, Redirect) into our local database. We removed the redirect articles and categories (recorded in Redirect table) from the Page and Category tables, then replaced the links to redirected articles (categories) in other tables, by the corresponding non-redirect articles (categories). We also removed the administrative categories in Wikipedia, using simple name patterns. The original Wikipedia category graph contains cycles. We ran a depth-first search to detect and remove around 600 cycles in order to make it a DAG.

Categorization: We believe a faceted interface is only meaningful for a set of homogeneous objects, i.e., articles within the same domain. In our implementation, we exploited the DBpedia ontology[1] for assigning articles to about 80 pre-determined domains (e.g., People, Places, etc). This is done offline and the categorization result of all the Wikipedia articles is stored in a database table.

When the user issues a keyword search query, the query is sent to an external search engine which returns a ranked list of Wikipedia articles. The returned articles are most likely from different domains, thus Facetedpedia asks the user to select one particular domain of her interest. An alternative approach is to let Facetedpedia select the dominant or largest domain of result articles automatically. However, there indeed exist situations when the user would like to select from other domains of her interest. A k-facet interface is then discovered for the top-s search result articles belonging to the chosen domain. (In our implementation, we use Google.com as the external search engine. A typical value of s that we used is 400. The value of k is usually set to 20.)

Facet Discovery: The facet discovery component is a multithread background daemon program. The main process creates a new thread for each user session. The main process pre-loads all the preprocessed tables (1.2GB in total) into memory. After the user chooses a target domain, a new thread is created to run the facet ranking and search algorithms and generate the resulting faceted interface.



us action film

What kind of entities are you looking for: Film

Facets

Living_people [1]

{Arnold Schwarzenegger} [1]
{Bruce Willis} [1]
{Bonnie Bedelia} [1]
{Alan Rickman} [1]
{John McTiernan} [1]

Events_by_year [1]

Years in television [1]
Establishments by year [1]
Military history by year [1]
Conflicts by year [1]
Introductions by year [1]

Selected Categories:

[remove] Cinema_of_the_United_States>American film actors >{Bruce Willis}
[remove] Media_companies_of_the_United_States>Film production companies of the United States >{20th Century Fox}

Wikipedia Articles

- 1 Articles Selected

Die Hard
Die Hard is a 1988 American ***action film*** and the first in the Die Hard ***film*** series. The ***film*** was directed by John McTiernan and written by Jeb Stuart and ...
http://en.wikipedia.org/wiki/Die_Hard

Figure 2: A faceted interface from Facetedpedia.

Facetedpedia Web GUI: The generated faceted interface, including information such as the category hierarchy of each facet and the articles reachable from each category in the hierarchy, is stored in a database. The GUI is a dynamic Web page implemented using Ajax. It reads the generated interface data from the database, displays the faceted interface, and updates the interface based on the user's navigation.

3. DEMONSTRATION PLAN

During the demo session, users can try arbitrary queries and play with the resulting faceted interfaces, through our online demo at `http://idir.uta.edu/facetedpedia`. Figure 2 is a screenshot of Facetedpedia in action. Region (B) is for showing the navigational paths that the user selected to assist her navigation. We only show one path in Figure 2, but there could be multiple paths. Region (C) shows the target articles reachable from the paths in region (B). Region (A) shows the facets for the articles in region (C).

Below we give the sketch of one example query scenario:
(1) The user types the keyword query "us action film" in the search box and presses the search button (above region (A)). The result articles are shown in region (C) in ranked order.
(2) The user chooses a domain, e.g., Film in Figure 2. Then a faceted interface is generated for the articles in that domain.
(3) The user further chooses the facet "*Cinema_of_the_United_States*" and then navigates through the path: *Cinema_of_the_United_States>American_film_actors>Bruce_Willis* (region (B)). There are 6 articles selected by this path.
(4) The user could continue the navigation with another facet, e.g, *Media_companies_of_the_United_States>Film_production_companies_of_the_United_States >20th_Century_Fox* (region (B)). There is 1 article selected by these two paths (region (C)). The faceted interface is updated so that only those facets and categories that can reach this article are shown in region (A).
(5) The user finds the articles that she is interested in and clicks the article titles in region (C). The corresponding Wikipedia article would be shown below. (This part of the interface is omitted due to space limitations.)

4. REFERENCES

[1] R. Hahn, C. Bizer, C. Sahnwaldt, C. Herta, S. Robinson, M. Bürgle, H. Düwiger, and U. Scheel. Faceted wikipedia search. In *BIS*, 2010.

[2] M. A. Hearst. Clustering versus faceted categories for information exploration. *CACM*, 49(4):59–61, 2006.

[3] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das. Facetedpedia: Dynamic generation of query-dependent faceted interfaces for Wikipedia. In *WWW*, 2010.

[1] http://wiki.dbpedia.org/Ontology