Facetedpedia: Enabling Faceted Search for Wikipedia

Ning Yan, Chengkai Li, Senjuti B. Roy, Rakesh Ramegowda, Lekhendro Lisham, 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,lisham.singh}@mavs.uta.edu, gdas@uta.edu

ABSTRACT

In [4], we proposed Facetedpedia, a faceted search system for Wikipedia that dynamically discovers a query-dependent faceted interface given a set of Wikipedia articles. We designed ranking metrics for both individual facets and faceted interfaces. We developed faceted interface discovery algorithms that optimize the ranking metrics, given the large space of possible faceted interfaces. In this paper, we give an overview of the Facetedpedia system, present the system architecture, and introduce the implementation of the individual components in the system. We also elaborate on a demonstration scenario for the system.

1. INTRODUCTION

Faceted search is a useful technique for information exploration and discovery, especially when a user needs to browse through a long list of articles or objects, which could be time consuming and painstaking without any auxiliary facility. Faceted interface has become influential over the last few years and we have seen an explosive growth of interests in its application. Commercial faceted search systems have been adopted by vendors (e.g., Endeca, IBM, and Mercado), as well as E-commerce Websites (e.g., eBay.com, Amazon.com).

A faceted interface, or the so-called *hierarchical faceted categories* (HFC) [3], is a set of category hierarchies over a set of objects, where each hierarchy corresponds to an individual *facet* (dimension, attribute, property) of the objects. The user can navigate through the category hierarchy of an individual facet to reach the objects associated with the selected categories. She navigates multiple facets and the intersection of the chosen objects on individual facets are brought to the user's attention.

Facetedpedia [4] is a system aimed at enabling faceted search for Wikipedia. The following example shows a scenario when this interface is useful.

Example 1 (Motivating Example): Imagine a user is exploring Wikipedia articles about action films produced in the United States. If she uses a keyword query, say, "us action film", what she gets would be a long ranked list of articles, with no extra information to assist her for further refinement. By contrast, Facetedpedia would provide several facets to help her shrink the long list based on her interests. For instance, if there are two facets on actors and film production companies, respectively, she can navigate the facets to choose the action films starring "Tom Cruise" and produced by "20th Century Fox". We will further elaborate on this example in Section 4. ■

Facetedpedia focuses on the *dynamic* discovery of *query-dependent* faceted interfaces. Given the 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 could not be precomputed 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 (attributes), therefore a query-independent static faceted interface, either manually or automatically generated, may suffice. By contrast, the articles in Wikipedia 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 non-trivial research undertaking. The concept of faceted interface is built upon two pillars: facets (i.e., dimensions or attributes) and the category hierarchy associated with each facet. The definition of "facet" itself for Wikipedia does not arise automatically, leaving alone the discovery of a faceted interface. Therefore we must tackle the challenges in both *facet identification* and *hierarchy construction*. The essence of our approach in [4] is to build upon the collaborative vocabulary in Wikipedia, more specifically the intensive internal structures (hyperlinks) and folksonomy (category system). Given the sheer size and complexity of this corpus, the space of possible faceted interfaces is prohibitively large. We proposed metrics for ranking individual facet hierarchies by user's navigational cost, and metrics for ranking interfaces (each with k facets) by both their average pairwise similarities and average navigational costs. We thus developed faceted interface discovery algorithms that optimize the ranking metrics.

Existing research prototypes or commercial faceted retrieval 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. When building the two pillars in a faceted interface, namely the facet and the hierarchy, Facetedpedia is both automatic and dynamic. On this aspect,

VLDB '10, September 13-17, 2010, Singapore

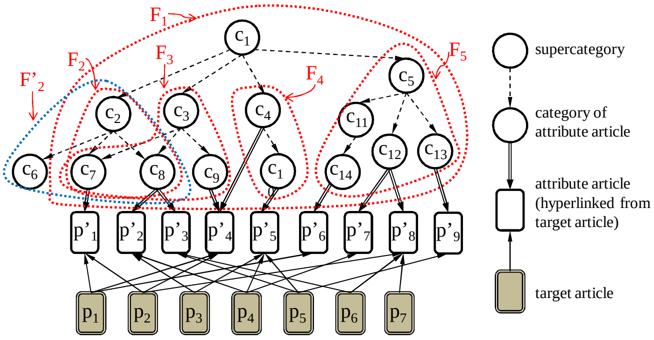


Figure 1: The concept of facet.

none of the existing systems could be effectively applied in lieu of Facetedpedia, because none is fully dynamic in both facet identification and hierarchy construction.

To the best of our knowledge, Facetedpedia is the first query-dependent faceted search system for Wikipedia. CompleteSearch [2] supports query completions and query refinement in Wikipedia by three special "facets": query completions matching the query terms; category names matching the query terms; and categories of result articles. Recently, a faceted Wikipedia search interface came out of the DBPedia [1] project around the same time as our work. Their faceted interface appears to be largely query independent and may be generated from the infobox of Wikipedia articles, although the underlying method remains to be proprietary at this moment.

2. AN OVERVIEW OF FACETEDPEDIA

In this section, we give a brief overview of our key methods in modeling facets and discovering faceted interfaces. The interested reader is referred to [4] for details.

2.1 The Concept of Facet

In building the two pillars of a faceted interface, the facet and the hierarchy, the basis of our approach is to make use of the *collaborative vocabulary* in Wikipedia, including the hyperlinks in articles, the categories of articles, [1] and the hierarchical relationships between different categories. The collaborative vocabulary represents the collective intelligence of many users and rich semantic information, and thus constitutes the promising basis for faceted interfaces.

With regard to the concept of facet, the Wikipedia articles hyperlinked from a search result article are exploited as its attributes. The fact that the authors of an article collaboratively made hyperlinks to other articles is an indication of the significance of the linked articles in describing the given article. We call the articles in the search result *target articles* and the hyperlinked articles *attribute articles*. With regard to the concept of category hierarchy, the Wikipedia category system provides the category-subcategory relationships between categories, allowing users to go from general to specific when specifying conditions.

We further use Figure 1 to explain the concepts. Given a keyword query q, the top-s ranked Wikipedia articles are returned as the target articles (e.g., $p_1, ..., p_7$). Given a target article p, each Wikipedia article p' that is hyperlinked from p is an attribute article of p, which is represented as $p' \leftarrow p$ (e.g., $p'_1, ..., p'_9$). We model the category hierarchy in Wikipedia as a rooted directed acyclic graph.[2] The category-subcategory relationships are indicated by the edges between them, e.g, $c_2 \dashrightarrow c_7$ indicates c_7 is a subcategory of c_2.

We define a *facet* as a rooted and connected subgraph of the category hierarchy. If every category in a facet can at least reach one target article, it is a *safe reaching facet*. In Figure 1, $\mathcal{F}_1, \ldots, \mathcal{F}_5$ are safe reaching facets, while $\mathcal{F}'_2$ is not. The path that a user follows to navigate through the categories in a facet and arrive at a target article is called a *navigational path*, e.g., $c_1 \dashrightarrow c_2 \dashrightarrow c_7 \Rightarrow p'_1 \leftarrow p_1$. A *faceted interface* is a set of safe reaching facets, e.g., $\{\mathcal{F}_2, \mathcal{F}_5\}$.

2.2 Faceted Interface Discovery

The search space of the faceted interface discovery problem is prohibitively large. Given the set of s target Wikipedia articles to a keyword query, $\mathcal{T}$, there are a large number of attribute articles which in turn have many categories associated with complex hierarchical relationships. By definition, any rooted and connected subgraph of category hierarchy that safely reaches $\mathcal{T}$ is a candidate facet, and any combination of k facets would be a candidate faceted interface. Given the large space, we need ranking metrics for measuring the "goodness" of facets, both individually and collectively as interfaces.

Single-Facet Ranking: Since a faceted interface is for a user to navigate through the associated category hierarchies and ultimately reaching the target articles, it is natural to rank the interfaces by the user's navigational cost. Based on our user navigational model [4], we compute the navigational cost of an individual facet as the average cost of its navigational paths. Intuitively a low-cost path should have a small number of steps and at each step only require the user to browse a small number of choices. Therefore, the cost of a navigational path is defined as the summation of the fan-outs (i.e., the number of choices) at every step, in logarithmic form. Moreover, a facet may not fully reach all the target articles, which presents an unsatisfactory user experience. Therefore, to incorporate into the cost formula the notion of "coverage", we penalize a facet by associating a high-cost *pseudo path* with each unreachable article.

Given the target articles, the categories in all the safe reaching facets form a relevant category hierarchy (RCH), which is a subgraph of the Wikipedia category hierarchy. The RCH can be established by first getting the attribute articles of target articles and the categories of attribute articles, and then recursively obtaining the super-categories of those categories. In Facetedpedia, we only consider RCH-induced facets [4] as the candidate facets. With RCH, a recursive algorithm calculates the navigational costs of all the safe reaching facets by only one pass depth-first search of RCH.

Multi-Facet Ranking: The best k-facet interface may not be simply found by getting the cheapest k facets according to single-facet ranking. The reason is that when the user navigates multiple facets, the selection made at one facet has impact on the available choices on other facets. In general an interface should not include two facets that overlap much. Based on this intuition, we propose to capture the overlap of the k facets by their *average pair-wise similarity*. The pair-wise similarity of two facets is the degree of overlap of their category hierarchies and associated attribute articles, defined by Jaccard coefficient.

[1] A Wikipedia article may belong to one or more categories. These categories are listed at the bottom of this article.

[2] The original Wikipedia categories hierarchy contains cycles. We discuss cycle removing in Section 3.

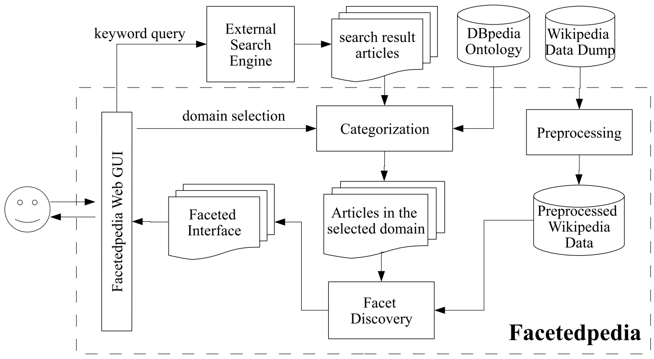


Figure 2: The architecture of Facetedpedia.

Due to the large search space of possible multi-facet interfaces, we design a hill-climbing based heuristic algorithm to look for a local optimum. To further tackle the challenge of modeling the interactions of multiple facets in measuring the cost of an interface, the hill climbing algorithm optimizes for both the average navigational cost and the pair-wise similarity of the facets.

3. SYSTEM IMPLEMENTATION

Facetedpedia is implemented in C++. It consists of four major components, as shown in Figure 2. The first component is the preprocessing of Wikipedia data. The original Wikipedia data dump is imported into a MySQL database, and then preprocessed and stored back into the database. The second component is the categorization of Wikipedia articles. 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. Facetedpedia categorizes the top-s articles into various domains, such as People, Places, etc. (In our implementation, we use Google.com as the external search engine. A typical value of s that we used is 200.) The user will then choose the domain of her interest and a k-facet interface is discovered for the target articles belonging to the chosen domain, by the third component–facet discovery. We believe a faceted interface is only meaningful for a set of homogeneous objects, i.e., the articles in the same domain. The generated faceted interface is stored in a database. The last component, Facetedpedia GUI, reads the interface data, displays the interface to the user, and updates the interface data based on the user's navigation actions. Below we elaborate on these components in detail.

Preprocessing Wikipedia Data Dump

The Wikipedia data set we used is a Wikimedia MySQL data dump generated at July 24th 2008.[3] We imported four tables (Page, Pagelinks, Categorylinks, Redirect) into our local database. The original Page table records both articles and categories (by using a *page_namespace* field to differentiate), therefore we separately stored them in new Page and Category tables. Similarly, the Pagelinks table records both the hyperlinks between two articles and the ones from articles to their categories, differentiated by a *pl_namespace* field. We accordingly divided Pagelinks into a new Pagelinks table and a PageCategory table.

We removed the redirect articles and categories (recorded in Redirect table) from the Page and Category tables. We also replaced the links to redirected articles (categories) by the corresponding non-redirect articles (categories), in Pagelinks, Categorylinks, and PageCategory.

We then removed the administrative categories in Wikipedia, using category name patterns including:
`_Wikiproject_, _Templates_, _unknown, _related_, Redirects_, Non-article_, Unassessed_, Wikipedia_, Wikipedians_, Automatically_assessed_, Unstable_, Unreferenced_, _needed_articles, Stubs_, List_of_`

In Section 2 we assumed the Wikipedia category hierarchy is a directed acyclic graph (DAG). However, in reality it contains cycles, although Wikipedia guidelines in general are against cycles.[4] Therefore one significant task in preprocessing the data was to remove the cycles in the category hierarchy. We used depth-first search (DFS) to traverse the entire category hierarchy starting from the "root" category, Category:Fundamental.[5] The DFS maintains a stack of visited categories in the path from the root to the current category. Whenever the DFS encoutners a category that is already in the stack, a cycle is detected and this edge is recorded as a *back edge*. The DFS continues, without visiting that category again. We detected about 600 cycles in total. We eliminated the cycles by just removing all such back edges.

Categorization

Facetedpedia asks the user to select the domain that she is interested in and the faceted interface produced is for the articles in the selected domain. In our implementation, we exploited the DBpedia ontology[6] for assigning articles to a set of pre-determined domains. DBpedia ontology itself is a subsumption hierarchy with about 200 classes (domains) and the dataset contains the articles belonging to each class. We used this dataset because it is generated based on Wikipedia's infobox and has a manually verified high accuracy. However, one main drawback of using this dataset is that it only covers the Wikipedia articles that have infoboxes, which account for 30% of all the Wikipedia articles. Therefore Facetedpedia is limited to the 30% as well. A more general method for categorizing Wikipedia articles is in our future agenda.

Facet Discovery

The facet discovery component is a multi-thread 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, for providing fast interactive user response. After the user chooses a target domain, a new thread is created to run the single-facet ranking and multi-facet ranking algorithm and generate the resulting faceted interface.

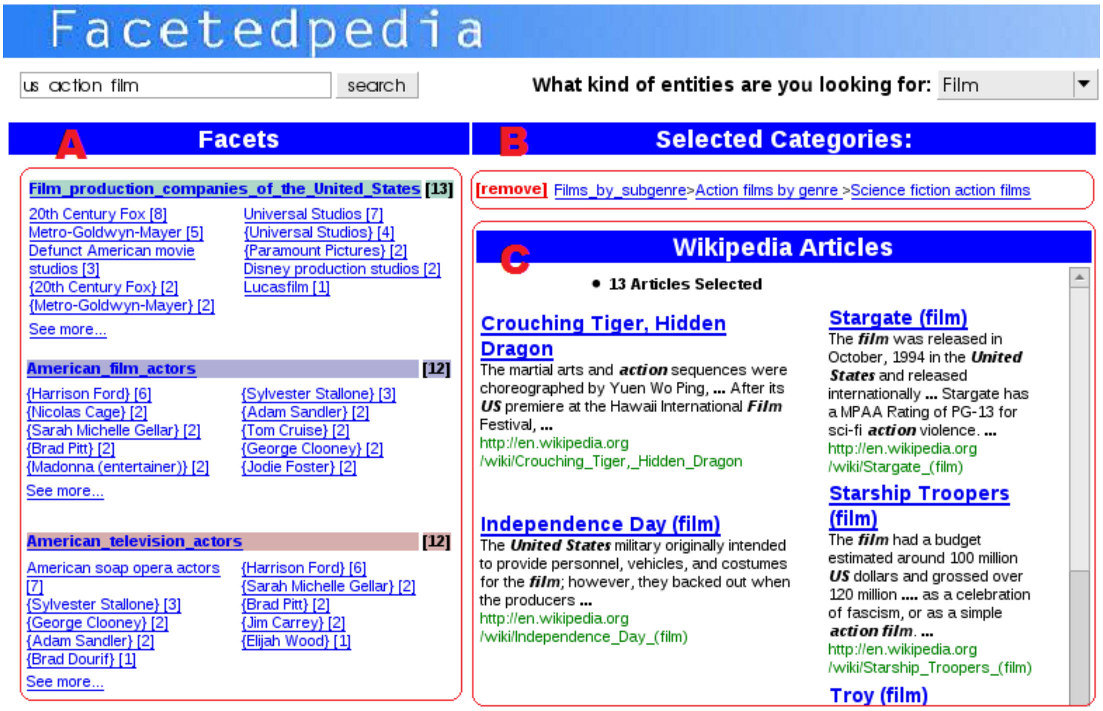
In multi-facet ranking, the algorithm must compute the pairwise similarity for many facets. For that purpose, the algorithm needs to efficiently compute the articles and categories that can be reached by each facet, which relies on the construction of the relevant category hierarchy (RCH). Since the facets are query-dependent, the RCH has to be dynamically built from the preprocessed tables in memory. The edges in the RCH are represented by adjacency list. To compute the set of articles (categories) that each internal node (i.e., the root of a candidate facet) covers, we can simply use a depth-first search to traverse the RCH. The set of articles (categories) covered by a node is the set-union of the sets covered by the children nodes. Directly implementing

[3] http://download.wikimedia.org/enwiki/20080724/.

[4] http://en.wikipedia.org/wiki/Wikipedia:Categorization

[5] http://en.wikipedia.org/wiki/Category:Fundamental

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



Facetedpedia

us action film

search

What kind of entities are you looking for: Film

Facets

Selected Categories:

[remove] Films_by_subgenre>Action films by genre >Science fiction action films

Film_production_companies_of_the_United_States [13]

20th Century Fox [8]
Metro-Goldwyn-Mayer [5]
Defunct American movie studios [3]
{20th Century Fox} [2]
{Metro-Goldwyn-Mayer} [2]
Universal Studios [7]
{Universal Studios} [4]
{Paramount Pictures} [2]
Disney production studios [2]
Lucasfilm [1]
See more...

American_film_actors [12]

{Harrison Ford} [6]
{Nicolas Cage} [2]
{Sarah Michelle Gellar} [2]
{Brad Pitt} [2]
{Madonna (entertainer)} [2]
{Sylvester Stallone} [3]
{Adam Sandler} [2]
{Tom Cruise} [2]
{George Clooney} [2]
{Jodie Foster} [2]
See more...

American_television_actors [12]

American soap opera actors [7]
{Sylvester Stallone} [3]
{George Clooney} [2]
{Adam Sandler} [2]
{Brad Dourif} [1]
{Harrison Ford} [6]
{Sarah Michelle Gellar} [2]
{Brad Pitt} [2]
{Jim Carrey} [2]
{Elijah Wood} [1]
See more...

Wikipedia Articles

- 13 Articles Selected

Crouching Tiger, Hidden Dragon
The martial arts and ***action*** sequences were choreographed by Yuen Wo Ping, ... After its ***US*** premiere at the Hawaii International ***Film*** Festival, ...
http://en.wikipedia.org/wiki/Crouching_Tiger,_Hidden_Dragon

Independence Day (film)
The ***United States*** military originally intended to provide personnel, vehicles, and costumes for the ***film***; however, they backed out when the producers ...
http://en.wikipedia.org/wiki/Independence_Day_(film)

Stargate (film)
The ***film*** was released in October, 1994 in the ***United States*** and released internationally ... Stargate has a MPAA Rating of PG-13 for sci-fi ***action*** violence. ...
http://en.wikipedia.org/wiki/Stargate_(film)

Starship Troopers (film)
The ***film*** had a budget estimated around 100 million ***US*** dollars and grossed over 120 million .... as a celebration of fascism, or as a simple ***action film***. ...
http://en.wikipedia.org/wiki/Starship_Troopers_(film)

Troy (film)

Figure 3: A faceted interface generated by Facetedpedia.

the union operation over the sets of article (category) IDs is costly, given the large size of Wikipedia. Therefore, instead of using verbatim IDs, we associate a bit vector with each node, where each bit in the vector indicates if the corresponding article (category) is reachable from the node. Thus the set-union can be fulfilled by an efficient bit-wise AND operation. Our experiments showed this reduces the time to discover a faceted interface from several minutes to several seconds.

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 Facetedpedia GUI reads the interface data and displays the faceted interface in a dynamic Web page, implemented by AJAX technique. It thus does not load all the interface data at once. Instead, when the user selects one category, the GUI will send a request to our server to load the subcategories of the selected category, saving both time and space for the Web browser client. The look and feel of the Facetedpedia GUI is inspired by an avant-garde faceted interface Flamenco.[7]

4. DEMONSTRATION

The screenshot shown in Figure 3 is a faceted interface generated during the process of a user's navigation. The interface is divided into three regions. Region (B) is for showing the navigational paths that the user selected to assist her navigation. We only show one path in Figure 3, but there could be multiple paths. Region (C) is for showing the target articles reachable from the paths in region (B). Region (A) shows the facets for the articles in region (C).

The scenario of the demonstration is as follows:

(1) The user types the keyword query "us action film" in the search box and presses the search button. The system shows the result articles in ranked order.

(2) The user chooses a domain, e.g., Film in Figure 3. Then a faceted interface is generated for the result articles belonging to that domain.

(3) The user further chooses the facet "*Films_by_subgenre*" and then navigates through the path: *Films_by_subgenre*> *Action_films_by_genre*>*Science_ fiction_action_films* (region (B)). There are 13 articles selected by following the path (region (C)). The faceted interface is updated so that only those facets and categories that can reach some of the 13 articles will be shown (region (A)). Figure 3 shows 3 of the remaining facets– *Film_production_companies_of_the_United _States*, *American_film_actors*, and *American_television_actors*.

(4) The user could continue the navigation with another facet, e.g, *American_film_actors*> *Tom_Cruise*.

(5) The user finds the articles that she is interested in and clicks an article title in region (C). The corresponding Wikipedia article would be shown. (This part of the interface is omitted due to space limitations.)

[7] http://flamenco.berkeley.edu

5. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *6th Int.l Semantic Web Conf.*, 2007.

[2] H. Bast and I. Weber. The CompleteSearch engine: Interactive, efficient, and towards IR & DB integration. In *CIDR*, pages 88–95, 2007.

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

[4] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das. Facetedpedia: Dynamic generation of query-dependent faceted interfaces for Wikipedia. In *To appear in the Proceedings of the 19th International World Wide Web Conference (WWW)*, 2010.