Facetedpedia: Enabling Query-Dependent Faceted Search for Wikipedia

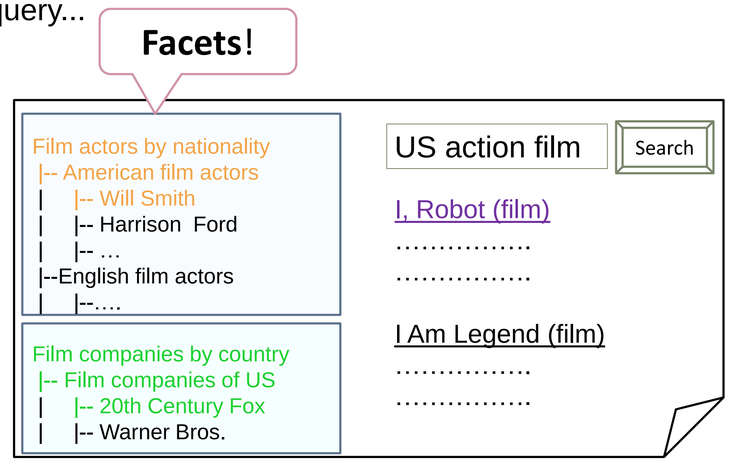
http://idir.uta.edu/facetedpedia

Ning Yan, Chengkai Li, Senjuti B. Roy, Rakesh Ramegowda, Gautam Das

The University of Texas ARLINGTON

1. Motivation

A user is exploring Wikipedia articles using keyword query...



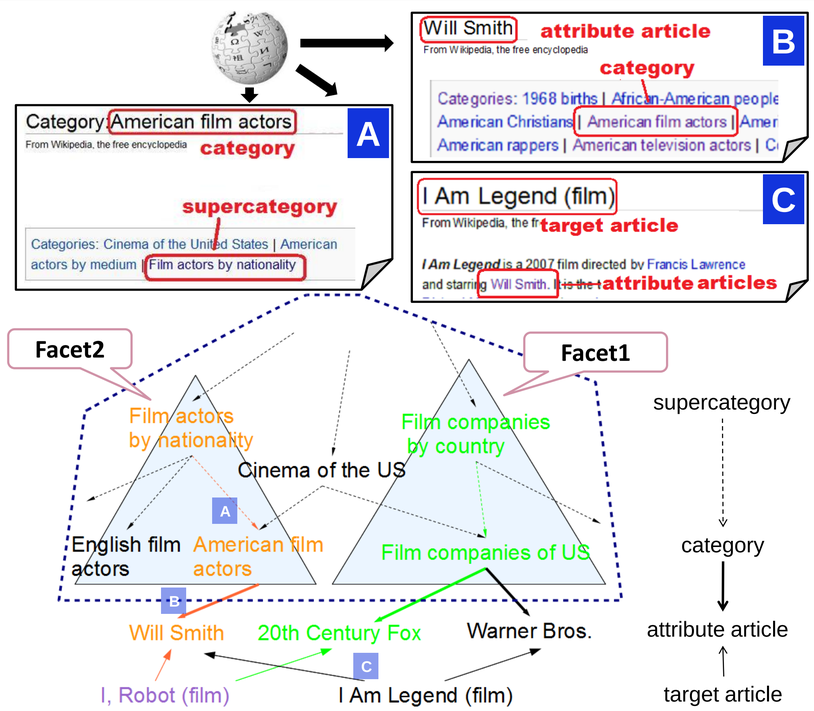
Then using two facets intersection to find Wikipedia article:

Film actors by nationality>American film actors>Will Smith

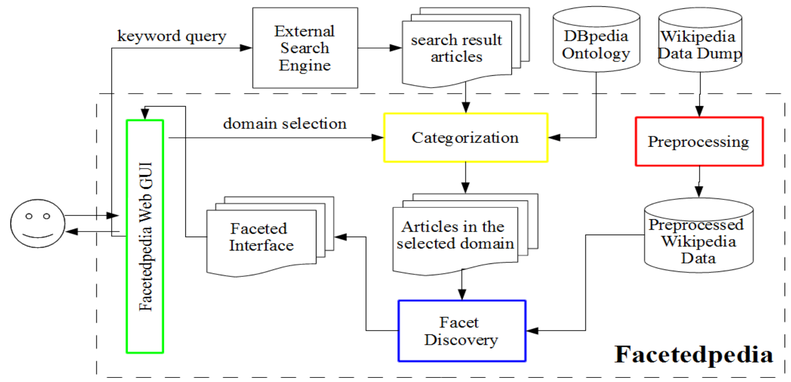
Film companies by country>Film companies of US>20th Century Fox

→ I, Robot (film)

2. Facets in Wikipedia



3. Facetedpedia Architecture



Preprocessing
- remove redirect pages and categories
- remove cycles from Wikipedia category hierarchy

Facet Discovery
- multi-thread background daemon loads preprocessed data into memory
- discover faceted interface per user session

Categorization
- assign Wikipedia articles to 80 domains

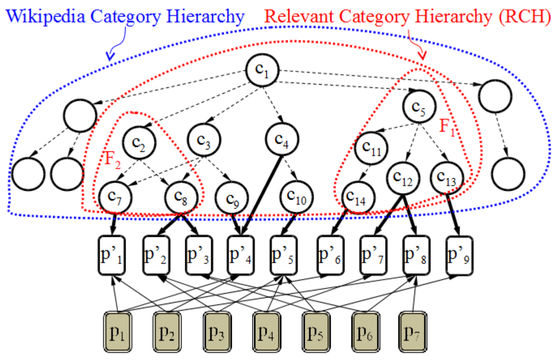
Facetedpedia Web GUI
- AJAX based Web GUI

4. Facet Discovery

The problem: discover the best *k*-facet interface

Facet: an induced subgraph of RCH (e.g. $F_2=\{c_2,c_7,c_8\}$)

Faceted interface: a set of k facets (e.g. $\{F_1,F_2\}$)

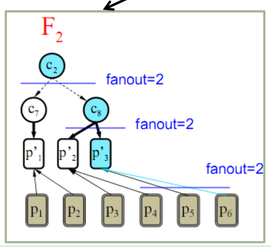


Step1: Retrieve attribute articles and build RCH
Step2: Rank single facets based on navigational cost
Step3: Select multiple facets using hill-climbing method

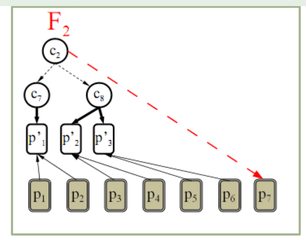
5. Facet Ranking

Single-Facet Ranking

$$cost(F_2) = \frac{1}{7} \times (\sum cost(\text{path } L_{p\in\{p_1,\ldots,p_6\}}) + penalty \times \#L'_{p\in\{p_7\}})$$



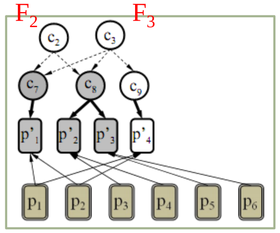
penalty path: $L' = c2 \dashrightarrow p_7$



navigational path: $L = c_2 \dashrightarrow c_8 \Rightarrow p'_3 \leftarrow p_6$

Multi-Facet Selection

1. Sum of single facet cost is small
2. Overlapping of *k*-facet is small (Jaccard Similarity)



$$sim(F_2, F_3) = \frac{|\{c_7, c_8\}| + |\{p'_1, p'_2, p'_3\}|}{|\{c_2, c_7, c_8, c_3, c_9\}| + |\{p'_1, p'_2, p'_3, p'_4\}|}$$

6. System Configuration

Data Set:
- Wikipedia data dump July 24th 2008
-2.4M target articles, 110M attribute article hyperlinks
-330K categories, 730K category links in category hierarchy

References

[1] 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.

[2] N. Yan, C. Li, S. B. Roy, R. Ramegowda, and G. Das. Facetedpedia: Enabling Query-Dependent Faceted Search for Wikipedia. In CIKM 2010.