Facetedpedia: Dynamic Generation of Query-Dependent Faceted Interfaces for Wikipedia

Chengkai Li, Ning Yan, Senjuti B. Roy, Lekhendro Lisham, Gautam Das

Department of Computer Science and Engineering
University of Texas at Arlington

June 18, 2010
WHAT IS A FACETED INTERFACE
A search engine result list looks like this:

us action film site:en.wikipedia.org
In contrast, a faceted interface over the result list looks like this:

<table>
<thead>
<tr>
<th>Facets</th>
<th>Selected Categories:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Living people</td>
<td>[remove] American film actors&gt; [Will Smith]</td>
</tr>
<tr>
<td>(Will Smith) [4]</td>
<td></td>
</tr>
<tr>
<td>(Bill Pullman) [1]</td>
<td></td>
</tr>
<tr>
<td>(Jeff Goldblum) [1]</td>
<td></td>
</tr>
<tr>
<td>(Dean Devlin) [1]</td>
<td></td>
</tr>
<tr>
<td>(Roland Emmerich) [1]</td>
<td></td>
</tr>
<tr>
<td>Events by year</td>
<td></td>
</tr>
<tr>
<td>Establishments by year [4]</td>
<td></td>
</tr>
<tr>
<td>Military history by year [2]</td>
<td></td>
</tr>
<tr>
<td>Years in television [1]</td>
<td></td>
</tr>
<tr>
<td>People by city in the United States by state</td>
<td></td>
</tr>
<tr>
<td>People by city in Pennsylvania [4]</td>
<td></td>
</tr>
<tr>
<td>People by city in Illinois [2]</td>
<td></td>
</tr>
<tr>
<td>People by city in Missouri [1]</td>
<td></td>
</tr>
<tr>
<td>People by city in New York [2]</td>
<td></td>
</tr>
<tr>
<td>People by city in California [1]</td>
<td></td>
</tr>
<tr>
<td>California counties</td>
<td></td>
</tr>
</tbody>
</table>

What kind of entities are you looking for: [Film]

**Wikipedia Articles**

- **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...

- **Enemy of the State (film)**

- **I, Robot (film)**
  I, Robot is a 2004 science-fiction action film directed by Alex Proyas. ... Produced with a budget of US$120 million, the film grossed US$144801023 ...

- **I Am Legend (film)**
  I Am Legend was released on December 14, 2007, in the **United States**, and opened to ... and that there was “a pretty heavy screenplay for an action film, ...
  [http://en.wikipedia.org/wiki/I_Am_Legend_(film)]
Clusters are **partitions** of articles by topics

Facets provides **multiple dimensional views** for a set of articles
A system generates **faceted interfaces** for a set of Wikipedia articles (target articles)

- Facets are **query-dependent**
  - Query="us action film", Facets=actors, directors, film production companies,...
  - Query="Texas university", Facets=cities in Texas, alumni, football teams, ...

- Generation is **automatic** compared with manual or predefined approach
Our Goal - Facetedpedia

- A system generates **faceted interfaces** for a set of Wikipedia articles (target articles)
- Facets are **query-dependent**
  - Query = "us action film", Facets = actors, directors, film production companies, ...
  - Query = "Texas university", Facets = cities in Texas, alumni, football teams, ...
- Generation is **automatic** compared with manual or predefined approach
Our Goal - Facetedpedia

- A system generates **faceted interfaces** for a set of Wikipedia articles (target articles)
- Facets are **query-dependent**
  - Query="us action film", Facets=actors, directors, film production companies,...
  - Query="Texas university", Facets=cities in Texas, alumni, football teams, ...
- Generation is **automatic** compared with manual or predefined approach
Our Goal - Facetedpedia

- A system generates **faceted interfaces** for a set of Wikipedia articles (target articles)
- Facets are **query-dependent**
  - Query="us action film", Facets=actors, directors, film production companies,...
  - Query="Texas university", Facets=cities in Texas, alumni, football teams, ...
- Generation is **automatic** compared with manual or predefined approach
Our Goal - Facetedpedia

- A system generates **faceted interfaces** for a set of Wikipedia articles (target articles)

- Facets are **query-dependent**
  - Query="us action film", Facets=actors, directors, film production companies,...
  - Query="Texas university", Facets=cities in Texas, alumni, football teams, ...

- Generation is **automatic** compared with manual or predefined approach
Challenge 1

What are our facets?

⇐ dimensions formed by attribute articles of target articles
CHALLENGE 1

What are our facets?
⇐ dimensions formed by attribute articles of target articles
**Challenge 1**

What are our facets?

⇐ dimensions formed by attribute articles of target articles

I, Robot (film)  I Am Legend (film)  
{Warner Bros.}  {20th Century Fox}  {Will Smith}  (Target articles)

{Will Smith}  {20th Century Fox}  {Warner Bros.}  (Attribute articles)

I Am Legend (film)  I, Robot (film)  (Target articles)
CHALLENGE 1

How to group attribute articles?
⇐ by their categories
CHALLENGE 1

How to group attribute articles? ⇐ by their categories

Will Smith is an American actor, film producer, and rapper. He has enjoyed success in music, television, and film. In April 2007, Newsweek called him the most powerful actor on the planet. Smith has been nominated for four Golden Globe Awards, two Academy Awards, and has won multiple Grammy Awards.

In the late 1980s, Smith achieved modest fame as a rapper under the name The Fresh Prince. In 1990, his celebrity increased dramatically when he starred in the popular television series The Fresh Prince of Bel-Air. The show ran for nearly six years (1990–1996) on NBC and has been syndicated consistently on various networks since then. In the mid-1990s, Smith transitioned from television to film, and ultimately starred in numerous blockbuster films that received broad box office success. In fact, he is the only actor in history to have eight consecutive films gross over $100 million in the domestic box office as well as being the only actor to have eight consecutive films in which he starred open at the #1 spot in the domestic box office tally.

Fourteen of the 19 fiction films he has acted in have accumulated a worldwide grosses of over $100 million, and 4 of them took in over $500 million in global box office receipts. His most financially successful films have been Bad Boys, Bad Boys II, Independence Day, Men in Black, Men in Black II, I, Robot, The Pursuit of Happyness, I Am Legend, Hancock, Wild Wild West, Enemy of the State, Shark Tale, Hitch and Seven Pounds. He also earned critical praise for his performances in Six Degrees of Separation, Ali and The Pursuit of Happyness, receiving Best Actor Oscar nominations for the latter two.
How to group attribute articles?
⇔ by their categories

American film actors

- {Will Smith}
- I, Robot (film)

Film production companies of United States

- {20th Century Fox}
- I Am Legend (film)

- {Warner Bros.}

(Target articles) (Attribute articles) (Categories)
**CHALLENGE 1**

How to organize categories?

⇐ by category hierarchy
**CHALLENGE 1**

How to organize categories?

⇐ by category hierarchy
**CHALLENGE 1**

How to organize categories?
⇐ by category hierarchy

- **Film actors by nationality**
  - American film actors
    - {Will Smith}
    - I, Robot (film)
    - I Am Legend (film)
- **Film production companies by country**
  - Film production companies of United States
    - {20th Century Fox}
    - {Warner Bros.}
  - Film production companies by country
    - (Supercategories)
  - Film production companies of United States
    - (Categories)
  - {20th Century Fox}
    - {Warner Bros.}
    - I Am Legend (film)
  - {Warner Bros.}
    - I, Robot (film)
  - (Attribute articles)
  - (Target articles)
establish the concept of facet by \textit{"structures"} from Wikipedia

- \textbf{RCH: Relevant Category Hierarchy}
The Concept of Facet - Cont.

- **facet**: an induced subgraph of a RCH, e.g., $F_2$ is a facet with $\{c_2, c_7, c_8\}$
The Concept of Facet - Cont.

- **faceted interface**: an interface consists of \( k \) facets, e.g., \( I = \{F_2, F_5\} \) is a faceted interface with 2 facets \( (k=2) \)
**Challenge 2**

1. What are our facets?
2. How to pick good facets and faceted interfaces?

**Single-Facet Ranking & Multi-Facet Ranking**
Intuitions:

1. How much user effort is needed when navigating through a single facet.

2. How many target articles a single facet could reach or cover.
SINGLE-FACET RANKING

Intuitions:

1. How much user effort is needed when navigating through a single facet.

2. How many target articles a single facet could reach or cover.
Cost of Navigational Path

capture first intuition (user effort) by cost of navigational path:

**Example**

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

![Diagram of navigational path](image-url)
Cost of Navigational Path

- Cost of navigational path: sum of logarithmic fanout

**Example**

\[
\text{cost}(L) = \sum_{\text{step}=c_2,c_8,p_3} \log_2(\text{fanout}(\text{step}))
\]

\[
= \log_2(2) + \log_2(2) + \log_2(2) = 3.
\]
capture second intuition (coverage) by adding a high cost pseudo **penalty path** for unreachable targeted articles:

**Example**

penalty path: $L' = c2 \rightarrow p_7$
**Cost of Single Facet**

- **cost of facet**: average cost of navigational paths + **penalty**
  (penalty value is determined empirically)

**Example**

\[
cost(F_2) = \frac{1}{7} \times \left( \sum \text{cost(path } L_p \in \{p_1, \ldots, p_6 \} \text{)} + \text{penalty} \times \#L_p' \in \{p_7 \} \right)
\]
MULTI-FACET RANKING

Intuitions:

1. overall cost of $k$ facets should be small
**MULTI-FACET RANKING**

Intuitions:

1. overall cost of $k$ facets should be small
2. $k$ facets should NOT overlap a lot
Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...
Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...
Example

Given a faceted interface with 2 non-overlapping facets \( I = \{F_2, F_5\} \), a user navigates like...

![Diagram showing user navigation model with nodes and edges representing navigation paths and facets.](image-url)
Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...
Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...
**Example**

Given a faceted interface with 2 non-overlapping facets $I = \{ F_2, F_5 \}$, a user navigates like...
Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...
Example

Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...
Given a faceted interface with 2 non-overlapping facets \( I = \{F_2, F_5\} \), a user navigates like...
Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...

![Diagram showing user navigation model]

---

**Example**

Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...

![Diagram showing user navigation model]
**Example**

Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...

![Diagram](image-url)
Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...
Given a faceted interface with 2 non-overlapping facets $I = \{F_2, F_5\}$, a user navigates like...
capture second intuition (overlap) by Jaccard similarity between 2 facets:

**Example**

\[
sim(F_2, F_3) = \frac{|\{c_7, c_8\}| + |\{p'_1, p'_2, p'_3\}|}{|\{c_2, c_7, c_8, c_3, c_9\}| + |\{p'_1, p'_2, p'_3, p'_4\}|} = 5/9.
\]
and by **average pair-wise similarity** (AVG) between $k$ facets, e.g., a faceted interface with 3 facets $\{F_2, F_3, F_5\}$:

**Example**

$$AVG(\{F_2, F_3, F_5\}) = \frac{1}{3} \times (sim(F_2, F_3) + sim(F_2, F_5) + sim(F_3, F_5))$$
**Challenge 3**

1. What are our facets?
2. How to pick good facets and faceted interfaces?
3. How to do the rankings efficiently for Wikipedia?

**Facet Discovery Algorithm**
Step 1: Retrieve attribute articles & Build RCH
Input: target articles
Output: attribute articles, RCH
Step 1: Retrieve attribute articles & Build RCH
Input: target articles
Output: attribute articles, RCH


**Step 1: Retrieve attribute articles & Build RCH**

**Input:** target articles  
**Output:** attribute articles, RCH
Step 1: Retrieve attribute articles & Build RCH
Input: target articles
Output: attribute articles, RCH
Step 1: Retrieve attribute articles & Build RCH
Input: target articles
Output: attribute articles, RCH
Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs
Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs
Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs
Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs
Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs
Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs
Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs

The Concept of Facet Facet Ranking Facet Discovery Algorithm
Facet Discovery Algorithm

Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs
Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs
Step 2: Rank single facets based on their cost
Input: target articles, attribute articles, RCH
Output: a ranked list of single facets and their costs
Step 3: Select multi-facet using hill-climbing method

Input: top-\(n\) facets from the ranked list and their costs \(C\)

Output: a faceted interface with \(k\) facets

1. randomly pick \(k\) facets from \(n\) facets
2. compute \(cost\), \(AVG\) of \(k\) facets
3. repeat
   - switch facet \(i\) in \(k\) facets with facet \(j\) from remaining
   - compute \(cost_{new}\)
   - compute \(AVG_{new}\)
   - if \((cost_{new} \downarrow \land \land AVG_{new} \downarrow)\) then replace facet \(i\) with \(j\)

until \(k\) facets are stable

4. return \(k\) facets
## Related Works

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Types of Facets</th>
<th>Facet Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rich</td>
<td>Structured</td>
<td>Predetermined</td>
</tr>
<tr>
<td></td>
<td>Text</td>
<td>Automatically Discovered</td>
</tr>
<tr>
<td>Shallow</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Hierarchy examples:**

- **Shallow:** People → Will Smith
- **Rich:** People → Actor → Film Actor → American Film Actor → Will Smith
**Castanet** [StoicaHearst07]

- not for general articles
  - short
  - domain-specific (e.g. recipes...)
- limited vocabulary
- hierarchy is shallow
Faceted Wikipedia Search [http://dbpedia.neofonie.de/browse/]
### Related Works

**Faceted Wikipedia Search**

<table>
<thead>
<tr>
<th>Facet Identification</th>
<th>Hierarchy Construction</th>
</tr>
</thead>
<tbody>
<tr>
<td>predefined</td>
<td>automatically discovered</td>
</tr>
<tr>
<td>Facetedpedia</td>
<td>Facetedpedia</td>
</tr>
</tbody>
</table>

**Facets:**
- query independent for specific domain
- probably from infobox attributes, similar to relational data, no facet identification
- only for articles with infobox
- no hierarchy
**Experiment Settings**

- **Raw Data**: Wikipedia dump 20080724, 2.4M articles, 110M hyperlinks, 330K categories, 730K category links in category hierarchy
- **Programs**: Preprocessing of Wikipedia dump (removing irrelevant articles and categories, removing cycles in category hierarchy)
- **Parameters**: \#target articles=200, \#facets=10
A User Study

- select 20 queries from different domains
- generate faceted interfaces for each query by Facetedpedia and Castanet
- each interface has been explored and evaluated by 8~10 user

<table>
<thead>
<tr>
<th>Q1</th>
<th>action film</th>
<th>Q2</th>
<th>country singer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q3</td>
<td>philosophers</td>
<td>Q4</td>
<td>Texas universities</td>
</tr>
<tr>
<td>Q5</td>
<td>Turing Award winner</td>
<td>Q6</td>
<td>missile</td>
</tr>
<tr>
<td>Q7</td>
<td>Ivy League schools</td>
<td>Q8</td>
<td>NBA players</td>
</tr>
<tr>
<td>Q9</td>
<td>historic landmarks</td>
<td>Q10</td>
<td>cartoon characters</td>
</tr>
<tr>
<td>Q11</td>
<td>Microsoft acquired game companies</td>
<td>Q12</td>
<td>stand up comedian</td>
</tr>
<tr>
<td>Q13</td>
<td>graph theorists</td>
<td>Q14</td>
<td>lakes in North America</td>
</tr>
<tr>
<td>Q15</td>
<td>American presidents</td>
<td>Q16</td>
<td>battle far east</td>
</tr>
<tr>
<td>Q17</td>
<td>waterfall national park</td>
<td>Q18</td>
<td>Chinese cuisine</td>
</tr>
<tr>
<td>Q19</td>
<td>premier league clubs</td>
<td>Q20</td>
<td>PS3 game</td>
</tr>
</tbody>
</table>
Average user ratings of Facetedpedia vs. Castanet
(Rating scale: 2 Very useful; 1 Useful; 0 Useful to some extent; -1 Not Very Useful; -2 Useless)
**QUALITY EVALUATION**

Measurement of coverage, average width, depth, pairwise similarity of faceted interfaces.

<table>
<thead>
<tr>
<th></th>
<th>Coverage</th>
<th>average width</th>
<th>average height</th>
<th>average pair-wise similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random-k</td>
<td>72.3%</td>
<td>53.8</td>
<td>8.6</td>
<td>0.108</td>
</tr>
<tr>
<td>Top-k</td>
<td>73.9%</td>
<td>10.2</td>
<td>5.5</td>
<td>0.187</td>
</tr>
<tr>
<td>Hill-climbing</td>
<td>68.9%</td>
<td>9.8</td>
<td>5.7</td>
<td>0.072</td>
</tr>
</tbody>
</table>
Run time varies for different number of target articles almost linearly.
CONCLUSIONS

1. Facetedpedia System
2. Facet Ranking Metrics
3. Faceted Interface Discovery Algorithm
4. Evaluation