Ranking Categories for Faceted Search

Gianluca Demartini

L3S Research Seminars
Hannover, 09 June 2006
Outline

- Introduction
- Basic Concepts
- Rankings Algorithms considered
- Experimental Setup
- Results
- Conclusions
Introduction

■ Search the Web: Ranked list or Categories based organization?

■ Clustering Search vs Faceted Search

  ▪ **Clustering**: Grouping documents according to some measure of similarity computed using associations among features (typically words and phrases)

    Result - 1 big hierarchy

  ▪ **Faceted**: Creating a set of category hierarchies each of which corresponds to a different facet (dimension or feature type) relevant to the collection to be navigated

    Result - a set of category hierarchies each of which corresponds to a different facet

■ Supporting Vector Machines Classifiers
A linear SVM is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin.
SVM text classification

- The formula for the output of a linear SVM is
  \[ u = \bar{w} \cdot \bar{x} - b. \]

- Where \( \bar{w} \) is the normal vector to the hyperplane, and \( \bar{x} \) is the input vector

- Given training examples labeled either "yes" or "no", a maximum-margin hyperplane is identified which splits the "yes" from the "no" training examples
Clustering Search Engines

- SVM better than Bayesian for Text Classification
- Many clustering algorithms proposed
- How to rank the resulting Categories?
  - Algorithm independent
- We analyze 9 different metrics used to order the clusters
Outline

- Introduction
- Basic Concepts
- Rankings Algorithms considered
- Experimental Setup
- Results
- Conclusions
Category Ranking Algorithms

- 9 different ranking algorithms considered:
  - Rank based metrics
  - Text Similarity metrics
  - Other Metrics
Category Ranking Algorithms - Rank Based Metrics

- PageRank computation: \( PR_v = x^{-2.1} \)  
  \( p \) at position \( x \)

- Average PageRank
  \[ \text{AvgPR}(C) = \frac{1}{n} \sum_{p=1}^{n} PR_v(p), \forall \text{ page } p \in C \]

- Total PageRank
  \[ \text{SumPR}(C) = \sum_{p=1}^{n} PR_v(p), \forall \text{ page } p \in C \]

- Average Rank
  \[ \text{AvgRank}(C) = \frac{1}{n} \sum_{p=1}^{n} \text{Rank}(p), \forall \text{ page } p \in C \]

- Minimal Rank
  \[ \text{MinRank}(C) = \min_{p} \text{Rank}(p), \forall \text{ page } p \in C \]
Category Ranking Algorithms - Text Similarity Metrics

- Similarity between pages and categories (title + description)
  - Values returned by the SVM classifiers

- Average Similarity Score (AvgValue)
  - Over all the pages that belong to a category

- Maximum Similarity Score (MaxValue)
  - Over all the pages that belong to a category
Category Ranking Algorithms - Other Metrics

- **Order by Size**: using the number of docs belonging to the category
  - Used by most of the Clustering Search Engines (Vivisimo)

- **Alphabetical Order**
  - Used in Faceted Search (Flamenco)

- **Random Order**
  - To compare the other metrics
Outline

- Introduction
- Basic Concepts
- Rankings Algorithms considered
- Experimental Setup
- Results
- Conclusions
Experimental Setup

- 9 algorithms, 18 people
- Supporting Vector Machines (SVM) as Text Classifiers
- ODP categories (top 3 levels)
- 50,000 most frequent terms in DMOZ titles and descriptions of web pages
- 5,894 English categories

- Each user evaluated each algorithm once
- We measure the time spent for search the relevant result and the position of the results
You searched for...

Keywords: hot dog

Results:

Recreation_Food_Meat

National Hot Dog And Sausage Council | www.hot-dog.org

Snippet: Conducts scientific research to benefit hot dog and sausage manufacturers. Brochures, facts and trivia, news, and recipes.

Recreation_Pets_Travel

Hot Dog Holidays - Pet Friendly hotels and holiday homes in

Snippet: Offers hotels, castles, mansions, and other pet friendly accommodations in Europe.

Sports_Softball_News_and_Media

Hot Dogs As America | Baseball As America | American Museum

Snippet: Baseball As America, the first major exhibition to explore baseball and American culture, will open at the American Museum of Natural History on March 16, ...
Experimental Results

- **Time to find the relevant result:**

![Bar chart showing average time to find relevant result for different methods.](chart.png)
Experimental Results

- Average of the position of the algo for each user:
Experimental Results

- Average Rank of the Result and of the Cluster
Conclusions & Future Work

- MaxValue seems to be the best way to rank the clusters in a Clustering Search Engine

- Alphabetical and Size Ranking are not so good

- We want to test other algorithms
  - Using query-based metrics (similarity between q and p)
  - Click-thorough data
Thanks for your attention!

Q&A