Chapter IR:VII

VII. Result Presentation

- Showing the Results

Showing the Results

Snippet Generation

Tropical Fish

One of the U.K.s Leading suppliers of **Tropical**, Coldwater, Marine **Fish** and Invertebrates plus.. . next day **fish** delivery service ...

www.**tropicalfish**.org.uk/**tropical_fish**.htm Cached page

- Query-dependent document summary
  - Link to web page and cached version
  - Title and URL
  - Short text summary (snippet)
    - Sometimes full sentences, sometimes not
  - Some query-independent features may be used
- Simple summarization approach
  - First proposed by Luhn in 50's
    1. Rank each sentence in a document using a significance factor
    2. Select the top sentences for the summary

Showing the Results

Sentence Selection

- Significance factor for a sentence is calculated based on the occurrence of significant words
  - Significant words are of medium frequency
  - If $f_{d,w}$ is the frequency of word w in document d, then w is a significant word if it is not a stopword and

$$f_{d,w} \geq \begin{cases} 7 - 0.1(25 - s_d) & \text{if } s_d < 25 \\ 7 & \text{if } 25 \leq s_d < 40 \\ 7 + 0.1(s_d - 40) & \text{otherwise} \end{cases}$$

    where s_d is the number of sentences in d
  - Example: $s_d = 20$ yields $f_{d,w} \geq 7 - 0.1(25 - 20) = 6.5$

Showing the Results

Sentence Selection

- Text is *bracketed* by significant words
  - Limit on number of non-significant words between two significant ones
    - Usually 4 non-significant words
- Significance factor for bracketed text spans is computed by dividing the square of the number of significant words in the span by the total number of words
  - Initial sentence: `w w w w w w w w w w w.`
  - Significant words: `w w s w s s w w s w w.`
  - Bracketed: `w w[s w s s w w s]w w.`
  - Significance factor = $4^2/7 = 2.3$
- Significance factor for entire text is maximum significance factor for any bracket

Showing the Results

Snippet Generation

- Improvements based on better selection of significant words and sentence fragments
  - In particular: Query dependent
- Involves more features than just significance factor
- E.g. for a news story, could use
  - whether the sentence is a heading
  - whether it is the first or second line of the document
  - the total number of query terms occurring in the sentence
  - the number of unique query terms in the sentence
  - the longest contiguous run of query words in the sentence
  - a density measure of query words (significance factor)
- Weighted combination of features used to rank sentences

Showing the Results

Snippet Generation

- Web pages are less structured than news stories
  - Can be difficult to find good summary sentences
- Snippet sentences are often selected from other sources
  - Metadata associated with the web page
    - E.g., <meta name="description" content= . . . >
  - External sources such as web directories
    - E.g., Open Directory Project, http://www.dmoz.org
- Certain pages, such as Wikipedia have better structure
  - Snippet generation easier

Showing the Results

Snippet Guidelines

- Derived from analysis of clickthrough data
  - All query terms should appear in the summary, showing their relationship to the retrieved page
  - When query terms are present in the title, they need not be repeated
    - Allows snippets that do not contain query terms
  - Highlight query terms in URLs
  - Snippets should be readable text, not lists of keywords
    - Sentences or contiguous sentence fragments
- Snippet generation should be fast
  - Local document store

Showing the Results

Advertising

- Two kinds of advertising
  - *Sponsored search* – advertising presented with search results
  - *Contextual advertising* – advertising presented when browsing web pages
- Both involve finding the most relevant advertisements in a database
  - An advertisement usually consists of a short text description and a link to a web page describing the product or service in more detail
  - Special case of text search

Showing the Results

Searching Advertisements

- Factors involved in ranking advertisements
  - Similarity of text content to query
  - Bids for keywords in query
  - Popularity of advertisement
- Who defines factors and weighting?
  - Payment model
  - Economics and game theory

- Small amount of text in advertisement
  - Dealing with vocabulary mismatch is important
  - Expansion techniques are effective
    - Both for query and for document (=advertisement)

Showing the Results

Searching Advertisements

- Query reformulation based on search sessions
  - 50% of queries are reformulations
    - I.e., rich repository of associations
  - Learn associations between words and phrases based on co-occurrence in search sessions
    - `aquarium` followed by `fish tank` in same session

- Pseudo-relevance feedback
  - Expand query and/or document using the Web
  - Use ad text or query for pseudo-relevance feedback
  - Effective ranking order
    1. Exact matches
    2. Stem matches
    3. Expansion matches

Showing the Results

Example Advertisements

fish tanks at Target
Find **fish tanks** Online. Shop & Save at Target.com Today.
www.target.com

Aquariums
540+ Aquariums at Great Prices.
fishbowls.pronto.com

Freshwater **Fish** Species
Everything you need to know to keep your setup clean and beautiful
www.FishChannel.com

Pet Supplies at Shop.com
Shop millions of products and buy from our trusted merchants.
shop.com

Custom **Fish Tanks**
Choose From 6,500+ Pet Supplies. Save On Custom **Fish Tanks**!
shopzilla.com

- Advertisements retrieved for query `fish tank`
  - Second one not obvious, but relevant
  - Fourth one probably based on keyword bid

Showing the Results

Clustering Results

- Result lists often contain documents related to different aspects of the query topic
  - `jaguar`
- Clustering is used to group related documents to simplify browsing (cf. course on Machine Learning)

- Example clusters for query `tropical fish`

Pictures (38)
Aquarium Fish (28)
Tropical Fish Aquarium (26)
Exporter (31)
Supplies (32)
Plants, Aquatic (18)
Fish Tank (15)
Breeding (16)
Marine Fish (16)
Aquaria (9)

Showing the Results

Clustering Results – Requirements

- Efficiency
  - Must be specific to each query and are based on the top-ranked documents for that query
  - Typically based on snippets, not full text
    - Snippets focus on query-relevant part of text, not on entire text
- Easy to understand
  - Can be difficult to assign good labels to groups
  - Monothetic vs. polythetic classification

Showing the Results

Types of Classification

- Monothetic
  - Every member of a class has the property that defines the class
  - Typical assumption made by users
  - Easy to understand, because easy to explain
- Polythetic
  - Members of classes share many properties but there is no single defining property
  - Most clustering algorithms (e.g., K-means) produce this type of output

Showing the Results

Classification Example

- $d_1 =$ `a b c`
- $d_2 =$ `a d e`
- $d_3 =$ `d e f g`
- $d_4 =$ `f g`
- Possible monothetic classification
  - Not necessarily disjoint
  - $\{d_1, d_2\}$ (labeled using `a`), $\{d_2, d_3\}$ (labeled `e`), and $\{d_3, d_4\}$ (labeled `f g`)
- Possible polythetic classification
  - Based on term overlap
  - $\{d_2, d_3, d_4\}, d_1$
  - No single term in common
  - Labels?

Showing the Results

Result Clusters

- Simple algorithm
  - Group based on words in snippets
  - Use all non-stop-terms that appear in at least two snippets
    - aquarium (5) (documents 1, 3, 4, 5, 8)
    - freshwater (4) (1, 8, 9, 10)
    - species (3) (2, 3, 4)
    - hobby (3) (1, 5, 10)
    - forums (2) (6, 8)
- Refinements
  - Use phrases
  - Use more features
    - whether phrases occurred in titles or snippets
    - length of the phrase
    - collection frequency of the phrase
    - overlap of the resulting clusters

Showing the Results

Faceted Classification

- A set of categories, usually organized into a hierarchy, together with a set of facets that describe the important properties associated with the category
  - Document can have value in every facet
- Manually defined
  - Potentially less adaptable than dynamic classification
  - Tedious
- Easy to understand
  - Commonly used in e-commerce

Showing the Results

Example Faceted Classification

- Categories for `tropical fish`

Books (7,845)
Home & Garden (2,477)
Apparel (236)
Home Improvement (169)
Jewelry & Watches (76)
Sports & Outdoors (71)
Office Products (68)
Toys & Games (62)
Everything Else (44)
Electronics (26)
Baby (25)

DVD (12)
Music (11)
Software (10)
Gourmet Food (6)
Beauty (4)
Automotive (4)
Magazine Subscriptions (3)
Health & Personal Care (3)
Wireless Accessories (2)
Video Games (1)

Showing the Results

Example Faceted Classification

- Subcategories and facets for “Home & Garden”

Home & Garden
Kitchen & Dining (149)
Furniture & Décor (1,776)
Pet Supplies (368)
Bedding & Bath (51)
Patio & Garden (22)
Art & Craft Supplies (12)
Home Appliances (2)
Vacuums, Cleaning & Storage (107)

Brand
<brand names>

Seller
<vendor names>

Discount
Up to 25% off (563)
25% - 50% off (472)
50% - 70% off (46)
70% off or more (46)

Price
$0-$24 (1,032)
$25-$49 (394)
$50-$99 (797)
$100-$199 (206)
$200-$499 (39)
$500-$999 (9)
$1000-$1999 (5)
$5000-$9999 (7)