Search in the Clothing Domain

Sara Stolbach

NLP For the Web, COMS 6998-03

Spring 2008

Motivation

- General:
  - Focus of the query can change depending on the domain
- Specific:
  - A single source for searching multiple websites in the clothing domain.
  - Exact matching is important (in search for skirt don't want shirt)

Data & Resources

- Extracted pages from 3 popular clothing sites – Men, Women, and Children
- There are **4987** documents
- There are **28515** unique terms with stemming
- Lucene (http://lucene.apache.org) is used as the core search engine.

Sample Document

```
<document>
<id>
LE721703891
</id>
<url>
<![CDATA[http://www.landsend.com/pp/BicolorHopsackBlazer~162596_59.html]]>
</url>
<contents>
<![CDATA[
women

shirts &amp;amp; sweaters

jackets &amp;amp; blazers
view similar items

women&amp;apos;s regular bi-color hopsack blazer
e-mail to a friend
overview
more info
kicked-up classic.
•••

try it on
|
get a size recommendation
div id=&amp;quotvalueaddedservice_prd_x192825&amp;quot class=&amp;quotvalueaddedservice
]]>
</contents>
</document>
```

Domain Specific Score Improvement

The entire phrase is queried ordinarily. Then, the score of the document increases based upon:

- Important features
  - Domain specific terms
- Focus of query (Future Work)
  - Using Part of Speech

Important Features

- There are currently **241** important features in the clothing domain.
- Examples: women, pocket, navy, black, shirt
- Each word has a score associated with it defining how much it should boost the document's score

Future Work:

- Obtain more features using synonyms via Wordnet
- Sets of features – if shirt is in the query, don't returns skirts

Focus Of Query (Future Work)

- Use part-of-speech to find the focus of the query.
- This is useful when there are no important features. (There may be domain specific terms that are not in the knowledge base).
- Documents that contain the term that is the focus of the query will have a boost to their score.

Rescoring

- The documents for the query are received. The scores for each document are increased based on features.

- Ex:
  - Query: D1 => 1.4, D2 => .78, D3 => .62
  - "Term1" => +1.5 (D1,D2)
  - "Term2" => +2 (D2,D3)
  - "Term3" => 0 (D1,D3)

  Final Scores:
  - D2 => 4.28, D1 => 2.9, D3 => 2.62

Evaluation Method

- Method 1:
  - Create list of results wanted
  - Search for result using general words
    - How many correct results appeared in the top 10?
- Method 2:
  - Create a list of queries.
  - Run each query on this search engine and Lucene search engine
  - Have two users compare the results for each search engine.
    - Which has the better result has the first answer?
    - How many results rankings improved correctly?

Difficulties in domain search

- Numbers
  - sizes
  - quantities
- Obtaining and cleaning up data
  - html
  - menu's
- Duplicate term inconsistency

Demo

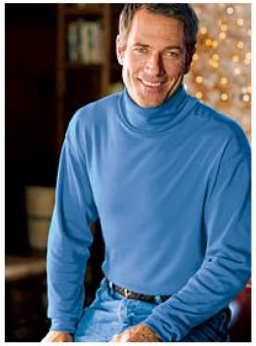
- Queries
  - black sleeveless dress
  - Pink turtleneck

Black Sleeveless Dress

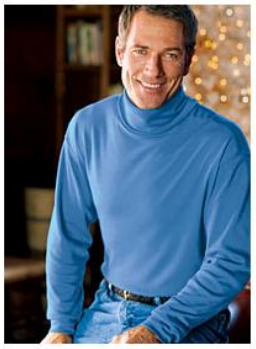
Which row is better?

Pink Turtleneck







Which row is better?

Questions?