Social Tagging in Query Expansion: a new Way for Personalized Web Search

Claudio Biancalana, Alessandro Micarelli, Francesco Saverio Profiti

Department of Computer Science an Automation
Artificial Intelligence Laboratory
Rome Tre University
and
LAzio Innovazione Tecnologica S.p.A.
Motivations

- Information overload
- Human Factors
- Semantics
- Vocabulary problem
- Ineffectiveness of Short Queries
Query Expansion (1/2)

- The process of expanding a user query with additional related words and phrases

- In the context of web search engines, query expansion involves evaluating a user input typed into the search query area and expanding the search query to match additional documents
Personalized Web Search Architecture

User

Query
a,b

Personalization

Search Engine

Visited pages

Implicit Feedback

User Model

a,b,c,d,e

a,b,c,d,e
Query Expansion (2/2)

Original Query: \( Q = \{q_1, q_2, \ldots, q_k, q_{k+1}, \ldots, q_n\} \)

Terms to add: \( Q^+ = \{e_1, e_2, \ldots, e_m\} \)

Terms to remove: \( Q^- = \{q_{k+1}, \ldots, q_n\} \)

Expanded Query

\[
EQ = (Q \cup Q^+) - Q^-
\]

\[
\{q_1, q_2, \ldots, q_k, e_1, e_2, \ldots, e_m\}
\]
User modeling and personalization in web

- Personalization is tailoring a consumer product, electronic or written medium to a user, based on personal details or characteristics that the user provides.
- In web search engines, personalization is tailoring search results based on the interests of a user.
Pre-processing

- HTML Tag Elimination
- Semantic Analysis: Monty POS tagger
  - adjective, noun, proper noun, preposition
- Stop Word Elimination
- Stemming
User Model and Co-occurrence

- Co-occurrence is the extent of which two terms tend to appear simultaneously in the same context
- User Model: Co-occurrence terms matrix

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>t2</td>
<td>1.0</td>
<td>0.0</td>
<td>3.0</td>
<td>2.0</td>
<td>0.0</td>
</tr>
<tr>
<td>t3</td>
<td>0.0</td>
<td>3.0</td>
<td>0.0</td>
<td>9.0</td>
<td>0.0</td>
</tr>
<tr>
<td>t4</td>
<td>2.0</td>
<td>2.0</td>
<td>9.0</td>
<td>0.0</td>
<td>4.0</td>
</tr>
<tr>
<td>t5</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>4.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Personalization and Query Expansion

- **Method I**
  - Bigrams
- **Method II**
  - Hyperspace Analogue to Language
- **Method III**
  - Page Level co-occurrence
- **Method IV**
  - Page Level co-occurrence and term proximity
Method I - Bigrams

- The user model is built around the concept of bigrams, namely a pair consisting of two adjacent terms in the text of a web page. Two terms are considered co-occurring only if adjacent.
- The context of a term is thus exclusively limited to the term that is directly next to it, either to the left or to the right;
Method II - HAL

- Given a window of N terms, that can be scrolled inside a page text, two terms are considered co-occurring only if they are within such window.
- The co-occurrence value will be inversely proportional to the distance between the two terms within the window;

Example N=5

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/5</td>
<td>1/4</td>
<td>1/3</td>
<td>1/2</td>
<td>1/1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Method III – Page Level co-oc (1/2)

- Within this method, the context of a term is expanded to the entire page considered.
- Two terms are then deemed co-occurring only if they are both present, simultaneously, in the same page;
Method III – Page Level co-oc (2/2)

- For each document, a co-occurrence matrix is generated and then summed up in a single matrix.
- POS tagger extracts the nouns, proper nouns and adjectives.
- Only the first k keyword are used, following an order based on tf*idf.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.61</td>
<td>0.40</td>
<td>0.61</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>b</td>
<td>0.61</td>
<td>0.40</td>
<td>0.61</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>c</td>
<td>0.61</td>
<td>0.40</td>
<td>0.61</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>d</td>
<td>0.61</td>
<td>0.40</td>
<td>0.61</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>f</td>
<td>0.61</td>
<td>0.40</td>
<td>0.61</td>
<td>0.20</td>
<td>0.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.3721</td>
<td>0.244</td>
<td>0.3721</td>
<td>0.122</td>
<td>0.122</td>
</tr>
<tr>
<td>b</td>
<td>0.244</td>
<td>0.16</td>
<td>0.244</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>c</td>
<td>0.3721</td>
<td>0.244</td>
<td>0.3721</td>
<td>0.122</td>
<td>0.122</td>
</tr>
<tr>
<td>d</td>
<td>0.122</td>
<td>0.08</td>
<td>0.122</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>f</td>
<td>0.122</td>
<td>0.08</td>
<td>0.122</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Co-occurrence matrix

Weighted co-occurrence matrix
Method IV – Page Level co-oc and term proximity

Co-occurrence at page level (method III) +

Term Proximity

\[
\begin{array}{cccccc}
1/5 & 1/4 & 1/3 & 1/2 & 1/1 \\
t1 & t2 & t3 & t4 & t5 & t6
\end{array}
\]
Personalization and Query Expansion

- In this research we implement method III

- Method III has better performance compared to others as we present in:

Query Expansion

- In Query Expansion process, we select the rows representing original query terms (ex. t2, t3).

![Table]

\[ Q = t2, t3 \]
Query Expansion

- Sum up selected rows
- Select the first $N$ terms (high values) of new vector.

$$N=1, \ Q = t_2, t_3$$

\[
\begin{array}{cccccc}
  & t_1 & t_2 & t_3 & t_4 & t_5 \\
 t_1 & 0.0 & 1.0 & 0.0 & 2.0 & 1.0 \\
 t_2 & 1.0 & 0.0 & 3.0 & 2.0 & 0.0 \\
 t_3 & 0.0 & 3.0 & 0.0 & 9.0 & 0.0 \\
 t_4 & 2.0 & 2.0 & 9.0 & 0.0 & 4.0 \\
 t_5 & 1.0 & 0.0 & 0.0 & 4.0 & 0.0 \\
\end{array}
\]
Co-occurrence matrices limits

- Semantic aspects:
  - In particular polisemey and homonimy

- For example:
  - Possible results:
    - http://www.amazon.com/
    - http://en.wikipedia.org/wiki/Amazon_River
  - User query : “amazon”
  - Expanded query : “amazon buy river”
Our solution for Co-occurrence matrices limits

- **Extension** of Co-occurrence matrix:
  - Introduction of metadata as third dimension of the matrix

- Use of **Social Bookmarking** services for metadata retrieval:
  - del.icio.us
  - stumbleupon.com
  - ...
Three-dimensional co-occurrence matrix structure
## Example

- **User query:** *amazon*

<table>
<thead>
<tr>
<th>Category</th>
<th>Expanded Query</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>e-commerce</td>
<td>amazon AND buy AND (book OR books)</td>
<td><a href="http://www.amazon.com/">http://www.amazon.com/</a> <a href="http://www.amazon.co.uk/">http://www.amazon.co.uk/</a> ...</td>
</tr>
<tr>
<td>nature</td>
<td>amazon AND (river OR rivers)</td>
<td><a href="http://en.wikipedia.org/wiki/Amazon_River">http://en.wikipedia.org/wiki/Amazon_River</a> <a href="http://www.mbarron.net/Amazon/">http://www.mbarron.net/Amazon/</a> ...</td>
</tr>
</tbody>
</table>
Experimentations

- The employed benchmark: *Lazio Region Portal Data* (LRDP)
- An example of topics is:
  - Top/Sala Stampa/Presidente/Biografia;
    - Level I: Sala Stampa;
    - Level II: Presidente;
    - Level III: Biografia
- Given the large quantity of links contained in LRPD, we decided to consider only level III links
Lazio is a region situated in the central of Italy, whose largest city is Rome.

Lazio Region is also the name of public administration that governs citizens of this region.

Like most Italian public administrations, Lazio has a web portal through which provides e-government services to its citizens.
Experimentations

- Each topic’s links were then subdivided in a training set, corresponding to 25% of the links, and set of tests, corresponding to 75% of the links.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Test links</th>
<th>Training links</th>
<th>Information Needs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricoltura</td>
<td>15</td>
<td>5</td>
<td>yes</td>
</tr>
<tr>
<td>Ambiente_e_Cooperazione</td>
<td>27</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Attivita_Produttive</td>
<td>74</td>
<td>18</td>
<td>yes</td>
</tr>
<tr>
<td>Protezione_Civile</td>
<td>52</td>
<td>14</td>
<td>yes</td>
</tr>
<tr>
<td>Bilancio_Economia_Partecipata</td>
<td>100</td>
<td>27</td>
<td>yes</td>
</tr>
<tr>
<td>Casa</td>
<td>35</td>
<td>7</td>
<td>no</td>
</tr>
<tr>
<td>Cultura_Spettacolo_Sport</td>
<td>25</td>
<td>6</td>
<td>no</td>
</tr>
<tr>
<td>Demanio_Personale_Patrimonio</td>
<td>26</td>
<td>7</td>
<td>no</td>
</tr>
<tr>
<td>Sanita</td>
<td>13</td>
<td>5</td>
<td>no</td>
</tr>
<tr>
<td>Protezione_Civile</td>
<td>15</td>
<td>5</td>
<td>no</td>
</tr>
<tr>
<td><strong>Tot.</strong></td>
<td><strong>382</strong></td>
<td><strong>101</strong></td>
<td></td>
</tr>
</tbody>
</table>
Experimentations

- We use for experimentations:
  - **Page Level Co-oc** metric for the Co-occourence matrix construction
  - del.icio.us Social Bookmarking service
  - F1-measure as performance indicator:

\[
\pi(t) = \frac{n_t}{50} \quad \rho(t) = \frac{n_t}{N_t} \quad F_1 = \frac{2 \times \pi \times \rho}{\pi + \rho}
\]

where \( n_t \) stands for the number of returned links belonging to topic t, only the first 50 pages are taken in consideration for our tests, and \( N_t \) the overall number of test links belonging to topic t present in the index.
Experimentations

- We have compared our system with:

  - a system based on a traditional content-based user-modeling approach, where documents are represented in the Vector Space Model and without Query Expansion (no QE)

  - system focuses on the update of the user model by means of Relevance Feedback (RF) techniques (no Social Bookmarking)
Experimentation

- The following table shows the results obtained:

<table>
<thead>
<tr>
<th>Topic</th>
<th>no QE</th>
<th>RF</th>
<th>Our System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiente_e_Cooperazione</td>
<td>0.05</td>
<td>0.08</td>
<td>0.16</td>
</tr>
<tr>
<td>Agricoltura</td>
<td>0.09</td>
<td>0.13</td>
<td>0.09</td>
</tr>
<tr>
<td>Protezione_Civile</td>
<td>0.10</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td>Attivita_Produttive</td>
<td>0.19</td>
<td>0.14</td>
<td>0.19</td>
</tr>
<tr>
<td>Bilancio_Economia_Partecipata</td>
<td>0.05</td>
<td>0.14</td>
<td>0.57</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td><strong>F_1</strong></td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.13</td>
<td><strong>0.24</strong></td>
</tr>
</tbody>
</table>
Thank you

claudio.biancalana@dia.uniroma3.it
micarel@dia.uniroma3.it
francescosaverio.profiti@laitspa.it