Extracting user interests from search query logs: a clustering approach

Lyes Limam, David Coquil, Lionel Brunie, Harald Kosch

Presented by: David Coquil
David.Coquil@uni-passau.de

DEXA-TIR Workshop, 30.08.2010
Introduction (1)

- User-centric systems
  - Design stage
  - Production stage

- Needs of online user-centricism
  - Gain knowledge from user interactions

- User logs analysis

Extracting user interests from search query logs: A clustering approach
Query logs analysis

Semantic analysis

Textual search queries analysis
  ◦ Semantically: identifying user interests
  ◦ Technically: a query terms clustering problem
Extracting user interests from search query logs: A clustering approach
What do we need in our method?

- Restructure the query logs to enable quantifying terms relationships
  - External source of semantic information
- Query terms clustering algorithm
- Semantic distance
WordNet as external source of semantics

- (English) WordNet
  - Large number of synsets
  - Hyponymy/(IS–A) relations

- Representation of the logs as a hierarchical structure
Preliminary phases

- Preprocessing
  - Elimination of unusable queries
  - Stop words

- Taxonomy construction process
  - Vocabulary
  - Hypernymy paths
  - Virtual nodes
Query term classification (Keywords Taxonomy)

- Global semantic representation of the log
- Defines a metric that measures the semantic distance between the terms
- A base for analysis
  - query terms clustering process
The distance function is defined as follows:

- \( G(V,E) \) a tree structure
  - \( V \) the set of terms
  - \( E \) the set of edges that models the relationships \textit{term1 is-a term2}

- Let “\( L \)” be a function which returns the level of an element
- The weight function “\( W \)” is defined on “\( E \)” as:

\[ \forall (u,v) \in E \: u \text{ is } \textit{a}\: v : W(u,v) = 1/ L(v) \]

- Let \( P = \{e_1, \ldots, e_n\} \) the set of edges in the path (unique) between \( x \) and \( y : (x,y) \in V^2 \)
- The distance function “\( D \)” is defined on \( V^2 \) as:

\[ \forall (x, y) \in V^2 : D(x, y) = \sum_{i=1}^{n} W(e_i) \]
Clustering Algorithm

- Groups terms whose all the distances are less than a threshold
- The clusters are constructed by pruning
  - The construction starts from the bottom
- The algorithm:
  - Is deterministic
  - Its complexity is $O(n)$, where $n$ the number of nodes

QUERY TERMS CLUSTERING ALGORITHM:

```
T    // Taxonomy with weighted links
E = {e0, e1...}  // set of query terms (nodes)
C = {}    // set of clusters
ci =     // ci C
D     // distance function
ts = Value    // threshold
While Not (empty(E))
  ed = deepest(E ) // find the deepest term
  ci = ci U {ed } // init. ci with the deepest term
  cluster_up(ed , parentOf(ed))
  C = C U {ci}
  E=E-{ci}
End

function cluster_up(predecessor, e)
  If  D(ed ,e) $\leq$ ts
    While has_children(e)
      if childOf(e) $\cdot$ predecessor
        cluster_down(pull_childOf(e))
      endif
      ci=ci U {e}
    end
  endif
  cluster_up(e , parentOf(e))
End

function cluster_down(e)
  If  D(ed ,e) $\leq$ ts
    While (has_children(e))
      cluster_down(pull_childOf(e))
      ci=ci U {e}
    endif
End
```

Extracting user interests from search query logs: A clustering approach
Clustering Algorithm

Cluster 1

Cluster 2

Cluster 3

Level $i=0$

Extracting user interests from search query logs: A clustering approach
Evaluation: test dataset

- AOL search logs
- 20 millions of queries collected over 650k users (USA) in a period of 3 months

<table>
<thead>
<tr>
<th>AnonID</th>
<th>Query</th>
<th>QueryTime</th>
<th>ItemRank</th>
<th>ClickURL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2771158</td>
<td>california hospital association</td>
<td>19.03.2006 23:16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2771158</td>
<td>glendale adventist medical center</td>
<td>19.03.2006 23:16</td>
<td>1</td>
<td><a href="http://www.glendaleadventist.com">http://www.glendaleadventist.com</a></td>
</tr>
<tr>
<td>2771158</td>
<td>free electronic greeting card</td>
<td>20.03.2006 22:47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2771158</td>
<td>csun webct</td>
<td>21.03.2006 08:01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2771158</td>
<td>the bodega</td>
<td>22.03.2006 01:29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2771158</td>
<td>the bodega pasadena</td>
<td>22.03.2006 01:29</td>
<td>1</td>
<td><a href="http://losangeles.citysearch.com">http://losangeles.citysearch.com</a></td>
</tr>
<tr>
<td>2771158</td>
<td>the bodega pasadena</td>
<td>22.03.2006 01:29</td>
<td>2</td>
<td><a href="http://www.pasadenacitycenter.com">http://www.pasadenacitycenter.com</a></td>
</tr>
<tr>
<td>2771158</td>
<td>el paseo mall pasadena</td>
<td>22.03.2006 01:35</td>
<td>2</td>
<td><a href="http://www.EngleKirk.com">http://www.EngleKirk.com</a></td>
</tr>
<tr>
<td>2771158</td>
<td>el paseo mall pasadena</td>
<td>22.03.2006 01:35</td>
<td>8</td>
<td><a href="http://www.rubios.com">http://www.rubios.com</a></td>
</tr>
<tr>
<td>2771158</td>
<td>the bodega el paseo mall</td>
<td>22.03.2006 01:37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2771158</td>
<td>the bodega el paseo mall</td>
<td>22.03.2006 01:37</td>
<td>13</td>
<td><a href="http://www.apa.udel.edu">http://www.apa.udel.edu</a></td>
</tr>
<tr>
<td>2771158</td>
<td>mapquest</td>
<td>22.03.2006 01:39</td>
<td>1</td>
<td><a href="http://www.mapquest.com">http://www.mapquest.com</a></td>
</tr>
<tr>
<td>2771158</td>
<td>hollywood fitness private trainers</td>
<td>22.03.2006 01:44</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation

- Objective cluster quality measures
- Manual study of cluster semantics
- Influence of threshold on cluster distribution
The threshold is determined experimentally by tuning: it balances small clusters and too general clusters.
Conclusion... Next step

- Efficient and fast user interests identification
- The threshold could be determined experimentally by tuning
- Clusters are inputs to the user communities discovery and resource aggregation processes

Next...
- Improvements/cluster quality evaluation
- Users profiles/similarity (overlap), resource aggregation
- Discover other potential applications in the “black box”
Thank you for your attention

Any questions?
Users community and resource aggregation

- Depending on the adopted approach (global or local) the users grouping process is realized as:
  - Global: two users are considered to be in one group if they share the same clusters
  - Local: two users are in the same group if their corresponding clusters overlaps
Experimentation (users–clusters)

1. Users with the same Nb of clusters Ts=0.6

2. Users with the same Nb of clusters Ts=1.0

3. Clusters with the same Nb of users 0.6

4. Clusters with the same Nb of users Ts=1.0

Extracting user search interests using a pruning algorithm
Outline

- Issue
- Framework for usage analysis
- Query terms clustering algorithm
- Experimentations
- Users community and resources aggregation
- Conclusion & Next step

Extracting user interests from search query logs: A clustering approach
In the context of clustering several improvement have been proposed:

- Include the co-occurrence relationship in the distance function:
  \[ D'(x,y) = \frac{D(x,y)}{C[x,y]} \]

- Include the terms frequency as it reflects the term importance
How to measure the efficiency of a distance/similarity measure?

- Use of human judgment/similarity measure correlation proposed by Miller and Charles, the MC correlation
  - 30 pairs of nouns rated (0–4) by 38 native English speakers
Existing algorithms for clustering

- Hierarchical algorithms
  - Single linkage
  - Complete linkage
  - Average linkage

- Partitioning algorithms
  - K-means

- Graph algorithms
  - Neighborhood graph algorithm (spanning tree)
  - B-coloring

Extracting user interests from search query logs: A clustering approach