An API-based Search System for One Click Access to Information

Dan Ionita
University of Twente
Enschede, The Netherlands

Niek Tax
University of Twente
Enschede, The Netherlands

Djoerd Hiemstra
University of Twente
Enschede, The Netherlands

Introduction

Next-generation search engines should be able to directly address user’s information needs more often by providing the right information on the result page for certain types of information needs. Chilton et al. [1] showed the current one click search implementation in Bing search to be regularly used, but also shows not every user’s intent is always addressed in current one click systems.

This gave birth to our research question: What is the best approach to design a One Click Access Information Retrieval system?

For one-click search engines, the two most important features are being able to correctly identify the query’s information need and using it to select the most relevant information sources. Web API’s, if selected properly, can provide consistent, parse-able responses which can easily be used to gain information about the query. Such information from multiple specialized API’s might then be used to classify the query as well as to construct a response.

In order to get insights in the best performing approaches we conducted a small literature study, focusing on the 1CLICK@NTCIR-9 participants. Based on findings in literature we implemented a prototype one click system.

Approach

We implement a 3-tier, API-based approach:
1. Retrieve features by parsing query and analyzing results returned by various APIs;
2. Classify into pre-defined categories based on retrieved features;
3. Extract information from various APIs depending on category;

Design and fine-tuning

Manually created a set of queries used to for initial training. Baseline user study was then conducted with two-fold purpose: (1) get an impression of the performance, (2) gather more training queries.

Resulting set of 240 queries (30 of each query category) used to evaluated the performance of various machine learning algorithms using the WEKA toolkit [2] and a stratified 10-fold cross validation. The Random Forest learner showed the highest classification accuracy, as can be seen in the confusion matrix below.

![Confusion Matrix]

Table 1: Confusion matrix for Random Forest (with bagging) classifier

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>&lt;= classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>22</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>b = ARTIST</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>c = POLITICIAN</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>d = ATHLETE</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>e = FACILITY</td>
</tr>
<tr>
<td>e</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>f = GEO</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>g = DEFINITION</td>
</tr>
<tr>
<td>g</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>h = QA</td>
</tr>
</tbody>
</table>

Feature retriever initially based on 1CLICK@NTCIR participants’ approaches. After user study, the performance evaluation was used to iteratively improve feature-set used for classification and APIs used for information extraction.

Results

The system was evaluated using the S#-measure metric as part of the 1CLICK@NTCIR evaluation (described in [2]). Figure 2 below maps the results for English-Desktop runs (character limit 1000).

![Results Map]

Both UT runs (implementing slightly different feature retrievers) scored highest amongst participant runs.

References


Acknowledgement

This publication was supported by the Dutch national program COMMIT and the Netherlands Organization for Scientific Research (NWO, project 639.022.809).