Topic specific search engines in civil engineering

Predrag Petronijević, Miloš Kovačević & Branislav Bajat
Faculty of Civil Engineering, University of Belgrade, Serbia

Keywords: search engines, standard queries, evaluation, construction equipment, geostatistics

Professionals in the construction sector make an intensive use of information in their decision-making processes, but only make limited use of the abundant information that is potentially available to them, particularly on the Web. The motivation of this research is the awareness that general purpose search engines (SE) such as Google® are not well suited for the information needs of the construction and building community (Zhang et al., 2004; Kovačević et al., 2008). A topic-specific search engine (TSE) could better respond to the needs of a specific community (Kovačević et al., 2008; Kovačević and Davidson, 2008). It is capable of finding, collecting and organizing only the domain specific information by eliminating the noise contained in Web pages. In this paper we evaluated the quality of the searching process from the experts’ point of view, comparing Grain and Google in a series of equivalent searching sessions.

The proposed approach is evaluated on two case studies (Construction equipment and Geostatistics) using the same methodology. There are three types of standard queries defined: query with one keyword (Q1), phrase query consisting of two keywords inside the quotes (Q2) and phrase query with two keywords + one additional keyword joined together with Boolean AND operator (Q3). We defined three types of query attributes: standard (std), category (cat) and cluster (cls). The attributes describe the particular usage of additional Grain functionality. Queries with the standard attribute use neither predefined subcategories filter, nor the clustering possibility. Queries with the category attribute use only predefined subcategories filter. Finally, the cluster attribute assumes that results produced by a query are evaluated using only the clustering functionality. The combination of query type and query attribute determines 6 different experiments that are performed in the research: (Q1, std), (Q1, cat), (Q2, std), (Q2, cat), (Q3, std) and (Q2, cls). Each experiment consists of several iterations depending on the number of queries of the related type to be posed. In all experiments experts posed queries both on Grain and Google and evaluated first 10 results per each query on both search engines. Unlike Grain, Google does not have a possibility to specify the predefined subcategory information. In the experiments with the cat attribute, standard queries of type Q1 and Q2 issued on Google are expanded with the additional keyword description of the related subcategory. For example in (Q2, cat) experiment, Google actually received two phrases (three or four keywords depending on the subcategory keyword description) compared to only one phrase issued on Grain. In the last experiment (Q2, cls), expert posed n queries of type Q2 and compared first 10 Google answers with the answers from Grain chosen in a following manner: when a query of type Q2 is issued, expert selects the most promising cluster by its name and evaluates 10 links. If the cluster does not contain enough links, then expert chooses another promising cluster. The procedure repeats until 10 links from Grain are selected.
Experts judged the corresponding page and assigned a relevance mark from 0 to 3. Mark 0 denotes irrelevant page to the topic of interest. If a page is in the topic but it is not informative with respect to the actual meaning of the query, then expert assigns 1. Pages that are on the topic and that contain many links to informative pages, but are not informative themselves are assigned 2 (i.e. portals). Such pages are named as hubs. Finally, mark 3 is assigned pages that are informative.

We defined three metrics for the evaluation of our TSE approach: for each experiment, expert judgments are recorded for all links produced using all queries participating in the experiment. Let \((Q, attr)\) be an experiment, where \(Q \in \{Q1, Q2, Q3\}\) and \(attr \in \{std, cat, cls\}\). Let \(q_1, q_2, \ldots, q_n\) be all the queries of type \(Q\). If the number of links produced from \((q_i, attr)\) is \(L^{(Q, attr)}(q_i)\) and the number of links that are evaluated as \(r \in \{x, 0, 1, 2, 3\}\) is \(L^{(Q, attr)}(r)\), then the percentage of relevance mark \(r\) when comparing to all given marks is given with Eq. 1. A query recall for the experiment \((Q, attr)\), is defined according to Eq. 2. Finally, the rank relevance is measured for each experiment. The idea of this metric is to take into account the position of the link in the answer list, giving the advantage to relevant links that are placed higher in the list. Let the experiment \((Q, attr)\) consists of \(n\) queries of the form \((q_i, attr)\) and consider the particular \((q_i, attr)\) which produces \(k\) links placed in the ordered list \((1 \leq k \leq 10)\) with assigned marks \(r_1, r_2, \ldots, r_k\). First we define the rank relevance of the particular answer list (Eq. 3). In Eq. 3 weights are assigned from 1 to 0.1 to correspond to list positions from 1 to 10. These weights are multiplied with the corresponding relevance marks. After summing weighted marks and normalizing the sum with the maximal score, we obtain \(rr(q_i, attr) \in [0, 1]\). The overall rank relevance is now given with Eq. 4.

\[
 r_{r_i}(Q, attr) = \frac{100 \cdot L^{(Q, attr)}(r)}{\sum_{i=1}^{n} L^{(Q, attr)}(q_i)} 
\]

\[
 qr(Q, attr) = 1 - \frac{L^{(Q, attr)}(x)}{\sum_{r=x, 0, 1, 2, 3} L^{(Q, attr)}(r)} 
\]

\[
 rr(q_i, attr) = \sum_{i=1}^{k} (1.1 - 0.1 \cdot i) \cdot r_i 
\]

\[
 RR(Q, attr) = \frac{\sum_{q_i} rr(q_i, attr)}{n} 
\]

In Construction Equipment community training process included 400 links (relevant pages) organized into 7 subcategories according to ISO/TR 1263:1966 Building construction machinery and equipment – Classification, published by the International Organization for Standardization (i.e. earth works). Many of the examples belong to more than one category. After the training, a crawling session lasted 3 days and the system collected nearly 200,000 potentially relevant pages. The expert from the field formulated 15 Q1, 35 Q2 and 15 Q3 queries (available upon request).

In Geostatistics community training process included 100 links organized into 5 subcategories according to the type of the source of information (i.e academic, commercial...). After the training, a crawling session lasted one day and the system collected nearly 21,000 potentially relevant pages. The expert from the field formulated 25 Q1, 25 Q2 and 25 Q3 queries (available upon request).

The obtained results showed significant advantage of the proposed approach in both case studies, especially in \(RR\) which is, by our opinion, the most important metric. However, it remains to future work to improve the query recall of our system which is now optimized for precision.

References

