EVALUATION OF THE PROPOSED TOPIC SPECIFIC SEARCH ENGINE IN GEOSTATISTICS

Miloš KOVAČEVIĆ1 Branislav BAJAT2,

[1]*Chair of managament and technology of building, Faculty of Civil Engineering, University of Belgrade, Bul. Kralja Aleksandra 73, 11000 Belgrade, Serbia,* milos@grf.rs

[2]*Department of geodesy and geoinformatics, Faculty of Civil Engineering, University of Belgrade Bul. Kralja Aleksandra 73, 11000 Belgrade, Serbia,* bajat@grf.rs

Abstract

The Internet is not used simply for general collecting of information, but also for obtaining the information which are essential for specialists in many scientific fields. This paper introduces a topic specific search engine (TSE) designed for geostatistics, the discipline which is mostly used in geoinformation and spatial modeling community. TSE is able to learn the desired topic using a machine learning approach, based on a small number of examples i.e. links that point to relevant pages. The solution for the search engine proposed in this work could better respond to the needs of a specific community in finding, collecting and organizing only domain specific information by eliminating the noise in Web pages. The quality of the searching process was evaluated from the experts's point of view, comparing proposed search engine and Google® in series of equivalent searching sessions. For that purpose three types of standard queries were defined: query with one keyword, phrase query consisting of two keywords inside the quotes and phrase with two keywords plus one additional keyword joined together with Boolean AND operator. A comparison between the proposed solution and standard searching procedure using Google® shows that our search engine outperforms standard search engine in almost all predefined evaluation metrics.

Keywords: search engines, standard queries, evaluation, geostatistics

INTRODUCTION

The internet is one of the most popular sources for gathering information and knowledge that facilitates the work of experts in many fields. It is especially important for scientists and academic community, since it is a fast and reliable tool for obtaining required information. Geoscientists and professionals use common methods and procedures in diverse geo-disciplines, and one of them is geostatistics. It is specific for geostatistics that it is aimed more to academic network of experts than to commercial sector. The nature of retrieved data is based on the general information about the geostatistics, that are available on the web pages (text content search), not to the spatial data sources. The information contained in the web pages is not structured systematically, lots of them are unreliable and there is a lot of "noise". Indeed, unlike the library, information on the Web is not accompanied by metadata.

In this research we propose a new searching paradigm which relies on the concept of a topic specific search engine. A topic specific search engine (TSE) could better respond to the needs of a specific community, such as geosciences sector. It is capable of finding, collecting and organizing only the domain specific information by eliminating the noise contained in pages on the Web. In our previous work (Kovačević et al., 2008; Kovačević and Davidson, 2008) we developed a construction-oriented search engine using artificial intelligence techniques and evaluated the intelligent acquisition component in terms of information retrieval measures. As in the previous case, the geostatistical TSE is focused on the text content of Web pages.

The main objective of this research is to estimate the quality of provided answers (pages) from the expert's point of view. For that purpose we created TSE –for the specific topic of *Geostatistics*. After formulating a set of standard queries in field of geostatistics, experts tested answers obtained both from the related TSE and Google®. The initial results are very promising indicating that our approach usually gives more precise results. Despite the majority of web pages from geostatistics in fact based on scientific work we have tried to develop a search engine that will be useful and enable fast and efficient browsing of pages that are designed for this area whether for research or commercial purpose.

PROBLEM STATEMENT

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 specific professional community (Zhang et al., 2004; Kovačević et al., 2008). They exhibit a lot of noise in the provided pages since many keyword combinations that form a user query are contained in different professional domains.

However, the pages are not logically organized and the searching process often fails after inspecting the list of first n provided links (usually n<10). The presence of many irrelevant pages and the lack of information organization lead to the increased searching time, or even abandoning the search process at all.

The problem could be partially solved by approaching it in a completely different way than modern search engines do. We developed a topic-specific search engine that is capable of learning the information requirements of the geosciences community. After providing a small number of positive examples (links that point to relevant pages), our tool is able to learn the desired topic using a machine learning approach. It then collects the similar pages performing the acquisition (or crawling) sessions over the web graph independently on other search engines. The basic idea is shown in Figure 1, taken from our previous work (Kovačević and Davidson, 2008).

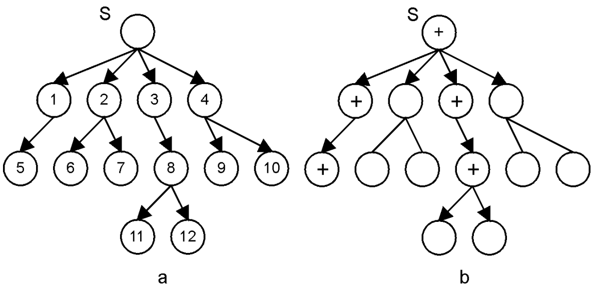

Fig. 1. Starting from a seed set S, ordinary SE collects all pages that can be reached from S. It visits the web graph following the order defined with numbers in the circles – pages (part a). Part b: TSE starts from the seed set of relevant examples. It then follows only promising links in order to collect relevant pages denoted with $\oplus$.

TSE uses the fact that the Web is a social network in which page owners tend to reference similar pages (Menczer, 2004). It collects only relevant pages by applying a classifier previously trained only on the positive examples, pruning branches that lead from irrelevant ones (classifier decides whether a visited page is relevant or not). This way TSE could maintain a highly relevant database of links, with low level of noise. Relevant pages are further classified into predefined topic subcategories, since the examples used for training can be labeled to belong to certain predefined classes. The existence of subcategories enables a user to narrow the search to a predefined subset of the main topic. Finally, the advantage of the TSE approach is the capacity to find new pages in the community faster and keep the index database more current when compared to standard SE, since it revisits only a small subset of the whole Web.

Technical details about the design and the architecture of our TSE named GeoSPACE are described in (Kovačević et al., 2008; Kovačević and Davidson, 2008). In this paper we evaluated the quality of the searching process from the experts' point of view, comparing GeoSPACE and Google in a series of equivalent searching sessions.

PROPOSED APPROACH

GeoSPACE is a complete solution for building an independent TSE. It is a distributed software application implemented in Java and related open-source technologies. It consists of 5 main parts: teacher (T), intelligent crawler (IC), indexing engine (IE), clustering engine (CE) and user web interface (WI). The simplified scheme of the system is shown on Figure 2.

Teacher implements the logic for training relevance and subcategories classifiers. Assume that one creates a Geostatistics TSE. After providing links to pages from the desired community, teacher downloads positive examples and trains the relevance classifier. Relevance classifier will lead the intelligent crawler during the crawling session on the web graph (Figure 1). Provided links can be accompanied with subcategory labels (i.e. academic, application …) in order to train the subcategories classifier.

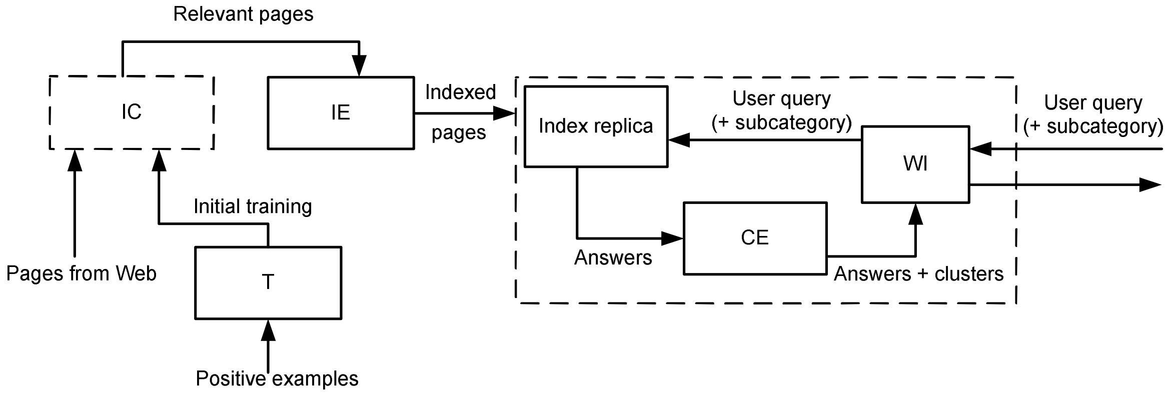

Fig. 2. Simplified view of the system architecture. The whole system can be installed on one machine; dashed boxes represent components which can be replicated on more than one machine.

After the training phase, intelligent crawler performs the crawling session, trying to collect only relevant pages. Relevant pages are further classified into predefined categories and transferred to the indexing engine for indexing. Indexer is built using Lucene framework (http://lucene.apache.org).

Users pose standard keyword queries using web interface. Queries can be accompanied with the desired subcategory filter in order to narrow the search. When a query is issued, it is processed against the index and the answers are clustered in the real-time by the clustering engine. It is built using the Carrot2 framework (http://www.carrot2.org). Finally, a user is presented with the list of answers (link + summary) on the right side of the user window; on the left side, the same answers are grouped by similarity into logical clusters. The clusters are not known in advance like predefined subcategories. They are created depending on the query and furthermore help a user to accelerate the search.

The underlying methods for training relevance and subcategories classifier are based on the Support Vector Machines (SVM) method (Vapnik, 1995) and its variants. According to (Caldas et al., 2002), Support Vector Machines is the best performing method compared to other commonly used approaches when dealing with diverse construction project documents. Our system learns only from the small set of positive examples. A Positive Examples Based Learning approach (Yu et al., 2004) is used to extract the negative class description from unlabelled data freely available on the Web. In our setting we used 10,000 unlabelled pages downloaded from the Open Directory Project® archive (http://dmoz.org). Clustering engine uses Lingo clustering algorithm (Osinski and Weiss, 2005) which is independent of the language of the collected pages.

EXPERIMENTAL EVALUATION

The proposed approach is evaluated on the "case study" using the following 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). It is worth mentioning that the teaching component of our system can generate significant phrases of length *n* using Information Gain method (Yang and Pedersen, 1997) in which the unlabeled set is used as a negative class. A list of 100 most informative phrases of length two is given to the related expert in each case study, to help in formulating Q2 queries.

We defined two types of query attributes: *standard* (*std*), and *category* (*cat*). The attributes describe the particular usage of additional GeoSPACE 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. The combination of query type and query attribute determines 5 different experiments that are performed in the research: (Q1, *std*), (Q1, *cat*), (Q2, *std*), (Q2, *cat*), and (Q3, *std*). We defined five different categories, namely: *academic*, *commercial*, *software*, *research* and *application*. The category academic relates to web pages intended for students or people who want to get more information about the possibilities for education in the field of geostatistics (university home pages, etc). Commercial category primarily refers to the institutions whose services are based on the application of geostatistics, as well as commercial software that contains geostatistical modules. The category software refers both to open-source or commercial geostatistical software. The research category is mainly focused on web pages containing scientific articles, and finally, category application covers web pages which describe geostatistical applications in different fields of human activities. The examples of all predefined queries are illustrated in the Table 1. The choice of the combination of the words that compose quires is based on the expert knowledge on the frequency of their occurrence within the predefined categories. 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 GeoSPACE and Google and evaluated first 10 results per each query on both search engines. Unlike GeoSPACE, Google does not have a possibility to specify the predefined subcategory information. Hence, in 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 GeoSPACE.

Table 1. Examples of queries.

query	category				
	Academic	Commercial	Software	Research	Application
Q1	Kriging, regression, error,…	estimation, analysis, expertise,…	mapping, variance, validation,…	pedometrics, topography, accuracy,…	geology, maps, DTM,…
Q2	regression analysis, variogram modeling, error propagation,…	environmental modelling, data analysis, geospatial applications,…	contour maps, multivariate analysis, open source,…	conditional simulation, variogram analysis, trend analysis,…	soil mapping, risk analysis, groundwater modeling,…
Q3	variogram modeling AND residual, semivariance analysis AND error,…	environmental modelling AND assessment, data analysis AND evaluation,…	spatial regression AND correlation, cross validation AND variance,…	conditional simulation AND fuzzy, regression kriging AND drift,…	risk analysis AND mapping, applied geostatistics AND geology,…

Expert 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 to the page that is informative with respect to the issued query. Sometimes, GeoSPACE does not provide enough answers (<10) since it is currently optimized for the precision, and not for the recall (coverage of the complete community). In such cases expert assigns mark *x* (missing).

Three metrics were defined 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∈{Q1,Q2,Q3} and *attr*∈{*std*, *cat*}. Let q_1, q_2,…, 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∈{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 in Eq.1.

$$r_{\%}(Q,attr)=\frac{100\cdot L^{(Q,attr)}(r)}{\sum_{i=1}^{n}L^{(Q,attr)}(q_i)} \qquad (1)$$

A query recall for the experiment (Q, *attr*), is defined with Eq. 2.

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

In addition, we measured the *rank relevance* for each experiment. The idea of this measure 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 ($1\le k\le 10$) with assigned marks r_1, r_2,…, r_k. First we define the rank relevance of the particular answer list (Eq. 3).

$$rr(q_i,attr)=\frac{\sum_{i=1}^{k}(1.1-0.1\cdot i)\cdot r_i}{3\cdot\sum_{i=1}^{k}(1.1-0.1\cdot i)} \qquad (3)$$

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 over all links and normalizing the sum with the maximal score, we obtain $rr(q_i,attr)\in[0,1]$. In order to obtain the overall rank relevance it is needed to average $rr(q_i,attr)$ over all queries in the experiment (Eq. 4).

$$rr(q_i,attr)=\frac{\sum_{i=1}^{k}(1.1-0.1\cdot i)\cdot r_i}{3\cdot\sum_{i=1}^{k}(1.1-0.1\cdot i)} \qquad (4)$$

Training process included 100 links (relevant pages) organized into 5 categories according to the type of the source of information (i.e *academic*, *commercial*…). Many of the examples belong to more than one category. 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 percentage of Q2 queries that are suggested by the system and accepted by the expert is 44%. The results for the first five experiments are presented in Tables 2, 3 and 4.

Table 2. Distribution of $r_{\%}(Q,attr)$. Google(G) and GeoSPACE (P).

	Q1, Standard				Q1, Category				Q2, Standard				Q2, Category					Q3, Standard				
	0	1	2	3	0	1	2	3	0	1	2	3	x	0	1	2	3	x	0	1	2	3
G	64	27	5	4	68	21	8	3	16	51	22	11	0	30	35	20	15	0	22	45	24	9
P	14	35	34	17	9	25	35	31	1	29	31	39	12	2	23	18	45	23	3	20	19	35

x- denotes missing pages (TSE returned less than 10 web sites)

Table 3. Query recall *qr.*

	Q1, Standard	Q1, Category	Q2, Standard	Q2, Category	Q3, Standard
Google	1	1	1	1	1
GeoSPACE	1	1	1	0.88	0.77

Table 4. Rank relevance.

	Q1, Standard	Q1, Category	Q2, Standard	Q2, Category	Q3, Standard
Google	0.18	0.16	0.44	0.42	0.42
GeoSPACE	0.50	0.63	0.68	0.74	0.72

The GeoSPACE shows better distribution of relevance marks in all experiments since the recall is much higher in this case (Table 3). The higher recall could be explained with the usage of only one keyword for the subcategory description when comparing to the previous case study (two keyword phrase) GeoSPACE provides higher relevance marks for all types of queries, which is by our opinion the most important metric. A reader could try the system on the following address http://geospace.grf.bg.ac.rs.

CONCLUSION

In this paper we developed a set of metrics for evaluation of previously designed intelligent, topic specific search engine, specialized for the needs of the geosciences communities. A comparison between the proposed approach and the standard searching procedure using Google showed the significant advantage of topic-specific search in conducted case study –*Geostatistics*. The presented approach could be successfully used in other professional domains. The main advantage of our geostatistics search engine is reflected in noise reduction (non-geostatistics pages) and hierarchical presentation of the results (predefined categories filter and real-time clustering of similar answers), thus leading to more efficient search for geostatistics information contained in web pages. The TSE utilized for geostatistics is the first step in project that embraces subsequent developing of TSEs specialized for other geoscientific fields (GIS, web cartography, geomatics, etc). The agglomeration of specific TSEs dedicated to particular geosciences, organized in a unique web portal, should serve to provide broad information in geosciences. The portal could be more suited to the particular needs of geoscientists, when compared to standard, general purpose search engines.

REFERENCES

Caldas, C., Soibelman, L., and Han, J., 2002. Automated Classification of Construction Project Documents. *Journalof Computing in Civil Engineering*, 16(4), 234-243

Kovačević, M, Nie, J-Y, and Davidson, C., 2008. Providing Answers to Questions from Automatically Collected Web Pages for Intelligent Decision-Making in the Construction Sector. *Journal of Computing in Civil Engineering*, 22(1), 3 – 13

Kovačević, M., and Davidson, C., 2008. Crawling the Construction Web – A Machine Learning Approach without Negative Examples. *Applied Artificial Intelligence*, 22(5), 459 – 482

Menczer, F., 2004. Lexical and semantic clustering by web links. *Journal of the American Society for Information Science and Technology*, 55(4), 1261-1269

Osinski, S. and Weiss, D., 2005. A Concept-Driven Algorithm for Clustering Search Results. *IEEE Intelligent Systems*, 20(3) 48-54.

Vapnik,V., 1995. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Yang, Y. and Pedersen, J. O., 1997. A comparative study on feature selection in text categorization. *In*: *14th international Conference on Machine Learning, 1997, Nashville, USA*

Yu, H., Han, J. and Chang, K., 2004. PEBL: Web page classification without negative examples, *IEEE Transactions on Knowledge and Data Engineering*. 16(1), 1-12

Zhang, Z., Da Sylva, L., Davidson, C.H., Lizarralde, G., and Nie, J.-Y. 2004. Domain-Specific Q-A for the Construction Sector. *In: Information Retrieval for Question Answering*, *SIGIR 2004 Workshop*, *Sheffield, U.K*