A Supervised Learning Algorithm for Binary Domain Classification of Web Queries using SERPs

Alexander C. Nwala and Michael L. Nelson
Old Dominion University, Department of Computer Science
Norfolk, Virginia, 23529
{anwala, mln}@cs.odu.edu

ABSTRACT

General purpose Search Engines (SEs) crawl all domains (e.g., Sports, News, Entertainment) of the Web, but sometimes the informational need of a query is restricted to a particular domain (e.g., Medical). We leverage the work of SEs as part of our effort to route domain specific queries to local Digital Libraries (DLs). SEs are often used even if they are not the "best" source for certain types of queries. Rather than tell users to "use this DL for this kind of query", we intend to automatically detect when a query could be better served by a local DL (such as a private, access-controlled DL that is not crawlable via SEs). This is not an easy task because Web queries are short, ambiguous, and there is lack of quality labeled training data (or it is expensive to create). To detect queries that should be routed to local, specialized DLs, we first send the queries to Google and then examine the features in the resulting Search Engine Result Pages. Using 400,000 AOL queries for the "non-scholar" domain and 400,000 queries from the NASA Technical Report Server for the "scholar" domain, our classifier achieved a precision of 0.809 and F-measure of 0.805.

CCS Concepts

•**Information systems** → *Clustering and classification;*

Keywords

Search Engines, Web queries, Query Understanding.

1. INTRODUCTION

In this paper we focus on domain classification of queries which targets two classes - the Scholar domain and the non-Scholar domain. The Scholar domain targets queries associated with academic or research content. For example, queries such as "*fluid dynamics*", "*stem cell*", "*parallel computing*" belong to the Scholar domain, while queries such as "*where to find good pizza*", "*bicycle deals*", and "*current weather*" belong to the non-Scholarly domain. In this work, we propose a novel method which does not rely on processing the actual query. Instead, we trained a classifier based on the features found in a Google SERP (Search Engine Result Page). The classifier was trained and evaluated (through 10-fold cross validation) with a dataset of 600,000 SERPs evenly split across both classes and the results were validated on 200,000 SERPs evenly split across both classes yielding a classification precision of 0.806 and F-measure of 0.805. We targeted a binary class, however our method could be scaled to accommodate other classes if the right features are found.

JCDL '16 June 19-23, 2016, Newark, NJ, USA

ACM ISBN 978-1-4503-4229-2/16/06.

DOI: http://dx.doi.org/10.1145/2910896.2925449

2. RELATED WORK

The problem of domain classification has been studied extensively. Jingbo et al. [4] built a domain knowledge base from web pages for query classification. Gravano, et al. [1] built a classifier targeting the geographical locality domain. Dou Shen et al. [3] built an ensemble-search based approach for query classification in which queries are enriched with information derived from Search Engines.

The query domain classification problem is not new and since queries are short, ambiguous, and in constant flux, maintaining a labeled training dataset is expensive. Therefore, we use Google SERPs instead of processing the query directly.

3. LEARNING ALGORITHM FOR DOMAIN CLASSIFICATION

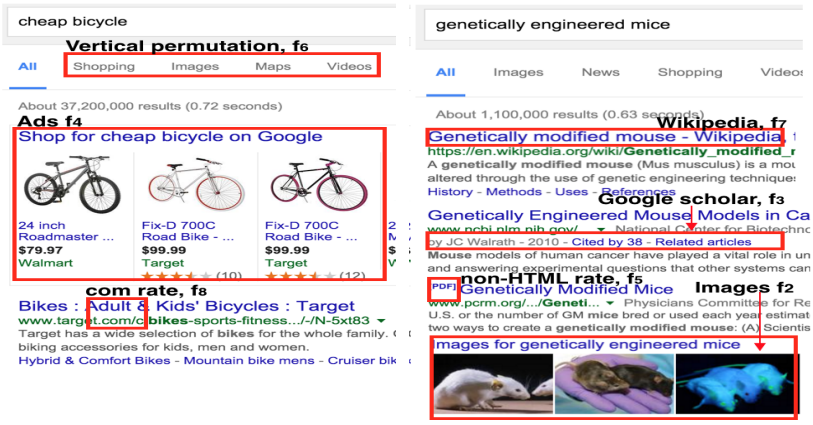
Our solution can be summarized in two stages:

Stage 1. Building the classifier: First, identify the discriminative features. Second, build a dataset for the Scholar domain class and non-Scholar domain class. Third, train a classifier. Fourth, evaluate the classifier using 10-fold cross validation.

Stage 2. Classifying a query: First, issue the query to Google and download the SERP. Second, extract the features. Finally, use the classifier hypothesis function to make a prediction.

Feature Identification and Dataset Building: After extensive study, we identified 10 features to be extracted from the Google SERP:

The binary features ***Knowledge Entity*** - f_1 and ***Images*** - f_2 (Fig. 1b) represent the presence or absence of the Google knowledge entity and images respectively. The binary features ***Google Scholar*** - f_3 (Fig. 1b) and ***Wikipedia*** - f_7 (Fig. 1b) represent the presence or absence of a citation and a Wikipedia page respectively. The feature ***Ad ratio*** - $f_4 \in [0, 1]$ (Fig. 1a) represents the proportion of



(a) **non-Scholar SERP** (b) **Scholar SERP**

Figure 1: Example of features in the non-Scholar SERP (Left) and Scholar SERP (right). Non-Scholar queries typically feature "shopping" in the Vertical permutation (f_6), Ads (f_4), and are dominated by *com* TLDs (f_8). Scholar queries typically feature results from Wikipedia (f_7) and Google Scholar (f_3), as well as *non-HTML* types (f_5).

ads on the SERP. The feature ***com rate*** - $f_8 \in [0,1]$ (Fig. 1a) represents the proportion of *com* links on the SERP. The feature ***non-HTML rate*** - $f_5 \in [0,1]$ (Fig. 1b) represents proportion of non-*html* types on the SERP. The feature ***Vertical permutation*** - $f_6 \in [1,336]$ (Fig. 1a) represents the ${}^8P_3$ possible page order of the SERP. The feature ***Maximum Title Dissimilarity*** - $f_9 \in [0,1]$, represents the maximum dissimilarity value between the query and all SERP titles: Given a query q, with SERP title t_i, and longest SERP title T, and Levinshtein Distance function LD, $f_9 = \max_{t_i} \frac{LD(q,t_i)}{|T|}$. Finally, the feature ***Maximum Title Overlap*** - $f_{10} \in \mathbb{Z} \in [0, \max_{t'_i} |q' \cap t'_i|]$, represents the cardinality of the maximum common value between the query and all SERP titles: Given a query set q with SERP title set t'_i, $f_{10} = \max_{t'_i} |q' \cap t'_i|$

Figures 1a and 1b present a subset of the features for both classes. Note the ads present in the non-Scholar SERP and the absence of PDF documents. For the Scholar SERP, note the presence of a Wikipedia page, the PDF document and the Google Scholar article. Therefore, at scale, we could learn from not just the presence of a feature, but also its absence. After feature identification, we downloaded the Google SERPs for 400,000 AOL 2006 queries [2] and 400,000 NTRS (NASA Technical Report Server) 1995-1998 queries for the non-Scholar and Scholar datasets, respectively. Our method is not without limitations. For example, the datasets are presumed "pure." But this is not the case since Scholar queries exist in the non-Scholar dataset and vice versa, thus contribute to classification errors.

Classifier Training and Evaluation: Using Weka, we built a logistic regression model (Eqn. 1) on a 600,000 dataset evenly split across both classes. The model was evaluated using 10-fold cross validation yielding a classification precision of 0.809 and F-Measure of 0.805.

$$g_q = 2.7585 + \sum_{i=1}^{10} c_i f_i \tag{1}$$

The coefficients matrix C^T contains the coeffients c_i for each feature f_i: $C^T = [c_1 \ldots c_{10}]$, C^T = [0.8266, -1.1664, -2.7413, -1.7444, 6.2504, -0.0017, -1.0145, -1.5367, 1.8977, -0.1737]

Classifying a Query: To find the domain $d \in$ {Scholar, non-Scholar} of a query q,

First: issue the query to Google and download SERP.

Second: initialize all feature values $f_1 \ldots f_{10}$,

Third: use Eqn 1. to estimate g_q

Fourth: use logistic regression hypothesis (Eqn 2.) to estimate the class probability.

$$p(q) = \frac{e^{g_q}}{1 + e^{g_q}} \tag{2}$$

Fifth: If $p(q) \geq 0.5$ predict $d = scholar$, else predict $d = non - scholar$

4. CONCLUSIONS

We define a set of features in SERPs that indicate if the domain of a query is scholarly or not. Our classifier which has a precision of 0.809 and F-measure of 0.805 can be further applied to other domains once discriminative and informative features are identified.

5. REFERENCES

[1] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 325–333, 2003.

[2] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*, 2006.

[3] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Q2C@UST: our winning solution to query classification in KDDCUP 2005. *ACM SIGKDD Explorations Newsletter*, 7(2):100–110, 2005.

[4] J. Yu and N. Ye. Automatic web query classification using large unlabeled web pages. In *Web-Age Information Management, 2008. WAIM'08. The Ninth International Conference on*, pages 211–215, 2008.