Semantic Features for Classifying Referring Search Terms

Chandler May, Michael Henry, Liam McGrath, Eric Bell, Eric Marshall, and **Michelle Gregory**
Pacific Northwest National Laboratory
{chandler.may, michael.j.henry, liam.mcgrath, eric.bell, eric.marshall, michelle}@pnnl.gov

1 Introduction

When an internet user clicks on a result in a search engine, a request is submitted to the destination web server that includes a referrer field containing the search terms given by the user. Using this information, website owners can analyze the search terms leading to their websites to better understand their visitors' needs. This work explores some of the features that can be used for classification-based analysis of such referring search terms. We present initial results for the example task of classifying HTTP requests' countries of origin. A system that can accurately predict the country of origin from query text may be a valuable supplement to IP lookup methods which are susceptible to the obfuscation of dereferrers or proxies. We suggest that the addition of semantic features improves classifier performance in this example application. We begin by looking at related work and presenting our approach. After describing initial experiments and results, we discuss paths forward for this work.

1.1 Prior work

To the best of our knowledge, there does not exist a body of work specifically in the field of country identification from text features in referring search terms. However, similar work has been conducted in identifying speaker origin from text (Sehgal, 2004) and corpus profiling (Granger and Rayson, 1998).

In (Sehgal, 2004), the researcher attempted to detect a speaker's native language from their pattern of speaking English. Text transcriptions of speakers from four European and two Asian languages were analyzed using n-grams to determine if a speaker's country of origin could be determined from text features alone. High accuracy was obtained in this task, and bi-grams were the most useful features.

In (Granger and Rayson, 1998), researchers sought to identify terms in a corpus that distinguish it from other corpora. This approach was applied to the comparison of the written text from native and non-native English speakers. The researchers found that the text from the non-native English speakers more closely compared to spoken transcriptions from native English speakers, showing that a differentiation can be made between speakers of different languages based on text features.

Referring search terms have been used to analyze general-purpose (Silverstein et al., 1999) and site-specific (Chau et al., 2005) search engine logs. Time series analysis has also been applied to search engine logs to predict user behavior (Zhang et al., 2009). Finally, referring search terms have been used to evaluate user intent (Stolz et al., 2006) and to investigate "Self-Googling" (Nicolai et al., 2008).

2 Approach

We extract HTTP requests with referring search terms from a web traffic log, group them into "sessions," and try to classify the sessions by their respective countries of origin using combinations of the browser's accepted languages from the `Accept-Language` HTTP header (Casado and Freedman, 2007), the referring search terms, and semantic features of the referring search terms as the explanatory variable. Before each experiment, the inputs are converted to a word vector representation. Ten-fold cross-validation is employed. The gold standard is an IP lookup, so the experiments are not intended to show whether the selected features can overcome IP lookup's shortcomings, but to compare the features' usefulness in classification.

The distribution of sessions by country in our dataset exhibits faster than exponential decay. As a workaround, we perform our experiments on a subset of 25 countries whose session counts are on a similar order of magnitude.

2.1 Entity, concept, and domain analysis

A given session's referring search terms are concatenated and processed by Zymurgy, a tool that identifies entities, concepts, and domains in a text document (Schone et al., 2009). The concepts identified come from the Omega ontology (Philpot et al., 2005). Zymurgy also computes a weight associated with each entity, concept, or domain that measures the importance of the feature in the text. This weight is used as the value of the entity, concept, or domain in the feature vector.

For example, Zymurgy outputs five features (with weights) for "thomas bayes publications": one domain, `publishing-subject` (0.2841); three concepts, `publication<work` (1.0), `work<product` (0.2998), and `product` (0.2034); and one entity, `thomas_bayes` (1.0), whose type is `person`.

2.2 Sessions

The HTTP requests containing referring search terms are partitioned into "sessions" in order to provide more context-rich documents to Zymurgy. A definition of "session" has not been settled upon in the literature (Gayo-Avello, 2009). In our study, the sequence of requests that forms a session is the maximal sequence of requests sharing the same IP address and user agent and for which adjacent requests are separated by at most ten minutes, a time interval found effective previously (He and Goker, 2000).

3 Results

We now present preliminary results for three country classification experiments. We first use the browser language word vector as the explanatory variable. We then add the word vector of referring search terms. Lastly, we add the referring search terms and entities (domains and concepts prove less useful).

For each experiment, average precision, recall, F_1-measure, and area under the ROC curve are shown in Table 1. The increase in ROC area due to adding referring search terms to the browser language suggests that the referring search terms have some utility in predicting country of origin. The further increase due to adding semantic features indicates that those semantic features may also be useful in prediction. ROC area improved for 17 of the 25 countries when semantic features were added, which suggests there is some robustness across classes.

Experiment	P	R	F_1	A_{ROC}
L	0.707	0.611	0.624	0.910
L+T	0.692	0.632	0.636	0.922
L+T+S	0.693	0.629	0.642	0.924

Table 1: Average precision (P), recall (R), F_1-measure, and area under ROC curve (A_{ROC}) for the three experiments: browser language only (L), browser language with referring search terms (L+T), and browser language with referring search terms and semantic features thereof (L+T+S). The third experiment demonstrates the best performance, suggesting that referring search terms and their semantic features may both have predictive value.

4 Conclusion

Here we have investigated the effectiveness of semantic features for classifying search terms from referring search terms, in an initial experiment to categorize sessions by country. We have found that including semantic features improves overall classification performance, although additional tests are needed to demonstrate statistical significance.

4.1 Future work

Other potential applications of classifying sessions include analyzing user intent (Stolz et al., 2006), predicting user behavior (Zhang et al., 2009), and investigating search term relevance as it relates to website content (Chau et al., 2005).

A more sophisticated session identification procedure (Gayo-Avello, 2009), perhaps through consideration of semantic features, could improve accuracy of classification. Better word-sense disambiguation, possibly through incorporation of text from the website visited after a query, could improve accuracy of the application.

Report number: PNNL-SA-86963.

References

Martin Casado and Michael J. Freedman. 2007. Peering through the shroud: the effect of edge opacity on ip-based client identification. In *Proceedings of the 4th USENIX conference on Networked systems design & implementation*, NSDI'07, pages 13–13, Cambridge, MA. USENIX Association.

Michael Chau, Xiao Fang, and Oliva R. Liu Sheng. 2005. Analysis of the query logs of a web site search engine. *Journal of the American Society for Information Science and Technology*, 56(13):1363–1376, August.

Daniel Gayo-Avello. 2009. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(1):1822–1843.

S. Granger and P. Rayson. 1998. *Automatic Profiling of Learner Texts*. Longman, London.

D. He and A. Goker. 2000. Detecting session boundaries from web user logs. In *Proceedings of the 22nd Annual Colloquium on Information Retrieval Research*, pages 57–66. Cambridge, April.

Thomas Nicolai, Lars Kirchhoff, Axel Bruns, Wilson Jason A., and Barry J. Saunders. 2008. Google yourself! measuring the performance of personalized information resources. In *Assocation of Internet Researchers 2008: Internet Research 9.0: Rethinking Community, Rethinking Place*. Queensland University of Technology.

A. Philpot, E. Hovy, and P. Pantel. 2005. The omega ontology. In *Proceedings of the ONTOLEX Workshop at IJCNLP 2005*.

P. Schone, A. Goldschen, C. Langley, S. Lewis, B. Onyshkevych, R. Cutts, B. Dawson, J. MacBride, G. Matrangola, C. McDonough, C. Pfeifer, and M. Ursiak. 2009. TCAR at TAC-KBP 2009. In *Proceedings of the Second Text Analysis Conference*, TAC-KBP 2009, Gaithersburg, MD. National Institute of Standards and Technology.

Siddharth Sehgal. 2004. Identification of speaker origin from transcribed speech text. Master's thesis, University of Sheffield.

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, September.

Carsten Stolz, Michael Barth, Maximilian Viermetz, and Klaus D. Wilde. 2006. Searchstrings revealing user intent: A better understanding of user perception. In *Proceedings of the 6th International Conference on Web Engineering*. ACM.

Ying Zhang, Bernard J. Jansen, and Amanda Spink. 2009. Time series analysis of a web search engine transaction log. *Information Processing and Management*, 45(2):230–245, March.