Annotation of URLs: More than the Sum of Parts

Max Hinne
Dept. Computer Science,
Radboud University Nijmegen
mhinne@sci.ru.nl

Wessel Kraaij
Dept. Computer Science,
Radboud University Nijmegen
TNO, Delft
kraaijw@acm.org

Stephan Raaijmakers
TNO, Delft
stephan.raaijmakers@tno.nl

Suzan Verberne
CLST,
Radboud University Nijmegen
s.verberne@let.ru.nl

Theo van der Weide
Dept. Computer Science,
Radboud University Nijmegen
tvdw@cs.ru.nl

Maarten van der Heijden
Dept. Computer Science,
Radboud University Nijmegen
m.vanderheijden@cs.ru.nl

ABSTRACT

Recently a number of studies have demonstrated that search engine logfiles are an important resource to determine the relevance relation between URLs and query terms. We hypothesized that the queries associated with a URL could also be presented as useful URL metadata in a search engine result list, e.g. for helping to determine the semantic category of a URL. We evaluated this hypothesis by a classification experiment based on the DMOZ dataset. Our method can also annotate URLs that have no associated queries.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]

General Terms: Experimentation, Human Factors

1. INTRODUCTION

In this study, we investigate the applicability of semantic annotation of web pages by creating short document descriptions (term lists) extracted from associated queries. We assume that, when presented to a user, these term lists may help in the disambiguation of a URL and/or identifying whether the URL corresponds to the user's query intent [2]. In previous work, we conducted a pilot experiment that demonstrated that document descriptors extracted from the queries associated with URLs provide useful semantic information about documents in addition to descriptors extracted from the full text of the web pages [8].

In the current paper, we explore which level of URL generalization (exact URL, domain, or separate URL terms) yields the most informative associations from the query logs. We use the most salient terms extracted from the (weighted) set of query terms associated with an URL as a description for that URL. Specifically, we aim to find out whether the query terms associated with the individual URL terms can be used to construct a generative model to generate a query term description for any URL based on a combination of the individual URL terms. To answer these questions we conduct a URL classification experiment, comparing different query term based representations.

2. RELATED WORK

It has been shown that queries and click information can serve as a means for *implicit tagging* [3]. Several studies have investigated whether a collection of queries and click information can serve as an (improved) model of the semantic contents of a web page e.g. [1]. A document representation based on queries has been compared with a traditional document content vector space representation for a clustering task [5]. The query-based representation resulted in a better clustering than the content-based representation. Working with the site access logs of a portal means that log files are complete. Our experiments differ since they comprise a much larger set of web pages, are based on a search engine log file (and therefore query logs are incomplete) and classify URLs without any specific associated queries. A similar comparison study has been carried out between a content-based document representation, a tag-based representation (*del.icio.us*) and an anchor text representation [6]. The tag-based representation outperformed the text and anchor-based representation in a clustering task. Web page classification experiments regularly use the DMOZ Open Directory RDF Dump[1] as a resource. Individual URL elements have been suggested as features for web page classification [4]. Our work differs since it explicitly models the relation between URL elements and associated query terms.

3. METHOD

The objective of our experiments is to evaluate whether terms derived by aggregating query log data provide useful hints for disambiguation or identifying query intent. To evaluate the value of query log data, we performed an URL classification experiment using different sets of terms extracted from the URL itself and the associated queries.

Data: The Microsoft 2006 RFP dataset consists of approximately 14 million queries from US users entered into the Microsoft Live search engine in the spring of 2006. For each query the following details are available: a query ID, the query itself, the user session ID, a time-stamp, the clicked URL, the rank of that URL and the number of results.

SIGIR'09, July 19-23,2009,Boston,Massachusetts,USA.

[1] `http://rdf.dmoz.org/`

For the classification task we used the DMOZ Dump, consisting of URLs and their class labels (e.g. bikeriderstours.com — Top/ Sports/ Cycling/ Travel/ Tour_Operators). We restricted the data to DMOZ level 2 labels (e.g Top/ Sports). We discarded the URLs labeled *Top/Regional*, since *Regional* is the top node of a different hierarchy (a regional classification). The intersection of the MSN and DMOZ collections (with the above restriction) consists of 109,493 URLs, divided over 15 classes.

Experiments: To determine the most descriptive and discriminating query terms associated with a given URL, we calculated the pointwise Kullback-Leibler divergence between each pair of a query term and associated URL:

$$\delta_w(p||q) = p(w)\log(p(w)/q(w))$$

where $p(w) = P(w|url)$ is the probability of observing the query term w given the url, and $q(w) = P(w|C)$ is the probability w is observed in the collection of all queries.

We proceeded by deriving the top fifty associated query terms for a URL in three different ways. Each different approach is an increasing generalization of the URL.

M1: $P(w|url)$ estimated on the bag of query terms associated with the URL.

M2: $P(w|url)$ estimated on the bag of query terms associated with the domain of the URL.

M3: $P(w|url) = \sum_{u \in URL} P(w|u)P(u|URL)$: A simple association model between query terms and individual URL tokens in the URL was estimated on the click data by computing likelihoods $P(w|u)$ across all clicks.

For the last method, URL tokens u were extracted by splitting the URLs on non-alphanumeric characters. This resulted in raw term strings (e.g. *bikeriderstours*). We decompounded these raw strings using a script that subsequently looks up substrings in the CELEX lexicon[2]. It stores an URL term u if the lexicon contains the term and it lemmatizes it (e.g. resulting in the URL tokens *bike*, *rider* and *tour*). The URL token representation also served as a baseline for the classification experiment[3].

Next we investigated whether the obtained query terms provide significant discriminative information for the associated URL. To do so, we used the extracted query terms as features for a URL classification experiment. The DMOZ URL class label was used as ground truth.

For our classification experiments, Adaboost.MH [7] was used. The advantages are mainly its good generalization capacity, and its innate feature weighting capacity. Adaboost is particularly capable of dealing with short utterances with high lexical variation. Our data fits this description: lexically varied small-sized bags of terms.

4. RESULTS AND DISCUSSION

Table 1 shows the classification accuracy (proportion of instances that were classified with a correct DMOZ category). The majority class baseline is an artificial setting in which all urls are classified as the majority class ('Business'). All conditions involving (aggregates of) query tags outperform the URL tokens baseline, which is encouraging. Good performance was obtained for aggregation at the domain name level, which can be explained by the fact that more information is available which helps to boost frequent, more reliable terms. The model based on associations between individual URL tokens and query terms performs remarkably well. We did some preliminary tests on DMOZ URLs that did not contain clicks in the MSN dataset, but had a relation on the domain level. The results of these tests are shown in parentheses. The method that aggregated query terms across the domain scored worse than the baseline (0.26), probably indicating that many URLs with a similar domain name do have different categories. Our annotations based on the probabilistic url token to query term association dictionary $p(w|u)$ scored slightly worse than the baseline.

Table 1: Classification accuracy for various experimental settings. Between parentheses are results obtained on DMOZ URLs that did not contain clicks in the MSN dataset. Significant improvement over the URL tokens baseline (2-sided T-test) is denoted with an asterisk (*) ($P < .05$).

Features (method number)	Accuracy	
Majority class baseline	0.15	
URL tokens (baseline)	0.38	(0.34)
URL associated query terms (M1)	0.43*	
domain associated query terms (M2)	0.45*	(0.26)
URL-tokens associated query terms (M3)	0.42*	(0.33)
baseline + M1	0.42*	
baseline + M2	0.46*	(0.35)
baseline + M3	0.42*	(0.35)

5. CONCLUSION AND FUTURE WORK

We constructed several document descriptions (term lists) using click data. The representations differed in the level of aggregation applied to overcome data sparseness. The most general model is a generative model based on a URL as a bag of tokens and an association dictionary trained on the search engine logfile to annotate URLs for which we have no associated queries. These models can serve as (a basis of) useful document descriptors, as shown by a URL categorization task.

6. REFERENCES

[1] I. Antonellis, H. Garcia-Molina, and J. Karim. Tagging with queries: How and why? In *ACM WSDM '09*, 2009.

[2] D. J. Brenes, D. G. Avello, and K. P. Gonzalez. Survey and evaluation of query intent detection methods. In *Proceedings of WSCD '09*, pages 1–7, Barcelona, Spain, 2009. ACM.

[3] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, 2005.

[4] M.-Y. Kan and H. O. N. Thi. Fast webpage classification using url features. In *CIKM '05*, pages 325–326, 2005.

[5] B. Poblete and R. Baeza-Yates. Query-sets: using implicit feedback and query patterns to organize web documents. In *Proceedings of WWW '08*, pages 41–50, New York, NY, USA, 2008. ACM.

[6] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *ACM WSDM '09*, pages 54–63, 2009.

[7] R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39:135–168, 2000.

[8] M. van der Heijden, M. Hinne, W. Kraaij, S. Verberne, and T. van der Weide. Using query logs and click data to create improved document descriptions. In *Proceedings of WSCD '09*, pages 64–67. ACM New York, NY, USA, 2009.

[2] `http://www.ldc.upenn.edu/`

[3] Unfortunately no full text crawl from the same time period is available for the URLs involved. The experimental focus is on dealing with sparseness.