Annotating URLs with query terms: What factors predict reliable annotations?

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

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

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

Eva D'hondt
CLST,
Radboud University Nijmegen
e.dhondt@let.ru.nl

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

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

ABSTRACT

A number of recent studies have investigated the relation between URLs and associated query terms from search engine log files. In [5], the query terms associated with the domain of a URL were used as features for a URL classification task. The idea is that query terms that lead to successful classification of a URL are reliable semantic descriptors of the URL content. We follow up on this work by investigating which properties of a URL and its associated query terms predict the classification success. We construct a number of URL and query properties as predictors and proceed to analyze these in-depth. We conclude that the classification success — and thus the reliability of the query terms as URL descriptors — cannot easily be predicted from properties of the URL and the queries.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

General Terms

Click data, URL classification, Human factors

1. INTRODUCTION

In previous work on the use of query log data [5], the authors investigated the applicability of semantic annotation of web pages by creating short document descriptions (term lists) extracted from associated queries. The assumption here is that, when presented to a user, these term lists may help in the disambiguation of a URL and/or identify whether the URL corresponds to the user's query intent [3]. The term lists in [5] were extracted from the (weighted) set of query terms that are associated with an URL. In order to find out whether the associated term lists provided good descriptions of the URL (and consequently, good clues in disambiguation), a classification experiment was conducted on a set of URLs, using the term lists as features. Depending on the level of query term aggregation, a classification accuracy of up to 45% was obtained [5].

These classification results are reasonably satisfying. We aim at a future implementation of semantic annotation of URLs in the user interface of a web search engine. In this implementation, the search engine not only has the query term descriptors for a URL available, but also the indexed content of the web page. Previous work on semantic annotations from URL content and query logs [11] showed that the query terms associated to a URL provide useful additional information to terms extracted from the content of the web page. Thus we can expect the classification accuracy to go up if we not only use the query terms but also the content terms.

However, a classification accuracy of 45% means that URL classification based on query terms only is unsuccessful for more than half of the URLs. In order to prevent the semantic annotation of URLs to be negatively influenced by unreliable query term descriptors, we aim to predict the reliability of the query term annotations given a URL and its associated queries. Therefore, in the current paper, we study which properties of the URL and the associated term list can predict the reliability of the query term annotations. Following [5], we consider the classification accuracy as indicator for the informativeness and reliability of the query terms annotation: the better a set of query terms describes a URL, the higher the chance that the URL is classified correctly. Thus, we investigate the relation between URL and query properties on the one hand and the classification quality on the other hand.

2. RELATED WORK

Other studies have shown that query–URL associations and click information can serve as a means for implicit feedback [4, 6, 7] and for learning to rank e.g.[1]. Several others have investigated whether a collection of queries and click information can be used as a model of the semantic contents of a web page [2, 8]. A document representation based on queries has been compared with a traditional document content vector space representation for a clustering task [9]. The query based representation resulted in a better clustering than the document content based representation. A similar experiment has been conducted where human assessors were asked whether they preferred a document description

SIGIR '09, July 19-23, 2009 Boston, United States.
ACM X-XXXXX-XX-X/XX/XX.

based on queries or on a vector space representation [11]. The assessors tended to prefer the query-based representation.

Most related work uses site access logs of a portal page, which means that the log files show a complete picture of all queries leading to that site: query terms can be derived from the referring URL in the HTTP-header. In our case, we use a much larger collection of web pages and click data that is based on the query log of one single search engine (See Section 3.1). Our data do not comprise the page content of the URLs because the page content was not available in the query log data and recrawling would give many inconsistencies due to web pages changing significantly in the course of a few years. Our work differs from [5] since we explicitly aim to explain *in which cases* (for which types of URLs) query log data can be used to inform the user about the semantic contents of a web document, or for disambiguation of a URL or identification of query intent.

3. EXPERIMENTS

The objective of our experiments is to identify a set of key factors that predict whether a URL can be classified correctly using query log data. With the aid of such factors, a search engine can use terms from query log data as document descriptors, which help the user in disambiguating URLs or finding URLs that match the user's query intent.

3.1 Data

RFP: The Microsoft 2006 RFP[1] 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 URL of the clicked document, the rank of that URL in the result list and the number of results.

DMOZ: The DMOZ Open Directory RDF Dump[2] is a set of URLs and their class labels according to the Open Directory Project DMOZ. 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 labelled `Top/Regional`, since `Regional` is the top node of a different hierarchy (a regional classification). The intersection of the RFP and DMOZ collections (with the above restriction) consists of 245.742 URLs, distributed over 15 classes.

3.2 Properties of URL and query

In [5] the classification features for URLs in the RFP-DMOZ intersection were extracted by finding the query terms that were most strongly associated with the URL[3]. These features were aggregated at the level of the URL, the domain of the URL and the individual words in the URL. Using these features, the URLs were classified with Adaboost.MH [10]. The highest classification accuracy was achieved when the query terms were aggregated at the level of the domain of the URL. In the current paper, we therefore focus on query terms associated with URLs on the domain level, aggregating queries over all URLs from our data collection that belong to the same domain[4].

In order to investigate what properties of the URL and the associated query terms play a role in the correct classification of some URLs and the incorrect classification of others, we extracted the following properties for each URL in the RFP-DMOZ intersection:

D: The domain of the URL.

DL: The number of terms the domain was compounded of (the domain length)[5]

NC: The number of clicks in the RFP dataset that were associated with the domain.

NUQ: The number of unique query terms associated with the domain.

AQC: The average number of query terms per click on a URL from the domain.

MCP: The position that the clicked URL had in the result list of the search engine, averaged over all clicks that led to the domain.

PN: The proportion of navigational queries in the total number of queries that led to the domain of this URL. We consider a query to be navigational if the concatenated query terms are a substring of the URL string (e.g. "bike riders tours" is a navigational query for the URL `www.bikeriderstours.com`).

TWE: The token-wise entropy of the domain, as the sum of all the terms the domain is compounded of, i.e.: $H(D) = -\sum_{t \in D} P(t) \cdot \log P(t)$, with $P(t)$ the probability of observing term t in any domain in the RFP-DMOZ intersection.

KLD: The Kullback-Leibler divergence of the associated query term probability distribution, relative to the distribution of all query terms in the RFP dataset, i.e.: $D_{KL}(P\|Q) = \sum_{t \in D} P(t) \cdot \log \frac{P(t)}{Q(t)}$ with $P(t)$ the probability of observing t in all queries associated with D and $Q(t)$ the probability of observing t in all queries in the RFP collection.

For all but the first of these properties, we investigated their relation to classification success. Our hypothesis was that especially NC and NUQ would have a positive predictive value for the classification success. We expect that more clicks (higher NC) and more unique query terms (higher NUQ) for a domain result in a better representation of the domain and therefore in a better classification accuracy. The details of our analyses are in Section 4 below.

[1] `http://research.microsoft.com/en-us/um/people/nickcr/wscd09/`

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

[3] Strength of association was calculated using Kullback-Leibler divergence with the total query collection as background model.

[4] As a consequence, we can only investigate URL properties that generalize to the domain level. Moreover, aggregating on the domain level has the risk of grouping together heterogenous URLs from large domains. We come back to this in Sections 4 and 6.

[5] We decompounded the domains using a script that subsequently looks up substrings in the CELEX lexicon (`http://www.ldc.upenn.edu/`) and greedily splits the domain string into lemmatized lexicon entries. E.g. the domain `bikeriderstours.com` was decompounded into the lemmas *bike*, *rider* and *tour*).

4. RESULTS

We considered three different strategies for finding the relevance of each of the predictors for the success of the classification: calculating the correlation coefficient ρ in order to get an indication of the strength and the direction of the relation between each predictor's value and the classification outcome. However, this coefficient assumes a linear relation that is independent of other predictors. Our data seemed more complicated than that. Therefore, we assessed the possibility of using a logistic regression model (LRM) for predicting the classification outcome based on the predictor values (normalized to their z-score). Unfortunately, the LRM outcome was difficult to interpret: We did get positive and negative predictor coefficients that significantly contributed to the prediction model but the model fit on the data was relatively poor.

These preliminary results suggested that there is no linear relationship between any of the predictors that we investigated and the classification success. We felt however that some tendency could be discerned from the individual predictors' values and the classification accuracies for specific ranges of these values. In order to assess this hypothesis, we created 10 bins for the values range of each predictor. Subsequently, we derived the classification accuracy for each bin, together with the number of domains in this range. We plotted these numbers in bar charts in order to visualize the relation between the value ranges of the predictors and the classification accuracy.

Unfortunately, we did not find very strong tendencies for most of the predictors that would support the idea that the classification success can be predicted from these properties. Most of the bar charts appeared to be relatively flat, confirming that the classification accuracy is relatively stable, only slightly dependent of the value of the predictor. As an example, Figure 1 shows the classification accuracy as a function of the token-wise entropy of the domain. The only bar that is rising above the others is the right-most one, representing the domains with the maximum entropy value. However, this bar only represents a small number of domains (3,423) and the classification accuracy for this range is still mediocre (60%).

In the next sub-section, we discuss the results for the two predictors that we had expected to give the most promising results (see Section 3.2).

4.1 Analysis of the NC and NUQ predictors

NC: When we look at the number of domains in relation to (a range of) the number of clicks on those domains (Figure 2), we first notice that most domains in our data collection have a small number of associated clicks (1 to 4). At the same time we see that domains with the lowest numbers of clicks are the domains with the lowest classification accuracy. This confirms our earlier assumption that many clicks result in a better representation of the domain and therefore a better classification accuracy. The maximum classification accuracy is 58% (for the range of 33–64 clicks). However, Figure 2 also shows that for a higher number of clicks, the classification accuracy starts to decrease again.

We suspect that this behavior can be explained from the heterogeneity of the domains that have a large number of associated clicks. For example, portal web sites such as `ebay.com` or `amazon.com` contain many URLs that may be very diverse in their semantic content. Consequently, these

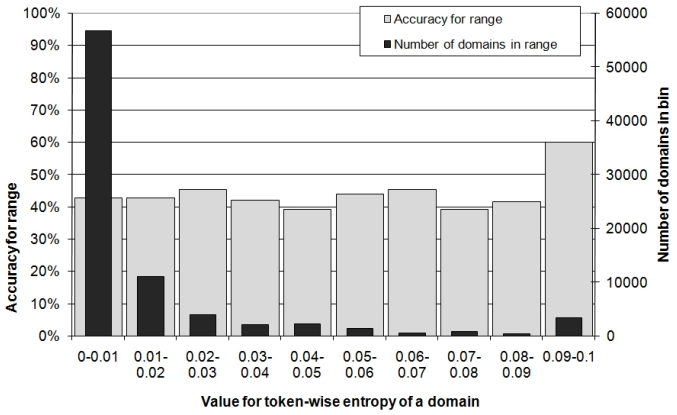

Figure 1: Classification accuracy as a function of the token-wise entropy of the domain. The token-wise entropy values have been grouped in ranges of i to $i + 0.01$ for $i \in \{0, \ldots, 0.09\}$

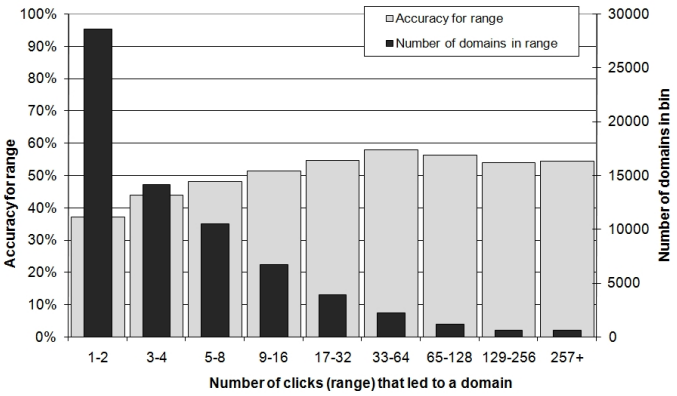

Figure 2: Classification accuracy as a function of the number of clicks that led to a domain. The numbers of clicks have been grouped in ranges of 2^i to 2^{i+1} terms for $i \in \{0, \ldots, 9\}$

URLs are harder to classify, since in the aggregated term set for the corresponding domain there are many terms for semantically unrelated URLs from the same domain.

NUQ: When we look at Figure 3, we see that the number of domains in a given range of unique query terms decreases much less sharply than for the number of clicks. However, we see a similar pattern in the classification accuracies for these ranges. For domains with 17–32 unique associated terms, the accuracy is optimal. Figure 3 shows that classification accuracy sharply increases initially for an increasing number of unique query terms, starting at 30% for domains with only 1–2 unique terms, up to 60% for domains with 17–32 unique terms. After that point, the accuracy decreases again.

Domains with very few unique terms apparently provide a too sparse classification vector to be classified correctly. At the other end of the spectrum, domains with too many unique terms are hard to classify as well. We again attribute this to the heterogeneity of the domains with a large number of unique query terms: it is very difficult to classify them as belonging to a single class.

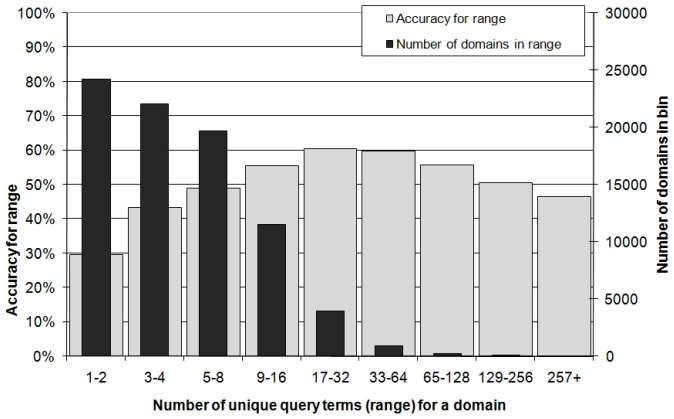

Figure 3: Classification accuracy as a function of the number of unique query terms per domain. The numbers of unique query terms have been grouped in ranges of 2^i to 2^{i+1} terms for $i \in \{0, \ldots, 9\}$

5. DISCUSSION

After analyzing our predictors in detail, we found that many of them cannot predict the classification success. The two most promising predictors (number of clicks and number of unique query terms) showed interesting tendencies but do not provide ranges of high accuracies (optimal ranges give 60% classification accuracy). It is clear that the success of classifying URLs based on query terms depends on many different factors. In the previous section, we mentioned the heterogeneity of the domain as a potentially important factor.

If we want to adapt our strategy for the heterogeneity of domains (for example, by not providing query-based descriptions for very heterogenous domains), the question that rises here is how we can identify domains as being heterogenous. Two of the factors that we saw in Section 4 are the number of clicks and the number of unique query terms that are associated with a domain. A third factor may be the domain size: the more URLs a domain contains, the larger the heterogeneity of the domain probably is. Part of our future work will be to estimate the domain heterogeneity based on these factors.

6. CONCLUSION AND FURTHER WORK

We continued the work of [5] and investigated which factors are relevant for the success of URL classification based on associated query terms. We created a series of classification success predictors and subsequently analyzed their relation to the classification success. None of the predictors we investigated can fully predict the classification success. We found however that a couple of predictors show interesting tendencies: the number of clicks on URLs (**NC**) and the number of unique terms associated with a URL (**NUQ**). In both cases, the predictors initially correlate positively with the classification accuracy, but after a certain saturation point this correlation becomes negative. We suggest that this is caused by heterogeneous domains (domains that contain URLs from different semantic categories). We argue that our suggested approach of providing query terms as document descriptors for disambiguation is particularly useful for URLs from homogenous domains.

An important point for further research is to determine the heterogeneity of a domain using query log data. Another direction is to investigate what factors predict classification accuracy when query terms are not aggregated on domain level, but on the level of individual URLs. As these cannot be heterogenous, it will be worthwhile to see the performance of the predictors in this situation.

We are currently experimenting with different types of classifiers in order to see whether we can improve the classification accuracy of our data. We also study our data in more detail in order to see whether removing a subset of the click data from the training set can increase the classification performance. This subset can be either category-based (remove noisy categories), feature-based (remove instances with too few query terms) or based on overall consistency (remove instances that have very similar term sets but contradictory classes).

In the somewhat more distant future, we aim to investigate the possibilities of implementing our URL descriptor approach in a user interface. Following the results obtained by [11], we will combine salient terms from the URL's content and the queries associated with the URL into a semantic annotation of the URLs in the result list. One challenge that we foresee for this experiment is the evaluation: User judgments are time-consuming but essential for this kind of implementation.

7. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA, 2006. ACM.

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

[3] 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. ACM New York, NY, USA, 2009.

[4] 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.

[5] M. Hinne, W. Kraaij, S. Raaijmakers, S. Verberne, T. van der Weide, and M. van der Heijden. Annotation of URLs: more than the sum of parts. In *SIGIR '09: Proceedings of the 32th ACM SIGIR international conference on Information Retrieval*, New York, NY, USA, 2009. ACM.

[6] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference*, pages 154–161. ACM New York, NY, USA, 2005.

[7] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[8] B. Krause, R. Jäschke, A. Hotho, and G. Stumme. Logsonomy - social information retrieval with logdata. In *Hypertext*, pages 157–166, 2008.

[9] 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. ACM New York, NY, USA, 2008.

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

[11] 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.