Web-Centric Language Models

Jaap Kamps[1,2]
[1] Archives and Information Studies, University of Amsterdam, Amsterdam, The Netherlands
[2] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
kamps@science.uva.nl

ABSTRACT

We investigates language models for informational and navigational web search. Retrieval on the web is a task that differs substantially from ordinary ad hoc retrieval. We perform an analysis of prior probability of relevance for a wide range of non-content features, shedding further light on the importance of non-content features for web retrieval. Language models can naturally incorporate multiple document representations, as well as non-content information. For the former, we employ mixture language models based on document full-text, incoming anchor-text, and document titles. For the latter, we study a range of priors based on document length, URL structure, and link topology. We look at three types of topics—distillation, home page, and named page—as well as for a mixed query set. We find that the mixture models lead to considerable improvement of retrieval effectiveness for all topic types. The web-centric priors generally lead to further improvement of retrieval effectiveness.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval

General Terms: Measurement, Performance, Experimentation.

Keywords: Web retrieval, non-content information, language models, priors.

1. INTRODUCTION

For the public at large, the field of Information Retrieval (IR) is synonymous with Internet search engines and web search. Yet at the same time, IR researchers broadly agree that web search is different from traditional ad hoc search [2]. This immediately prompts the question: what are then the appropriate retrieval models for web search? In this paper, we aim to shed further light on this question, and our approach strives to create maximum transparency. That is, rather than tuning or training the many parameters in a retrieval model, we try to isolate specific, web-centric features of retrieval models and to understand their role. For this type of analysis, the language modeling framework is a natural candidate to work in.

In line with related work for navigational web search [4, 5], we employ mixture language models with a range of non-content priors. For the web tasks we use a specific mixture language model. Given a query q_i and a document d, we employ three document models: $P_{\text{text}}(q_i|d)$ (based on the full-text index), $P_{\text{anchor}}(q_i|d)$ (anchor-texts), and $P_{\text{title}}(q_i|d)$ (titles). The three models are combined as follows:

$$P(q|d) = P(d) \cdot \prod_{i=1}^{n} ((1 - \lambda_1 - \lambda_2 - \lambda_3) \cdot P(q_i|C) + \lambda_1 \cdot P_{\text{text}}(q_i|d) + \lambda_2 \cdot P_{\text{anchor}}(q_i|d) + \lambda_3 \cdot P_{\text{title}}(q_i|d)),$$

where each of the document models, $P_x(q_i|d)$, is estimated using a maximum likelihood estimate. All runs on which we report below use equal weights for all three document models. There, we use the full text index as the collection model, and investigate how the prior probability of a document, $P(d)$, can be used to incorporate non-content features into the scoring mechanism.

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

2. DOCUMENT REPRESENTATIONS

First, we look at the effectiveness of mixture language models with a uniform prior. All experiments are based on the test suite of the TREC 2004 Web Track [1], having a stream of 225 topics consisting of three types: topic distillation, home-page finding, and named-page finding. Our first set of experiments investigates various document representations. The best scores with the associated values of λ are reported in Table 1, where we calculate the MAP for distillation topics, and the MRR for known-item topics.

Some observations present themselves. First, the scores for distillation topics are much lower than for the known-item topics, with a particularly impressive score for the named page topics. Second, compact indexes of titles and anchor are very effective for known-item topics, and outperform the massive full text index. Third, for topic distillation the full text index is substantially more effective. We now turn to mixture language models covering all three indexes. Again, we make some observations. First, we see that the mixture model leads to improvement for all topic types. Second, the differences over the smoothing parameter are small, although again the known-item topics prefer little smoothing whereas the distillation topics prefer more smoothing.

3. WEB-CENTRIC PRIORS

We use the mixture model run with $\lambda = 0.3$ in our further experiments and investigate how the prior probability of a document can be used to incorporate non-content features

Table 1: Results for web-centric document representations.

Index	TD	HP	NP	Mixed
Text	0.0797/.25	0.1620/.15	0.3633/.95	0.1985/.55
Titles	0.0620/.20	0.3475/.10	0.4390/.15	0.2828/.10
Anchors	0.0519/.30	0.2767/.75	0.4106/.10	0.2416/.60
Mixture	**0.0971**/.15	**0.4394**/.30	**0.6788**/.20	**0.4047**/.30

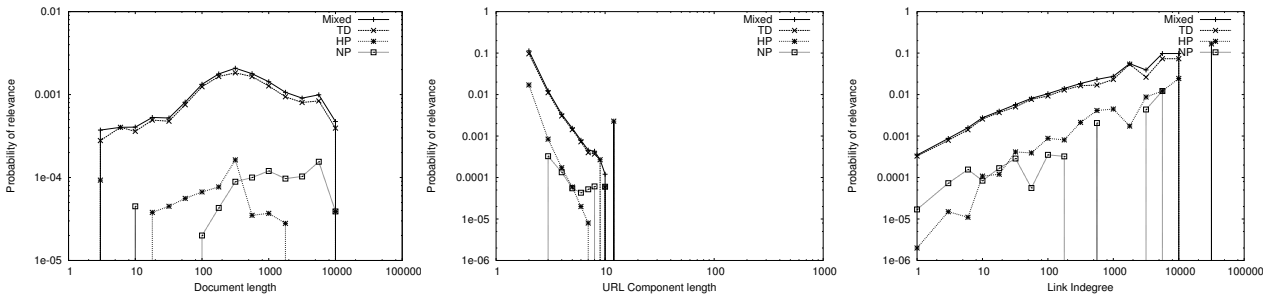

Figure 1: Prior probability of relevance against document length (left), URL length (middle), and number of incoming links (right).

into the scoring mechanism. We analyze a range of non-content features, such as document length, the page's URL, and link topology, and investigate their usefulness to boost retrieval effectiveness.

Document length Let us focus on document length first. Figure 1 (left) shows the prior probability of relevance against the length of a document. Unlike in standard adhoc retrieval, for the web-centric tasks there appears to be no marked effect of length on relevance.

URL We will now focus on the uniform resource locator (URL) as a non-content feature, independent of the particular query at hand. We investigated three measures of the length of the URL: (1) the number of occurrences of '/' in the URL, (2) its number of characters, or (3) the number of 'components'. [3]. We determine the number of components as follows: split the URL in the *domain name* and *file path*, and count the number of '.' separated components in the domain name, and the number of '/' separated components in the file path. E.g., `trec.nist.gov/act_part/act_part.html` has length 5. Figure 1 (middle) shows the prior probability of relevance for the URL component length. The length of a URL has a clear reciprocal relation with relevancy: the shorter the URL, the more likely the page is to be relevant. We looked at a number of operationalization of URL component length, and decided to use a *URL prior* that is proportional to $(\frac{1}{component\_length})^2$.

Link Topology Next, we focus on the link topology. We look at the number of pages linking to a document (indegree), or the number of pages to which a document links (outdegree). Figure 1 (left) shows the prior probability of relevance over indegree. The degree of a page has a clear relation with relevancy: the more links a pages receives the more likely it is that the page is relevant. We looked at a number of operationalizations of the indegree, and decided to use an *indegree prior* that is proportional to the indegree.

We conduct experiments investigating the effectiveness of various language model priors, based on the URL prior, the indegree prior, and use their product as a combined URL/indegree prior. The results are reported in Table 2. Before we discuss our results for the mixed query task, we present the results for a breakdown of the set of topics into the three subtasks, i.e., topic distillation, home page finding, and named page finding.

Table 2: Results for web-centric priors.

Prior	TD	HP	NP	Mixed
Uniform	0.0965	0.4394	**0.6783**	0.4047
URL	0.1121	0.5769	0.6752	0.4547
Indegree	0.1313	**0.6724**	0.6625	**0.4887**
URL/indegree	**0.1440**	0.6617	0.6145	0.4734

Topic distillation We see that all priors (URL, indegree, and combined prior) pay off, leading to impressive improvements over the uniform prior scores. In particular, the indegree prior makes a substantial difference.

Home page finding We see, again, that the priors pay off. There is a substantial improvement for both the URL and indegree prior. The best MRR score is for the indegree prior.

Named page finding Here, the results for the priors are negative, the result for the URL prior is neutral and the indegree prior leads to a small loss. Of course, the uniform prior run sets a very high baseline.

Mixed query task Overall, the priors help to improve retrieval effectiveness. The indegree only prior is the most effective and gets the highest scores.

4. CONCLUSIONS

Our findings highlight that web retrieval is unlike standard ad hoc retrieval. Whereas document-length is a useful indicator for relevance in the general ad hoc case, it is not for the case of web retrieval. Specific web-centric techniques, such as using the URL structure or using the link topology, turn out to be useful indicators of relevance for the mixed query task. Our findings extend earlier results on the effectiveness of language model priors for web retrieval. Kraaij et al. [4] established the effectiveness of web-centric priors for the home-page finding task. Ogilvie and Callan [5] extended these results to the other navigational task of named-page finding. Our findings extend these results to the informational task of topic distillation.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 612.066.302 and 640.001.501.

References

[1] N. Craswell and D. Hawking. Overview of the TREC 2004 web track. In *TREC 2004*. NIST, 2005.

[2] D. Hawking and N. Craswell. Very large scale retrieval and web search. In *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[3] J. Kamps, G. Mishne, and M. de Rijke. Language models for searching in Web corpora. In *TREC 2004*. NIST, 2005.

[4] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR 2002*, pages 27–34. ACM Press, 2002.

[5] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *SIGIR 2003*, pages 143–150. ACM Press, 2003.