Identifying Primary Content from Web Pages and its Application to Web Search Ranking

Srinivas Vadrevu
Yahoo! Labs
Sunnyvale, CA, USA
svadrevu@yahoo-inc.com

Emre Velipasaoglu
Yahoo! Labs
Sunnyvale, CA, USA
emrev@yahoo-inc.com

ABSTRACT

Web pages are usually highly structured documents. In some documents, content with different functionality is laid out in blocks, some merely supporting the main discourse. In other documents, there may be several blocks of unrelated main content. Indexing a web page as if it were a linear document can cause problems because of the diverse nature of its content. If the retrieval function treats all blocks of the web page equally without attention to structure, it may lead to irrelevant query matches. In this paper, we describe how content quality of different blocks of a web page can be utilized to improve a retrieval function. Our method is based on segmenting a web page into semantically coherent blocks and learning a predictor of segment content quality. We also describe how to use segment content quality estimates as weights in the BM25F formulation. Experimental results show our method improves relevance of retrieved results by as much as 4.5% compared to BM25F that treats the body of a web page as a single section, and by a larger margin of over 9% for difficult queries.

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

General Terms
Algorithms, Experimentation

Keywords
page structure, segmentation, content quality models, search

1. INTRODUCTION

A retrieval function, that treats different segments of a web page equally, may surface irrelevant query matches. Traditional Information Retrieval methods attempt to deal with structure indirectly, and with limited success. For example, long documents may have disadvantage due to document length normalization of the term frequencies. [3] addresses this issue by finding a better document length normalization factor, however, it does not exclude noise segments before attempting to do so. Long documents can also have multiple topics rather than just being verbose. In these cases, fixed size windows of continuous terms is used to inform the ranking function about the proximity of query terms. While, this is an attractive method, it does not guarantee that the windows will not span boundaries of content segments that should be treated separately.

In this paper, we describe methods of improving the retrieval function performance by paying attention to the structure of the web pages. This is achieved by segmenting the web pages into semantically coherent blocks and identifying content segments from noise segments in the page by learning a statistical predictor of segment content quality and apply to all web pages. We propose ranking features based on these PAge Segment Content quALity (PASCAL) models that can be directly incorporated in the ranking function.

2. WEB PAGE SEGMENTATION AND CLASSIFICATION

Our segmentation algorithm uses the DOM tree structure of HTML document to segment the page into visually coherent blocks. The algorithm starts by treating each DOM node as a segment and proceeds by merging the segments using heuristic rules similar to [1, 5].

The central argument of our paper is that including the segment content quality scores directly in the ranking function improves relevance of retrieved results. Naturally, the content quality models are expected to be reasonably accurate. For this, we rely on a supervised learning setting leveraging a rich feature space similar to [4]. We define two editorial classes for Web page segments: content segment, which usually corresponds to the primary content within a web page and noise segment that contains no immediate relevant content that meets or supports information need.

Our feature space for machine learned segment classification contained both visual and content properties of the segments. We used gradient boosted trees with logistic loss to learn a model of content vs. noise binary classification. A target value of 0 is used for noise segments and 1 is used for primary content segments. We measured the classifier accuracy as 0.84 precision and 0.88 recall for primary content segments, in cross-validated experiments.

3. PASCAL BASED RANKING FEATURES

We begin with BM25F [2], which is a useful feature in the literature of retrieval models, as our baseline text matching feature and devise a segment weighted version of it to test our PASCAL's application to ranking models. In the BM25F formulation, the normalized term frequency in each of the fields is given by the following formula:

$$\overline{g}(t, D, f) = \frac{g(t, D, f)}{(1 + B_f(\frac{|f|}{avgfl} - 1))} \quad (1)$$

where t is term in the field f in the document D, $|f|$ is the

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

length of the field in words, $avgfl$ is the average field length in the text collection from which documents are drawn. B_f is a field-independent free parameter that can be tuned, similar to B parameter in BM25 formulation [2]. $g(t, D, f)$ is the raw term frequency of term t in the field f of document D and $\overline{g}(t, D, f)$ is the normalized term frequency, normalized with average field length.

We propose an alternate version of BM25F formulation that treats each segment inside a web page as a separate document field in the original formulation and weighs each segment in the web page according to the PASCAL score of the segment.

$$\overline{g}(t, D, f_b) = \sum_{i=1}^{k} w(S_i).\overline{g}(t, D, f_b, S_i) \qquad (2)$$

$$\overline{g}(t, D, f_b, S_i) = \frac{g(t, D, f_b, S_i)}{(1 + B_{f_b}(\frac{|S_i|}{avgsl} - 1))} \qquad (3)$$

where k is the number of segments in document D and $w(S_i)$ is the PASCAL score of the segment S_i in document D. The higher $w(S_i)$ is, the more likely it is to be a content segment as opposed to be noise. $|S_i|$ is the length of the segment S_i in words, $avgsl$ is the average length of the segment in the collection. $g(t, D, f_b, S_i)$ is the raw term frequency of t in the segment S_i in the body field f_b in the document D and $\overline{g}(t, D, f_b, S_i)$ is the average segment length normalized term frequency.

In addition, we propose a set of nonlinear PASCAL text matching features based on the page segment content quality predictions. These nonlinear text matching features capture exclusive noise match of the query where all of the query terms appear in the noise segments, and make sure none of the query terms appear in the noise segments of the Web page.

Another set of nonlinear features can be formulated on the segment based structure of the document that address the proximity issues with the queries. To describe these nonlinear features, we introduce *segment cover* π that can be defined as a set of segments, such that it is a tree with the segment edges and that it should span across all query terms, at least one. Intuitively this could be thought of as a spanning tree across query terms. Note that there could be multiple sets of segments that satisfy the definition of segment cover that have the same minimal size. From these sets of segment covers, the following nonlinear features can be derived: (1) the size of the minimal segment cover. This feature *query match segment cover* corresponds to the minimum number of segments required to cover all query terms. (2) $argmax(\sum_{s \in \pi} l(s))$, for each minimal segment cover π, i.e., maximum of the sum of PSCQ weights of the segments in each of the segment covers. (3) for each minimal segment cover π_i, let μ_i be the minimum node weight, i.e., $\mu_i = min(l(s))$, for each segment s in π. The nonlinear feature can now be defined as min (μ_i), i.e., minimum of the minimum segment weight in each of the segment covers.

4. EXPERIMENTAL RESULTS

We used two evaluation data sets. The first *random* data set contains 10000 queries randomly sampled from the logs of a search engine, and is split into two equal size parts; one for training and one for testing. The second *difficult* data contains 4000 queries that include a product, person or location name as part of the query. We use original BM25F as the baseline for evaluating ranking with PASCAL features. We also test ranking models developed in other works in literature utilizing segmentation information. We adopt Discounted Cumulative Gain (DCG) as the evaluation metric.

Features	DCG-5 Gain Random Qrys	DCG-5 Gain Difficult Qrys
BM25F-PrimContent	3.14%	3.83%
BM25F-Interpolated	1.89%	4.82%
BM25F-SegWeight	3.58%	3.17%
PASCAL-All	4.65%	9.08%

Table 1: DCG gains of various ranking models in comparison to baseline *BM25F-Original*, over the random and difficult test sets. All DCG gains are statistically significant.

Table 1 shows the performance of the ranking models in comparison with the baseline *BM25F-Original*, over the random and difficult test sets. The best performing model is *PASCAL-All*, which includes all the PASCAL features, by a significant margin. It is followed by *BM25F-SegWeight*, the ranking based on the PASCAL segment weighted version of the BM25F. The BM25F formulation with just primary content segments, *BM25F-Primcontent* yields DCG-5 gain close to *BM25F-SegWeight*, mainly because the former does contain some nonlinearity in the scoring due to the thresholding of segment weights. The interpolated BM25F, which is a linear combination of original BM25F and maximum Bm25 score for a segment, yields the a moderate gain over the baseline.

We conducted two additional experiments to test the effect of the length of the query and the rank position. In query length experiments, it was evident that the larger the length of the query, the DCG improvements with our system are. In the rank position experiments, it was clear that the DCG gain is larger in higher rank positions, i.e., lower in the search results ranking. This indicates that the PASCAL system works well on the difficult and longer queries and the PASCAL-All ranker is able to achieve higher gains in those rank positions, where the relevance is usually low.

5. CONCLUSIONS

In this paper, we described how to utilize content quality models for web page segments in a retrieval function. Our method is based on segmenting a web page into semantically coherent parts and learning a predictor of segment content quality. The experimental results showed that our method improves relevance of retrieved results, especially for difficult queries. Overall our method is able to achieve 9.08% statistically significant DCG-5 gain on difficult queries.

6. REFERENCES

[1] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Extracting content structure for web pages based on visual representation. In *Asia Pacific Web Conference*, 2003.

[2] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *CIKM*, pages 42–49, 2004.

[3] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *SIGIR*, 1996.

[4] R. Song, H. Liu, J. Wen, and W. Ma. Learning block importance models for web pages. In *WWW*, 2004.

[5] S. Vadrevu, F. Gelgi, and H. Davulcu. Semantic partitioning of web pages. In *WISE*, 2005.