A Densitometric Analysis of Web Template Content

Christian Kohlschütter
L3S / Leibniz Universität Hannover
Appelstr. 9a, 30167 Hannover
Germany
kohlschuetter@L3S.de

ABSTRACT

What makes template content in the Web so special that we need to remove it? In this paper I present a large-scale aggregate analysis of textual Web content, corroborating statistical laws from the field of Quantitative Linguistics.
I analyze the idiosyncrasy of template content compared to regular "full text" content and derive a simple yet suitable quantitative model.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval; G.3 [**Probability and Statistics**]: Distribution Functions

General Terms

Theory, Experimentation, Measurement

Keywords

Content Analysis, Template Detection, Template Removal, Web Page Segmentation, Noise Removal

1. INTRODUCTION

In contrast to plain text, documents in the Web expose some specific structural properties. A Web page's text is not limited to the actual "main content" but usually consists of several additional segments. These provide further information (site maps, headline lists, tables, "related article" links, text-ads etc.) and are basically meant to augment the full-text. This additional text provided seems only be partially useful or probably even counterproductive for search and classification; the common solution to the problem is simply erasing template content or at least ignoring it. However, current approaches identify templates only heuristically or by machine learning.

In this paper, the textual Web content of several large corpora were subjected to a quantitative analysis. By deriving a densitometric text model based upon techniques from the field of Quantitative Linguistics, it can be shown that the text corpus exposes two fuzzy classes of text, covering full-text and navigational information respectively. The proportions of the two classes (in tokens) roughly is $1:2$, whereas "template text" can be divided in equal shares into these classes (noisy, short navigational text hints vs. frequently used full-text).

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. THE BETA DISTRIBUTION MODEL

To understand the distinction between templates and main content, a large-scale statistical classification was performed on the level of intra-document segments, under the assumption that the segments are sufficiently homogeneous (i.e., either template or main content). The analysis was conducted on the representative Webspam UK-2007 dataset (on the *ham* part, 356,437 out of 106 million crawled pages). As a manual segmentation appears infeasible at corpus-scale, the state-of-the-art BlockFusion segmentation algorithm was employed, which utilizes the text density measure $\varrho(b)$, relating the number of tokens in a particular text block to the occupied text "area" determined by word-wrapping the character data at a fixed line width $w_{\max}$. It was shown that the resulting block structure closely resembles a manual segmentation [3].

Text density is a particularly useful measure when analyzing the Web's quantitative structure. It does not depend on the notion of "sentence", which we could hardly define for the Web's content – many portions of text simply do not contain sentences, nor anything meaningful that could be separable by full stop (this is especially true for template text). To reduce the impact of errors caused by a too fine-grained segmentation, the amount of text (= number of tokens) contained in segments of a particular text density ϱ is examined at corpus-level. We can model this histographically by rounding: $\varrho'(b) = [\varrho(b)]$.

Figure 1 depicts the retrieved token-level count/density distribution for the whole corpus. Apparently, two modal scores are visible, at $\varrho' = 2$ and $\varrho' = 12$ respectively. This indicates at least two classes of text within the corpus. The superimposition of different classes ("strata") of text is known in linguistics; from a theoretical perspective it may even be the normal case [1]. To confirm the presence of multiple classes we need to find a corresponding distribution function whose compound functions are already known in the theory, e.g. the Beta distribution, which is the conjugate prior of the binomial distribution.

In fact, an almost perfect fit was achieved by combining two beta distributions with a normal distribution ($R^2 = 0.998$, RMSE $= 0.0021$ for $a_1 = 68.03$, $b_1 = 132.5$; $a_2 = 4,034$, $b_2 = 54,49$; $c = 0.015$, $d = 0.64$; $e = 78.87$, $f = 7.834$, $\mu = 28.65$, $\sigma^2 = 6.489$; x scores (densities) have been normalized by $x_{\text{norm}} = 36$ to $[0:1]$ before fitting):

$$f(x) = c \cdot (d \cdot f_{beta}(x, a_1, b_1) + (1-d) \cdot f_{beta}(x, a_2, b_2)) + (1-c) \cdot \varphi_{\mu,\sigma^2}(e \cdot x + f) \quad (1)$$

As the combination of *three* beta distributions, while adding another parameter, resulted in a far less accurate fit ($R^2 = 0.944, RMSE = 0.0031$), I conclude that the distribution of text densities can be divided into two *fuzzy* classes C_1 and C_2; the transition from C_1 to C_2 follows the normal distribution, which means that for blocks with particular densities it is rather undetermined to which class the contained text belongs. The distribution parameter d reveals that C_1 roughly covers one third of the tokens enclosed in the corpus and C_2 covers two thirds; from Figure 1 we see that for $5 \leq \varrho' \leq 10$ the normal distribution dominates.

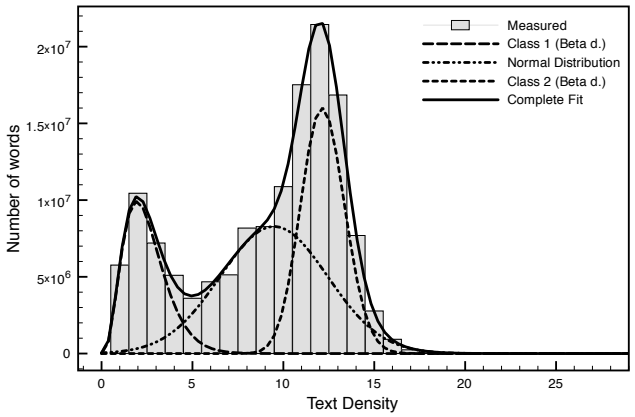

Figure 1: Density Distribution Model

3. TERM TYPICALITY

To make a statement on the meaning of the determined two classes, the content of these classes, i.e., the *term vocabulary*, needs to be analyzed. If the two classes are different, then the contained token vocabulary should also expose noticeable differences. As we want to understand the peculiarities of the two classes C_1 and C_2, which are roughly represented by the two partitions π_1 ($\varrho' \leq 8$) and π_2 ($\varrho' \geq 9$), the partition-specific term document frequencies are compared. We may expect that terms that are *typical* for C_1 appear much more often in π_1 than in π_2, and vice versa. I examine this relationship by computing the corresponding document frequency ratios. The normalized ratio follows a power law distribution of the form $y = c\,(x/(1-x))^{-a_1}$ with $a_1 = 0.39$ and $c = 0.01$ ($R^2 = 0.9468, \mathrm{RMSE} = 0.0034$). This type is a generalization of Zipf's law [4]. In our case, we can interpret the ratio $x/(1-x)$ as the combination of two Zipfian subsets, a top-ranked and a bottom-ranked one, which mutually influence the curve. In fact the frequencies of the considered terms apparently are Zipfian, too, and for both partitions enough typical terms exist. To avoid over-interpreting the impact of rarely occurring terms, the analysis is limited to terms with a collection-wide document frequency $w_{1\cup 2}$ of at least 100. For these terms, I compute the *term typicality* $\varepsilon(t)$, which I define as the logarithmic ratio of the corresponding document frequencies w_1, w_2 of the examined term t in the two partitions. The ratio is normalized by the logarithm to base $N+1$ with N being the number of documents in the corpus (i.e., the maximum document frequency):

$$\varepsilon(t) = \log_{N+1} \frac{w_2(t)+1}{w_1(t)+1} \qquad (2)$$

Rank	Term	ε	Term	ε
1	sitemap	-0.33	spelled	0.51
2	bookmark	-0.29	thousands	0.36
3	accessibility	-0.29	temporarily	0.35
4	misc	-0.29	gave	0.34
5	skip	-0.28	tried	0.33
6	shipping	-0.28	aimed	0.33
7	polls	-0.28	seem	0.32
8	affiliates	-0.27	eventually	0.31
9	username	-0.27	unfortunately	0.31
10	thu	-0.27	obvious	0.31

Table 1: The top-10 terms for π_1 and π_2

The resulting values are in the range of $[-1;+1]$; the absolute score is the degree of typicality, the sign indicates the direction of typicality (-1 means the term clearly belongs to C_1, $+1$ states that the term clearly belongs to class C_2). In our setup, of the 2938 terms with $w_{1\cup 2} \geq 100$, 589 terms (20%) expose a term typicality $\varepsilon \leq -0.05$ (i.e., C_1) and 1255 terms (42.7%) a term typicality of $\varepsilon \geq +0.05$ (i.e., C_2). Table 1 shows the top-10 typical terms for C_1 and C_2 respectively. As one can see, C_1 terms are very likely to appear in template blocks, whereas C_2 terms are more likely for full-text.

4. FULL STOP AND COMMON BLOCKS

To further confirm the observed dichotomy of Web text, two other features of "full text" and "template text" were contrasted with the text density distribution: 1. the presence of full stop characters in the segment and 2. the frequency of the segment. As full stops indicate complete sentences, the number of tokens contained in segments with full-stop should be much higher for C_2 than for C_1 (and vice versa). Indeed, text density has a fairly high information gain for predicting the occurrence of a full stop (0.711), which is substantiated by a classification accuracy of 91.4% using a simple linear classifier. Finally, the token-level distribution of frequent templates [2] was analyzed (38,634 segments occuring at least 10 times, representing 28% of all tokens). A majority of the tokens in segments with $\varrho'(b) \leq 5$ are contained frequent templates (63%). With high chance the other 37% tokens in few-worded segments do also not describe the "main" conteent, and could be considered "template" (23% of all tokens in the corpus). On the other hand, many boilerplate templates contain full sentences ($\varrho'(b) \geq 9$; 20% of all tokens). While detecting these full-text templates obviously requires global information (such as segment fingerprints etc.), identifying the low-density templates can be performed on-the-fly without corpus-level statistics, which makes density-based template detection a clear choice as an upstream filter within a sophisticated template removal strategy.

5. REFERENCES

[1] Gabriel Altmann. Das Problem der Datenhomogenität. In *Glottometrika 13*. Brockmeyer, 1992.

[2] D. Gibson, K. Punera, and A. Tomkins. The volume and evolution of web page templates. In *WWW'05*.

[3] Chr. Kohlschütter and W. Nejdl. A Densitometric Approach to Web Page Segmentation. In *CIKM 2008*.

[4] D. Lavalette. A general purpose ranking variable with applications to various ranking laws. In *Exact Methods in the Study of Language and Text*. 2007.