Chapter XI
Supporting Text Retrieval by Typographical Term Weighting

Lars Werner
University of Paderborn, Germany

Stefan Böttcher
University of Paderborn, Germany

ABSTRACT

Text documents stored in information systems usually consist of more information than the pure concatenation of words, i.e., they also contain typographic information. Because conventional text retrieval methods evaluate only the word frequency, they miss the information provided by typography, e.g., regarding the importance of certain terms. In order to overcome this weakness, we present an approach which uses the typographical information of text documents and show how this improves the efficiency of text retrieval methods. Our approach uses weighting of typographic information in addition to term frequencies for separating relevant information in text documents from the noise. We have evaluated our approach on the basis of automated text classification algorithms. The results show that our weighting approach achieves very competitive classification results using at most 30% of the terms used by conventional approaches, which makes our approach significantly more efficient.

INTRODUCTION

Text documents combine textual and typographical information. However, since Luhn (1958), information retrieval (IR) algorithms use only term frequency in text documents for measuring the text significance, i.e., typographic information also contained in the texts is not considered by most of the common IR methods. Typographic information includes the employment of different character fonts, character sizes and styles, the choice of line length, text alignment and the type-area within the paper format.
Authors use typographical information in their texts to make them more readable. Therefore, we follow the arguments of Apté et al. (1994), Cutler et al. (1997), Kim and Zhang (2000), and Kwon and Lee (2000) that typographical information may help to classify or to better understand the meaning of texts, which results in the following hypothesis that can be regarded as an extension to Luhn’s thesis:

*The justification of measuring word significance by typography is based on the fact that a writer normally uses certain typographic styles to clarify his argumentation and the description of certain facts.*

In order to verify our hypothesis, we have implemented our ideas within the VKC document management system. For an evaluation of the classification quality of our approach, we have used two public data sets of the World Wide Knowledge Base (Web-Kb) project, which contains HTML documents with typographical information and our own selection of publications in PDF format from the ACM Digital Library. The evaluation result is that classification algorithms that consider typography information allow reducing the considered term set, thereby significantly improving the efficiency of the automated document classification.

The remainder of the article is organized as follows. The second section describes some related works. The third section outlines our previous HTML tag-based typographical weighting approach and the fourth section describes our catalogue evaluation scenario and summarizes the performance results of the tag based approach. Within the fifth section we describe our new general typography-based weighting approach, which we evaluate in the sixth section. The seventh section outlines a summary and the conclusions.

**RELATED WORKS**

Apté, Damerau and Weiss (1994) presented the first typographic term weighting approach for text classification. They measured the classification quality of the “Reuters-21578 text categorization test collection” and demonstrated that by counting the terms of the news titles twice, an improvement of nearly 2% (precision recall break even point) could be achieved.

Cutler, Shih and Meng (1997), for the first time, suggested an absolute weighting scheme for HTML tags. By weighting words enclosed in tags depending on the tag weight (c.f. Table 1) the average preci-

<table>
<thead>
<tr>
<th>HTML Tag</th>
<th>Tag Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;a&gt;</td>
<td>1</td>
</tr>
<tr>
<td>&lt;h1&gt;, &lt;h2&gt;</td>
<td>8</td>
</tr>
<tr>
<td>&lt;h3&gt;, &lt;h4&gt;, &lt;h5&gt;, &lt;h6&gt;</td>
<td>1</td>
</tr>
<tr>
<td>&lt;strong&gt;, &lt;b&gt;, &lt;em&gt;, &lt;i&gt;, &lt;u&gt;, &lt;dl&gt;, &lt;ol&gt;, &lt;ul&gt;</td>
<td>1</td>
</tr>
<tr>
<td>&lt;title&gt;</td>
<td>0</td>
</tr>
<tr>
<td>Remaining tags and normal text</td>
<td>1</td>
</tr>
</tbody>
</table>
Related Content

Customer Orientation Based Multi-Agent Negotiation for B2C e-Commerce
[www.igi-global.com/article/customer-orientation-based-multi-agent/43867?camid=4v1a](www.igi-global.com/article/customer-orientation-based-multi-agent/43867?camid=4v1a)

ALiSR: Agent Based Link Stability Aware Routing Protocol for Ad Hoc Networks
[www.igi-global.com/article/alisr-agent-based-link-stability/43866?camid=4v1a](www.igi-global.com/article/alisr-agent-based-link-stability/43866?camid=4v1a)

Autonomous Agents Adopting Organizational Rules
[www.igi-global.com/chapter/autonomous-agents-adopting-organizational-rules/21105?camid=4v1a](www.igi-global.com/chapter/autonomous-agents-adopting-organizational-rules/21105?camid=4v1a)

Agent-Based Modelling of Socio-Ecosystems: A Methodology for the Analysis of Adaptation to Climate Change
[www.igi-global.com/article/agent-based-modelling-socio-ecosystems/47414?camid=4v1a](www.igi-global.com/article/agent-based-modelling-socio-ecosystems/47414?camid=4v1a)