Chapter 2

Exploiting Semantics to Improve Classification of Text Corpus

Hammad Majeed  
*National University of Computer and Emerging Sciences (NUCES), Pakistan*

Firoza Erum  
*National University of Computer and Emerging Sciences (NUCES), Pakistan*

**ABSTRACT**

Internet is growing fast with millions of web pages containing information on every topic. The data placed on Internet is not organized which makes the search process difficult. Classification of the web pages in some predefined classes can improve the organization of this data. In this chapter a semantic based technique is presented to classify text corpus with high accuracy. This technique uses some well-known pre-processing techniques like word stemming, term frequency, and degree of uniqueness. In addition to this a new semantic similarity measure is computed between different terms. The authors believe that semantic similarity based comparison in addition to syntactic matching makes the classification process significantly accurate. The proposed technique is tested on a benchmark dataset and results are compared with already published results. The obtained results are significantly better and that too by using quite small sized highly relevant feature set.

**INTRODUCTION**

Web page classification (WPC) also known as web page categorization is the identification of membership of a web page. Choi and Yao (Choi & Yao, 2005) defined it mathematically as:

Let $C$ represents predefined categories $C = \{c_1, c_2, \ldots, c_k\}$ and $D = \{d_1, d_2, \ldots, d_n\}$ represents the number of web pages or documents need to be classified. The decision matrix be $Z = D \times C$ where each entry represents either belonging to a set $\{0, 1\}$ where 1 indicates the document $d_i$ belonging to category $c_j$, and 0 indicates not belonging to the category. A document can belong to more than one category.

Web page classification means approximating function $f: D \times C \rightarrow \{0, 1\}$ by a learned function called a classifier $f': D \times C \rightarrow \{0, 1\}$ both the functions closely match each other. The $f'$ is acquired by machine.

DOI: 10.4018/978-1-4666-9767-6.ch002
learning over training examples; each training example has a label of category to which it belongs. The function \( f \) is used during training and the classification of web pages. The decision matrix is given in Table 1.

**BASIC APPROACHES OF WEB PAGE CLASSIFICATION**

Web page classification is a supervised machine-learning problem in which a web page is categorized using a trained classifier. The web pages are written in HTML and are semi-structured in nature. They are connected through hyperlinks forming a directed graph. The data on web is frequent, non-homogeneous and vigorously changing.

Basic approaches of web page classification are subject classification, (also known topic-based classification), functional classification (also known genre-based or style-based classification), sentiment classification, binary classification, multi class classification, hard classification, soft classification, flat classification and hierarchical classification.

In subject classification, web pages are classified on the basis of subject or content of web page. This means a page can belong to many categories, for example the “news” domain in yahoo.com is world, business, entertainment, sports, tech, politics etc.

Functional or genre-based classification considers the role of web page based on the content. For example it decides whether the page is a “homepage”, “product catalogue” or an “advertisement page”.

Sentiment classification takes into account the opinion of the author of the page, i.e. how has author presented his views on a given topic. It deals with the emotions of the writer of the web page.

If there are exactly two categories in a given problem then binary classification is used to classify the pages. For example 0 represents not belonging to a class and 1 represents belonging to it. Multiclass classification has many class labels, and it decides the class of the web page. For example, in case of yahoo pages the classifier has multi-labels business, sports, entertainment etc.

On the basis of type of class assignment, classification can be divided into hard and soft classification. Hard classification means a page exclusively belongs to one class or not and has no transitional state, whereas soft classification classifies a page probabilistically.

Organization of categories leads to two types of classification, flat and hierarchical classification. The categories in flat classification are parallel i.e. no category surpasses other. For example in yahoo news domain entertainment, sports, business are alike categories. While in entertainment category movies, television, celebs are sub categories.

Table 1. Decision matrix for classifying n documents in k classes

<table>
<thead>
<tr>
<th>D1</th>
<th>C1</th>
<th>C2</th>
<th>...</th>
<th>Ck</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dk</td>
<td>Zn1</td>
<td>Zn2</td>
<td>...</td>
<td>Znk</td>
</tr>
</tbody>
</table>

24
Related Content

Cloud Computing for Earth Observation
www.igi-global.com/chapter/cloud-computing-for-earth-observation/119904?camid=4v1a

Security for Hybrid Mobile Development: Challenges and Opportunities
www.igi-global.com/chapter/security-for-hybrid-mobile-development/174763?camid=4v1a

Evaluation of Cloud System Success Factors in Supply-Demand Chains
www.igi-global.com/chapter/evaluation-of-cloud-system-success-factors-in-supply-demand-chains/129707?camid=4v1a

Multi-Layer Token Based Authentication Through Honey Password in Fog Computing
www.igi-global.com/article/multi-layer-token-based-authentication-through-honey-password-in-fog-computing/198412?camid=4v1a