CALA: An unsupervised URL-based web page classification system

Hernández I.

Rivero C.R.

Ruiz D.

Corchuelo R.

Unsupervised web page classification refers to the problem of clustering the pages in a web site so that each cluster includes a set of web pages that can be classified using a unique class. The existing proposals to perform web page classification do not fulfill a number of requirements that would make them suitable for enterprise web information integration, namely: to be based on a lightweight crawling, so as to avoid interfering with the normal operation of the web site, to be unsupervised, which avoids the need for a training set of pre-classified pages, or to use features from outside the page to be classified, which avoids having to download it. In this article, we propose CALA, a new automated proposal to generate URL-based web page classifiers. Our proposal builds a number of URL patterns that represent the different classes of pages in a web site, so further pages can be classified by matching their URLs to the patterns. Its salient features are that it fulfills all of the previous requirements, and it has been validated by a number of experiments using real-world, top-visited web sites. Our validation proves that CALA is very effective and efficient in practice. © 2013 Elsevier B.V. All rights reserved.

Enterprise web information integration

URL classification

URL patterns

Web page classification

Web page clustering