Summary: As the Web expands exponentially, the need to put some order to its content becomes apparent. Hypertext categorization, that is the automatic classification of web documents into predefined classes, came to elevate humans from that task. The extra information available in a hypertext document poses new challenges for automatic categorization. HTML tags and metadata provide rich information for hypertext categorization that is not available in traditional text classification. This paper looks at (i) what representation to use for documents and which extra information hidden in HTML pages to take into consideration to improve the classification task, and (ii) how to deal with the very high number of features of texts. A hypertext dataset and four well-known learning algorithms (Naive Bayes, K-nearest neighbor, support vector machines and C4.5) were used to exploit the enriched text representation along with feature reduction. The results showed that enhancing the basic text content with HTML page keywords, title and anchor links improved the accuracy of the classification algorithms.