Abstract

Large amount of insights can be drawn from the articles that are published online. Instead of manually reading all the articles and assigning relevant tags to them satisfying the content, it will be highly efficient if there exists an automated process for performing the task. In this paper, an unsupervised approach for the automated tagging of articles in Chinese language has been implemented. The input is an article and output is the tags to that article. The major challenge is the segmentation of the Chinese characters, which do not make use of separators unlike the English characters. To overcome this, different approaches are combined together in order to get accurate results. Efficient tagging of articles is required, which can be used for many applications in the analysis, one of which is in Recommendation Engine. The tagging process should consider all the aspects of the article and assign the most relevant tags accordingly. The proposed algorithm was implemented for a Chinese Publication House and relevant tags were assigned to its articles of different categories. At the end of the project, the results were manually checked for, in a corpus of 10000 Chinese articles, which reflected the attainment of overall accuracy of around 85%, greater than that obtained through different traditional
Unsupervised Tagging of Chinese Articles

methods.

References

5. “Automatic free-text-tagging of online news archives”.
7. Tf-idf weighting - Stanford NLP Group

Index Terms

Computer Science
Artificial Intelligence

Keywords

Text, articles, automated tagging, tags, unsupervised approach, Recommendation Engine, Chinese, segmentation, corpus.