International Journal of Innovative Computing, Information and Control
Volume **4**, Number **4**, April **2008**

ICIC International ©2008 ISSN 1349-4198
pp. 905–913

SINGLE DOCUMENT KEYWORD EXTRACTION FOR INTERNET NEWS ARTICLES

David B. Bracewell[1], Jiajun Yan[1] and Fuji Ren[1,2]

[1]Department of Information Science and Intelligent Systems
The University of Tokushima
Tokushima 770-8506, Japan
{ davidb; yanjj; ren }@is.tokushima-u.ac.jp

[2]School of Information Engineering
Beijing University of Posts and Telecommunications
Beijing 100876, P. R. China

Received May 2007; revised October 2007

Abstract. *Keywords are a fundamental part of information retrieval (IR) and as such they have been studied extensively. They are used for everything from searching to describing a document. A Keyword extraction algorithm can be defined as a combination of a keyword representation and a selection/weighting scheme. The most common selection/weighting schemes are based on collection statistics or using supervised machine learning algorithms. In these cases, keywords can, typically, only be extracted from documents that belong to a collection or using a large amount of annotated training data. The importance of extracting keywords without a document collection has been gradually increasing due to the Internet. In this paper, a keyword extraction algorithm designed with news in mind that requires neither a document collection or training data is presented. It uses noun phrases as its keyword representation and takes in document statistics to derive its weighting scheme. Through experimentation it is shown that the quality of the keywords extracted from the proposed algorithm are better than standard algorithms for both information retrieval and humans.*

Keywords: Keyword extraction, Indexing, news, Information retrieval, Natural language processing

1. **Introduction.** Keywords are the important words of a document that help describe the content. They are an integral part of information retrieval systems. Keywords have been a part of IR from its earliest days. Most systems represent documents and queries as a set of keywords. Additionally, keywords are used to help in such tasks as classification, summarization, document similarity, and relevance ranking. Because of this, an effective algorithm for extracting keywords is necessary for a successful IR system.

Until now, the prominent paradigm has been to build IR systems on top of preexisting document collections. These systems then allow users to query the document collection for desired information. In this situation, keywords are, typically, defined as single words (unigrams) and are assigned to a document based on statistical information from the words in the document and in the document collection, such as TFIDF (term frequency inverse document frequency) [9].

Over the past few years more and more research has used the Internet as their information source. The web has a wealth of information and it is very desirable for IR systems to harness this information. When using the Internet as the source of information, there is no tangible preexisting document collection and keywords have to be extracted from a single document.