

## عنوان مقاله:

HWS: A Hierarchical Word Spotting Method for Farsi Printed Words Through Word Shape Coding

## محل انتشار:

مجله بین المللی ارتباطات و فناوری اطلاعات, دوره 7, شماره 2 (سال: 1394)

تعداد صفحات اصل مقاله: 12

## نویسندگان:

Mohammadreza Keyvanpour

Reza Tavoli

Saeed Mozaffari

## خلاصه مقاله:

Word shape coding (WSC) is a method of document image retrieval (DIR) based on keyword spotting. By using this method, a word can be recognized in the document image, only by identifying some of the features of the word. In this paper, a hierarchical word spotting method, namely HWS, is presented for Farsi document image retrieval through WSC. In HWS method, document images are retrieved by using a new indexing method. In HWS, at first the words in the document images are shape coded based on topological properties. These features include number of sub-words, ascenders, descenders, and holes. A new feature that has been used for this paper is dot's position in word. Six features are obtained which are one top dot, two top dots, three top dots and one bottom dot, two bottom dots, and three bottom dots. Precision of retrieval increases by using these features. Then, all of the shape codes are indexed by building a tree. Retrieval is done based on keyword query in the tree. The results show that the proposed technique is very fast for large volumes of documents. Time complexity for successful and non-successful searching is  $O(\log kn)$ . This value is better than values in ordinal method. Also, time complexity for indexing is  $O(\log kn)$ . The HWS method is tested on Bijankhan database. ۸۷۸۶۷ common words from this database are used for building the dictionary. Test results show that average of precision is ۰.۸۳ and average recall is ۰.۹۴.

## کلمات کلیدی:

Tree indexing, Information Retrieval, Document Image, word shape coding, Farsi document

## لینک ثابت مقاله در پایگاه سیویلیکا:

<https://civilica.com/doc/1425578>

