पेटेंट कार्यालय
शासकीय जर्नल

OFFICIAL JOURNAL
OF
THE PATENT OFFICE

निर्गमन सं. 08/2014	शुक्रवार	दिनांक: 21/02/2014
ISSUE NO. 08/2014	FRIDAY	DATE: 21/02/2014

पेटेंट कार्यालय का एक प्रकाशन
PUBLICATION OF THE PATENT OFFICE

(12) PATENT APPLICATION PUBLICATION
(19) INDIA
(22) Date of filing of Application :11/02/2014

(21) Application No.627/CHE/2014 A
(43) Publication Date : 21/02/2014

(54) Title of the invention : A NOVEL METHOD AND SYSTEM FOR DETECTING NEAR DUPLICATORS FOR BOTH WEB DOCUMENTS AND NORMAL DOCUMENTS

(51) International classification :G06F17/00
(31) Priority Document No :NA
(32) Priority Date :NA
(33) Name of priority country :NA
(86) International Application No :NA
Filing Date :NA
(87) International Publication No : NA
(61) Patent of Addition to Application Number :NA
Filing Date :NA
(62) Divisional to Application Number :NA
Filing Date :NA

(71)**Name of Applicant :**
1)V A NARAYANA
Address of Applicant :H.NO.3-5-187/1, PLOT NO.357, ROAD NO.7F, KRISHNA NAGAR COLONY, MOULA-ALI, HYDERABAD - 500 040 Andhra Pradesh India
(72)**Name of Inventor :**
1)V A NARAYANA
2)DR. P. PREMCHAND
3)DR. A. GOVARDHAN

(57) Abstract :
Exemplary embodiment of the present disclosure is directed towards a novel method and system for detecting near duplicates for both web and normal documents. The system includes a document parsing unit interfaced with the database analyze the extracted web documents to provide duplicate web documents by removing language scripts and tags, a stop words is used to detect stop words in extracted web documents for removing the stop words, a stemming algorithm unit retrieves the web documents restricted by a common root, a keyword representation unit counts keywords in each extracted web document and sorted in descending order in a table, and a score calculation unit calculates the similarity score measure and a duplication detection unit then compares the calculated similarity score measure with predefined threshold value and finds whether pairs of documents are near duplicate or near duplicates. The similarity score measure value less than the predetermined threshold value to consider the extracted web document to be a near-duplicate web document. Similarly, the similarity score measure value greater than the predetermined threshold value to consider the web document to be not near duplicate web document and further added to the repository.

No. of Pages : 18 No. of Claims : 10