(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 July 2002 (11.07.2002)

PCT

(10) International Publication Number
WO 02/054279 A1

(51) International Patent Classification[7]: G06F 17/21

(21) International Application Number: PCT/SG01/00023

(22) International Filing Date: 4 January 2001 (04.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **KENT RIDGE DIGITAL LABS** [SG/SG]; 21, Heng Mui Keng Terrace, Singapore 119613 (SG).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KANAGASABAI, Rajaraman** [IN/SG]; Block 331, #06-140, Clementi Ave 2, Singapore 120331 (SG). **PAN, Hong** [CN/SG]; Block 11, #07-62, Ghim Moh Road, Singapore 270011 (SG).

(74) Agent: **HELEN YEO & PARTNERS**; #33-00 UOB Plaza 1, 80 Raffles Place, Singapore 048624 (SG).

(81) Designated States *(national)*: SG, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

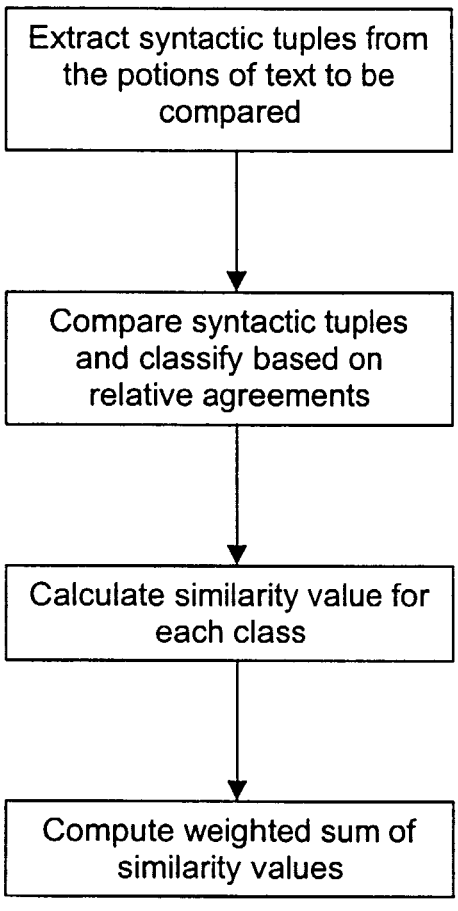
Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: IMPROVED METHOD OF TEXT SIMILARITY MEASUREMENT



(57) Abstract: In one aspect, the present invention provides a method for estimating the similarity between at least two portions of text including the steps of forming a set of syntactic tuples, each tuple including at least two terms and a relation betweeen the two terms; classifying the relation between the terms in the tuples according to a predefined set of relations; establishing the relative agreement between syntactic tuples from the portions of text under comparison according to predefined classes of agreement; calculating a value representative of the similarity between the portions of text of each of the classes of agreement; and establishing a value for the similarity between the portions of text by calculating a weighted sum of the values representative of the similarity between the portions of text for each of the classes of agreement. Preferaly, the step of calculating a value representative of the similarity between the portions of text for each of the classes of agreement includes a weighting based upon the number of matched terms occurring in particular parts of speech in which the text occurs. It is also preferred that the step of calculating a value representative of the similarity between the portions of text for each of the classes of agreement include the application of a weighting factor to the estimate of similarity for each of the classes of agreement and the parts of speech in which matched terms occur.

WO 02/054279 A1