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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 December 2004 (29.12.2004)

PCT

(10) International Publication Number
WO 2004/114157 A1

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number:
PCT/SG2004/000179

(22) International Filing Date: 14 June 2004 (14.06.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/465,567 20 June 2003 (20.06.2003) US

(71) Applicant *(for all designated States except US)*: **AGENCY FOR SCIENCE, TECHNOLOGY AND RESEARCH** [SG/SG]; 20 Biopolis Way, #07-01 Centros, Singapore 138668 (SG).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **JI, Donghong** [CN/SG]; c/o Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 (SG). **YANG, Lingpeng** [CN/SG]; c/o Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 (SG). **NIE, Yu** [CN/SG]; c/o Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 (SG).

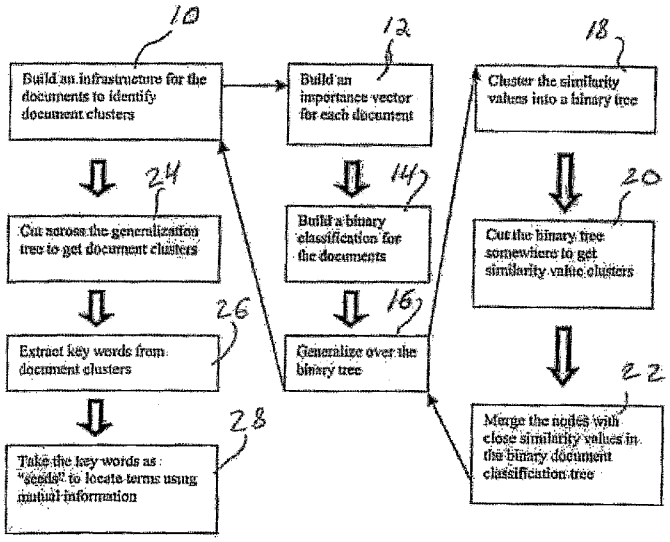
(74) Agent: **GREEN-KELLY, James, Patrick**; Lloyd Wise, Tanjong Pagar, P.O. Box 636, Singapore 910816 (SG).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

*[Continued on next page]*

(54) Title: METHOD AND PLATFORM FOR TERM EXTRACTION FROM LARGE COLLECTION OF DOCUMENTS




(57) Abstract: A method and platform for statistically extracting terms from large sets of documents is described. An importance vector is determined for each document in the set of documents based on importance values for words in each document. A binary document classification tree is formed by clustering the documents into clusters of similar documents based on the importance vector for each document. An infrastructure is built for the set of documents by generalizing the binary document classification tree. The document clusters are determined by dividing the generalized tree of the infrastructure into two parts and cutting away the upper part. Statistically significant individual key words are extracted from the clusters of similar documents. Key words are treated as seeds and terms are extracted by starting from the seeds and extending to their left or right contexts.

WO 2004/114157 A1

**Published:**
— *with international search report*

*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.*