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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 August 2008 (14.08.2008)

PCT

(10) International Publication Number
WO 2008/097194 A1

(51) International Patent Classification:
*G06F 7/04* (2006.01) *G06F 17/30* (2006.01)
*G06F 7/00* (2006.01) *G06F 17/28* (2006.01)
*G06F 17/21* (2006.01)

(21) International Application Number:
PCT/SG2007/000044

(22) International Filing Date: 9 February 2007 (09.02.2007)

(25) Filing Language: English

(26) Publication Language: English

(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)*: **BAI, Shuanhu** [CN/SG]; Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 (SG). **LI, Haizhou** [SG/SG]; Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 (SG).

(74) Agent: **MCLAUGHLIN, Michael, Gerard**; 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, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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, IS, IT, LT, LU, LV, 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).

**Declaration under Rule 4.17:**
— *of inventorship (Rule 4.17(iv))*

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

(54) Title: KEYWORD CLASSIFICATION AND DETERMINATION IN LANGUAGE MODELLING

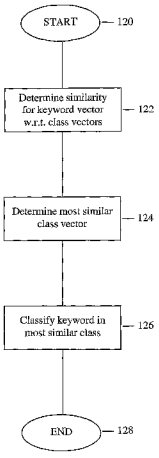



Figure 4

(57) Abstract: A computer-implemented method and apparatus defines a keyword class vector. A set of seed keywords is determined from a set of keywords and first and second most similar keywords from the set of seed keywords are then determined. A class vector is determined from first and second keyword vectors associated with the first and second most similar keywords. The method and apparatus also classifies a keyword in a keyword class. A similarity for a keyword vector associated with the keyword is determined with reference to a plurality of class vectors, each class vector having an associated class and determines a most similar class vector of the plurality of class vectors from the similarity determination. The keyword is then classified in a most similar class associated with the most similar class vector.

WO 2008/097194 A1