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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 June 2005 (02.06.2005)

PCT

(10) International Publication Number
WO 2005/050475 A1

(51) International Patent Classification[7]: G06F 17/27, 17/28

(21) International Application Number:
PCT/SG2004/000373

(22) International Filing Date:
19 November 2004 (19.11.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
200307192-5 21 November 2003 (21.11.2003) SG

(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)*: **LAI, Fon Lin** [MY/SG]; c/o Institute For Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 (SG). **TAN, Ah Hwee** [SG/SG]; c/o Institute For Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 (SG).

(74) Agent: **ELLA CHEONG SPRUSON & FERGUSON (SINGAPORE) PTE LTD**; P.O. Box 1531, Robinson Road Post Office, Singapore 903031 (SG).

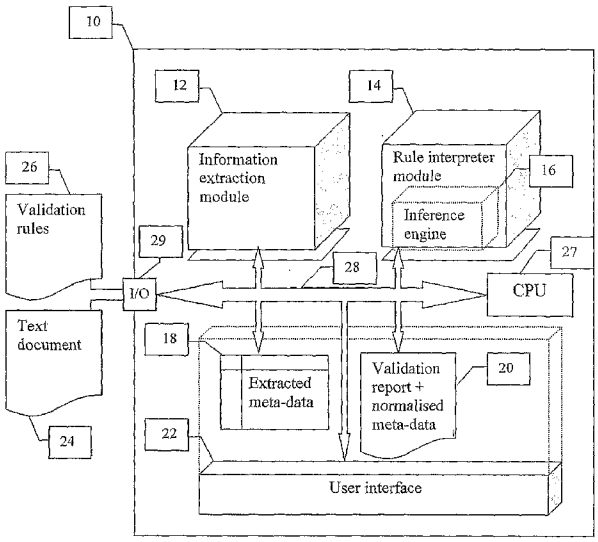
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, 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, IS, 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).

Published:
— *with international search report*

[Continued on next page]

(54) Title: METHOD AND SYSTEM FOR VALIDATING THE CONTENT OF TECHNICAL DOCUMENTS



(57) Abstract: An automatic document validation system that can be trained to extract domain-specific entities and their linguistically-associated physical, abstract or relational properties, as described within an electronic document. Training of the system can be achieved through the provision of a set of example documents representative of the domain and that have been manually tagged by a domain expert in such a way as to identify the various types of entities and their associated set of recordable properties. Together with a domain-specific vocabulary (e.g.. a dictionary), the trained system is then able to automatically process new documents belonging to the same domain and to test the extracted information on any number of content-conditional rules that have been specified by the domain expert as necessary to confirm the completeness and validity of the new documents.

WO 2005/050475 A1

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.