

REVISED VERSION

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
18 September 2014 (18.09.2014)



(10) International Publication Number
WO 2014/140977 A9

(51) International Patent Classification:
G06F 17/28 (2006.01)

(21) International Application Number:

PCT/IB2014/059310

(22) International Filing Date:

27 February 2014 (27.02.2014)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

13/843,377 15 March 2013 (15.03.2013) US

(71) Applicant: **INTERNATIONAL BUSINESS MACHINES CORPORATION** [US/US]; New Orchard Road, Armonk, New York 10504 (US).

(71) Applicants (for MG only): **IBM UNITED KINGDOM LIMITED** [GB/GB]; PO Box 41, North Harbour, North Harbour, Portsmouth Hampshire PO6 3AU (GB). **IBM**

(CHINA) **INVESTMENT COMPANY LIMITED** [CN/CN]; 25/F, Pangu Plaza, No.27, Central North 4th Ring Road, Chaoyang District, Beijing 100101 (CN).

(72) Inventors: **GERKEN III, John, Kenyon**; Ibm Corporation, 4205 S Miami Blvd, Durham, North Carolina 27703-9141 (US). **ZBOICHYK, Fiodar**; Ibm Corporation, 4205 S Miami Blvd, Durham, North Carolina 27703-9141 (US). **PRAGER, John, Martin**; Ibm Corporation, 1101 Kit-chawan Rd, Route 134, PO Box 218, Yorktown Heights, New York 10598 (US).

(74) Agent: **SHAW, Anita**; IBM United Kingdom Limited, Intellectual Property Law, Hursley Park, Winchester Hampshire SO21 2JN (GB).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,

[Continued on next page]

(54) Title: IMPROVING ENTITY RECOGNITION IN NATURAL LANGUAGE PROCESSING SYSTEMS

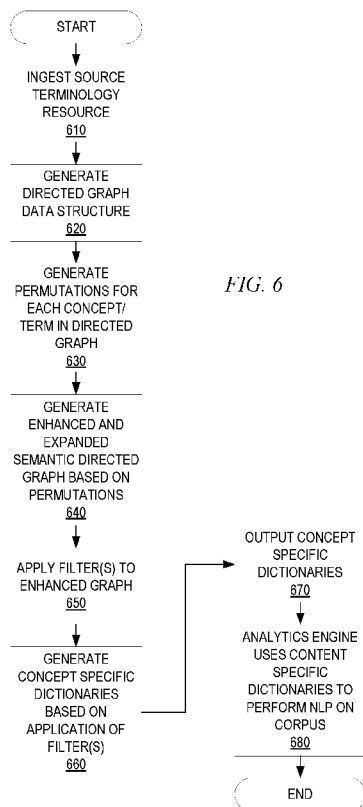


FIG. 6

(57) Abstract: Mechanisms are provided for generating a dictionary data structure for analytical operations. A source terminology resource is ingested to generate a hierarchical representation of the source terminology resource comprising nodes for terms related to concepts in the source terminology resource. For a node of the nodes in the hierarchical representation of the source terminology resource, a permutation of a corresponding term associated with the node is generated. An expanded hierarchical representation of the source terminology resource is generated based on the generated permutation. An enhanced dictionary data structure is generated based on the expanded hierarchical representation and output to an analytics engine to perform analysis of a corpus of information using the enhanced dictionary data structure.



HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU,

(88) Date of publication of the revised international search report:

18 December 2014

(15) Information about Correction:
see Notice of 18 December 2014

INTERNATIONAL SEARCH REPORT

International application No.

PCT/IB2014/059310

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/28(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE:natural language processing, NLP, dictionary, hierarchical, expand, node, tree, semantic, syntax, order.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 7596485 B2 (MICROSOFT CORPORATION) 29 September 2009 (2009-09-29) claims 1-17, col.2, line 6 - col.18, line 67, figures 1-9	1-34
A	CN 1226327 A (MICROSOFT CORPORATION) 18 August 1999 (1999-08-18) the whole document	1-34
A	CN 101251841 A (EAST CHINA NORMAL UNIVERSITY) 27 August 2008 (2008-08-27) the whole document	1-34

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A”	document defining the general state of the art which is not considered to be of particular relevance	“T”	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“E”	earlier application or patent but published on or after the international filing date	“X”	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“L”	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“Y”	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“O”	document referring to an oral disclosure, use, exhibition or other means	“&”	document member of the same patent family
“P”	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

08 June 2014

Date of mailing of the international search report

30 June 2014

Name and mailing address of the ISA/

STATE INTELLECTUAL PROPERTY OFFICE OF THE
P.R.CHINA(ISA/CN)
6,Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088 China

Authorized officer

DU,Jingzi

Facsimile No. (86-10)62019451

Telephone No. (86-10)88231606

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/IB2014/059310

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	7596485	B2	29 September 2009	Non e			
CN	1226327	A	18 August 1999	EP	0907923	B1	13 February 2002
				DE	69710458	T2	27 June 2002
				HK	1022027	A1	14 May 2004
				JP	4668363	B2	13 April 2011
				US	5966686	A	12 October 1999
				WO	9800793	A1	08 January 1998
CN	101251841	A	27 August 2008	Non e			