



- (51) **International Patent Classification:**
G06F 15/18 (2006.01)
- (21) **International Application Number:**
PCT/US2012/029260
- (22) **International Filing Date:**
15 March 2012 (15.03.2012)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
13/050,726 17 March 2011 (17.03.2011) US
- (71) **Applicant (for all designated States except US):**
AMAZON TECHNOLOGIES, INC. [US/US]; Po Box 8102, Reno, NV 89507 (US).
- (72) **Inventor; and**
- (75) **Inventor/Applicant (for US only):** **GERSHNIK, Eugene** [US/US]; 410 Terry Avenue North, Seattle, WA 98109-5210 (US).
- (74) **Agents:** **LEE, Lewis, C.** et al.; Lee & Hayes, PLLC, 601 W. Riverside Ave, Suite 1400, Spokane, WA 99201 (US).
- (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, 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, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, 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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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.

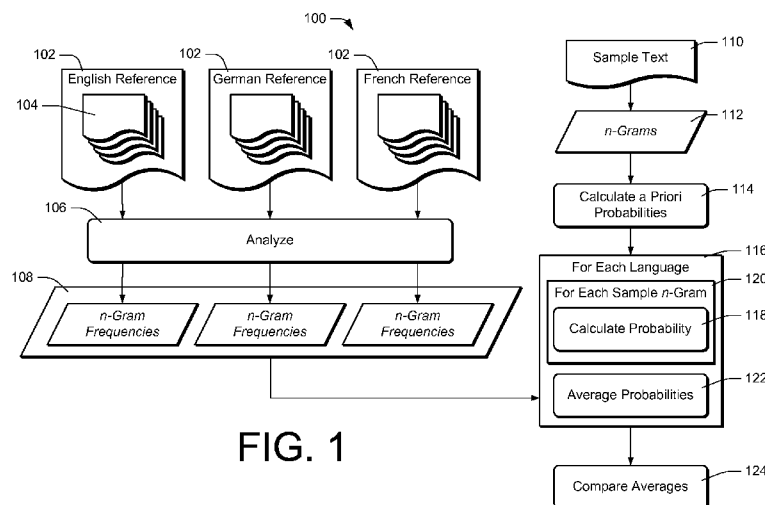
- (84) **Designated States (unless otherwise indicated, for every kind of regional protection available):** ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR,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

Published:

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

- (88)
- Date of publication of the international search report:**

8 May 2014

(54) **Title:** N-GRAM-BASED LANGUAGE PREDICTION

(57) **Abstract:** Techniques are described for predicting the language of a text excerpt. The language prediction is accomplished by comparing w-grams of the text excerpt with n-grams of different language references. A probability is calculated for each w-gram of the text excerpt with respect to each of the language references. The calculated probabilities corresponding to a single language are then averaged to yield an overall probability corresponding to that language, and the resulting overall probabilities are compared to find the most likely language of the sample text.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 12/29260

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 15/18 (2012.01)

USPC - 706/12

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)

USPC: 706/12

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

USPC: 706/50, 52, 59; 715/262, 264, 265; 707/706, 723, 999.003, E17.058, E17.078, E17.109; 704/1, 9 (keyword limited - see search terms below)

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

PubWEST (PGPB, USPT, USOC, EPAB, JPAB); Google Web; Google Scholar

Terms: language, predict, n-gram, probability, bayesian, frequency, occurrences, string, average, foreign, unique, cluster, decompose.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2007/0219777 A1 (CHU et al.) 20 September 2007 (20.09.2007) entire document, especially abstract; para [0007], [0016], [0027], [0031], [0040], [0043], [0044], [0047], [0049], [0050], [0052], [0055].	1-15
A	US 2011/0035656 A1 (KING et al.) 10 February 2011 (10.02.2011) entire document, especially abstract; para [0047], [0048], [0145], [0252], [0275], [0276].	1-15
A	US 2010/0306139 A1 (WU et al.) 02 December 2010 (02.12.2010) entire document, especially abstract; para [0003], [0005], [0012], [0030], [0036], [0038].	1-15

☐ Further documents are listed in the continuation of Box C.


* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"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)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"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

"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

"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

"&" document member of the same patent family

Date of the actual completion of the international search

30 May 2012 (30.05.2012)

Date of mailing of the international search report

08 JUN 2012

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-3201

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774