



(51) International Patent Classification:

G06F 17/27 (2006.01) G06F 17/30 (2006.01)

(21) International Application Number:

PCT/IB2018/000472

(22) International Filing Date:

12 April 2018 (12.04.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/485,428 14 April 2017 (14.04.2017) US
15/950,537 11 April 2018 (11.04.2018) US

(71) Applicant: **NOVABASE BUSINESS SOLUTIONS, S.A.**
[PT/PT]; Av. D. Joao II, N 34, Parque Das Nacoes,
1998-031 Lisboa (PT).

(72) Inventors: **LEAL, Joao**; Rua Francisco Ferreira Neves, N
4 - 4 Esq., 3800-510 Aveiro (PT). **DE FATIMA MACHA-**
DO DIAS, Maria; Rua Das Pocas N3, 5450-203 Capeludos
De Aguiar (PT). **PINTO, Sara**; Rua Vale Da Barrega N68

Costa Do Valado, 3810 806 Oliveira (PT). **VERRUMA, Pedro**; Rua Princesa Cindazunda, N61 1f, 3030-503 Coimbra (PT). **ANTUNES, Bruno**; Rua Antonio Goncalves N 65, Fraccao E, 3040-375 Coimbra (PT). **GOMES, Paulo**; Rua Alvaro Correia, 42, 3020-493 Coimbra (PT).

(74) Agent: **PELLY, Jason, Charles**; Boulton Wade Tennant LLP, Verulam Gardens, 70 Gray's Inn Road, London WC1X 8BT (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, 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.

(54) Title: DOCUMENT CLASSIFICATION USING MACHINE LEARNING

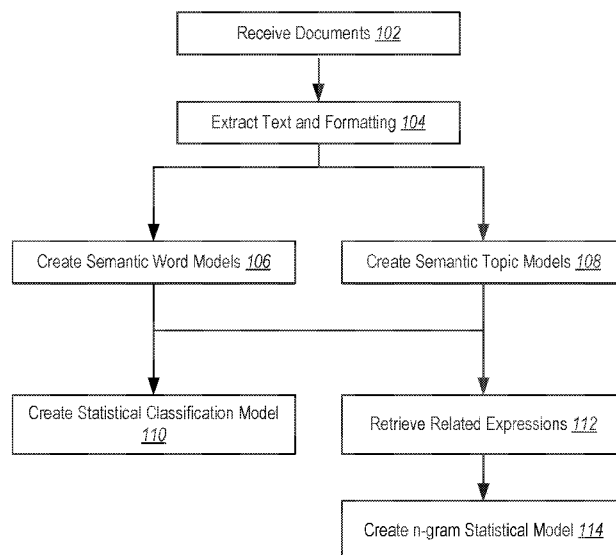


FIG. 1A

(57) Abstract: Disclosed herein are embodiments of systems, devices, and methods automated document analysis and processing using machine learning techniques. In one embodiment, systems and methods are disclosed for automatically classifying documents. In another embodiment, systems and methods are disclosed for identifying new tags for untagged documents. In another embodiment, systems and methods are disclosed for identifying documents related to a target document.

(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, ST, 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, 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))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

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

29 November 2018 (29.11.2018)

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2018/000472

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/27 G06F17/30
ADD.

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)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 2 624 149 A2 (XEROX CORP [US]) 7 August 2013 (2013-08-07) paragraphs [0013] - [0021], [0025] - [0033]; claims 1-5; figures 1-7 -----	1-5, 17-21
X	Stéphane Clinchant ET AL: "Aggregating Continuous Word Embeddings for Information Retrieval", Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, 9 August 2013 (2013-08-09), pages 100-109, XP055495645, Retrieved from the Internet: URL:http://wing.comp.nus.edu.sg/~antho/W/W 13/W13-3212.pdf [retrieved on 2018-07-26] sections 1-3 ----- -/--	1-5, 17-21



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

"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

26 July 2018

Date of mailing of the international search report

02/10/2018

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Alt, Susanne

INTERNATIONAL SEARCH REPORT

International application No

PCT/IB2018/000472

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2016/110343 A1 (KUMAR RANGARAJAN SRIDHAR VIVEK [US]) 21 April 2016 (2016-04-21) paragraphs [0052] - [0064]; figure 8 -----	1-5, 17-21

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2018/000472

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-5, 17-21

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-5, 17-21

Computer-implemented classification of documents

2. claims: 6-11, 22-27

Computer-implemented suggestion of tags for documents

3. claims: 12-16, 28-32

Computer-implemented detection of related documents

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/IB2018/000472

Patent document cited in search report	Publication date	Patent family member(s)	Publication date	
EP 2624149	A2	07-08-2013	EP 2624149 A2	07-08-2013
		US 2013204885 A1		08-08-2013

US 2016110343	A1	21-04-2016	US 2016110343 A1	21-04-2016
		US 2017116178 A1		27-04-2017
		US 2018157639 A1		07-06-2018
