

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

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
25 August 2022 (25.08.2022)



(10) International Publication Number
WO 2022/177449 A1

(51) International Patent Classification:

G06F 16/28 (2019.01) G06N 5/02 (2006.01)
G06F 16/22 (2019.01) G06N 20/00 (2019.01)

(21) International Application Number:

PCT/NZ2021/050135

(22) International Filing Date:

19 August 2021 (19.08.2021)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

2021900421 18 February 2021 (18.02.2021) AU

(71) Applicant: **XERO LIMITED** [NZ/NZ]; 19-23 Taranaki Street, Te Aro, Wellington, 6011 (NZ).

(72) Inventors: **WU, Yu**; 19-23 Taranaki Street, Te Aro, Wellington, 6011 (NZ). **FAKHOURI, Salim**; 19-23

Taranaki Street, Te Aro, Wellington, 6011 (NZ). **KHODEIR, Mohamed**; 19-23 Taranaki Street, Te Aro, Wellington, 6011 (NZ). **GLEYZES, Jerome**; 19-23 Taranaki Street, Te Aro, Wellington, 6011 (NZ). **CHEAH, Soon-Ee**; 19-23 Taranaki Street, Te Aro, Wellington, 6011 (NZ).

(74) Agent: **FB RICE**; Level 33, 477 Collins Street, Melbourne, Victoria 3000 (AU).

(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, IT, 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,

(54) Title: SYSTEMS AND METHOD FOR GENERATING LABELLED DATASETS

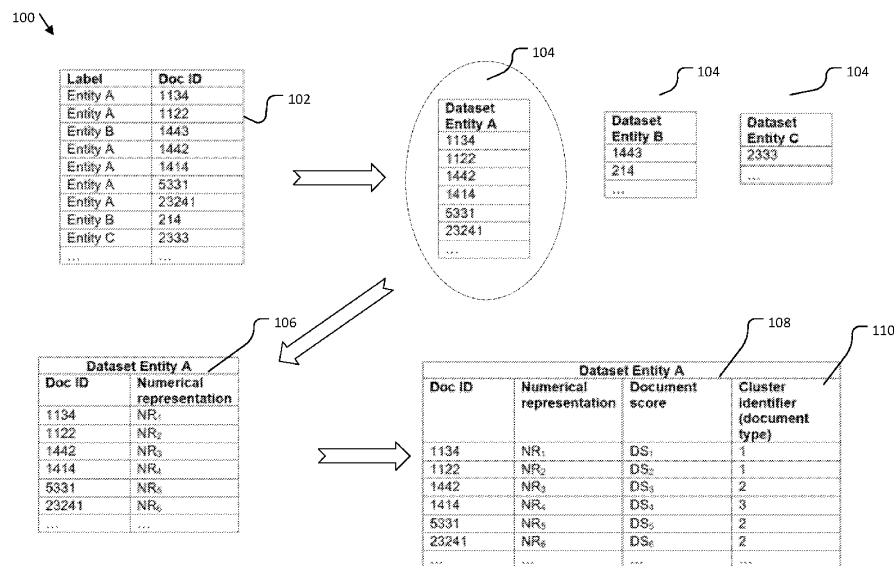


Figure 1

(57) Abstract: A method comprises determining a plurality of documents and for each document of the plurality of documents: (i) providing the document to a numerical representation generation model; (ii) generating, by the numerical representation generation model, a numerical representation of the document; and (iii) determining a document score for the document based on the numerical representation. The method further comprises c) providing the document scores to a clustering module; d) determining, by the clustering module, one or more clusters, each cluster being associated with a class of the documents; e) outputting, by the clustering module, a cluster identifier indicative of the class of each document; and f) associating each document with its respective cluster identifier.

[Continued on next page]

SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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, 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))*
- *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*