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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 November 2021 (25.11.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/236027 A1

(51) International Patent Classification:
*G06F 16/30* (2019.01)

(21) International Application Number:
PCT/TR2020/050440

(22) International Filing Date:
22 May 2020 (22.05.2020)

(25) Filing Language: English

(26) Publication Language: English

(72) Inventor; and
(71) Applicant: **TEKİN, Yaşar** [TR/TR]; Cumhuriyet Mah. 3ncü Şube Sok. No:23/1, Uşak (TR).

(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, 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))*
— *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))*

WO 2021/236027 A1

(54) **Title:** PARAMETER OPTIMIZATION IN UNSUPERVISED TEXT MINING

(57) **Abstract:** The present disclosure provides a method for parameter optimization in unsupervised text mining techniques. The method comprises: a) generating a parameter pool composed of a plurality of parameter vectors; b) generating a model for each parameter vector in the parameter pool; c) calculating pairwise semantic relatedness scores between representative texts in clusters of the models; d) calculating scores of the clusters by averaging the scores of the representative texts; e) calculating scores of the models by averaging the scores of the clusters; f) comparing the scores of the parameter vectors which are the scores of the corresponding models; g) updating the parameter pool; h) repeating the steps b through g until termination condition is met. The method increases the accuracy of the unsupervised text mining techniques by effectively and efficiently optimizing their parameters.

# PARAMETER OPTIMIZATION IN UNSUPERVISED TEXT MINING

## TECHNICAL FIELD

The present disclosure relates to text mining field, and more particularly relates to a method for parameter optimization in the unsupervised text mining techniques.

## BACKGROUND ART

Text Mining is about discovering patterns from textual data. The techniques used in this field can be grouped in two main categories: supervised and unsupervised. While supervised text mining uses labelled text for training, unsupervised text mining uses unlabelled text.

Performance of a model in an unsupervised text mining technique depends on its parameter settings. The performances of the models generated with different parameter values vary greatly. Despite their broad use in many different fields, the unsupervised text mining techniques have an unresolved problem: how to optimize parameters. Examples of the parameters may include, but are not limited to, the number of topics, a Dirichlet prior on document-topic distributions and a Dirichlet prior on topic-word distributions in Latent Dirichlet Allocation topic model, and the number of clusters in K-means clustering.

Parameter optimization problem prevents the unsupervised text mining techniques from obtaining accurate results. If the parameters are not optimized in an appropriate manner, the results become meaningless and can be effective neither in the intrinsic nor in the extrinsic tasks. Thus, there is a need to develop an effective and efficient method for parameter optimization.

## DETAILED DESCRIPTION

As used herein, the singular forms "a," "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. Additionally, the plural forms are intended to include that the item is one or more, including both singular and plural forms of the term it modifies.

The terms “comprises”, “comprising”, or any other variations thereof, are intended to cover a non-exclusive inclusion, such that a process or method that comprises a list of

steps does not include only those steps but may include other steps not expressly listed or inherent to such a process or method.

References throughout this specification to “one embodiment”, “an embodiment”, “another embodiment”, “such embodiment”, “some embodiment”, “an example”, “another example”, “a specific example”, “an example embodiment”, etc., indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Furthermore, the particular feature, structure, or characteristic may be combined in any suitable manner in one or more embodiments or examples.

Embodiments described and descriptions made in this specification are explanatory, illustrative, and used to make the present disclosure understandable. The embodiments and descriptions shall not be construed to limit the present disclosure. Other embodiments are possible, and modifications and variations can be made to the embodiments without departing from spirit, principles and scope of the present disclosure.

It would also be apparent to one of skill in the relevant art that the embodiments described in this specification can be implemented in many different embodiments of the unsupervised text mining techniques, the optimization techniques and the semantic relatedness measures. Various working modifications can be made to the method in order to implement the inventive concept taught in this specification.

Unless otherwise defined, all technical and scientific terms used in this specification have the same meaning as commonly understood by those skilled in the relevant art to which this disclosure belongs. The system, methods, and examples provided herein are only illustrative and not intended to be limiting.

Embodiments of the present disclosure relate to a method for optimizing parameters in the unsupervised text mining techniques. The method includes the following steps:

**At step a**, a parameter pool is generated composed of a plurality of parameter vectors. A parameter vector is a collection of parameter values which have the same size with the number of parameters being optimized. A parameter vector may be any kind of

collection that has a value for each of the parameters. In some embodiments, parameter vectors may be initialized randomly within a range between the parameters' predefined minimum and maximum values, while in another embodiment, they may be initialized using a braced initializer list.

**At step b**, a model is generated with each parameter vector in the pool by using the selected unsupervised text mining technique.

In one embodiment, the technique and the model may be the topic modeling and a topic model respectively, while in another embodiment, they may be the clustering and a cluster model.

Moreover, in one embodiment, the model may be a single model, while in another embodiment, it may be a plurality of replicated models generated with the same parameter vector. Average score of the replicated models may be used as the score of the parameter vector with which the replicated models are generated to alleviate the effects of the model instability.

**At step c**, the pairwise semantic relatedness scores are calculated between the representative texts in the clusters of the models.

In one embodiment, the cluster may be a topic of a topic model, while in another embodiment, it may be a cluster of a clustering model.

Moreover, in one embodiment, the representative texts may be top words of a topic, while in another embodiment, they may be top n-grams.

Furthermore, in one embodiment, the semantic relatedness score may be calculated by a distributional semantic similarity measure, while in another embodiment, it may be calculated by a knowledge-based semantic similarity measure.

**At step d**, the scores of the clusters are calculated by averaging the scores of the representative texts. For each cluster, the score is calculated by averaging the scores of its representative texts. In one embodiment, the measure used to average the scores may be the mean, while in another embodiment, it may be the median.

**At step e**, the scores of the models are calculated by averaging the scores of the clusters. For each model, the score is calculated by averaging the scores of its clusters.

**At step f**, the scores of the parameter vectors are compared to choose the next candidates. The score of a parameter vector is the score of the model generated with this parameter vector.

In one embodiment, the aim of the comparison may be to select the parameter vectors with higher scores, while in another embodiment, there may also be situations where the parameter vectors with lower scores are selected.

**At step g**, the parameter pool is updated based on the rules determined by the selected optimization technique. In one embodiment, the rules may be determined by the mutation and crossover strategies of the Differential Evolution algorithm.

**At step h**, the steps b through g are repeated until the termination condition is met. In one embodiment, the termination condition may be the maximum number of iterations, while in another embodiment, it may be a pre-specified threshold between the best and the worst scores of the parameter vectors.

Additionally, in one embodiment, the method given in this specification may be implemented as a distributed system.

**WHAT IS CLAIMED IS:**

1. A method for optimizing parameters in unsupervised text mining techniques, the method comprising:

a) generating a parameter pool composed of a plurality of parameter vectors;

b) generating a model for each parameter vector in the parameter pool;

c) calculating pairwise semantic relatedness scores between representative texts in clusters of the models;

d) calculating scores of the clusters by averaging the scores of the representative texts;

e) calculating scores of the models by averaging the scores of the clusters;

f) comparing the scores of the parameter vectors, which are the scores of the corresponding models;

g) updating the parameter pool; and

h) repeating the steps b through g until termination condition is met.

2. The method of Claim 1, wherein the model is a topic model, the cluster is a topic and the representative text is a top word.

3. The method of Claim 1, wherein the model comprises a single model or a plurality of replicated models generated with the same parameter vector, the score of which is calculated by averaging the scores of the replicated models.

# INTERNATIONAL SEARCH REPORT

International application No.
**PCT/TR2020/050440**

**A. CLASSIFICATION OF SUBJECT MATTER**

G06F 16/30 (2019.01)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 16/30

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

TURKPATENT Patent Databases

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

EPO Abstract and Fulltext Databases and keywords: text mining, text analysis, unsupervised, parameter, vector, pool, set, iterative, semantic relatedness, termination condition

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | US 2016299955 A1 (MUSIGMA BUSINESS SOLUTIONS PVT LTD [IN]) 13 October 2016 (2016-10-13)<br>The whole document | 1-3 |
| A | US 2011208709 A1 (KINKADEE SYSTEMS GMBH [DE] (B2)HOLTHAUSEN KLAUS [DE]; KINKADEE SYSTEMS GMBH [DE]) 25 August 2011 (2011-08-25)<br>The whole document | 1-3 |
| A | US 2004117336 A1 (IBM [US]) 17 June 2004 (2004-06-17)<br>The whole document | 1-3 |

☐ 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

"D" document cited by the applicant in the international application

"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 | Date of mailing of the international search report |
|---|---|
| **16 September 2021** | **16 September 2021** |

| Name and mailing address of the ISA/TR | Authorized officer |
|---|---|
| **Turkish Patent and Trademark Office (Turkpatent)**<br>**Hipodrom Caddesi No. 13**<br>**06560 Yenimahalle**<br>**Ankara**<br>**Turkey** | **Ahmet Kayakoku** |
| Telephone No. **(90-312) 303 11 82**<br>Facsimile No. **+903123031220** | Telephone No. |

**INTERNATIONAL SEARCH REPORT**
**Information on patent family members**

International application No.
**PCT/TR2020/050440**

| Patent document cited in search report | | | Publication date (day/month/year) | Patent family member(s) | | | Publication date (day/month/year) |
|---|---|---|---|---|---|---|---|
| US | 2016299955 | A1 | 13 October 2016 | WO | 2016162879 | A1 | 13 October 2016 |
| | | | | AU | 2015204283 | A1 | 27 October 2016 |
| | | | | CN | 106055545 | A | 26 October 2016 |
| | | | | KR | 20160121382 | A | 19 October 2016 |
| | | | | SG | 10201506472V | A | 29 November 2016 |
| | | | | TW | 201638803 | A | 01 November 2016 |
| | | | | ZA | 201504892 | B | 27 July 2016 |
| US | 2011208709 | A1 | 25 August 2011 | WO | 2009068072 | A1 | 04 June 2009 |
| | | | | US | 8396851 | B2 | 12 March 2013 |
| | | | | EP | 2215567 | A1 | 11 August 2010 |
| US | 2004117336 | A1 | 17 June 2004 | NONE | | | |