DOI: 10.20998/2079-0023.2023.01.08
UDC 004.94

***A. M. KOPP***, Doctor of Philosophy (PhD), Docent, National Technical University "Kharkiv Polytechnic Institute", Associate Professor at the Department of Software Engineering and Management Intelligent Technologies, Kharkiv, Ukraine, e mail: kopp93@gmail.com, ORCID: https://orcid.org/0000-0002-3189-5623
***D. L. ORLOVSKYI***, Candidate of Technical Sciences (PhD), Docent, National Technical University "Kharkiv Polytechnic Institute", Associate Professor at the Department of Software Engineering and Management Intelligent Technologies, Kharkiv, Ukraine, e mail: orlovskyi.dm@gmail.com, ORCID: https://orcid.org/0000-0002-8261-2988

# AN ALGORITHM FOR NLP-BASED SIMILARITY MEASUREMENT OF ACTIVITY LABELS IN A DATABASE OF BUSINESS PROCESS MODELS

Business process modeling is an important part of organizational management since it enables companies to obtain insights into their operational workflows and find opportunities for development. However, evaluating and quantifying the similarity of multiple business process models can be difficult because these models frequently differ greatly in terms of structure and nomenclature. This study offers an approach that uses natural language processing techniques to evaluate the similarity of business process models in order to address this issue. The algorithm uses the activity labels given in the business process models as input to produce textual descriptions of the associated business processes. The algorithm includes various preprocessing stages to guarantee that the textual descriptions are correct and consistent. First, single words are retrieved and transformed to lower case from the resulting textual descriptions. After that, all non-alphabetic and stop words are removed from the retrieved words. The remaining words are then stemmed, which includes reducing them to their base form. The algorithm evaluates the similarity of distinct business process models using similarity measures, including Jaccard, Sorensen – Dice, overlap, and simple matching coefficients, after the textual descriptions have been prepared and preprocessed. These metrics provide a more detailed understanding of the similarities and differences across various business process models, which can then be used to influence decision-making and business process improvement initiatives. The software implementation of the proposed algorithm demonstrates its usage for similarity measurement in a database of business process models. Experiments show that the developed algorithm is 31% faster than a search based on the SQL LIKE clause and allows finding 18% more similar models in the business process model database.

**Keywords:** business process model, database of business process models, natural language processing, similarity measurement algorithm, activity labels, software implementation of the algorithm.

**Introduction.** Business process modeling is the baseline technique of the Business Process Management (BPM) approach. It focuses on the depiction of organizational workflows in the form of visual diagrams similar to workflows but focused on business activities rather than programming tasks. Business process modeling helps to describe activities visually to train new employees, detect inefficient spots in the company operations, capture requirements to enterprise information systems, design new business processes, etc.

Today BPMN is the de-facto standard for business process modeling maintained by the Object Management Group (OMG) since 2005 and then updated to the BPMN 2.0 in 2011 [1]. This modeling notation has been extended to the XML-based (eXtensible Markup Language) language suitable not only for visual depiction of business process scenarios but also to exchange created diagram files between heterogeneous BPM suites and execute depicted workflows using BPM engines that can automate routine process scenarios.

Therefore, organizations at the higher levels of BPM maturity tend to continuously improve their business activities using BPMN modeling techniques. However, it does not mean enterprises should deal with BPM decisions only based on their resources and experience. Most organizations used so-called “reference models” – collections of typical business processes, generic or industry-specific. Such reference models accumulate proven industry standards, and best practices of multiple successful companies, and could be customized according to particular business needs [2]. Some of the most widely-spread collections of reference business process models are Process Classification Framework (PCF) by American Productivity & Quality Center (APQC) and Supply-Chain Operations Reference (SCOR). While APQC’s PCF is the cross-industry taxonomy of business processes [3], SCOR focuses on logistics process areas, such as supply, manufactory, and delivery [4].

Hence, many organizations may face the problem of searching for similar business process models in collections of business process models, such as APQC’s PCF or SCOR. The solution should provide the capability to find similar BPMN models to a given model or only by the textual description of a business process if an organization does not have a BPMN model yet.

**Literature review and problem statement.** The similarity measurement between business process models has been studied in many papers. Some of the most relevant studies ([5]; [6]; [7]; [8]; [9]; [10]) are described below.

Paper [5] introduces the similarity search problem, where the objects in a collection and a query object are business process models. The authors consider the similarity search as the comparison of the query object against a collection of objects to identify ones that are close to the query object [5].

Study [6] proposes three business process model similarity metrics: (i) “node matching” metric based on labels comparison, (ii) “structural similarity” metric based on topology comparison, and (iii) “behavioral similarity” based on causal relations [6].

In paper [7], authors mention that comparing business process models is a complex problem, performed mostly



**Research Article:** This article was published by the publishing house of ***NTU "KhPI"*** in the collection "Bulletin of the National Technical University "KhPI" Series: System analysis, management and information technologies." This article is distributed under a Creative Common Creative Common Attribution (CC BY 4.0). **Conflict of Interest:** The author/s declared no conflict of interest.

manually. This is why the authors propose an approach to measure the semantic similarity between business process models in an automated manner [7].

The authors of [8] have identified a linear search of similar business process models when a query model is compared to each model in a collection as inefficient and computationally complex. Hence, this paper suggests a fast similarity search algorithm based on the comparison of business process model features [8].

In paper [9] authors propose the "behavioral-based" comparison of business process models based on the causal footprints captured formally using Petri nets and informally using Event-driven Process Chain (EPC) notation [9].

A previous study in this field proposes the business process model similarity metric based on the graph structural characteristics [10]. It allows comparing models described using different notations and standards and takes into account not only workflow elements but also the secondary objects given in business process models (e.g., organizational units, data objects, information systems, etc.) [10].

According to reviewed studies ([5]; [6]; [7]; [8]; [9]; [10]), business process models were mostly compared by their structure or behavior, but the label comparison is the less elaborated approach. Due to the lack of studies that measure similarity between business process models using NLP (Natural Language Processing) techniques, this study should propose the respective approach to bridge the gap.

**Research objective and tasks.** The objective of this paper is the improvement of a search process for similar business process models within and across organizational collections and reference process libraries.

Therefore, the following tasks are considered:

- to propose the NLP-based algorithm to measure business process model similarity using activity labels given in business process models;
- to use the proposed algorithm to compare business process models given in different notations;
- to use the proposed algorithm compare textual descriptions to business process models.

However, compared business process models should be machine-readable, e.g. based on eXtensible Markup Language (XML), JavaScript Object Notation (JSON), or Yet Another Markup Language (YAML) formats.

**NLP-based similarity measurement of business process models.** The proposed algorithm for NLP-based similarity measurement of business process models includes the following steps:

- take two BPMN business process models $A$ and $B$ as the input and extract activity labels from these two models to obtain two collections $L_A$ and $L_B$:

$$\begin{aligned} L_A &= \left\{ l_{Ai}, i = \overline{1,n} \right\}, \\ L_B &= \left\{ l_{Bj}, j = \overline{1,m} \right\}, \end{aligned} \tag{1}$$

where $l_{Ai}$ – the $i$-th activity label extracted from the business process model $A$;

$l_{Bj}$ – the $j$-th activity label extracted from the business process model $B$;

$n$ – is the number of activity labels extracted from the business process model $A$;

$m$ – is the number of activity labels extracted from the business process model $B$;

- split labels (1) into words and change obtained words to lower case – two sets of words $W_A$ and $W_B$ will be obtained as the result:

$$\begin{aligned} W_A &= \left\{ w_{Ai}, i = \overline{1,p} \right\}, \\ W_B &= \left\{ w_{Bj}, j = \overline{1,q} \right\}, \end{aligned} \tag{2}$$

where $w_{Ai}$ – the $i$-th word extracted from labels of the business process model $A$;

$w_{Bj}$ – the $j$-th word extracted from labels of the business process model $B$;

$p$ – is the number of words extracted from labels of the business process model $A$;

$q$ – is the number of words extracted from labels of the business process model $B$;

- remove non-alphabetic and stop words from the previously obtained sets of words (2) to get cleansed sets of words $C_A$ and $C_B$ respectively:

$$\begin{aligned} C_A &= \left\{ c_{Ai}, i = \overline{1,r} \right\}, \\ C_B &= \left\{ c_{Bj}, j = \overline{1,s} \right\}, \end{aligned} \tag{3}$$

where $c_{Ai}$ – the $i$-th meaningful word that describes the business process model $A$;

$c_{Bj}$ – the $j$-th meaningful word that describes the business process model $B$;

$p$ – is the number of meaningful words that describe the business process model $A$;

$q$ – is the number of meaningful words that describe the business process model $B$;

- stem remaining words (3) to finally obtain sets of words $U_A$ and $U_B$:

$$\begin{aligned} U_A &= \left\{ u_{Ai}, i = \overline{1,x} \right\}, \\ U_B &= \left\{ u_{Bj}, j = \overline{1,y} \right\}, \end{aligned} \tag{4}$$

where $u_{Ai}$ – the $i$-th stemmed word that describes the business process model $A$;

$u_{Bj}$ – the $j$-th stemmed word that describes the business process model $B$;

$x$ – is the number of stemmed words that describe the business process model $A$;

$y$ – is the number of stemmed words that describe the business process model $B$;

- measure similarity between these two sets of words $U_A$ and $U_B$ (4) using Jaccard $J(U_A, U_B)$, Sorensen – Dice $SDC(U_A, U_B)$, overlap $overlap(U_A, U_B)$, and simple matching $SMC(U_A, U_B)$ coefficients [11]:

$$J(U_A,U_B)=\frac{|U_A\cap U_B|}{|U_A\cup U_B|}, \quad (5)$$

$$SDC(U_A,U_B)=\frac{2|U_A\cap U_B|}{|U_A|+|U_B|}, \quad (6)$$

$$overlap(U_A,U_B)=\frac{2|U_A\cap U_B|}{\min\{|U_A|,|U_B|\}}, \quad (7)$$

$$SMC(U_A,U_B)=\frac{|U_A\cap U_B|}{|U_A|+|U_B|}. \quad (8)$$

This algorithm is the improved algorithm for semantic quality analysis of business process models proposed earlier [12]. Whereas the earlier proposed algorithm was supposed to measure the closeness of a business process model to the textual description of a real business process, now the elaborated algorithm considers the comparison of business process models based on the semantic closeness of their activities (5) – (8).

An alternative algorithm takes a textual business process description $A$ and a BPMN model $B$ if the business process model that should be used as the search query does not exist. In this case, activity labels should be extracted only from the BPMN model, while the given textual description could be immediately split into a set of words turned into the lower case style (4).

The flowchart of the proposed algorithm is given below in fig. 1.

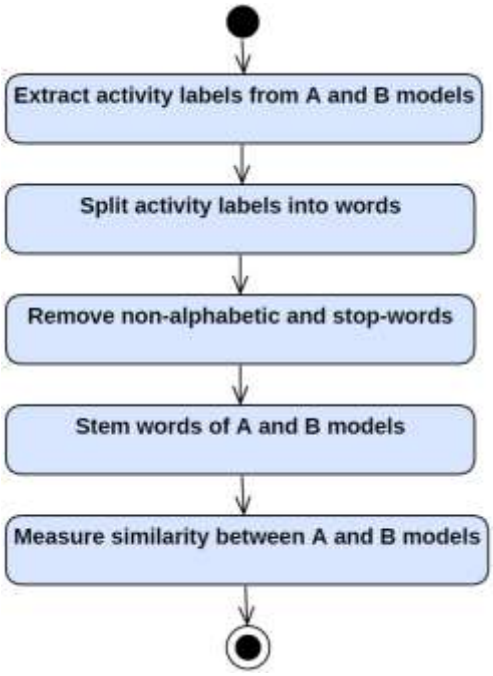


Fig. 1. The algorithm for NLP-based similarity measurement of business process models

Now it is necessary to verify the proposed algorithm. It can be implemented as a software component using the Python programming language [13] and Natural Language Toolkit (NLTK) [14]. The NLTK software platform is a leading solution for building Python programs that handle natural language processing [14].

**Experimental usage of the proposed algorithm.** In this section we demonstrate the experimental usage of the proposed algorithm (fig. 1).

The products supply process [4] according to the SCOR model involves scheduling product deliveries with the supplier, receiving the products at a specified location, verifying the goods to ensure they meet requirements, transferring the goods to the appropriate storage location, and finally authorizing payment to the supplier after the goods have been successfully delivered and inspected.

The BPMN model of described supply process [4] is demonstrated in fig. 2.

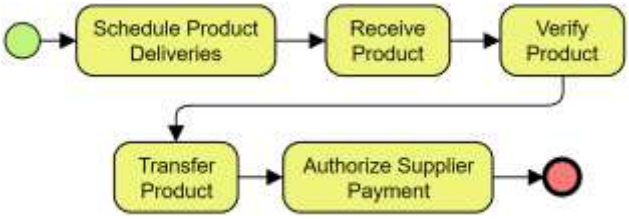


Fig. 2. The supply SCOR business process model [4]

The products delivery process [4] according to the SCOR model includes such steps as receiving, entering, and verifying an order, reserving inventory and setting a delivery date, preparing the products for delivery and loading them onto a vehicle, shipping the products, and invoicing the customer.

The BPMN model of described delivery process [4] is demonstrated in fig. 3.

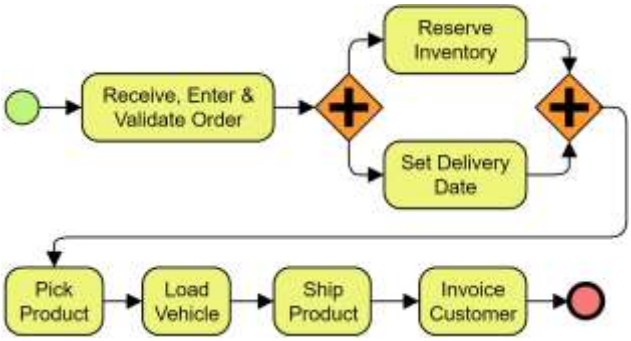


Fig. 3. The delivery SCOR business process model [4]

Two example models of supply (fig. 2) and delivery (fig. 3) business processes may be evaluated to illustrate the efficiency of the proposed algorithm in determining the similarity between such business process models.

The algorithm can construct and compare their relative textual descriptions by importing the activity labels from BPMN files that correspond these two business process models, providing for a more comprehensive understanding of their similarities and differences.

This data may be further used to discover areas for improvement and to enhance delivery and supply chain procedures for increased efficiency and profitability if comparing real company workflows toward reference models, such as SCOR [4] or APQC's PCF [3].

The original activity labels and obtained sets of words for supply and delivery business process models offered by the SCOR model [4] are demonstrated in Table 1.

Table 1 – The sets of words obtained for supply and delivery SCOR business process models

| Business process activity labels | Sets of words (4) |
|---|---|
| Schedule Product Deliveries<br>Receive Product<br>Verify Product<br>Transfer Product<br>Authorize Supplier Payment | schedul, product, deliveri, receiv, verifi, transfer, author, supplier, payment |
| Receive, Enter & Validate Order<br>Reserve Inventory & Set Delivery Date<br>Pick Product<br>Load Vehicle<br>Ship Product<br>Invoice Customer | receiv, enter, valid, order, reserv, inventori, set, deliveri, date, pick, product, load, vehicl, ship, invoic, custom |

The similarity measurement between the two sets of words demonstrated in Table 1 using coefficients (5) – (8) and the proposed algorithm (fig. 1) allowed us to obtain the following values:

- 0.14 using the Jaccard coefficient;
- 0.24 using the Sorencen – Dice coefficient;
- 0.33 using the overlap coefficient;
- 0.14 using the simple matching coefficient.

The comparison histogram of business process model similarities calculated using coefficients (5) – (8) is shown in fig. 4.

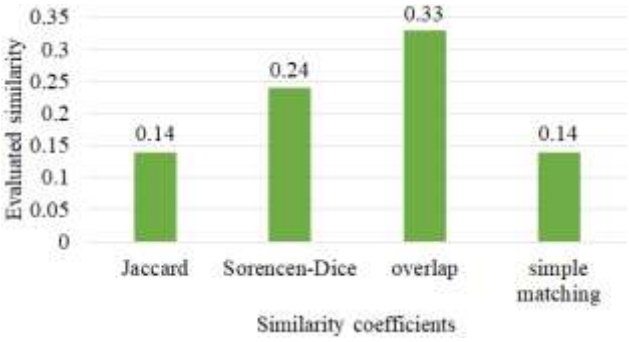


Fig. 4. The comparison of business process model similarities calculated using coefficients (5) – (8)

According to the obtained results, Jaccard and simple matching coefficients demonstrate equal similarity values of 0.14, since in this particular case the mutual absence of words in $U_A$ and $U_B$ is impossible. Sorensen – Dice and overlap coefficients show relatively low similarity values of 0.24 and 0.33 respectively for the given business process activity labels. Thus, further study in this direction may omit the simple matching coefficient, while focusing on the remaining ones.

**The software tool for similarity measurement in a database of business process models.** The software tool that implements the proposed algorithm and a database (DB) of business process (BP) models is demonstrated in fig. 5. The software tool, which implements the proposed algorithm, is a Python command-line application [13] that uses NLTK [14] and MySQL Connector [15]. Hence, the database of business process models is a relational schema created using MySQL database management system [16].

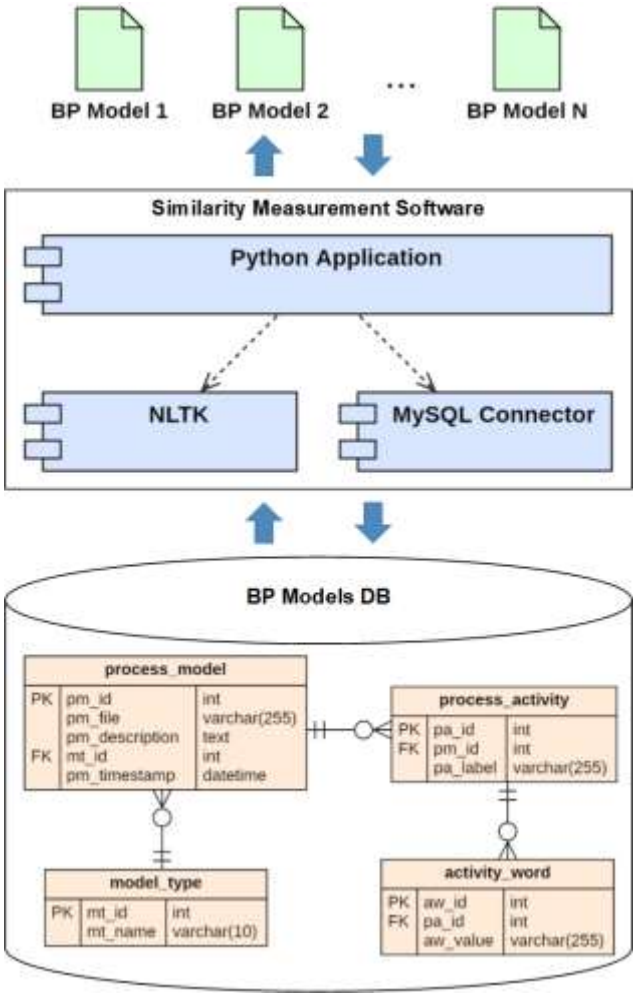


Fig. 5. Business process models database and software components that implement the proposed algorithm

According to fig. 5, the database of business process models stores text descriptions (i.e. the "pm_description" attribute) built from activity labels, activity label values (i.e. the "pa_label" attribute), and words extracted from activity labels (i.e. the "aw_value" attribute) for similarity measurement according to the proposed algorithm (fig. 1).

The software component "Python Application" that implements the proposed algorithm (fig. 1) uses Jaccard coefficient (5) for similarity measurement.

**Experimental usage of the software tool.** First of all, we have create the view "test_similarity" for querying the DB of BP models. The first query, which uses SQL LIKE clause, is shown in fig. 6.

```
sql = r"""SELECT
            DISTINCT file_name
        FROM
            test_similarity
        WHERE
            description LIKE
            '%invite logistic company%'"""
```

Fig. 6. The search query using SQL LIKE clause

The second query (fig. 7) uses prepared words for BP models according to the proposed algorithm (fig. 1) and the stemmed search words (we are using the Porter stemming algorithm [17]).

```
sql = """SELECT
            file_name,
            COUNT(file_name) / 3 AS similarity
        FROM
            test_similarity
        WHERE
            word IN (%s, %s, %s)
        GROUP BY
            file_name
        HAVING
            similarity = 1
        ORDER BY
            similarity DESC"""

val = ("invite", "logistic", "company")
val = tuple([porter_stemmer.stem(word)
                for word in val])
```

Fig. 7. The search query using the proposed algorithm

According to fig. 7, the similarity degree is calculated using Jaccard index (5) and then used to filter only 100% matches. Both queries in fig. 6 and 7 were looking for the BP models containing “invite logistic company” activities or similar.

The comparison of search results is given in Table 2.

Table 2 – The comparison of search results

| Similarity search | Seconds (average time) | Found BP models |
|---|---|---|
| SQL LIKE | 0.0124 | 2 (3%) |
| Proposed algorithm | 0.0085 | 14 (21%) |

According to Table 2, the proposed algorithm is 31% faster than SQL LIKE clause (we measured the average time of 10 executions), while allowing to find 18% more similar BP models (14 against 2, from a collection of 68 models of a goods dispatch process [18]).

These results are compared visually in fig. 8.

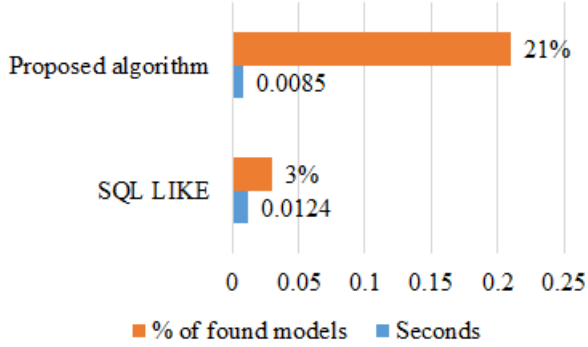


Fig. 8. The histogram of querying performance comparison

**Conclusions.** In this study, the algorithm for NLP-based similarity measurement of business process models is proposed.

1. The proposed algorithm (fig. 1) is based on the natural language processing techniques (such as tokenization, stemming, stop words elimination) used to compare textual activity labels of business process models or business process descriptions to measure their similarity.
2. This algorithm was verified using supply (fig. 2) and delivery (fig. 3) BPMN process models based on the SCOR reference model. Obtained results (fig. 4) demonstrate Jaccard and simple matching coefficients give the same values when comparing two sets with the impossible mutual absence of elements. Thus it is proposed to used Jaccard or simple matching coefficient in the further search for similar business process models.
3. The proposed algorithm is implemented using Python and NLTK to measure similarity in the database of BP models created using MySQL. Experimental results demonstrate that proposed algorithm is 31% faster than the SQL LIKE clause and allows to find 18% more similar BP models than the SQL LIKE clause.
4. However, the limitation of the proposed approach is the necessity of preliminary process of business process models to apply the proposed algorithm. This requires processing of large collections of BP models, which may require significant computing resources. Nevertheless, such a pre-processing of BP models takes place much less frequently than search for similar business process models.

Future work includes elaboration of computational techniques and software solutions for efficient similarity search in large collections of BP models.

**References**


1. Geiger M. et al. *BPMN 2.0: The state of support and implementation*. URL: https://doi.org/10.1016/j.future.2017.01.006 (access date: 01.04.2023).
2. Fettke P. et al. *Business Process Reference Models: Survey and Classification*. URL: https://doi.org/10.1007/11678564_44 (access date: 01.04.2023).
3. *APQC Process Classification Framework*. URL: https://www.signavio.com/reference-models/apqc-framework/ (access date: 01.04.2023).
4. *SCOR Model*. URL: https://scor.ascm.org/ (access date: 01.04.2023).
5. Dumas M. et al. *Similarity Search of Business Process Models*. URL: http://sites.computer.org/debull/A09sept/marlon.pdf (access date: 02.04.2023).
6. Dijkman R. *Similarity of business process models: Metrics and evaluation*. URL: https://doi.org/10.1016/j.is.2010.09.006 (access date: 02.04.2023).
7. Humm B. G., Fengel J. *Semantics-Based Business Process Model Similarity*. URL: https://doi.org/10.1007/978-3-642-30359-3_4 (access date: 02.04.2023).
8. Yan Z., Dijkman R. *Fast business process similarity search*. URL: https://doi.org/10.1007/s10619-012-7089-z (access date: 02.04.2023).
9. van Dongen B. et al. *Measuring Similarity between Business Process Models*. URL: https://doi.org/10.1007/978-3-540-69534-9_34 (access date: 02.04.2023).
10. Kopp A. M., Orlovskyi D. L. *Estimation and analysis of business process models similarity in enterprise continuum repository*. URL: https://doi.org/10.20535/SRIT.2308-8893.2018.4.04 (access date: 02.04.2023).
11. Verma V., Aggarwal R. K. *A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: empirical and theoretical perspective*. URL: https://doi.org/10.1007/s13278-020-00660-9 (access date: 04.04.2023).

12. Kopp A., Orlovskyi D. *The approach and the software tool to calculate semantic quality measures of business process models*. URL: http://dx.doi.org/10.20998/2079-0023.2022.02.12 (access date: 04.04.2023).
13. *Python*. URL: https://www.python.org/ (access date: 06.04.2023).
14. *NLTK*. URL: https://www.nltk.org/ (access date: 06.04.2023).
15. *MySQL Connector/Python Developer Guide*. URL: https://dev.mysql.com/doc/connector-python/en/ (access date: 06.04.2023).
16. *MySQL*. URL: https://www.mysql.com/ (access date: 07.04.2023).
17. *Porter Stemmer*. URL: https://tartarus.org/martin/PorterStemmer/ (access date: 08.04.2023).
18. *BPMN for research*. URL: https://github.com/camunda/bpmn-for-research (access date: 10.04.2023).

**References (transliterated)**

1. Geiger M. et al. *BPMN 2.0: The state of support and implementation*. Available at: https://doi.org/10.1016/j.future.2017.01.006 (accessed 01.04.2023).
2. Fettke P. et al. *Business Process Reference Models: Survey and Classification*. Available at: https://doi.org/10.1007/11678564_44 (accessed 01.04.2023).
3. *APQC Process Classification Framework*. Available at: https://www.signavio.com/reference-models/apqc-framework/ (accessed 01.04.2023).
4. *SCOR Model*. Available at: https://scor.ascm.org/ (accessed 01.04.2023).
5. Dumas M. et al. *Similarity Search of Business Process Models*. Available at: http://sites.computer.org/debull/A09sept/marlon.pdf (accessed 02.04.2023).
6. Dijkman R. *Similarity of business process models: Metrics and evaluation*. Available at: https://doi.org/10.1016/j.is.2010.09.006 (accessed 02.04.2023).
7. Humm B. G., Fengel J. *Semantics-Based Business Process Model Similarity*. Available at: https://doi.org/10.1007/978-3-642-30359-3_4 (accessed 02.04.2023).
8. Yan Z., Dijkman R. *Fast business process similarity search*. Available at: https://doi.org/10.1007/s10619-012-7089-z (accessed 02.04.2023).
9. van Dongen B. et al. *Measuring Similarity between Business Process Models*. Available at: https://doi.org/10.1007/978-3-540-69534-9_34 (accessed 02.04.2023).
10. Kopp A. M., Orlovskyi D. L. *Estimation and analysis of business process models similarity in enterprise continuum repository*. Available at: https://doi.org/10.20535/SRIT.2308-8893.2018.4.04 (accessed 02.04.2023).
11. Verma V., Aggarwal R. K. *A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: empirical and theoretical perspective*. Available at: https://doi.org/10.1007/s13278-020-00660-9 (accessed 04.04.2023).
12. Kopp A., Orlovskyi D. *The approach and the software tool to calculate semantic quality measures of business process models*. Available at: http://dx.doi.org/10.20998/2079-0023.2022.02.12 (accessed 04.04.2023).
13. *Python*. Available at: https://www.python.org/ (accessed 06.04.2023).
14. *NLTK*. Available at: https://www.nltk.org/ (accessed 06.04.2023).
15. *MySQL Connector/Python Developer Guide*. Available at: https://dev.mysql.com/doc/connector-python/en/ (accessed 06.04.2023).
16. *MySQL*. Available at: https://www.mysql.com/ (accessed 07.04.2023).
17. *Porter Stemmer*. Available at: https://tartarus.org/martin/PorterStemmer/ (accessed 08.04.2023).
18. *BPMN for research*. Available at: https://github.com/camunda/bpmn-for-research (accessed 10.04.2023).

*Received 05.05.2023*

УДК 004.94

***А. М. КОПП***, доктор філософії (PhD), доцент, Національний технічний університет «Харківський політехнічний інститут», доцент кафедри програмної інженерії та інтелектуальних технологій управління, м. Харків, Україна, e-mail: kopp93@gmail.com, ORCID: https://orcid.org/0000-0002-3189-5623
***Д. Л. ОРЛОВСЬКИЙ***, кандидат технічних наук (PhD), доцент, Національний технічний університет «Харківський політехнічний інститут», доцент кафедри програмної інженерії та інтелектуальних технологій управління, м. Харків, Україна, e-mail: orlovskyi.dm@gmail.com, ORCID: https://orcid.org/0000-0002-8261-2988

# АЛГОРИТМ ВИМІРЮВАННЯ ПОДІБНОСТІ МІТОК ДІЯЛЬНОСТЕЙ НА ОСНОВІ NLP У БАЗІ ДАНИХ МОДЕЛЕЙ БІЗНЕС-ПРОЦЕСІВ

Моделювання бізнес-процесів є важливою частиною організаційного управління, оскільки дозволяє компаніям отримати уявлення про свої операційні бізнес-процеси та знайти можливості для розвитку. Однак оцінити та кількісно виміряти схожість декількох моделей бізнес-процесів може бути складно, оскільки ці моделі часто сильно відрізняються за структурою та номенклатурою. Це дослідження пропонує підхід, який використовує методи обробки природної мови для оцінки схожості моделей бізнес-процесів, для розв'язку цієї задачі. Алгоритм використовує мітки діяльностей, наведені в моделях бізнес-процесів, як вхідні дані для створення текстових описів пов'язаних бізнес-процесів. Алгоритм включає декілька етапів попередньої обробки, щоб гарантувати, що текстові описи є коректними і послідовними. Спочатку з отриманих текстових описів вилучаються окремі слова і представляються у нижньому регістрі. Після цього з отриманих слів видаляються всі нелітерні та стоп-слова. Потім слова, що залишилися, піддаються стемінгу, тобто приведенню до їхньої базової форми. Після підготовки та попередньої обробки текстових описів алгоритм оцінює схожість різних моделей бізнес-процесів за допомогою мір схожості, включаючи коефіцієнти Жаккара, Соренсена – Дайса, перетину та простого коефіцієнту відповідності. Ці метрики забезпечують більш детальне розуміння подібності і відмінності між різними моделями бізнес-процесів, які потім можуть бути використані для впливу на прийняття рішень та ініціатив щодо вдосконалення бізнес-процесів. Програмна реалізація запропонованого алгоритму демонструє його використання для вимірювання подібності в базі даних моделей бізнес-процесів. Експерименти демонструють, що розроблений алгоритм є на 31 % швидшим за пошук на основі виразу SQL LIKE, а також дозволяє знайти на 18 % більше подібних моделей у базі даних моделей бізнес-процесів.

**Ключові слова:** модель бізнес-процесу, база даних моделей бізнес-процесів, обробка природної мови, алгоритм вимірювання подібності, мітки діяльностей, програмна реалізація алгоритму.

*Повні імена авторів / Author's full names*

**Автор 1 / Author 1:** Копп Андрій Михайлович, Kopp Andrii Mykhailovych
**Автор 2 / Author 2:** Орловський Дмитро Леонідович, Orlovskyi Dmytro Leonidovych