Towards Structuring of Electronic Marketplaces Contents: Items Normalization Technology

Olga Cherednichenko[1[0000-0002-9391-5220]], Olha Yanholenko[1[0000-0001-7755-1255]], Maryna Vovk [1[0000-0003-4119-5441]], Nataliia Sharonova[1[0000-0002-8161-552X]]

[1] National Technical University "Kharkiv Polytechnic Institute", 2, Kyrpychova str., 61002 Kharkiv, Ukraine
olha.cherednichenko@gmail.com, olga.yan26@gmail.com, marihavovk@gmail.com, nvsharonova@ukr.net

Abstract. The E-commerce industry is going strong and is bringing a great profit to its stakeholders. However, there is probably no buyer of the e-marketplace who has not faced the issues connected with inappropriate search results or inadequate filtering and recommendation of irrelevant products. Modern search and collaborative filtering algorithms of e-commerce systems do work well with the input data of high quality but the reality is that often items' description contains inaccuracies and incompleteness, which negatively affects the results. The given paper suggests the concept of e-marketplace items normalization which goal is to provide the unified and standardized patterns of items inside the system that can be used by search and filtering algorithms. Items normalization is implemented based on the algebra of predicates models specified in this work. The case study deals with constructing normalized models of knapsacks items from the online sports store. The developed models allowed to build 141 normalized item patterns with a unified set of attributes and their values.

Keywords: E-commerce Marketplace; Item Normalization; Item Attributes; Natural Language Processing; Predicate; Reference model.

1 Introduction

E-commerce positions in the global economy keep on strengthening. This is confirmed by the constant growth of the world online retail sales which increased by 15% in 2019 compared to 2018 [1]. The share of the world online sales in the total retail sales has also increased by 1% [1]. All the forecasts predict the future growth of these indicators. To be successful and to attract more clients, e-marketplaces have to support their buyers in the best possible way. This support should include efficient tools of product search, filtering, representation and comparison which will make the purchase process easy and comfortable. As the number of sellers and items being sold on the e-marketplaces is growing, the volume of data stored and processed by e-commerce information systems is increasing drastically. In this context, two situations can be considered. Firstly, in the case of global e-marketplaces that serve as a platform where a seller and a buyer meet each other, users can create multiple offers of the same product on the seller side. Thus,

a single real-world object can be presented in different ways in the offers of one or many sellers. Secondly, in the case of e-shop belonging to a single company that supposedly does not contain duplicate items of a single product, still there is a risk of having an incomplete and inaccurate description of the product. In both cases the arbitrary form of the item description stored by the e-commerce system sophisticates the processing of this data. This leads to negative buyers' experience due to bad search results.

To improve the quality of the data that is used as an input by filtering, clustering and other algorithms of the e-commerce systems it is suggested to develop a formalized model of item's description which will allow avoiding possible ambiguities and inaccuracies in its representation [2, 3]. The given study suggests calling this process as the item's normalization. Its goal is to represent the item in a unified way so that item's attributes with their values could be matched with the pattern view of the given type of product. Having the pattern model of a product, it will be easy to correct errors and fill in missed values reducing the degree of incompleteness of the initial data.

The rest of the paper is organized in the following way. Section 2 substantiates the problem statement and provides the general scheme of items normalization. Section 3 reviews the research in the given field. The reference model of items normalization are given in section 4. A case study of normalization of items of the sports online store is presented in section 5. Results of the experiment and conclusions are discussed in Sections 6 and 7 respectively.

2 Problem Statement

In the given paper the process of creating a full, accurate and unified form of the e-marketplace item is called normalization. Item normalization can be decomposed into several levels. Let's denote the set of items as I. Each item $i \in I$ is characterized by the set of attributes $X = (x_1, x_2, \dots, x_n)$, where n is the number of attributes. Each attribute takes values x_i^j, where $j = (1,2, \dots, m)$. On the lowest level of normalization, it is necessary to switch attribute's values x_i^j to the unified view. If an attribute is Weight, for example, then the normalized value would be the number complemented by the unit of measurement (e.g., 500 g). On the middle level of normalization, the ambiguity of attributes' names should be reached. For this purpose, it is necessary to conduct a semantic analysis and to substitute synonymous names with a single unified one. For example, if the item's attribute is called "name", "brand", "title", then one of the values should be selected as a uniform. On the highest level of normalization, the item description should be complemented with the missed values of attributes based on the data available from the quality sources.

Normalized representation of an item should be stored by the e-commerce system and used while performing its basic functions. The normalization process is aimed at: 1) creating a normalized item's model from data gathered from the item's description on the web site and 2) complementing this model with the missed attributes and their values, thus getting a full and unified item's representation. The detailed flow of actions that should be performed during normalization is shown on Fig. 1.

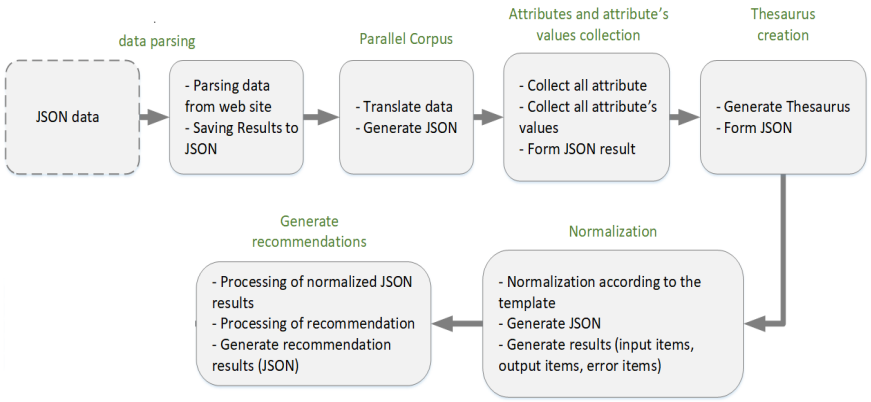

Fig. 1. Process of items normalization

So the goal of this paper is to improve search, filtering and other procedures of the e-commerce systems by means of items normalization based on mathematical models of the algebra of predicates. Normalized items are the unified internal representation of the products and are internally used by e-commerce algorithms.

3 Related Works

Big volumes of information that need to be gathered, processed and stored in the e-commerce area caused the intensive development of data mining methods. Electronic marketplaces with their infinite number of items have already been a subject of research for the paper authors [4, 5]. And we have the intention to follow up on our previous researches. Grouping similar products on the trading platforms according to their descriptions is studied in [4]. In order to study item similarity, researches [5] try to analyze item descriptions on e-commerce markets and it is found out that the k-means algorithm works well only for uniformly distributed data by categories, but this is not suitable for the segmentation of heterogeneous descriptions.

In the paper [6], it is explored how natural language processing methods can help to check contradictions in facts. The authors proposed an approach based on factual information systematization. As a result, it is proposed to use predicate algebra to create a model of searching and extracting factual data [7]. In the time when the size of databases increases, the complexity of the matching process becomes one of the major challenges for record normalization. Different indexing techniques have been developed for record normalization and deduplication [8, 9]. Such a problem belongs to the tasks of record linkage. Researchers [10, 20] solve this issue using a learning algorithm. The authors in the work [11] have developed a framework for solving the task of product record normalization. Paper [12] is devoted to studying and analyzing the problem of record normalization over a set of matching records.

The study [13] demonstrates a duplicate detection method for bio-informatics databases. The papers [14, 15, 16] explored a set of normalization techniques to achieve

better translation quality. Researchers in [17] suggest the flexible query-time record linkage and fusion framework. In the paper [18] authors described the rule-based method for deduplicating article records across databases and include an open-source script module that can be deployed freely.

Thus, we can conclude that a lot of authors worked on normalization on trading platforms and in other domains. Different approaches were developed. The study shows that there is substantial room for additional research on this topic. Our task is to research how the normalization of product description dimensions can be solved in order to provide complete information for a buyer on e-commerce marketplaces.

4 Reference Model of Items Normalization

The Intelligence Theory task is to designate the natural information processes that take place in human thinking. The Intelligence Theory assists logical mathematics, which covers the wider scope of questions [19]. It has such sections, which have not yet been used by informatization. The first stage of formalization of human intelligent processes is the construction of a thesaurus. Thesaurus contains words of the language that are used for normalization of both attributes' titles and their values. In information retrieval thesauruses, lexical units of text are replaced by descriptors. The general scheme of item's normalized view is shown in Fig. 2.

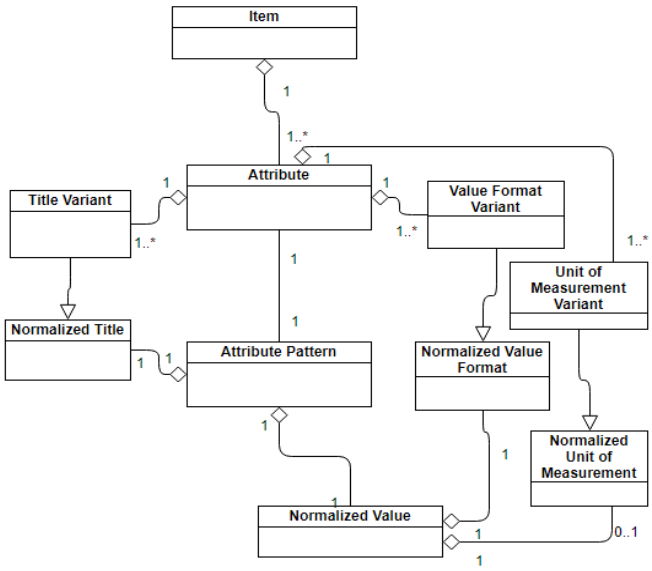

Fig. 2. Items normalization reference model

Figure 3 shows a data flow diagram (DFD) that shows total data flows when solving a normalization task.

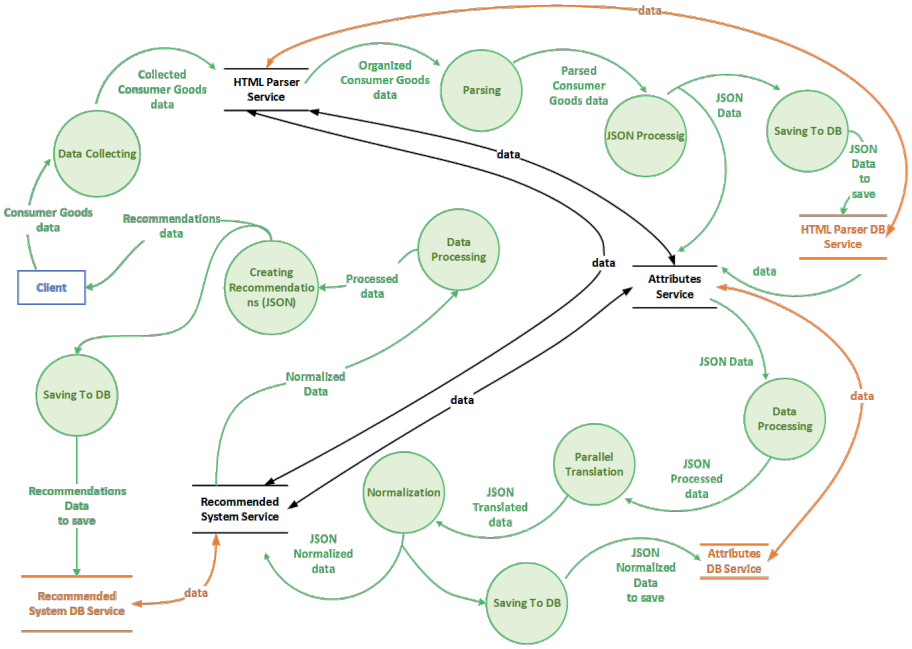

Fig. 3. Data flow diagram

The main notion of logical mathematics is a mathematical relation. A logical network accomplishes different operations on relationships. Relations show the attribute connections of the objects. Relations are general instruments for the object description. In order to demonstrate relationships, people use natural human language. Communicating with people, we express to them the sense of the sentence, which is an attitude. Defined relations can symbolize some notions. Each artifact and process of the outworld can be represented by relationships. We unrestricted select some non-empty set U and call its elements as objects. The set U as such is called the universe of objects. It can be either finite or infinite.

We suggest a model that is built on the comparator identification method. this method gives the opportunity for data and the template matching. The relation between the words and their location in the text are the main points of the approach. This method performs the process of extraction in that way as a human do it [19].

5 Case Study

The case study of the given work is based on the data of the online store Hervis Sports (https://www.hervis.at/store) that is specialized in sports clothes and equipment. The store belongs to a single company. The website of this e-shop is in German. The web crawler component launched on the website has gathered all web pages that contain knapsacks being sold. The number of items at the moment of the experiment is – 141. Let's introduce $Y = (y_1, y_2, \ldots, y_{141})$ objects of the real world.

Since there is a single seller (web site owner) in this e-commerce system, each knapsack model is present once on the site. So there are no duplicates of the same product on the site. However, the way of representing the same type of product (in our case - knapsack) differs from item to item. The example of the two knapsack item pages is shown on fig. 4.

Fig. 4. Items description (A- Deuter, B - Vaude)

From the preliminary analysis of the collected items, we can see that the description of knapsacks contains different attributes (Title, Technology/Material, Equipment, Volume, Dimensions, Weight, Load Range, etc.). Knapsack A has Weight attribute and doesn't have Load Range attribute while knapsack B does have it. Therefore, the description of items may contain different sets of attributes.

Additionally, the values of attributes are presented in a different way. Although Volume is commonly measured in liters, for example, knapsack A has Volume value followed by "Liter" and knapsack B – followed by "l". Among the collected items there are other variations of liter designation, like "L", "liter", "litre". Similarly, Weight attribute has values complemented with different units of measurement ("kg", "g", "G", "KG"). Dimensions attribute may have different forms of value representation as shown in Fig. 2 and its units of measurement are different as well ("cm", "mm"). Moreover, an attribute itself may have different names across items. For instance, Dimensions attribute has the following names: "Maße", "Dimension", "Abmessung", "Größe", "Grösse", "Maßen". The whole list of possible attributes' names extracted by the web crawler with their example values is given in Fig. 5.

Table 1 contains all 24 variants of attributes' names and their English translation since the normalized item's model is going to have its values in English. After normalizing attributes' names we have got 17 unique attributes $X = (x_1, x_2, \dots, x_{17})$ introduced.

```
{"Brand": "Kohla Zugspitze 26",
 "Price": "€ 69,99",
 "Technologie/ Material": "Surround Ventilationssystem",
 "Ausstattung": "Stretch- EInschubtasche an der Front, 2
Deckeltaschen, inkl. Regenhülle mit Reflektoren, 2 seit-
liche Trinkflaschenhalterungen, Hüft- und Brustgurt mit
Seitentasche und Fingerriemen",
 "Sonstiges": "Stocklhalterung",
 "Lastbereich": "0 - 4 kg",
 "Maße": "43 x 22 x 16 cm",
 "Volumen": "9,0 l",
 "Gewicht": "740 g",
 "Rückensystem": "MOTION V Frame™ Rückensystem, 2-Lagen
EVA-Rückenpolster, Rückenlänge: L (48,5 cm)",
 "Funktion": "Trinksystem kompatibel",
 "Ausstattug": "abnehmbare Kompressionsriemen, Deck-
eltasche, verstaubare Befestigungsschlaufen für Eispickel
oder Trekkingstöcke",
 "Material": "Dynajin 210, 30% Polyester / 70% Polyamid",
 "Dimension": "40 x 13 x 17 cm",
 "Technologie/Material": "Removable Airbag System 3.0",
 "Hinweis": "Kartusche ist nicht im Lieferumfang enthal-
ten",
 "Abmessung": "28 x 24 x 15 cm",
 "Gewich": "2,26 kg",
 "Füllung": "Stickstoff (nur Werkbefüllung möglich)",
 "Arbeitsdruck": "300 bar",
 "Größe": "75 x 36 x 30 cm",
 "Austattung": "Raincover für den ganzen Rucksack, easy
handle Zipper, hochwertige Qualitäts-Zipper von SBS",
 "Abmessungen": "500x142x280mm",
 "Liter": "30L",
 "Volumen/Gewicht": "30L / 1930g",
 "Grösse": "43 / 24 / 19 (H x B x T) cm",
 "Maßen": "45x31x25cm"
}
```


Fig. 5. Attributes' names

Table 1. Matching of German and English attributes' manes

German (DE)	English (EN)	Number of occurrences (DE, EN)	Normalized name of attribute (EN)	Designation
Marke	Brand	141	Brand	x_1
Preis	Price	141	Price	x_2
Technologie_Material	Technology_Material	106	Technology_Material	x_3
Ausstattung	Equipments	120	Equipments	x_4
Sonstiges	Other	94	Other	x_5
Lastbereich	LoadRange	2	LoadRange	x_6
Maße	Dimensions	52	Size	x_7
Volumen	Volume	101	Volume	x_8
Gewicht	Weight	76	Weight	x_9
Rückensystem	BackSystem	12	BackSystem	x_{10}
Funktion	Function	59	Function	x_{11}
Material	Material	30	Material	x_{12}
Dimension	Dimension	3	*Size*	x_7
Hinweis	Note	3	Note	x_{13}
Abmessung	Dimension	9	*Size*	x_7
Gewich	Weight	1	Weight	x_{14}
Füllung	Filling	1	Filling	x_{15}
Arbeitsdruck	WorkingPressure	1	WorkingPressure	x_{16}
Größe	Size	8	Size	x_7
Abmessungen	Dimensions	1	*Size*	x_7
Liter	Liter	1	*Volume*	x_8
Volumen_Gewicht	Volume_Weight	1	Volume_Weight	x_{17}
Grösse	Size	1	Size	x_7
Maßen	Size	1	Size	x_7

All these examples of different description of the same attributes/values/units of measurement allow concluding that information about the products in this e-commerce system is stored in a non-unified form. This leads to an inadequate work of search and filtering algorithms of the system. For example, if the knapsack was added to the system with the Volume equal to "9 Litres" and the system is able to process only items with Volume values ended by "L", then this specific knapsack will never be displayed in the

filtering results for all 9-liter knapsacks. Thus, to perform properly the system requires a normalized description of all items which will provide adequate and accurate results of search, filtering, and comparison.

From the other point of view, if a product doesn't contain Volume value at all, it does not mean that it does not have it. It was just missed while adding the item to the system. In this case, such particular knapsack also does not have many chances to be shown in the search results. Having a normalized form of such item will allow to define the missed values and to complement them with the information from the patterns. In the role of a pattern, we can consider official documents about the product, its quality certificates and specifications, description from official sites of the manufacturers, etc.

Assigning available values to attributes $X = (x_1, x_2, \dots, x_{17})$, we can define each item in a unique normalized way. For example, attribute x_1 can take values x_1^1="2117", x_1^2="ABS", x_1^3="APTEM", x_1^4="BCA", x_1^5="Babolat", x_1^6="Black Crevice", x_1^7="Deuter", x_1^8="Dynafit", x_1^9="Kilimanjaro", x_1^{10}="Kohla", x_1^{11}="Mammut", x_1^{12}="Salomon", x_1^{13}="Vaude", x_1^{14}="Wheel Bee". Attribute x_8 can take values x_8^1="≤10L", x_8^2=">10L and ≤20L", x_8^3=">20L and ≤30L", x_8^4=">30L and ≤50L", x_8^5=">50L and ≤70L", x_8^6=">70L". Having assigned all values to all attributes, it is possible to build the relation $L(X, Y)$ and define it unambiguously for each of 141 items. Normalization of items requires constructions of relations:

$$L(x_1, x_2, \dots, x_{17}, y_1) = 1,$$
$$L(x_1, x_2, \dots, x_{17}, y_2) = 1,$$
$$\dots$$
$$L(x_1, x_2, \dots, x_{17}, y_{141}) = 1.$$

The normalization of attributes' values was performed based on the comparator identification of the input values and units of measurement. For example, the comparator function for defining attribute units of measurement looks like:

$$f(a)=\begin{cases} \text{L, if } E(a,L) \vee E(a,l) \vee E(a,Litre) \vee E(a,litre) \vee E(a,Liter) \vee E(a,liter), \\ \text{kg, if } E(a,kg) \vee E(a,Kg) \vee E(a,K\ ), \\ \dots \\ \text{cm, if } E(a,cm) \vee E(a,Cm) \vee E(a,CM), \end{cases}$$

where E is a predicate of equivalence (identification) that defines one of the possible values of units of measurement entered to the system.

The results of normalization of Size attribute is shown on Fig. 6.

6 Discussion

As a result of the given research, we developed a reference model in order to give items descriptions from e-commerce marketplaces in the way of formal representation. The predicate representation of goods characteristics allows using any natural language for filing in items description by the seller. Thus, the seller is less obliged to be strict in the form of an item attribute description. The developed approach gives the opportunity to solve the issue of normalization in commodity designation. The given findings are the

basis of a two-layer information system. One layer presents how the product features are shown for a customer and the second layer of how the internal system sees them.

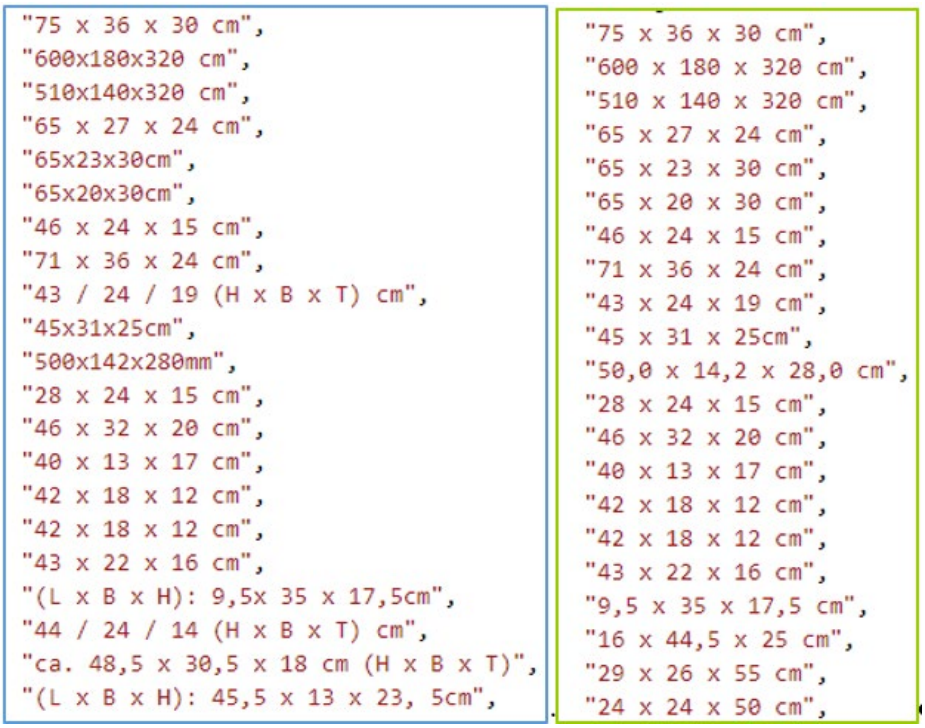

Fig. 6. Size attribute values normalization

7 Conclusions and Future Work

The main idea of the given research is that collaborative filtering, items search and matching processes of e-commerce business work well if the data they are dealing with is full and precise. But in the real world, the description of products on the e-marketplaces is far from the ideal. Thus, buyers may see irrelevant searching results while looking for some products. To improve this situation, the given work introduces the notion of items normalization as a process of constructing complete and accurate patterns of items being sold. Normalized items are treated as the high-quality input data for internal algorithms of e-commerce systems.

The presented models of items normalization allow: 1) to form the set of unique attributes of items; 2) translate attributes' values to a unified form; 3) build a relation between an item and attributes that uniquely defines a real-world product. The developed models were tested on the experimental set of knapsacks from the online sports store. The case study represents the results of attributes and their values normalization.

As a future direction of this research, it is planned to evaluate the performance of searching algorithms taking as an input row items' description and normalized patterns. Also the presented findings can be used for further development of items matching models. And finally, it would be interesting to explore the use of normalized items in the problem of e-marketplace localization.

8 References

1. How High Will E-Commerce Sales Go? http://www.cbre.us/real-estate-services/real-estate-industries/omnichannel/the-definitive-guide-to-omnichannel-real-estate/by-the-numbers/how-high-will-e-commerce-sales-go
2. Razia Sulthana, A., Ramasamy, S.: Ontology and context based recommendation system using Neuro-Fuzzy Classification. Computers & Electrical Engineering February (2018).
3. Ya, L. The Comparison of Personalization Recommendation for E-Commerce. International Conference on Solid State Devices and Materials Science, Physics Procedia 25, pp. 475-478 (2012).
4. Cherednichenko, O., Vovk, M., Kanishcheva, O., Godlevskyi, O.: Towards Improving the Search Quality on the Trading Platforms. In: S.Wrycza, J. Maslankowski(Eds): 11th SIGSAND/PLAIS 2018, LNBIP 333. pp. 21-30. Springer (2018).
5. Cherednichenko, O., Vovk, M., Kanishcheva, O., Godlevskyi, O.: Studying Items Similarity for Dependable Buying on Electronic Marketplaces. Proc. 2nd Int. Conf. On Computational Linguistics and Intelligent Systems (COLINS), Volume I: Main Conference CEUR-WS. Vol. 2136. pp.78-89. Lviv, Ukraine, (2018).
6. Sharonova, N., Doroshenko, A., Cherednichenko, O.: Issues of Fact-based Information Analysis. Proc. 2nd Int. Conf. On Computational Linguistics and Intelligent Systems (COLINS), Volume I: Main Conference CEUR-WS. Vol. 2136. pp. 11-19. Lviv, Ukraine, (2018).
7. Bondarenko, M. F., Shabanov-Kushnarenko, U. P.: Theory of intelligence: a Handbook SMIT Company, Kharkiv (2006).
8. Christen, P. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. IEEE Transactions on Knowledge and Data Engineering, 24(9), pp. 1537–1555. (2012).
9. Lusetti, M. Ruzsics, T., Gohring, A.: Encoder-Decoder Methods for Text Normalization. Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 18–28 Santa Fe, New Mexico, USA (2018).
10. Bilenko, M., Basu, S., & Sahami, M. (n.d.).: Adaptive Product Normalization: Using Online Learning for Record Linkage in Comparison Shopping. Fifth IEEE International Conference on Data Mining (2005).
11. Tak-Lam Wong, An Unsupervised Approach for Product Record Normalization across Different Web Sites. Proceedings of the 23rd national conference on Artificial intelligence - Volume 2, pp. 1249–1254 (2008).
12. Dong, Y., Dragut, E. C., & Meng, W.: Normalization of Duplicate Records from Multiple Sources. IEEE Transactions on Knowledge and Data Engineering. (2018).
13. Chen, Q., Zobel, J., Verspoor, K.: Evaluation of a Machine Learning Duplicate Detection Method for Bioinformatics Databases. Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics - DTMBIO '15. (2015).

14. Banerjee, P., Kumar Naskar, S., Roturier, J., Way A., Josef van Genabith. Domain Adaptation in SMT of User-Generated Forum Content Guided by OOV Word Reduction: Normalization and/or Supplementary Data? European Association for Machine Translation. (2012).
15. Clark, E., & Araki, K.: Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. Procedia - Social and Behavioral Sciences, 27, pp. 2–11. (2011).
16. Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Botsis, T.: Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of Biomedical Informatics, 73, pp. 14–29. (2017).
17. Rezig, E. K., Dragut, E. C., Ouzzani, M., Elmagarmid, A. K., & Aref, W. G.: ORLF: A flexible framework for online record linkage and fusion. 2016 IEEE 32nd International Conference on Data Engineering (2016).
18. Jiang, Y., Lin, C., Meng, W., Yu, C., Cohen, A. M., & Smalheiser, N. R.: Rule-based deduplication of article records from bibliographic databases. Database, (2014).
19. Bondarenko M. F., Shabanov-Kushnarenko U. P.: Brain-like structures: A reference book Naukova dumka, Kyiv (2011).
20. Vysotska, V., Burov, Y., Lytvyn, V., Oleshek, O.: Automated Monitoring of Changes in Web Resources. In: Advances in Intelligent Systems and Computing, 1020, pp.348–363. (2020).