УДК 004.8:004.9 **doi: 10.26906/SUNZ.2023.4.120**

С. Ф. Чалий, В. О. Лещинський

Харківський національний університет радіоелектроніки, Харків, Україна

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОЦІНКИ ПОЯСНЕНЬ В ІНТЕЛЕКТУАЛЬНІЙ ІНФОРМАЦІЙНІЙ СИСТЕМІ

Анотація. **Предметом** вивчення в статті є процеси побудови пояснень щодо процесів отримання рішення та отриманих результатів в інтелектуальній інформаційній системі. **Метою** розробка технології оцінки пояснень з урахуванням як чутливості цих пояснень до відмінностей у вхідних даних, так і можливостей використання пояснень користувачем згідно концепції використання рішення інтелектуальної системи.. **Завдання**: структуризація задач побудови пояснень у аспекті оцінки тлумачень; структуризація показників оцінки пояснень з урахуванням залежностей між цими показниками; розробка послідовності етапів інформаційної технології комплексного оцінювання пояснень в інтелектуальній системі.. Використовуваними **підходами** є: методи побудови пояснень, методи та підходи до оцінки пояснень в системах штучного інтелекту. Отримані наступні **результати**. Виконано структуризацію задач побудови пояснень з урахуванняv оцінки отриманих тлумачень. Виконано структуризацію показників оцінки пояснень з урахуванням обмежень доступу до процесу прийняття рішень в інтелектуальній системі. Показано, що в залежності від доступності даних про процес прийняття рішення в інтелектуальній системі доцільно використовувати показники точності або коректності. Показник чутливості дає можливість оцінити пояснення при категоризації знань щодо властивостей об'єктів або вхідних даних. Показник простоти визначає вплив кількості вхідних змінних на пояснення. **Висновки**. Наукова новизна отриманих результатів полягає в наступному. Запропоновано інформаційну технологію оцінки пояснень в інтелектуальній інформаційній системі. Технологія містить послідовність етапів із розрахунку показників чутливості, коректності та простоти пояснення, а також відбору підмножини пояснень за цими показниками з використанням взаємозалежностей між ними та можливості обмежень по показнику коректності. В практичному плані запропонована технологія створює умови для підбору пояснень за їх чутливістю та простотою для користувача з урахуванням особливостей вхідних даних та процесу використання рішення.

Ключові слова: інтелектуальна система, пояснення, процес прийняття рішення, причинно-наслідковий зв'язок, оцінка пояснень.

Вступ

Пояснення в інтелектуальних системах дають можливість користувачеві переконатись у коректності отриманих рішень[1, 2], оскільки задають прості причинно-наслідкові залежності між вхідними даними та отриманим результатом [3].

Дослідження в сфері психології пізнання показують, що людина зазвичай має отримати обґрунтування нових знань. Такі обґрунтування часто надаються у формі пояснень [4- 6].

Актуальність і важливість використання пояснень пов'язана із широкими застосуванням в сучасних інтелектуальних системах алгоритмів машинного навчання. На результати роботи таких алгоритмів суттєво впливають упередженість та викиди в даних. Відповідно, отримані рішення можуть не відповідати потребам користувача, що на практиці приводить до неефективного використання результатів системи штучного інтелекту і, як наслідок, до матеріальних збитків. Інший аспект роботи інтелектуальних систем пов'язаний із юридичними обмеженнями, що не дозволяють розкривати алгоритми їх роботи кінцевому користувачеві. Такі обмеження знижують довіру користувача до пропозицій системи штучного інтелекту і, знову ж, обмежують використання цих рішень [7].

Напрямок розробки пояснень в інтелектуальних системах інтенсивно розвивається останні 15 років. В перших роботах головна увага приділялась визначенню показників задоволеності клієнтів рекомендаційних систем [8]. Однак обчислення таких показників є достатньо трудомістким, оскільки потребує проведення інтерв'ю із клієнтами.

Швидкий розвиток методів побудови пояснень відбувся в останні роки в рамках «Explainable Artificial intelligence», що була розроблена DARPA [9].

Однак існуючі підходи до оцінки пояснень орієнтовані в першу чергу на перевірку каузальних та темпоральних зав'язків між вхідними даними та результатом інтелектуальної системи [10-13]. Розробці комплексної технології, яка була б орієнтована на вибір простих пояснень з перевіркою їх коректності не приділялось достатньо уваги.

Зазначене свідчить про актуальність задачі розробки інформаційної технології оцінки пояснень.

Метою статті є розробка технології оцінки пояснень з урахуванням як чутливості цих пояснень до відмінностей у вхідних даних, так і можливостей використання пояснень користувачем згідно концепції використання рішення інтелектуальної системи.

Для досягнення поставленої мети вирішуються такі задачі:

– структуризація задач побудови пояснень у аспекті оцінки тлумачень;

– структуризація показників оцінки пояснень з урахуванням залежностей між цими показниками;

– розробка послідовності етапів інформаційної технології комплексного оцінювання пояснень в інтелектуальній системі.

Показники оцінки пояснень в інтелектуальній системі

Задача оцінки пояснень вирішується в рамках комплексу задач побудови та застосування пояснень в інтелектуальних системах. Вона використовує в якості вхідних даних дані та результати задач

розуміння психології пояснень та розробки методів побудови пояснень. Загальна схема, що відображає зв'язок між цими задачами та передачу інформації для оцінки пояснень, представлена на рис. 1.

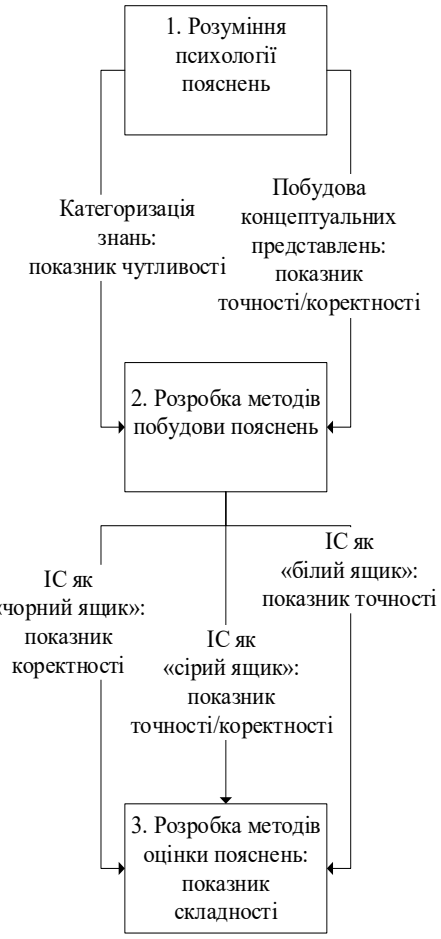

Рис. 1. Схема взаємодії задач формування та оцінки пояснень

Зальна послідовність формування пояснення передбачає виявлення психологічних характеристик користувача інтелектуальної системи. При вирішенні задачі розуміння психології пояснень використовуються теорії пізнання і формується відповідна ментальна модель, яка відображає процес пізнання людини [4-6]. Підтвердження виявлених закономірностей зазвичай реалізується через експерименти з навчанням дітей та наданням пояснень у процесі цього навчання.

Ментальна модель процесу пізнання дає можливість обґрунтувати такі аспекти пояснень в інтелектуальній інформаційній системі:

– набір показників, що використовується для оцінки пояснень;

– можливості та спосіб використання відомих властивостей вхідних даних для оцінки пояснень;

– узагальнену послідовність проведення оцінки пояснень в інтелектуальній інформаційній системі, яка враховувала б зв'язок між вибраними показникам та їх властивостями.

Виконаний аналіз психологічних теорій пояснення дає можливість зробити наступний висновок. Пояснення, яке буде зрозумілим для користувача інтелектуальної системи, має відповідати двом альтернативним шляхам пізнання [6]:

– пізнання шляхом побудови концепцій щодо стану та динаміки предметної області, в якій відбувається навчання;

– пояснення на основі категоризації знань щодо властивостей об'єктів предметної області; такі властивості можуть бути статичними та динамічними.

В першому випадку пояснення має забезпечити узгодження знань користувача із рішенням інтелектуальної інформаційної системи.

Таке узгодження може виконуватись за бінарною та числовою оцінкою. Числова оцінка представлена точністю пояснень. Точність визначає відхилення залежностей у поясненні від залежностей, пов'язаних із процесом прийняття та використання рішення.

З позицій власника інтелектуальної системи, точність визначається через відхилення залежностей пояснення та процесу отримання результату.

З точки зору користувача точність має відображати відповідність пояснення способу використання рішення інтелектуальної системи. Останнє ж залежить від знань про предметну область. І проблема оцінки точності полягає в тому, що фонові знання людини містить неявну складову.

Загалом сукупність знань користувача інтелектуальної системи доцільно розділити на явні (explicit), неявні описові (implicit) і неявні поведінкові (tacit) знання (рис. 2).

Явні знання можуть бути безпосередньо відображені у вигляді причинно-наслідкових, темпоральних та аналогічних залежностей. Тому явні знання можуть бути безпосередньо використані для побудови пояснень і також для оцінки точності пояснень з точки зору користувача.

Неявні описові знання можуть бути використані двома способами.

По-перше, вони можуть бути трансформовані в явні методами традиційної інженерії знань. Така екстерналізація знань виконується у процесі бесіди інженера знань із експертами в предметній галузі і широко використовується, наприклад, при побудові баз знань в експертних системах.

По-друге, неявні описові знання можуть бути отримані в результаті аналізу логів (журналів подій) інтелектуальної інформаційної системи. Лог зазвичай містить множину послідовностей подій, яка відображає процеси, що відбувається в інформаційній системі, а також процеси взаємодії користувача з цією системою. Кожен процес складається із виконаної послідовності дій. І кожна така дія представлена подією логу. Відповідно, при багаторазовому виконанні процесу прийняття рішень в інтелектуальній системі лог містить записи про альтернативні процеси прийняття рішень. Порівняльний аналіз цих процесів дає можливість виявити неявні знання.

Неявні поведінкові знання зазвичай складаються з навичок виконання окремих дій чи послідовностей дій у процесах вирішення задач у предметній галузі. Ці знання не вербалізуються внаслідок їх практичної направленості і тому не можуть бути екстерналізовані традиційними методами інженерії знань.

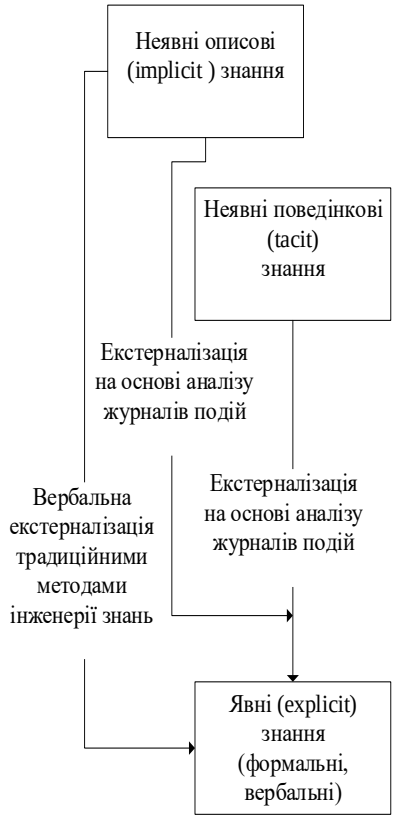

Рис. 2. Екстерналізація (перетворення в явну форму) неявних знань для оцінки пояснень в інтелектуальній системі

Для їх вилучення використовуються лише результати моніторингу процесів в інтелектуальній системі, представлені у вигляді журналів подій. Таким чином, неявні знання обох видів можуть бути отримані на основі аналізу логів, отриманих в результаті моніторингу процесів в інтелектуальній системі та процесі взаємодії з користувачем. Однак останній вид процесів не завжди може бути представлений в логах, що створює проблему з використання числової оцінки пояснень. У такому випадку доцільно використати бінарну оцінку пояснень виду: коректне/некоректне тлумачення. Дана оцінка може бути отримана з використання пар (вхідні дані, результат).

Узагальнену порівняльну оцінка показників точності та коректності пояснень наведено у табл.1.

Таким чином, при побудові пояснень на основі формування концепцій щодо предметної області, в якій використовуються результати інтелектуальної системи, доцільно використовувати оцінку «коректне/некоректне пояснення». Остання визначає відповідність рішення вхідним даним згідно процесу функціонування інтелектуальної системи (з позиції власника системи) або ж відповідність результату, отриманого за визначених вхідних даних знанням користувача щодо способів використання результату у визначеній предметній області. У другому випадку, при побудові пояснення на основі категоризації знань щодо об'єктів предметної області, необхідно, щоб пояснення було схожим для об'єктів однієї категорії і відрізнялось для об'єктів іншої категорії. Властивості об'єктів в інтелектуальні інформаційній системі задаються через набори вхідних даних. Тому в якості оцінки пояснення доцільно використовувати чутливість тлумачення до властивостей об'єктів. Така чутливість відображає ступінь відмінності рішення системи в залежності від відмінності значень вхідних даних. Відхилення у значеннях вхідних даних можна задати або безпосередньо через відмінності значень цих даних, або через відмінності їх ваг [14].

Таблиця 1 – **Порівняльна оцінка властивостей та обмежень показників точності та коректності пояснень**

Показник	Властивості	Обмеження та сфера застосування
Точність	1)Числова оцінка; 2) З позицій власника інтелектуальної системи: відповідність/відхилення залежностей пояснення та процесу отримання результату; 3) З точки зору користувача: відповідність пояснення способу використання рішення інтелектуальної системи.	1) Має бути наявним журнал подій інтелектуальної системи, що відображають процес прийняття рішення; 2) Наявність логів (журналів подій) взаємодії з користувачем; 3) Застосування переважно при категоризації знань.
Коректність	1)Бінарна оцінка; 2) З позицій власника системи – відповідність пояснення парі (вхідні дані, результат); 3) З позицій користувача системи – відповідність пояснення концепції використання у предметній області.	1)Наявність вхідних даних та результату інтелектуальної системи; Застосування переважно при побудові концептуальних представлень у предметній області

Перша ситуація може бути використана при числових значення вхідних даних. Тоді ці значення нормуються, а чутливість $S_j^{(i)}$ між $Expl_i$ та $Expl_j$ розраховується як відмінність співвідношень між різницею вхідних значень ΔV та відповідних рішень ΔR інтелектуальної системи за умови, що ці пояснення є близькими або співпадають:

$$S_j^{(i)} = \left| \frac{\Delta V_i \Delta R_j - \Delta V_j \Delta R_i}{\Delta R_i \Delta R_j} \right|, \quad iff\ Expl_i \approx Expl_j. \quad (1)$$

Якщо ж рішення співпадає в кількісному плані або є якісним чи бінарним (наприклад, погодження на виконання певної дії), то оцінка чутливості пояснення має вигляд:

$$S_j^{(i)} = \left| \Delta V_i - \Delta V_j \right| iff\ Expl_i = Expl_j. \quad (2)$$

Вирази (1) та (2) важко використати у випадку, якщо вхідні дані є якісними, наприклад, мають перелічуваний тип. Тобто дані мають кінцевий набір значень із заданої множини. Такі вхідні дані досить часто зустрічаються, наприклад, в рекомендаційних системах. Зокрема, при виборі комп'ютера може відбуватись порівняння між процесорами i3, i5, i7, i9, тощо. В даному випадку доцільно враховувати вагу значень. Наприклад, значення змінної «процесор», що дорівнює «i7», може бути компромісним між продуктивністю та ціною і може мати більшу вагу. Значення цієї ж змінної «i9» або «i3» - меншу внаслідок більшої ціни або меншої потужності відповідно. Тоді оцінка чутливості пояснення в рекомендаційній системі виду «ми запропонували Вам дану модель ноутбука, тому що вона використовує модель процесора …» базується на визначенні різниці ваг для пар значень змінних виду «i9» і «i7» або «i7» і «i3». Тобто

чим більше різниця у ціні або продуктивності, тим більшою має бути і чутливість пояснення.

З іншої сторони, при порівнянні процесорів з аналогічними можливостями від різних виробників чутливість має бути меншою для того, щоб пояснення відповідало потребам користувача.

Якщо для пояснення використовується декілька змінних, то ми маємо порівнювати їх сумарну вагу. За умови схожого результату оцінка чутливості пояснення має вигляд:

$$S_j^{(i)} = \left| \sum_i w_i - \sum_j w_j \right| iff\ Expl_i = Expl_j. \qquad (3)$$

Для того, щоб визначити межі чутливості, пропонується використовувати показник коректності. Тобто ми маємо досягти максимальної чутливості $S_j^{(i)}$ за умови коректності пояснень C_i та C_j. Відповідно, задач побудови пояснень з використанням їх оцінки на основі показників чутливості та коректності має вигляд:

$$\max\left(S_j^{(i)}\right) \left| C_i, C_j = true. \right. \qquad (4)$$

Таким чином, при оцінці пояснення з урахуванням результатів вирішення задачі розуміння психології пояснень доцільно використовувати комбінацію оцінок коректності та чутливості.

Друга задача – побудова пояснення має суттєві обмеження на обчислювальну складність. Дослідження в рамках програми XAI показали, що користувач має витрачати мінімум часу на розуміння пояснення. Інакше суттєво знижується ефективність використання результату роботи інтелектуальної системи. З урахуванням у загальному випадку відсутності повного доступу до процесу прийняття рішення в інтелектуальній системі складність пропонується оцінювати за кількістю вхідних змінних, що використані для побудови пояснення. Головна ідея даної оцінки полягає в тому, щоб визначити ключові вхідні змінні, значення яких дають можливість побудувати пояснення. Тоді задача зниження складності пояснення базується на визначенні ваг вхідних змінних (на відміну від чутливості, де можуть використовуватись ваги значень змінних у випадку перелічуваного типу даних). Таким чином, задача оцінювання пояснень в інтелектуальній системі полягає у послідовному визначенні коректності, чутливості та складності пояснення. Зазначимо, що в даному випадку ми розглядаємо коректність, а не точність, оскільки перша оцінка може бути використана навіть у випадку представлення системи як «чорний ящик».

Технологія оцінки пояснень

Послідовність етапів запропонованої інформаційної технології оцінки пояснень представлено на рис. 3.

При оцінці пояснень згідно запропонованої технології послідовно вирішується ряд підзадач, що враховують показники чутливості, коректності та простоти тлумачення.

1) Оцінка чутливості пояснень згідно виразів (1)-(3). Результатом даного етапу є множина можливих пояснень з різною чутливістю.

2) Оцінка коректності пояснення з урахуванням ваг значень вхідних змінних. Ваги значень вхідних даних можуть бути отримані такими способами: експертне оцінювання; статистична оцінка; машинне навчання.

Другий спосіб дає усереднене значення, а третій – забезпечує визначення ваг в залежності від категорії вхідних даних. Наприклад, визначення ваг конкретних моделей процесорів в рекомендаційній системі може бути отримано усередненням кількості покупок комп'ютерів з цими моделями процесорів. Однак при використанні машинного навчання ми можемо додатково коригувати ваги в залежності від регіону продаж, знань про користувача, тощо.

За результатами оцінки виконується відбір коректних пояснень згідно умови (4).

3) Оцінка складності пояснень виконується для множини коректних тлумачень, що були отримані в результаті виконання етапу 3.

4) Упорядкування пояснень виконується спочатку за оцінкою чутливості, а потім за оцінкою складності.

Такий підхід до упорядкування пов'язаний із тим, що більш чутливе (але коректне) пояснення відображає відмінності умов отримання рішення. В подальшому із пояснень зі схожою чутливістю можна вибрати більш просте.

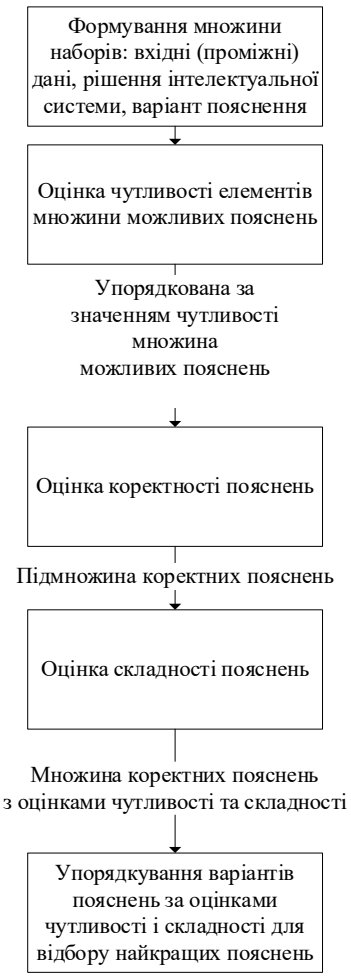

Рис. 3. Послідовність етапів технології оцінки пояснень з використанням показників чутливості, коректності та складності

Висновки

Виконано структуризацію задач побудови пояснень та зав'язків між цими задачами у аспекті оцінки отриманих тлумачень. Структуровано показники оцінки пояснень. Показано, що в залежності від підходу до вирішення задачі психології пояснень для оцінки останніх можна використовувати показники точності/ коректності та чутливості.

Показник точності є числовим і визначає ступінь відповідності пояснення і отриманого в інтелектуальній системі рішення. Другий показник є бінарним і може бути використаним як обмеження у процесі оцінки пояснень. Показник точності пояснення потребує додаткової інформації щодо процесу прийняття рішення в інтелектуальній системі. Тому його використання є обмеженим при представленні системи у вигляді чорного ящику. Показник чутливості задає оцінку пояснення через категоризацію знань щодо властивостей об'єктів. Показник простоти визначає вплив кількості вхідних змінних на пояснення. Використання даного показника дає можливість спростити пояснення для користувача. Запропоновано інформаційну технологію оцінки пояснень в інтелектуальній системі. Дана технологія передбачає послідовний розрахунок показників чутливості, коректності та простоти пояснення, а також відбір підмножини пояснень за цими показниками з використанням показника коректності в якості обмеження. В практичному плані запропонована технологія створює умови для підбору пояснення з урахуванням особливостей вхідних даних та процесу використання рішення, що відображає потреби типового користувача інтелектуальної інформаційної системи.

СПИСОК ЛІТЕРАТУРИ

1. Castelvecchi D. (2016), “Can we open the black box of AI?” Nature, Vol. 538 (7623), pp. 20-23.
2. Adadi, A., Berrada, M. (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138– 52160.
3. Чалий С. Ф., Лещинський В. О., Лещинська І. О. (2021) Контрфактуальна темпоральна модель причинно-наслідкових зв'язків для побудови пояснень в інтелектуальних системах,/ Вісник НТУ "ХПІ".– Харків : НТУ "ХПІ", № 2 (6), С. 41-46.
4. Chi, M., de Leeuw, N., Chiu, M., & LaVancher, C. Eliciting self-explanations improves understanding. Cognitive Science. 1994. Vol.18. P. 439–477.
5. Carey, S. The origin of concepts. New York, NY: Oxford University Press. 2009. 608 p.
6. Чалий, С., and Лещинська, І. (2023). Концептуальна ментальна модель пояснення в системі штучного інтелекту. *Вісник НТУ «ХПІ».* (1 (9), 70–75. https://doi.org/10.20998/2079-0023.2023.01.11
7. Miller T. (2019), “Explanation in artificial intelligence: Insights from the social sciences”, Artificial Intelligence, vol. 267, pp.1-38, DOI: https://doi.org/10.1016/j.artint.2018.07.007
8. Tintarev N., Masthoff J. (2012), “Evaluating the effectiveness of explanations for recommender systems”, User Model User-Adap Inter., Vol. 22, pp. 399– 439, https://doi.org/10.1007/s11257-011-9117-5.
9. Gunning i D. Aha, (2019) “DARPA’s Explainable Artificial Intelligence (XAI) Program”, AI Magazine, Vol. 40(2), pp.44-58, doi: 10.1609/ aimag.v40i2.2850.
10. Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, Phil Blunsom.(2019) Can I trust the explainer? Verifying post-hoc explanatory methods. arXiv:1910.02065.
11. M. Yang and B. Kim (2019). BIM: Towards quantitative evaluation of interpretability methodswith ground truth. arXiv:1907.09701
12. Chalyi, S., Leshchynskyi, V. (2020). Method of constructing explanations for recommender systems based on the temporal dynamics of user preferences. EUREKA: Physics and Engineering, 3, 43-50. doi: 10.21303/2461-4262.2020.001228.
13. Чалий С. Ф. Реляційно-темпоральна модель набору сутностей предметної області для процесу формування рішення в інтелектуальній інформаційній системі / С. Ф. Чалий, В. О. Лещинський, І. О. Лещинська // Вісник НТУ "ХПІ. *Серія: Системний аналіз, управління та інформаційні технології* – Харків : НТУ "ХПІ", 2022. – № 1 (7). – С. 84-89.
14. Chalyi, Sergii & Leshchynskyi, V.. (2023). Оцінка чутливості пояснень в інтелектуальній інформаційній системі. Системи управління, навігації та зв’язку. Збірник наукових праць. 2. 165-169. 10.26906/SUNZ.2023.2.165.

Received (Надійшла) 21.09.2023
Accepted for publication (Прийнята до друку) 15.11.2023

Information technology for evaluating explanations in an intelligent information system

Serhii Chalyi, Volodymyr Leshchynskyi

Abstract. The **article’s subject matter** is the process of constructing explanations regarding the decision-making process and the obtained results in the intellectual information system. The goal is to develop a technology for evaluating explanations, taking into account both the sensitivity of these explanations to differences in input data, and the possibility of using explanations by the user according to the concept of using an intelligent system solution. **Tasks**: structuring the tasks of constructing explanations in the aspect of evaluating interpretations; structuring of indicators of evaluation of explanations considering the dependencies between these indicators; development of a sequence of stages of information technology for comprehensive evaluation of explanations in an intelligent system. **The approaches** used are: methods of construction of explanations, methods and approaches to evaluation of explanations in artificial intelligence systems. The following **results** were obtained. The structuring of the tasks of constructing explanations taking into account the evaluation of the received interpretations was carried out. The structuring of the evaluation indicators of the explanations was carried out, taking into account the limitations of access to the decision-making process in the intelligent system. It is shown that, depending on the availability of data on the decision-making process in the intelligent system, it is advisable to use indicators of accuracy or correctness. The sensitivity indicator makes it possible to evaluate the explanation when categorizing knowledge about the properties of objects or input data. The simplicity index determines the effect of the number of input variables on the explanation. Conclusions. The scientific **novelty** of the obtained results is as follows. An information technology for evaluating explanations in an intelligent information system is proposed. The technology includes a sequence of stages for calculating indicators of sensitivity, correctness and simplicity of explanation, as well as selecting a subset of explanations based on these indicators using interdependencies between them and the possibility of restrictions on the indicator of correctness. In practical terms, the proposed technology creates conditions for the selection of explanations according to their sensitivity and simplicity for the user, taking into account the peculiarities of the input data and the process of using the solution.

Keywords: intellectual system, explanation, decision-making process, causal relationship, evaluation of explanations.