

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук _____
(повна назва)

Кафедра _____ Інформаційних управляючих систем _____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____ другий (магістерський) _____

Дослідження методів попередньої обробки даних у сільськогосподарських
інформаційних системах _____
(тема)

Виконав:
студент 2 курсу, групи ІУСТМ-19-1 _____
Орлов В.Г. _____
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки _____
(код і повна назва спеціальності)

Тип програми освітньо-професійна _____
(освітньо-професійна або освітньо-наукова)

Освітня програма Інформаційні управляючі системи та технології _____
(повна назва освітньої програми)

Керівник проф. Чалий С.Ф. _____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____
(підпис)

Петров К.Е. _____
(прізвище, ініціали)

2020р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Інформаційних управляючих систем
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Інформаційні управляючі системи та технології
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)

« ____ » _____ 20 ____ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Орлову Владиславу Геннадійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів попередньої обробки даних у сільськогосподарських інформаційних системах

затверджена наказом по університету від 27 листопада 2020 р. № 1455 Ст _____

2. Термін подання студентом роботи до екзаменаційної комісії _____ 2020 р.

3. Вихідні дані до роботи Науково-технічні публікації та інтернет джерела з тематики атестаційної роботи

4. Перелік питань, що потрібно опрацювати в роботі вступ, аналіз вимог до вхідних даних аналізу процесів, дослідження задач інтелектуальних процесів, перехід підприємства на процесне управління, аналіз вимог до вхідних даних аналізу процесів, дослідження методів структуризації вхідних даних у задачах process mining, конвертація вхідних даних у задачах інтелектуального аналізу процесів, постановка задачі атестаційної роботи, розробка методу підготовки даних у форматі хес, структуризація загального перетворення даних, розробка підсистеми проекту аналізу процесів, розробка проекту побудови процесних моделей на неструктурованих даних, статут проекту, планування проекту, висновки до розділу, впровадження та перевірка ефективності розробленого алгоритму, обґрунтування вибору платформи програмного забезпечення задачі, опис вимог до програмного забезпечення задачі, реалізація удосконаленого методу, експериментальна перевірка отриманих результатів, висновки.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Аналіз літератури та Інтернет-джерел	2.11.20-4.11.20	
2	Постановка задачі	5.11.20-6.11.20	
3	Обробка матеріалу	8.11.20-10.11.20	
4	Дослідження сучасного стану вирішення проблеми	11.11.20-13.11.20	
5	Дослідження моделей та методів попередньої обробки даних	14.11.20-20.11.20	
6	Дослідження особливостей реалізації методів попередньої обробки даних у сільськогосподарських інформаційних системах	21.11.20-25.11.20	
7	Експериментальна перевірка результатів дослідження	26.11.20-30.11.20	
8	Написання пояснювальної записки	1.11.20-5.11.20	
9	Підготовка презентації	6.12.20	
10	Надання роботи для перевірки на плагіат	7.12.20	
11	Надання роботи на підпис науковому керівникові	10.12.20	
12	Попередній захист	11.12.20	
13	Надання роботи на рецензію	12.12.20	
14	Надання роботи на підпис завідувачу кафедрою	14.12.20	
15	Надання підписаної завідувачем кафедрою роботи в ДЕК	16.12.20	
16	Захист	17.12.20	

Дата видачі завдання 02 листопада 2020 р.

Студент _____
(підпис)

Керівник роботи _____
(підпис)

проф. Чалий С.Ф.
(посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка до магістерської атестаційної роботи містить: 73 с., 4 розділи, 19 рис., 3 табл., 35 джерел.

МЕТОДИ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ, ІНФОРМАЦІЙНІ СИСТЕМИ СІЛЬСЬКОГО ГОСПОДАРСТВА, ФУНКЦІОНАЛЬНЕ УПРАВЛІННЯ, ПРОЦЕСНЕ УПРАВЛІННЯ, ЖУРНАЛ ПОДІЙ, КОНВЕРТАЦІЯ ВХІДНИХ ДАНИХ, СТАНДАРТ XES

У роботі виконано огляд методів попередньої обробки даних у сільськогосподарських інформаційних системах для побудови моделей процесів аграрних робіт. Проведено аналіз особливостей сільськогосподарських інформаційних систем. На підставі проведеного аналізу запропоновано удосконалення метод попередньої обробки вхідних даних для вирішення задачі інтелектуального аналізу при переході до процесного управління в сільськогосподарських системах, який зумовлює підвищення якості побудови процесних моделей на неструктурованих даних.

В ході дослідження отримані такі результати: визначені існуючі методи обробки даних у сільськогосподарських інформаційних системах; визначено методи Process mining; визначено особливості стандарту XES; розроблено удосконалений метод попередньої обробки вхідних даних при переході до процесного управління в сільськогосподарських системах; проведена експериментальна перевірка удосконаленого методу.

ABSTRACT

Explanatory Note to master certification work contains: 73 pages, 4 sections, 19 pictures, 3 tables, 35 sources.

DATA PROCESSING METHODS, AGRICULTURAL INFORMATION SYSTEMS, FUNCTIONAL MANAGEMENT, PROCESS MANAGEMENT, EVENT LOG, INPUT DATA CONVERSION, XES STANDARD.

The paper reviews methods of data preprocessing in Agricultural Information Systems for constructing models of agricultural work processes. The analysis of the features of Agricultural Information Systems is carried out. Based on the analysis, it is proposed to improve the method of preprocessing input data to solve the problem of intellectual analysis during the transition to process management in agricultural systems, which leads to an increase in the quality of building process models based on unstructured data.

In the course of the study, the following results were obtained: existing methods of data processing in agricultural information systems were determined; process mining methods were defined; features of the XES standard were defined; an improved method for preprocessing input data during the transition to process management in agricultural systems was developed; an experimental test of the improved method was carried out.

ЗМІСТ

Перелік скорочень, умовних позначень, символів, одиниць та термінів.....5

Вступ..... 6

.....

.....

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ10

1.1 Аналіз особливостей сільськогосподарських інформаційних систем.....10

1.2 Порівняльний аналіз функціонального та процесного підходів до управління.....19

1.3 Дослідження задач інтелектуального аналізу процесів.....23

1.4 Дослідження методів структуризації вхідних даних у задачах process mining.....25

1.5 Постановка задачі дослідження.....27

2 ДОСЛІДЖЕННЯ МЕТОДІВ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ У PROCESS MINING29

2.2 Перетворення вхідних даних у стандарт XES.....34

2.3 Удосконалений метод підготовки вхідних даних у форматі XES.....38

2.4 Конвертація вхідних даних при побудові процесної моделі.....39

3 ДОСЛІДЖЕННЯ ОТРИМАННЯ НАУКОВИХ РЕЗУЛЬТАТІВ.....42

3.1 Технологія підготовки вхідних даних.....42

3.2 Послідовність побудови процесної моделі44

Рисунок 3.2 – Алгоритм роботи.....45

4 ПРАКТИЧНЕ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ.....47

4.1 Програмна реалізація удосконаленого методу.....47

4.2 Експериментальна перевірка отриманих результатів.....53

ВИСНОВКИ.....54

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ.....55

ДОДАТОК А.....59

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ ТА ТЕРМІНІВ

АПК- агропромисловий комплекс;

ГІС- геоінформаційна система;

ІС – інформаційна система;

ProM- інструмент для вилучення та аналізу процесів;

SQL-мова програмування структурованих запитів;

XES- метод process mining.

ВСТУП

Сільське господарство-галузь господарства, спрямована на забезпечення населення продовольством (їжею) і отримання сировини для ряду галузей промисловості. Галузь є однією з найважливіших, представленої практично у всіх країнах.

До числа пріоритетних напрямків подальшого зростання і розвитку господарського виробництва, підвищення стійкості та ефективності діяльності підприємств належать розробка і впровадження прогресивних технологій, оптимізація використання ресурсів на основі накопиченого науково-технічного, виробничого, технологічного та організаційно-управлінського потенціалу, реалізація яких потребує збільшення капіталовкладень і залучення інвестицій в галузь.

У сільському господарстві слабкою ланкою залишається технологічна підготовка виробництва, у багатьох підприємствах відсутня чітка регламентація всього комплексу технологічних процесів, що призводить до всіляких зривів і відхилень від нормативів використання ресурсів, зниження якості виконуваних робіт і не дозволяє повною мірою контролювати хід виробництва.

Слабкий і недостатній контроль за виконанням процесів і недотримання встановлених технологічних вимог є причинами зниження показників виробничої діяльності і, в кінцевому підсумку, збитковості галузі.

Базою сучасних підходів до управління є процесний підхід, який передбачає визначення набору бізнес-процесів, що виконуються в організації, і подальшу роботу з ними.

Інтелектуальний аналіз даних - це процес отримання знань: він отримує дані з різних джерел даних і, нарешті, перетворює їх у знання, тим самим забезпечуючи розуміння своєї галузь застосування[1].

Мета попередньої обробки даних полягає в тому, щоб полегшити роботу з даними для моделей інтелектуального аналізу даних. Якість даних

може надавати істотний вплив на моделі інтелектуального аналізу даних. Це вважається що дані та об'єкти вже встановили верхню межу знань, які можуть бути отримані, і моделі інтелектуального аналізу даних тільки наближаються до верхньої межі.

Різні методи попередньої обробки винаходяться для того, щоб дані відповідали вхідним вимогам моделі, підвищували релевантність мети прогнозування та полегшували етап оптимізації модель. Як правило, необроблені дані, отримані з природного світу, мають погану форму. Ці проблеми включають в себе появу пропущених значень, дублювання, більш високі значення викидів, а також протиріччя в наборі даних[2].

Хоча існуючі методи попередньої обробки не гарантують рішення всіх цих проблем, вони можуть, принаймні, виправити деякі з них і поліпшити продуктивність моделей.

Тип даних і розподіл даних зазвичай перетворюються перед відправкою в моделі інтелектуального аналізу даних.

Мета перетворення даних включає в себе приведення даних у відповідність з вхідні вимоги моделей, що усувають шум даних і роблять розподіл даних більш відповідним для застосування алгоритмів оптимізації на етапі навчання моделі.

Вхідні дані для моделей інтелектуального аналізу даних можуть бути величезними: вони можуть мати занадто багато вимірювань або великий обсяг, що ускладнить навчання моделі інтелектуального аналізу даних або викличе проблеми при передачі і зберіганні даних[3].

Перехід від функціонального управління до процесного позитивно впливає на ефективність роботи процесів у сільському господарстві, але дані для переходу в інформаційній системі сільського господарства не можуть бути підготовленими сучасними методами.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Аналіз особливостей сільськогосподарських інформаційних систем

Агропромисловий комплекс функціонує в умовах постійно мінливого зовнішнього середовища, і від швидкості реакції на загрози і можливості залежить результат його діяльності.

Інформатизація АПК і сільськогосподарства-процес створення ефективних інформаційних систем і впровадження новітньої комп'ютерної техніки, програмного забезпечення з метою кардинального підвищення ефективності виробництва, багатопланових перетворень відповідно до світових тенденцій при максимально ефективному використанні всіх видів ресурсів і реалізується в сферах виробництва і його управління на різних рівнях; аграрної освіти і науки, а також соціальної сфери [5].

Інформатизація в сільськогосподарському виробництві поділяється на інформатизацію технологічних процесів та організаційно-економічного управління підприємствами та організаціями.

Інформатизація технологічних процесів передбачає створення системи засобів, за допомогою яких можна оцінювати і управляти процесами в рослинництві, тваринництві, при зберіганні і переробці продукції і здійснюється за такими напрямками:

- інформатизація контролю і управління окремими технологічними операціями шляхом використання комплексу приладів і обладнання для оснащення сільгоспмашин і вбудованими мікропроцесорними елементами;
- створення пакетів прикладних програм для фахівців для вирішення технологічних завдань;
- розробка комп'ютерних технологій управління повним процесом виробництва певного виду продукції [6].

В даний час реалізовані заходи державної програми «Розвиток сільського господарства і регулювання ринків сільськогосподарської

продукції, сировини і продовольства », де в розділі «створення загальних умов функціонування сільського господарства» було визначено створення єдиної системи інформаційного забезпечення АПК в якості комплексу заходів щодо формування системи державного інформаційного забезпечення у сфері сільського господарства, тому що своєчасну інформацію потребують і державні органи управління [8].

Інформаційна система АПК служить для створення однакового інформаційного обміну.

В Україні ведуться роботи з уніфікації надання електронних документів і бухгалтерської інформації, вирішення проблеми міжгалузевої взаємодії, організованого за допомогою юридично значущих документів, які відіграють важливу роль в структурі спільного ефективного функціонування територіальних інформаційних систем, консультативно-інформаційних центрів суб'єктів [9].

Автоматизована інформаційна система забезпечує інформаційну підтримку діяльності процесів управління галуззю, регулювання виробництва і агропродовольчих ринків[10].

Система призначена для забезпечення керівництва і фахівців достовірними та оперативними даними про результати оцінки стану та аналізу ситуацій, інформаційної підтримки процесів прийняття рішень в інтересах управління галуззю.

Система спирається на велику кількість баз даних і сховище відомостей, що агрегує основні інформаційні ресурси України в тимчасовому вимірі.

Важливим частиною, без якої неможливе функціонування інформаційної системи, є наявність єдиного підходу до формування та ведення системи нормативно-довідкової інформації та показників.

Практичною реалізацією такого підходу є структура, що об'єднує галузеві класифікатори і статистичні показники, статистичні показники і систему класифікаторів, а також міжнародні класифікатори [11].

Все це інтегровано в рамках галузевої системи , за коштами якої забезпечується ведення і публікація інформації з бази даних міжнародних класифікаторів, що застосовуються в сільському господарстві. Процедура ведення заснована на заявках підприємств [12].

За цим же принципом проводиться робота з організації ведення всієї нормативно-довідкової інформації агросфери, мета якої-методологічна цілісність єдиної системи, для чого забезпечується ведення класифікаторів і доступ до них через Інтернет, створюється база даних і показників АПК, покликана обслуговувати галузь, фіксуючи їх зміни, а також вирішуючи проблему зрозумілості та доступності статистичних і звітних даних.

Особливе значення має функціонування системи науково-технічної інформації АПК і сільського господарства, призначена для створення централізованого сховища на основі електронних версій текстів, забезпечення віддаленого формування повноформатних електронних версій текстових документів, а також для надання віддаленого доступу до центрального сховища із забезпеченням пошуку і перегляду повнотекстових документів, з урахуванням тематичних особливостей інформації АПК, довідників і словників.

У сучасних економічних умовах, коли сільськогосподарські організації діють в умовах розвиненого ринку, а органи держуправління АПК країни перетворилися в органи індикативного планування і регулювання ринкових відносин, для ефективної роботи необхідні дані про попит і пропозицію, умови поставок, розрахунків, що складається при цьому ціною товарів, тобто про аграрний ринок.

Це зумовило формування системи інформації про ринок АПК, що забезпечує збір, обробку, зберігання та надання ринкової інформації і суб'єктам агропродовольчого ринку, і населенню. Збір первинної ринкової інформації базується на відомчому моніторингу, здійснюваному фахівцями органів управління АПК всіх рівнів і ринковими репортерами установ в

муниципальних, регіональних і окружних філіях. Вимоги до наданої первинної інформації залежать від рівня споживача [13].

Аналіз тенденцій розвитку сільського господарства, вдосконалення його інфраструктури, розвитку виробництва свідчить про те, що поки фахівці не завжди можуть з необхідною точністю вирішувати завдання прогнозування врожайності сільгоспкультур, не володіють картографічним матеріалом загибелі посівів, ерозії ґрунтів і засоленості земель; не вирішена проблема контролю над використанням земель сільгосппризначення.

У зв'язку з цим в даний час розробляються інформаційні системи, що дозволяють приймати управлінські рішення по окремих територіях на основі оперативно-предметної інформації. Роботи здійснюються в рамках впровадження та адаптації в галузі географічних інформаційних систем – цифрового картографічного матеріалу сільськогосподарської тематики та оперативних даних супутникового моніторингу [14].

Так, система дистанційного моніторингу земель орієнтована на вирішення ряду завдань, пов'язаних з вивченням і аналізом їх стану, контролем врожаю з використанням даних дистанційного зондування. Але потрібно відзначити, що сільське господарство в Україні має свою специфіку, обумовлену слабо розвиненою методологічною базою, низьким рівнем інформаційної культури споживачів, слабкою формалізацією завдань. З урахуванням цього в рамках роботи з використання геоінформаційних систем розроблена загальна структура моделі галузевої ГІС, а також прототип системи супутникового моніторингу сільського господарства як компонент ієрархічної багаторівневої ГІС.

На федеральному рівні зібрана інформація, що дозволяє проводити макроаналіз всієї території України за матеріалами ґрунтових і ландшафтних карт, карт земельних угідь, узагальнених кліматичних даних. На рівні господарств дозволяє приймати рішення, засновані на аналізі різних інформаційних шарів: експлікації сівозмін, ґрунтово-ерозійна карта, карти вмісту мінеральних речовин та ін.

Сьогодні в індустріально розвинених країнах сільське господарство поставлено на промислову основу. Це означає не тільки використання потужної сільськогосподарської техніки, передових агрономічних методів і високоефективних хімікатів, а й залучення найсучасніших комп'ютерних технологій[15].

У промисловості комп'ютери не просто підвищили ефективність виробництва, вони, по суті, зробили другу промислову революцію, забезпечивши небувалі досі зростання ефективності виробництва, підвищення якості продукції і зниження собівартості.

У сільському господарстві впровадження комп'ютерної техніки дещо відстало від промисловості, проте сьогодні ми можемо спостерігати масове впровадження комп'ютерних технологій в сільське господарство Європи.

Самі по собі комп'ютери та програмне забезпечення, звичайно, не можуть замінити традиційні сільськогосподарські інструменти, але вони можуть дати дуже цінну інформацію щодо їх оптимального використання.

Вже класичними прикладами стали видаються спеціальними програмами підказки термінів, місць і обсягів внесення добрив, отрутохімікатів, управління іригацією і т.п. рішення таких завдань вимагає наявності досить великого обсягу вихідних даних, отримання яких становить основну частину вартості використання інформаційних технологій в сільському господарстві.

Обумовлено це тим, що необхідно збирати і вести інформаційні бази даних по великих площах на значному часовому протязі. Тільки актуальність, точність і повнота вихідних даних можуть забезпечити ефективне застосування інформаційних технологій в сільському господарстві [16].

Особливість використання інформаційних технологій в сільському господарстві полягає в тому, що практично всі використовувані дані мають просторову (географічну) прив'язку. І якщо ми необхідно, проаналізувати розподіл зволоження ґрунту спільно з врожайністю, то і ті, і інші дані

повинні знаходитися в одній системі координат і мати необхідну координатну точність. Обробляти ж такі дані можуть тільки програми, спеціалізовані на роботі з просторовою інформацією, а саме – географічні інформаційні системи[17]. Особливість цих систем в тому, що вони дозволяють інтегрувати, вести і спільно аналізувати найрізноманітніші види просторово розподілених показників і описових даних.

Ці системи використовуються для створення і ведення кадастрів земель і водних об'єктів, реєстрів власності, екологічного та погодного моніторингу, управління надзвичайними ситуаціями, оцінки виробничих ризиків, аналізу взаємозв'язків різних факторів, що впливають на врожайність сільськогосподарських культур і в багатьох інших додатках, заснованих на просторово розподіленій інформації [18].

ГІС - це об'єднання електронних карт, баз даних і засобів їх ведення та аналізу. Можливості і гнучкість цих систем забезпечують їх застосовність як в масштабах всієї країни, так і на рівні окремого фермерського господарства. І саме так ці системи використовуються сьогодні.

Дані для сільськогосподарських ГІС виходять різними способами. Основні джерела-це безпосередні виміри на полях з подальшою інтерполяцією і обробка знімків з літаків і космічних супутників. Безпосередні виміри використовуються головним чином на рівні окремого господарства або регіону.

Їх перевага-висока точність і надійність одержуваних результатів, можливості вимірювання самих різних показників при безпосередньому контакті з землею.

Недолік-висока вартість, особливо коли мова йде про великі площі. Дані космічної і висотної аерозйомки дозволяють контролювати обсяг біомаси, рівномірність росту рослин, зволоження ґрунту та інші показники.

Найважливіша перевага таких даних, особливо знімків з космічних апаратів, - їх низька вартість при регулярних обстеженнях великих територій.

Використання геоінформаційних систем можливо на різних рівнях управління сільським господарством. На федеральному рівні можливе ведення об'єднаної геоінформаційної системи, використовуваної для прогнозу врожайності в цілому по країні, оцінки сприятливих і несприятливих погодно-кліматичних факторів, ведення загального обліку і створення кадастру сільськогосподарських земель, стеження за довготривалими тенденціями в галузі і стратегічного планування.

Визначаються особливо охоронювані території, і, відповідно, ГІС цього рівня можуть використовуватися для планування заходів щодо захисту цих територій від несприятливого впливу аграрного сектора. Дані по подібних територіях можуть передаватися з федерального на регіональний рівень.

У ГІС можуть використовуватися як узагальнені дані з місць, так і дані космічної зйомки. Наявність незалежного та об'єктивного джерела інформації у вигляді даних дозволяє здійснювати контроль виконання Загальнодержавної аграрної політики, перевіряти відповідність інформації, що поставляється з місць, фактичному стану.

Розвинені можливості аналізу і висока потужність сучасних комп'ютерів призвели до виникнення в розвинених країнах так званого «точного» сільського господарства, при якому збираються дані і аналізується ефективність сільськогосподарського виробництва по дуже малих ділянках, а зібрана по ним статистика дозволяє найбільш повно врахувати варіації характеристик ґрунтів, гідрологічного режиму та інших показників.

На основі такого аналізу для кожного мікрополя пропонуються оптимальні режими іригації, внесення добрив і отрутохімікатів, проведення інших робіт, і навіть організації специфічної сівозміни.

Величезна популярність цієї методики обумовлена її високою ефективністю в той час, як всі інші методики вже практично вичерпали свій потенціал підвищення врожайності та якості продукції. Обробка ж і аналіз даних, використовуваних в даному підході просто не мислимі без комп'ютерів і

геоінформаційних систем, – тільки вони забезпечують на сьогодні необхідний рівень функціональних можливостей.

Традиційно структура сільського господарства включає дві основні галузі: рослинництво і тваринництво, які в свою чергу поділяються ще на цілий ряд галузей, що мають різноманітну складову.

Сільське господарство забезпечує населення продовольством, а також поставляє необхідну сировину різним галузям промисловості. Таким чином, дана галузь має дуже велике значення в усьому світі і від рівня її розвитку залежить добробут як окремо взятого суспільства, так і цілих держав.

Рослинництво являє собою галузь сільського господарства, спрямовану на вирощування різних культурних рослин для їх подальшого використання в якості джерел продуктів харчування, сировини для промисловості та інших цілей.

Рослинництво спрямоване на вивчення різноманіття сортів, форм і гібридів польових культур, особливостей біології та вдосконалення прийомів обробітку, що забезпечують досить високу якість і врожайність при мінімальних витратах[19].

Основними галузями рослинництва в залежності від виду вирощуваних рослин є: зернові культури; зернобобові культури; кормові культури; Технічні культури; овоче-баштанні культури; тонізуючі культури; олійні та ефіроолійні культури; хмелярство; виноградарство; садівництво; квітникарство; грибівництво і луговодство.

Тваринництво являє собою галузь сільського господарства, спрямовану на розведення і вирощування сільськогосподарських тварин з метою виробництва тваринницької продукції.

Основними галузями тваринництва є: скотарство, вівчарство, козівництво, кролівництво, Конярство, бджільництво, оленярство, звірівництво, птахівництво, свинарство[20].

Найважливішою економічною категорією є система господарювання, яка характеризує основні виробничі відносини всередині підприємства.

Система ведення сільського господарства - це сукупність соціально-економічних, організаційних, технічних і технологічних принципів побудови та ведення виробництва для конкретних умов з метою задоволення потреби суспільства в сільськогосподарських продуктах.

Раціональна система господарювання є найважливішою складовою організаційно-господарської плану будь-якого підприємства.

На основі регіональних систем розроблялися і системи ведення конкретних господарств.

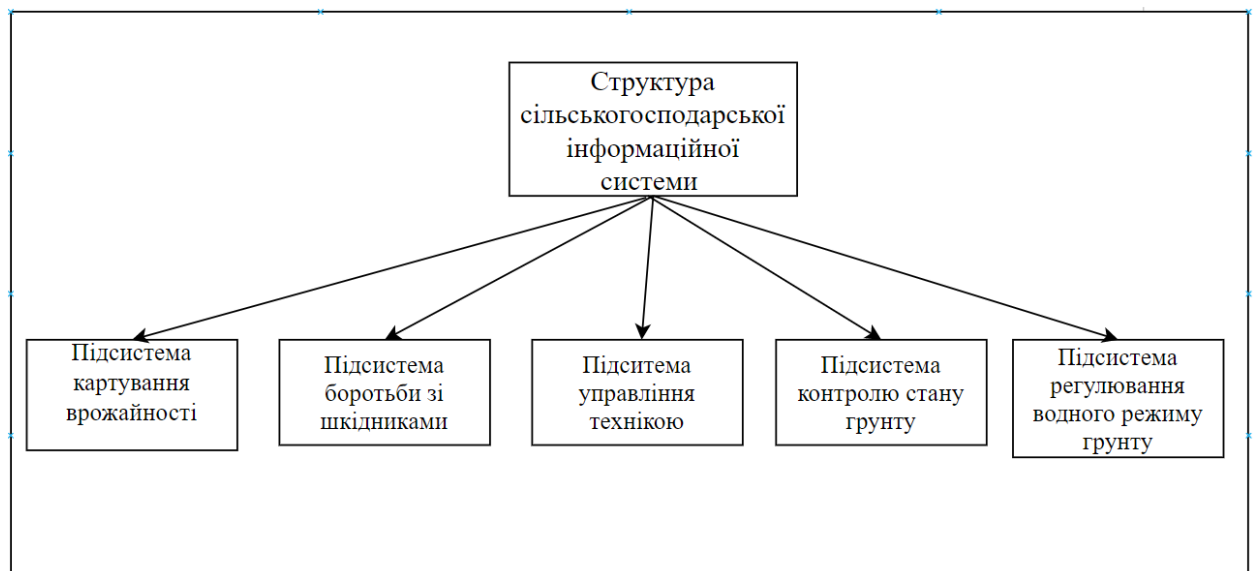


Рисунок 1.1 – Структура інформаційної системи сільського господарства

Основні принципи побудови раціональної системи господарювання є:

- цілісність системи. Означає, що її властивості не зводяться до механічної суми властивостей компонентів і елементів системи. У раціональній системі господарювання властивість кожного компонента, елемента залежить від його місця, функції всередині системи;
- структурність системи, тобто можливість представлення її через внутрішній зміст;

- ієрархічність системи. Кожен компонент або елемент може розглядатися як самостійна система;

- взаємозв'язок і співвідношення компонентів і елементів системи. Вони будуються з урахуванням найбільш раціонального використання факторів зовнішнього середовища[21].

1.2 Порівняльний аналіз функціонального та процесного підходів до управління

Управляти - означає, ставити конкретні цілі, планувати, здійснювати контроль на основі затверджених планів, аналізувати результати, зіставляючи їх з плановими показниками, виявляти причини відхилень і приймати рішення, що усувають ці розбіжності. У теорії управління цей замкнутий ланцюжок називається циклом управління, а її ланки утворюють етапи циклу.

Таким чином, функціонування кожної компанії завжди проходить в рамках управлінського циклу. Оптимальність циклу управління безпосередньо залежить від складових його елементів і ефективності їх взаємодії.

При цьому існує два основних підходи до управління: процесний і функціональний.

При функціональному підході підприємство розглядається як механізм, що володіє набором функцій, які розподіляються серед підрозділів і виконуються співробітниками підприємства. Вони виконують вузькоспеціалізовані завдання, не працюючи на досягнення місії підприємства.

Структурні підрозділи підприємства взаємодіють між собою і передають один одному керуючі впливи, що породжує різного роду

розбіжності: конфлікти інтересів, конфлікти бюджетів і т. п. Основні недоліки функціонального підходу полягають в наступному:

При функціонально структурованій організації відсутня зацікавленість співробітників в кінцевому результаті. Найчастіше бачення того, що відбувається працівниками не виходить за рамки своїх підрозділів, вони не орієнтовані на кінцеві цілі підприємства, на задоволення потреб покупця.

Значна частина реальних робочих процесів на підприємстві включає безліч функцій, тобто виходить за рамки окремих підрозділів.

Однак у функціонально орієнтованих структурах обмін інформацією між різними підрозділами надмірно ускладнений через її вертикальної ієрархічності, що призводить до великих накладних витрат, невиправдано тривалих термінів прийняття управлінських рішень [22].

Значна частина часу, необхідного для здійснення управлінського впливу на виробничий процес, витрачається на взаємовідносини служб і воно значно більше, ніж час на реалізацію самого рішення. Це призводить до того, що реакція на обурює вплив йде з невиправдано великим запізненням.

З цих причин загальновідомі способи вдосконалення функціональної системи управління підприємством, наприклад, зміна структури підприємства, скорочення чисельності співробітників, впровадження комп'ютерних інформаційних систем управління підприємством, спроби застосування систем якості на базі малоефективні, а в деяких випадках навіть шкідливі.

Тому кардинальна зміна ситуації на підприємстві без зміни принципів управління не представляється можливою.

Процесний підхід до управління підприємством можна вважати базовим при інтегруванні підприємства. При цьому підході об'єктом управління виступає певна діяльність на підприємстві – процес, який можна визначити як ряд взаємопов'язаних видів діяльності, що характеризуються споживанням ресурсів (вхід процесу) і дають певний результат (вихід процесу).

Процеси проходять через всі підрозділи, залучаючи всі служби підприємства, і орієнтуються на кінцевий результат – збільшення вартості бізнесу.

Керуючи процесами, що мають свої цілі, можна домогтися високої ефективної діяльності за допомогою добре налагоджених горизонтальних зв'язків у вертикальній структурі управління підприємством.

Процесний підхід - це вихід на продуктивну ідею внутрішніх постачальників і споживачів, так як реальна діяльність, що приносить додану вартість, не здійснюється окремими елементами функціональної ієрархії, а пронизує підприємство у вигляді сукупності процесів.

Процесний підхід дозволяє:

- враховувати такі важливі моменти, як орієнтація на кінцевий результат, зацікавленість кожного виконавця в підвищенні якості не тільки кінцевого продукту, але і своєї роботи;
- швидко реагувати на зміни умов зовнішніх і внутрішніх факторів;
- оптимізувати обмін інформацією між функціональними підрозділами;
- реалізувати важливу рису менеджменту якості вбудовування контролю якості в процес замість контролю якості готової продукції.

При реалізації процесного підходу виконавці володіють великими повноваженнями, збільшується їх роль, самостійність і, отже, віддача і задоволення працею. Керівники ж звільняються від поточних оперативних питань і повністю концентруються на вирішенні стратегічних, системних питань.

	Суть	Об'єкт	Недоліки
Функціональний	Підприємство розглядається як механізм, що володіє набором функцій, які розподіляються серед підрозділів	Відділ, організаційна структура	1) незацікавленість працівників 2) ускладненість обміну інформацією між підрозділами 3) потрібна велика кількість часу для реалізації

			управління на виробничий процес; 4) авторитарний управління
Процесний	Підприємство представляється набором процесів, що проходять через всі підрозділи і задіюють всі служби підприємства	Бізнес процес	1) взаємозалежність осіб, які приймають рішення

Таблиця 1.2 - Порівняння процесного і функціонального управління

1.3 Дослідження задач інтелектуального аналізу процесів

Головною задачею інтелектуального аналізу процесів є побудова моделей бізнес-процесів як є при переході до процесного управління від функціонального.

Існують такі типи Process Mining: Play-in, Play-out и Replay.

Одним з ключових елементів process mining є акцент на встановленні тісного зв'язку між моделлю процесу і "реальністю", відображеної у вигляді журналу подій.

Використовується терміни: Play-In, Play-Out, and Replay, щоб осмислити цей зв'язок.

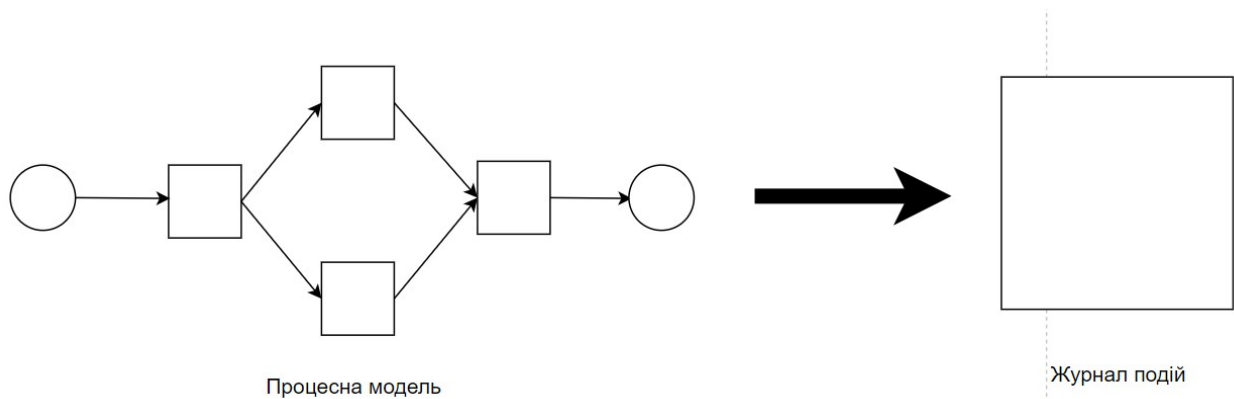


Рисунок 1.2 – Спосіб співвіднесення журналів подій Play-Out

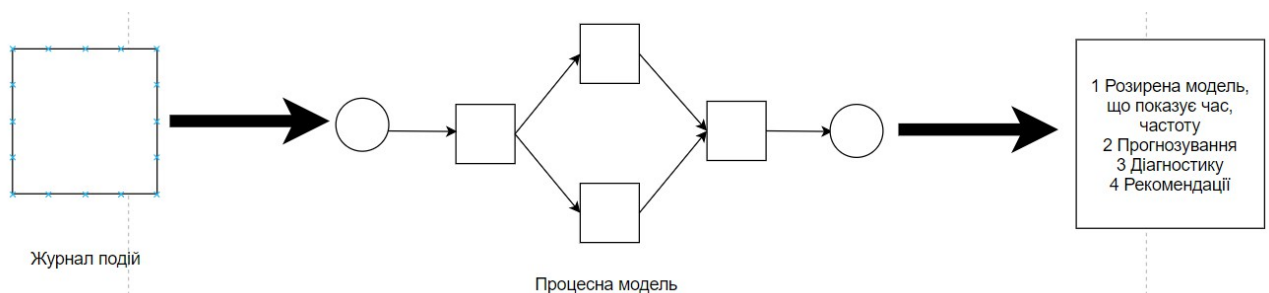


Рисунок 1.3 – Спосіб співвіднесення журналів подій Replay

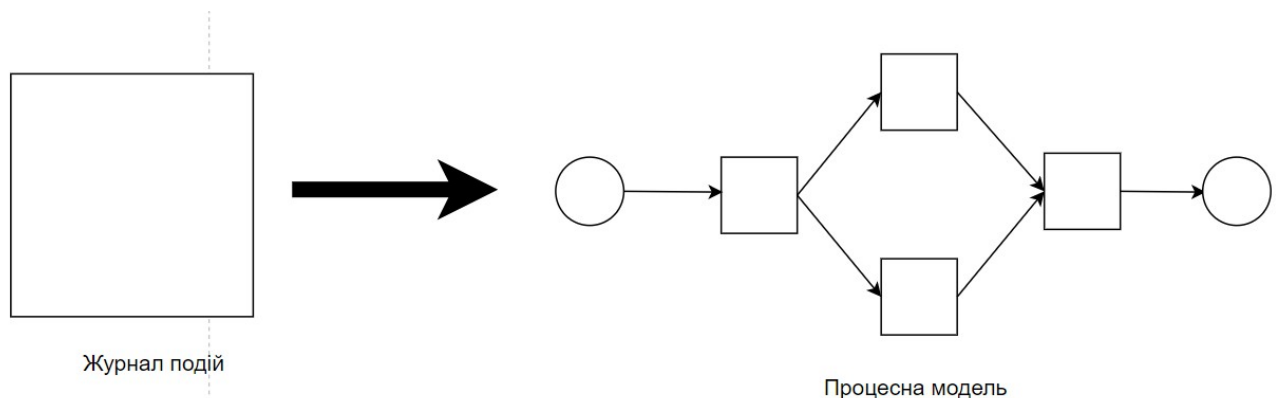


Рисунок 1.4 – Спосіб співвіднесення журналів подій Play-In

Play-Out відноситься до класичного використання моделей процесів. PlayOut можна використовувати як для аналізу, так і для реалізації бізнес-процесів. Механізм робочого процесу можна розглядати як "механізм відтворення", який управляє випадками, дозволяючи тільки "ходи", дозволені відповідно до моделі.

Отже, Play-Out може бути використаний для реалізації операційних процесів з використанням деякої виконуваної моделі. Інструменти моделювання також використовують механізм відтворення для проведення експериментів. Основна ідея моделювання полягає в багаторазовому запуску моделі і, таким чином, зборі статистики і довірчих інтервалів.

Механізм моделювання схожий на механізм документообігу. Основна відмінність полягає в тому, що механізм моделювання взаємодіє з середовищем, яке моделюється, в той час як механізм документообігу взаємодіє з реальним середовищем (робітники, клієнти тощо).

Крім того, класичні підходи до верифікації, що використовують вичерпний аналіз простору станів-часто званий перевіркою моделі - можна розглядати як методи відтворення.

Play-In-це протилежність Play-Out, тобто приклад поведінки береться в якості вхідних даних, і мета полягає в тому, щоб побудувати модель. Більшість

методів інтелектуального аналізу даних використовують Play-In, тобто модель вивчається на основі прикладів. Однак традиційно інтелектуальний аналіз даних не був пов'язаний з моделями процесів.

На жаль, неможливо використовувати звичайні методи інтелектуального аналізу даних для відтворення моделей процесів. Тільки недавно методи інтелектуального аналізу процесів стали легко доступні для виявлення моделей процесів, заснованих на журналах подій.

Replay використовує журнал подій і модель процесу в якості вхідних даних. Журнал подій "відтворюється" поверх моделі процесу. Журнал подій може відтворюватися для різних цілей:

- перевірка відповідності: розбіжності між журналом і моделлю можуть бути виявлені та кількісно оцінені шляхом відтворення журналу;
- розширення моделі частотної і тимчасової інформацією, при повторному перегляді журналу можна побачити, які частини моделі часто відвідуються;
- побудови прогностичної моделі, відтворюючи журнали подій, можна будувати прогностичні моделі, тобто для різних станів моделі можна робити певні передбачення;
- оперативна підтримка, повтор не обмежується історичними даними про події. Можна також відтворення часткових слідів справ, все ще запущених. Це може бути використано для виявлення відхилень під час виконання.

1.4 Дослідження методів структуризації вхідних даних у задачах process mining

Мережі Петрі-мережі в яких процеси можуть оброблятися одночасно. У стандартній моделі такої мережі передбачається, що час спрацьовування всіх переходів є однаковим. Для простоти будемо вважати, що цей час дорівнює одиниці. Також покладемо, що спрацьовування будь-якого переходу проводиться в цілочисельний момент часу.

Таким чином, час в моделі виявляється дискретним, його відлік будемо вести з нульового моменту. Для того, щоб описати паралельну поведінку мережі Петрі, необхідно визначити поняття конфлікту. Конфліктом в мережі Петрі називається ситуація, коли відразу кілька активних переходів претендують на одну мітку деякого місця. При послідовному спрацьовуванні переходів конфлікти ніяк не враховуються, однак при паралельній інтерпретації потрібно деякий спосіб їх вирішення.

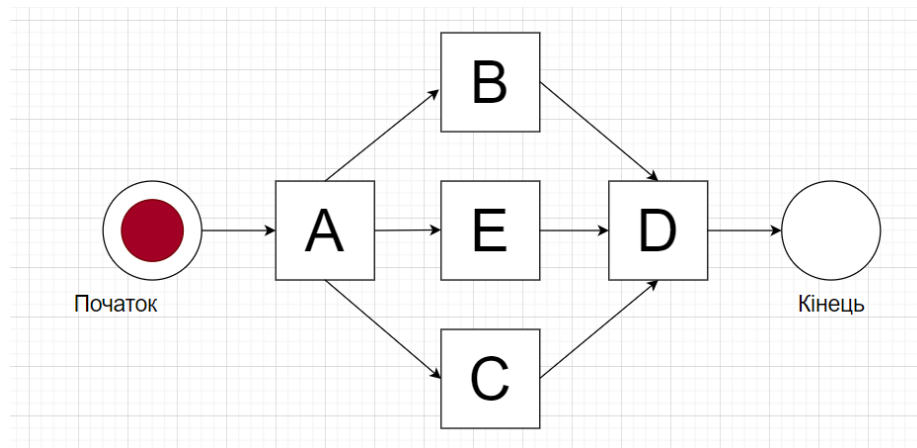


Рисунок 1.4 – Вимоги до вхідних даних аналізу процесів

Мережі Петрі-мережі в яких процеси можуть оброблятися одночасно. У стандартній моделі такої мережі передбачається, що час спрацьовування всіх переходів є однаковим. Для простоти будемо вважати, що цей час дорівнює одиниці. Також покладемо, що спрацьовування будь-якого переходу проводиться в цілочисельний момент часу. Таким чином, час в моделі виявляється дискретним, його відлік будемо вести з нульового моменту. Для того, щоб описати паралельну поведінку мережі Петрі, необхідно визначити поняття конфлікту. Конфліктом в мережі Петрі називається ситуація, коли відразу кілька активних переходів претендують на одну мітку деякого місця. При послідовному спрацьовуванні переходів конфлікти ніяк не враховуються, однак при паралельній інтерпретації потрібно деякий спосіб їх вирішення.

Для переходу до процесного управління необхідно використати методи process mining. Методи потребують вхідних даних у спеціалізованих форматах.

Це є основною проблемою, бо в сільському господарстві більшість процесів не надають інформацію у однакових форматах.

Вимоги до вхідних даних:

- дані необхідно представити у форматі логів журналу;
- дані повинні містити траси одного процесу;
- дані однієї моделі мають знаходитися у одному журналі подій;
- дані мають містити невелику кількість помилок.

1.5 Постановка задачі дослідження

Об'єктом дослідження в рамках даної роботи є процес попередньої обробки даних щодо сільськогосподарських робіт у аграрних системах.

Предметом дослідження являються методи інтелектуального аналізу процесів.

Метою даної роботи є дослідження методів попередньої обробки даних у сільськогосподарських інформаційних системах для побудови моделей процесів аграрних робіт.

Для досягнення мети потрібно вирішити такі задачі:

- аналіз особливостей сільськогосподарських інформаційних систем;
- дослідження методів process mining;
- дослідження методів структуризації вхідних даних у задачах process mining;
- удосконалення методу попередньої обробки вхідних даних для вирішення задачі інтелектуального аналізу процесів при переході до процесного управління в сільськогосподарських системах;
- експериментальна перевірка удосконаленого методу обробки вхідних даних.

Сільськогосподарські роботи мають процесний характер у відповідності до сезонів року. Сільськогосподарські системи використовують функціональний підхід до управління, тому у сільськогосподарському

виробництві відбувається перехід від традиційного функціонального до процесного управління.

Функціонування сільськогосподарських систем процесного управління потребує побудови процесних моделей аграрного виробництва.

Такі моделі формуються з використання методів process mining, з використанням вхідних даних у спеціальному процесно-орієнтованому форматі.

Однак існуючі дані про сільськогосподарське виробництво не відповідають цьому формату, що потребує реалізації попередньої обробки вхідних даних та їх трансформації в форму придатну для використання методів process mining.

Проблема: удосконалення методу побудови процесних моделей на неструктурованих даних у задачах інтелектуального аналізу процесів.

2 ДОСЛІДЖЕННЯ МЕТОДІВ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ У PROCESS MINING

2.1 Структуризації вхідних даних у задачах process mining

У процесі інтелектуального аналізу вхідні дані формуються у реляційну структуру наступного вигляду:

- події- найменша структурна одиниця. Належить процесу;
- процес- сукупність подій, що формують трасу;

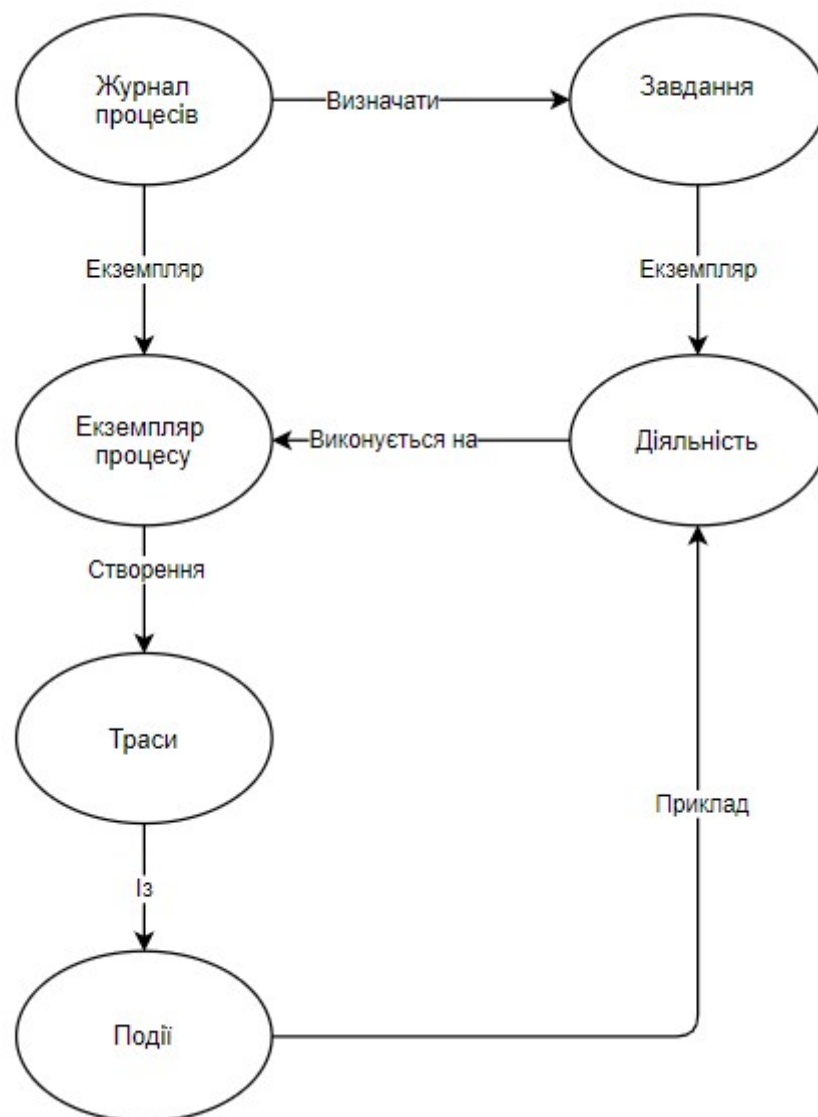


Рисунок 2.1 – Схема роботи

- журнал подій- логічне групування процесів із однаковою моделлю у єдине сховище;
- атрибути: мета-дані, що можуть характеризувати подію, або процес.

Дана структура визначена у стандарті XES. Стандарт журналів подій XES зберігає дані у деревовидній структурі згідно з їх фактичними відносинами[26].

На відміну від застарілого формату журналів MXML, він має гнучку систему атрибутів, що дозволяє найбільш ефективно характеризувати дані, що зберігаються.

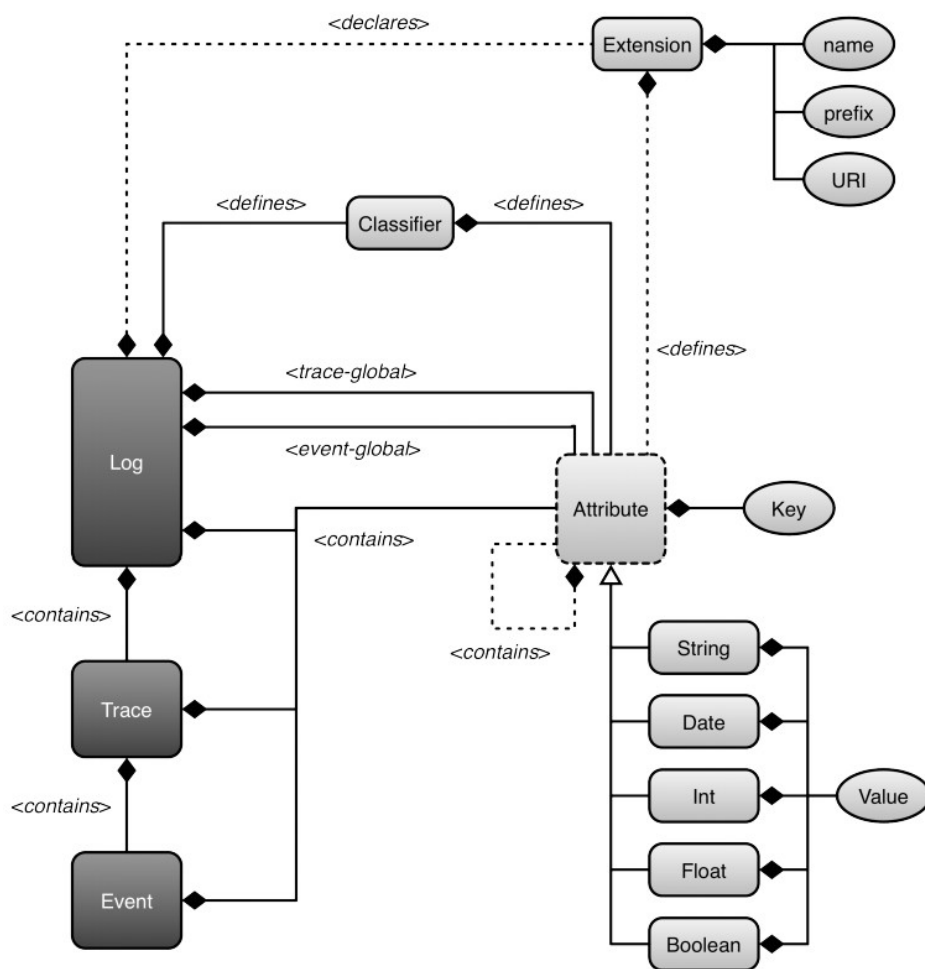


Рисунок 2.2 – Особливості стандарту XES

Базовий обов'язково має бути присутнім в будь-якому відповідному стандарту XES лозі подій. Базовий стандарт визначає основну структуру лога подій, в той час як всі дані, що стосуються процесу і зберігаються в Лозі подій зберігаються в атрибутах і визначені за допомогою розширення базового стандарту, що зображено на правій частині таблиці 2.2.

Таблиця 2.2 – Базовий та розширений стандарт XES

Базовий стандарт	Розширений стандарт
Xlog	XExtension
XTrace	Url
XEvent	Name
XElement	Prefix
XAttributable	
XAttribute	
Key	

Журнал містить трасування, і кожна трасування містить події. Журнал, трасування і події мають атрибути[27]. Розширення можуть визначати нові атрибути, і журнал повинен оголошувати використовувані у ньому розширення. Глобальні атрибути-це атрибути, оголошені обов'язковими.

Така атрибути знаходяться на рівні трасування або події. Атрибути можуть бути вкладеними. Класифікатори подій визначаються для журналу і присвоюють кожній події "мітку" (наприклад, назва дії). Класифікаторів може бути кілька.

На рисунку показано, що журнал оголошує набір використовуваних розширення. Кожне розширення може визначати атрибути, які вважаються стандартними при використанні розширення[28].

XES може заявити свої особливості, які повинні бути в обов'язковому порядку. Наприклад, можна вказати, що будь-який слід повинен мати ім'я або що будь-яка подія повинна мати мітку часу. Для цього журнал містить два списки глобальних атрибутів: один для трас і один для подій.

XES підтримує концепцію класифікатора. Журнал XES визначає довільне число класифікаторів. Кожен Класифікатор задається списком атрибут. Будь-які дві події, які мають однакові значення по відношенню до цих атрибутів вважаються рівними для цього класифікатора. Ці атрибути повинні бути обов'язковими атрибутами подій. Наприклад, якщо Класифікатор заданий як атрибутом імені, так і атрибутом ресурсу, то дві події зіставляються одному і тому ж класу, якщо їх атрибути імені та ресурсу збігаються.

В принципі, можливі багато серіалізації. Однак, щоб обмінятися фіксує документи, використовувати стандартний XML-серіалізації. Події мають три глобальні атрибути: атрибути `time: timestamp`, `concept: name` і `org: resource` обов'язкові для всіх подій.

У журналі XES, визначено три класифікатори. Ресурс класифікатора класифікує події на основі атрибута `org: resource`. Обидва класифікатори класифікують події на основі двох атрибутів: `concept: name` і `org: resource`. В контексті XES, були визначені п'ять стандартних розширень. Ці розширення описані в так званому XML-форматі XESEXT . Згадується лише підмножина стандартних атрибутів, визначених цими розширеннями.

Розширення поняття визначає атрибут `name` для трасувань і подій. Для трасувань атрибут зазвичай представляє певний ідентифікатор для випадку. Для подій, атрибута, як правило, являє собою ім'я дії. Розширення поняття також визначає атрибут екземпляра для подій. Це використовується для розрізнення різних екземплярів активності в одній і тій же трасуванні.

Розширення життєвого циклу визначає атрибут переходу для подій. При використанні стандартної моделі транзакційного життєвого циклу, можливими значеннями цього атрибутами є "розклад", "старт", "завершення", "автоскіп" тощо.

Організаційне розширення визначає три стандартні атрибути для подій: ресурс, роль і група. Атрибут `resource` відноситься до ресурсу, який викликав або виконав подію. Атрибути ролі і групи характеризують (обов'язкові) можливості ресурсу і його положення в організації. Задля наприклад, подія, що виконується менеджером з продажу, може мати пов'язану з ним роль "менеджер" і групу "відділ продажів".

Розширення `time` визначає атрибут `timestamp` для подій. Оскільки така тимчасова мітка має тип `xs: dateTime`, записуються як дата, так і час.

Семантичне розширення визначає атрибут `modelReference` для всіх елементів журналу. Це розширення натхненне SA-MXML. Посилання в журналі вказують на поняття в онтології[29].

Користувачі та організації можуть додавати нові розширення та ділитися ними з іншими. Наприклад, можуть бути додані загальні розширення, пов'язані з витратами, ризиками, контекстом і т. д. однак розширення також можуть бути специфічними для домену або організації.

В даний час встановлюється, підтримується інструментів, таких як випускний вечір, нітро, XESame, диско, Celonis, ті хвилиної, аналіз СНП бізнес-процесів, і посилання OpenXES реалізації. ProM, ймовірно, є найбільш широко використовуваним інструментом технологічного mining, що забезпечує широкий спектр методів технологічного mining. Який може використовуватися для швидкого перетворення журналів подій у формат XES (або MXML). XESame генерує файли XES з колекцій таблиць бази даних.

Ідея полягає в тому, що при заданому наборі таблиць можливі різні уявлення. Таким чином, файли XES служать в якості представлення даних про події. Відкриті вікна - це еталонна реалізація XES, бібліотека java з відкритим кодом для читання, зберігання та запису журналів XES. OpenXES може бути легко вбудований в інших інструментах і здатний ефективно (де)серіалізувати великі журнали подій в / з XML- файл. Це звільняє розробників програмного забезпечення від розробки виснажливого коду для імпорту та експорту даних подій.

Наявність високоякісних журналів подій має важливе значення для інтелектуального аналізу процесів. Крім того, хороші журнали подій можуть служити багатьом іншим цілям. Іноді термін " Походження бізнес-процесу " використовується для позначення систематичного збору інформації, необхідної для реконструкції того, що насправді відбулося в бізнес-процесі. Із точки зору аудиту систематична, надійна і заслуговує довіри запис подій має важливе значення.

Термін "Походження" походить від наукових обчислень. Тут Інформація про походження записується для забезпечення відтворюваності наукових експериментів. Високоякісні журнали подій, які не можна підробити, гарантують, що " історія не може бути переписана або прихована", і служать міцною основою для поліпшення процесу і аудиту.

Тому XES слід розглядати в контексті походження, який виходить за рамки виявлення процесів і включає такі теми, як перевірка відповідності [30].

2.2 Перетворення вхідних даних у стандарт XES

Мета process mining- аналіз даних про події з точки зору процесу. Process mining відповідає на питання про операційні процеси.

Наприклад:

- Що насправді сталося в минулому?
- Чому це сталося?
- Що може статися в майбутньому?
- Коли і чому організації та люди відхиляються від норми?
- Як краще контролювати процес?
- Як переробити процес, щоб підвищити його продуктивність?

Щоб відповісти на ці запитання потрібно спочатку зосередитися на необхідних даних про події.

Джерелом даних може бути простий плоский файл, електронна таблиця Excel, журнал транзакцій або таблиця бази даних. Однак не слід очікувати, що всі дані будуть знаходитися в одному добре структурованому джерелі даних. Реальність така, що дані про події зазвичай розкидані по різних джерел даних, і часто потрібні досить значні зусилля для збору відповідних даних. Дані можуть бути розкидані з технічних або організаційних причин[31].

Наприклад, можуть існувати застарілі системи зберігання найважливіших даних або інформаційні системи, використовуються тільки на рівні департаментів. Для міжорганізаційного аналізу процесів, дані можуть бути навіть розкидані по декількох організаціях.

Події також можуть бути зафіксовані шляхом прослуховування обміну повідомленнями (наприклад, soar-повідомлень) і запису дій читання і запису. Джерела даних можуть бути структуровані і добре описані за допомогою метаданих.

На жаль, у багатьох ситуаціях дані неструктуровані або важливі метадані відсутні. Дані можуть виходити з веб-сторінок, електронних листів, PDF-документів, відсканованого тексту, зскрібання екрану і т. д. навіть якщо дані структуровані і описуються метаданими, сама складність корпоративних інформаційних систем може бути переважною, немає сенсу намагатися вичерпно витягувати журнали подій з тисяч таблиць та інших джерел даних.

Витяг даних має ґрунтуватися на питаннях, а не на наявності великої кількості даних.

У контексті бізнес-аналітики та інтелектуального аналізу даних словосполученню "Витяг, перетворення і завантаження" використовується, щоб описати процес, який включає:

- витяг даних з поза джерел;
- перетворення його відповідно до оперативних потреб (робота з синтаксичними і семантичними проблемами при забезпеченні заданих рівнів якості);
- навантаження його в цільовій системі, наприклад, сховища даних, або реляційні бази даних. Сховище даних це єдине логічне сховище транзакційних і операційних даних організації[31].

Сховище даних не виробляє дані, а просто виводить дані з операційних система. Мета полягає в тому, щоб уніфікувати інформацію таким чином, щоб її можна було використовувати для звітності, аналізу, прогнозування, що дії можуть використовуватися для заповнення сховища даних.

Для створення загального представлення, необхідного для сховища даних, може знадобитися чимало зусиль. Різні джерела даних можуть використовувати різні ключі, Угоди про форматування й інше.

Одне джерело даних може використовувати формат дати "31-12-2010", в той час як інший використовує формат "2010/12/31".

Якщо сховище даних вже існує, воно, швидше за все, містить цінні вхідні дані для інтелектуального аналізу процесів. Однак багато організацій не мають надійного сховища даних.

Сховище може містити лише підмножину інформації, необхідної для наскрізного інтелектуального аналізу процесів. Більше того, якщо дані склад присутній, він не повинен бути орієнтований на процес.

Наприклад, типові дані сховища, що використовуються для оперативної аналітичної обробки, не містять великої кількості інформації, пов'язаної з процесом. Інструменти відмінно підходять для перегляду багатовимірних даних під різними кутами, деталізації і створення всіляких звітів. Однак ці інструменти не вимагають зберігання бізнес-подій та їх упорядкування.

Набори даних, що використовуються основними підходами інтелектуального аналізу даних, також не зберігають таку інформацію. Наприклад, вивчає дерево рішень може бути застосований до будь-якої таблиці, що складається з рядків (примірників) і стовпців (змінних). Процесний mining вимагає інформації про відповідні події та їх порядок.

Незалежно від того, чи є сховище даних чи ні, дані повинні бути витягнуті і перетворені в журнали подій. Тут визначення області має першорядне значення. Часто проблема полягає не в синтаксичному перетворенні, а в підборі відповідних даних.

Типовими форматами для зберігання журналів подій є XES (eXtensible Event Stream) і MXML (Розширюваний мову розмітки mining). На даний момент припускається, що один журнал подій відповідає одному процесу, тобто при визначенні обсягу даних на етапі вилучення повинні бути включені тільки події, що мають відношення до аналізованого процесу

Залежно від обраних питань і точки зору різні журнали подій можуть бути витягнутим з одного і того ж набору даних. Наприклад, дані в сільському господарстві [32]. Може бути, когось зацікавить оптимізація системи поливу або шляхів діагностики стану ґрунту.



Рисунок 2.3 – Особливості стандарту XES

Обидва підходи вимагають різних журналів подій, хоча деякі події можуть бути розділені між двома необхідними журналами подій. Один раз в створюється журнал подій, який зазвичай фільтрується. Фільтрація-це ітеративний процес.

При витяганні даних в журнал подій проводилася груба оцінка області видимості. Фільтрація відповідає дрібнозернистому визначенню області

видимості на основі вихідних результатів аналізу. Наприклад, для виявлення процесів можна вирішити зосередитися на 10 найбільш частих діях, щоб зберегти керованість модель.

На основі відфільтрованого журналу можуть бути застосовані різні типи інтелектуального аналізу процесів: виявлення, відповідність і поліпшення.

Слід зазначити, що результати інтелектуального аналізу процесів, швидше за все, викликають нові питання, і ці питання можуть призвести до дослідження нових джерел даних і більш детального вилучення даних.

Як правило, потрібно кілька ітерацій етапів вилучення, фільтрації і mining.

2.3 Удосконалений метод підготовки вхідних даних у форматі XES

Удосконалений метод підготовки вхідних даних виконується за 5 етапів, які виконуються поступово. Вхідні дані мають різний розмір, різний формат та беруться з різних процесів, які відбуваються в сільському господарстві.

Етап 1. Формування набору обмежень по об'єктам процесу.

Етап 2. Відбір подій журналу, що відображають вирішення операції із відібраними об'єктами процесу.

Етап 3. Побудова підмножин подій логу для інтервалів часу .

Етап 4. Упорядкування підмножин подій за часом та формування трас:

$$T = \{t_1, t_2, \dots, t_n\}, \quad (2.1)$$

де p – подія, t – траса, T – множина трас.

Етап 5. Уточнення множини трас за частотою подій.

Крок 5.1 Обчислення вірогідності появи події у процесі відповідає відношенню кількості трас, що містять подію, до кількості усіх процесів:

$$P(p) = \frac{|T_p|}{|T|}, \quad (2.2)$$

Крок 5.2 Відбір підмножини трас для яких виконується умова:

(2.3)

де – порогове значення частоти подій.

2.4 Конвертація вхідних даних при побудові процесної моделі

У якості вхідних даних використовуються наступні джерела:

- бази даних підприємства;
- файли логів виконаних задач.

Після попередньої обробки даних вони подаються на вхід до задачі аналізу процесів. Головною задачею попередньої обробки є підготування даних до інтелектуального аналізу. Процесні моделі містять перелік усіх подій а також їх коефіцієнти, що відображають вагу шляху.

Data transformation-це перетворення і узгодження наборів даних один з одним або з певною схемою. Воно відбувається після вилучення всіх наявних даних і забезпечує їх подальшу обробку, наприклад, для процесу інтеграції або завантаження в іншу інформаційну систему. У процесній аналітиці (process mining) перетворення даних – це компонент попередньої обробки даних[33].

Для подальшої обробки даних важливою умовою є їх стандартизованість. Відмінності в даних можуть бути викликані різними джерелами, таблицями або типами. Перетворення дуже важливо не тільки для аналізу даних, але і для збереження відносин між ними. Як правило, дані необхідно або "налаштовувати" один на одного, або на конкретний цільовий формат.

Якщо дані стандартизовані, тобто наведені до єдиного виду, то їх формат визначається як цільовий . Тоді всі дані, що надходять з інших джерел будуть приводитися саме і тільки до цільового.

У випадках, коли потрібна конкретна схема дані будуть наводитися відповідно до зазначеного цільового формату.

Щоб перетворити дані, зазвичай спочатку необхідно їх витягти. Виняток становлять дані з БД, які можуть бути перетворені в них самих за допомогою мови SQL.

Після вилучення даних задаються цільовий формат або схема. При проведенні перетворення необхідно знати специфікації вихідного формату і цільового формату, щоб було можливо провести операцію (обов'язкова умова).

Використовуючи фіксовані визначення та призначення, дані у вихідному файлі перетворюються та присвоюються певним значенням, щоб вони відповідали цільовому формату.

Окремо відбувається опрацювання нюансів порожніх значень. Вони з'являються, наприклад, якщо у об'єкта немає певного атрибута, тобто немає запису для нього.

Те, як обробляються такі значення, залежить від мети перетворення або цільової системи. Наприклад, в базах даних краще вводити NULL, оскільки порожні значення можуть привести до помилок під час перетворення або після обчислень.

Таким чином, можна виділити наступні етапи процесу data transformation:

- витяг даних;
- оцінка необхідного формату;
- визначення цільового формату;
- конвертація витягнутих даних;
- збереження перетворених даних в новий файл.

Можно виділити три основні переваги використання process mining в цьому процесі:

Прийняття рішень на основі реальних даних. Системи класу вибудовують карти процесів, ґрунтуючись на інформації з журналів подій інформаційних систем. Це 100% достовірна картина, виключені втручання і впливу людського фактора.

При цьому прийняття рішень про внесення змін до процесів приймається на підставі об'єктивної інформації, а не на припущеннях і теоріях.

Зниження витрат. Одна з цілей, до якої компанії зазвичай прагнуть, полягає в тому, щоб скоротити свої витрати і збільшити прибуток. Аналізуючи журнали подій, process mining може показати тривалість кожного етапу процесу, його вартість і способи поліпшення. Компанія може позбутися від непотрібних витрат і зробити відповідні кроки для їх зниження. Виявлення вузьких місць. У якийсь момент процеси можуть мати кілька етапів, які займають занадто багато часу і призводять до затягування процесів. Це знижує ефективність. За допомогою process mining компанії можуть визначити вузькі місця і виключити їх.

3 ДОСЛІДЖЕННЯ ОТРИМАННЯ НАУКОВИХ РЕЗУЛЬТАТІВ

3.1 Технологія підготовки вхідних даних

Стандарти Data Mining не випадково виділяють підготовку даних в окрему фазу. Data Preparation - вельми трудомісткий ітеративний процес, який займає до 80% всіх витрат ресурсів і часу в життєвому циклі Data Mining і включає наступні завдання обробки вихідних даних:

- вибірка даних-відбір ознак (features або предикторів) і об'єктів з урахуванням їх релевантності для цілей Data Mining, якості і технічних обмежень (обсягу і типу) ;

- очищення даних-видалення помилок, некоректних значень (наприклад, число в рядковому параметрі тощо), відсутніх значень (Missing values або NA), виключення дублів і різних описів одного і того ж об'єкта, відновлення унікальності, цілісності і логічних зв'язків ;

- генерація ознак-створення похідних ознак та їх перетворення у вектори для моделі Machine Learning, а також трансформація для підвищення точності алгоритмів машинного навчання ;

- інтеграція-злиття даних з різних джерел (інформаційних систем, таблиць, протоколів та ін.), включаючи їх агрегацію, коли нові значення обчислюються шляхом підсумовування інформації з безлічі існуючих записів ;

- форматування-синтаксичні зміни, які не змінюють значення даних, але потрібні для інструментів моделювання, наприклад, сортування в певному порядку або видалення непотрібних розділових знаків в текстових полях, обрізка «довгих» слів, округлення дійсних чисел до цілого і т.д. .

Для функціонування цього інструменту, як і будь-якого ІТ-засобу, необхідні чітко визначені вихідні дані та інструкції. Неможливо завантажити в алгоритм Machine Learning всі накопичені великі дані різних форматів і отримати на виході коректні результати. Крім того, вихідні дані часто спотворені і ненадійні: в них можуть бути присутніми значення, що виходять за

межі допустимих діапазонів (шуми), аномальні значення (викиди), а також пропуски (відсутність значень).

До того ж, часто виникає завдання попередньої підготовки вихідних даних. У географічних даних можуть зустрічатися помилки в адресах і помилки визначення координат через особливості місцевості, зокрема, в підвальних приміщеннях, серед пагорбів і т.д. . У числових рядах можуть зустрічатися значення, що виходять за межі можливого діапазону, наприклад, цифра 7 в п'ятибальній шкалі оцінок. Також числові значення вихідних даних можуть сильно варіюватися за абсолютними величинами: від декількох сотих відсотків до десятків тисяч одиниць. Подібні похибки спотворюють результати моделювання і не дозволяють отримати модель машинного навчання з прийнятною якістю.

Перед тим, як застосовувати алгоритми машинного навчання, дані необхідно перетворити в табличне представлення, яке найбільш поширене в Machine Learning і Data Mining. Отримавши файл з вхідними даними необхідно зрозуміти характер записів (рядків), а також сенс, тип і діапазон значень ознак (стовпців). Потім Data scientist формує вибірку – dataset - відбирає дані, які потенційно мають відношення до перевіряється гіпотези машинного навчання.

Для числових змінних застосовується нормалізація даних, щоб привести їх до однакової області зміни і використовувати їх разом в одній моделі Machine Learning. Як правило, нормалізація даних означає перетворення вихідних числових значень в нові в діапазоні від 0 до 1, ґрунтуючись на початковому мінімумі і максимумі.

Набір методів, що використовуються до застосування методу інтелектуального аналізу даних, називається попередньою обробкою даних для інтелектуального аналізу даних , це одна з найбільш значущих проблем в рамках відомого відкриття знань з процесу обробки даних , як показано на рисунку 3.1.

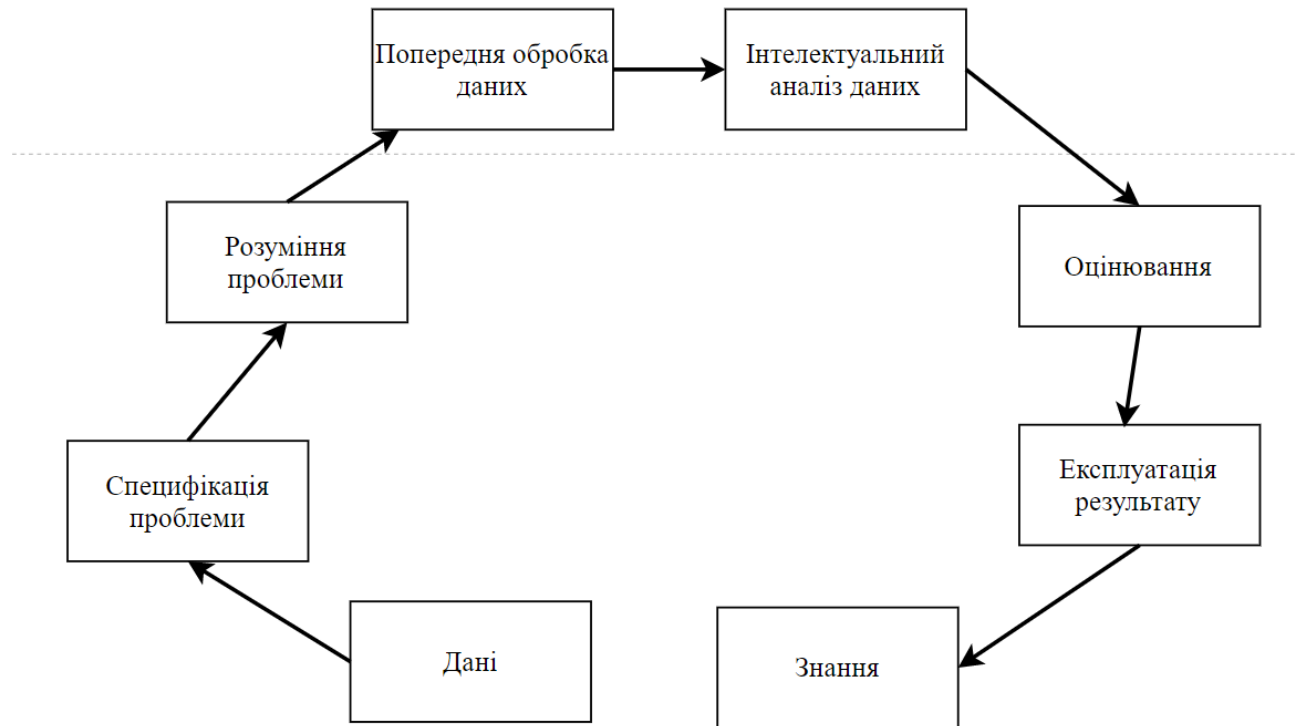


Рисунок 3.1 – Процес обробки даних

Оскільки дані, швидше за все, будуть недосконалими, будуть містити невідповідності не застосовані для запуску процесу інтелектуального аналізу даних.

Слід також зазначити швидке зростання темпів генерації даних та їх розмірів у бізнесі, промисловості, академічній та соціальній сферах.

Великі обсяги зібраних даних вимагають більш складних механізмів для їх аналізу. Попередня обробка даних дозволяє адаптувати дані до вимог, пропонованим кожним алгоритмом інтелектуального аналізу даних, що дозволяє обробляти дані різних типів та розмірів.

Не дивлячись на те, що попередня обробка даних є потужним інструментом, що дозволяє користувачеві обробляти складні дані, цей процес може займати велику кількість часу.

3.2 Послідовність побудови процесної моделі

Алгоритм зображено на рисунку 3.2. В ньому показано виконання етапу 5, а саме : формування набору підмножин подій журналу за класифікацією

частоти виникання у процесах. Основна ідея запропонованого етапу – агрегація та класифікація подій за відносною частотою вживання у процесі [34].

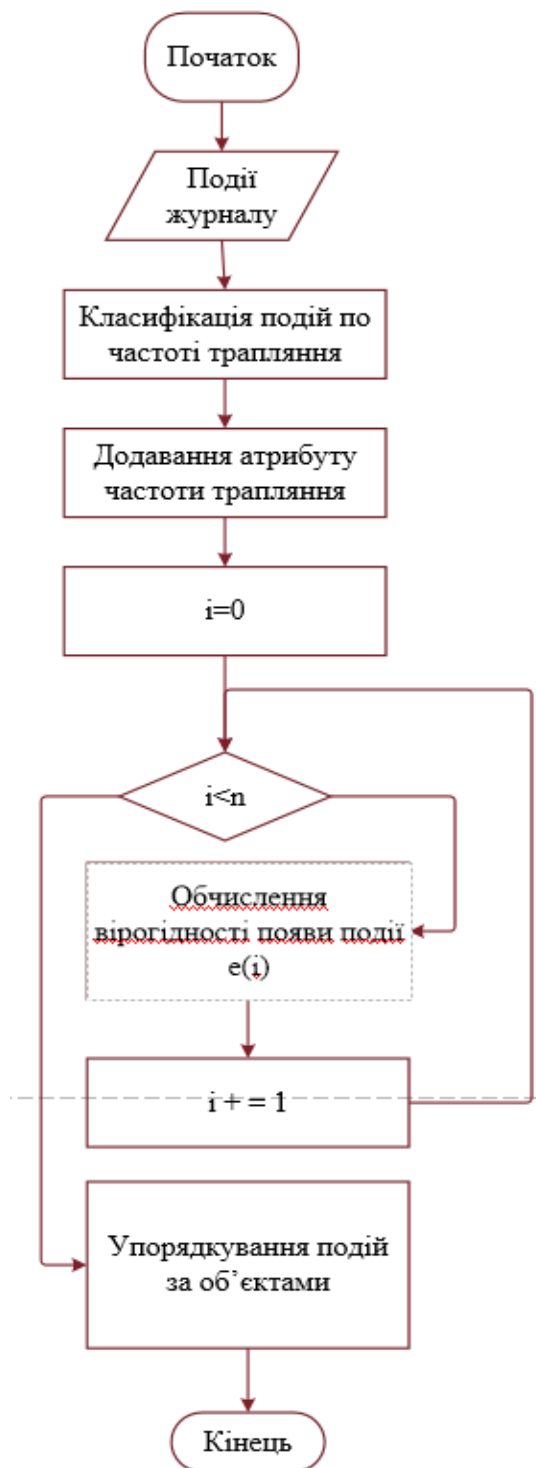


Рисунок 3.2 – Алгоритм роботи

При виконнанні алгоритму спочатку відбираються події журналу, що відображають вирішення завдань при заданих даних.

Додатково відбіраються підмножини подій логу із урахуванням обмеження за типами суб'єктів. Іноді використовується обмеження за розміром процесу.

Потім формуються набори підмножин подій обмежень вибраної предметної області.

Далі формування набори підмножин подій журналу за класифікацією частоти виникання у процесах. Основна ідея запропонованого етапу – класифікація подій за частотою вживання у процесі.

Потім додається атрибут частоти трапляння події, яка позначається "i"

Далі виконується підрахунок кількості спільних атрибутів, який позначається "e(i)"

Потім виконується сортування атрибутів по частоті вживання та кількості спільних атрибутів.

4 ПРАКТИЧНЕ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

4.1 Програмна реалізація удосконаленого методу

Для удосконаленого методу був написан код на мові Java і реалізован за допомогою Visual Studio cod. Частина програми показана на рисунку 4.1

Формування набору підмножин подій журналу за класифікацією частоти виникання у процесах. Основна ідея запропонованого етапу – класифікація подій за частотою вживання у процесі.

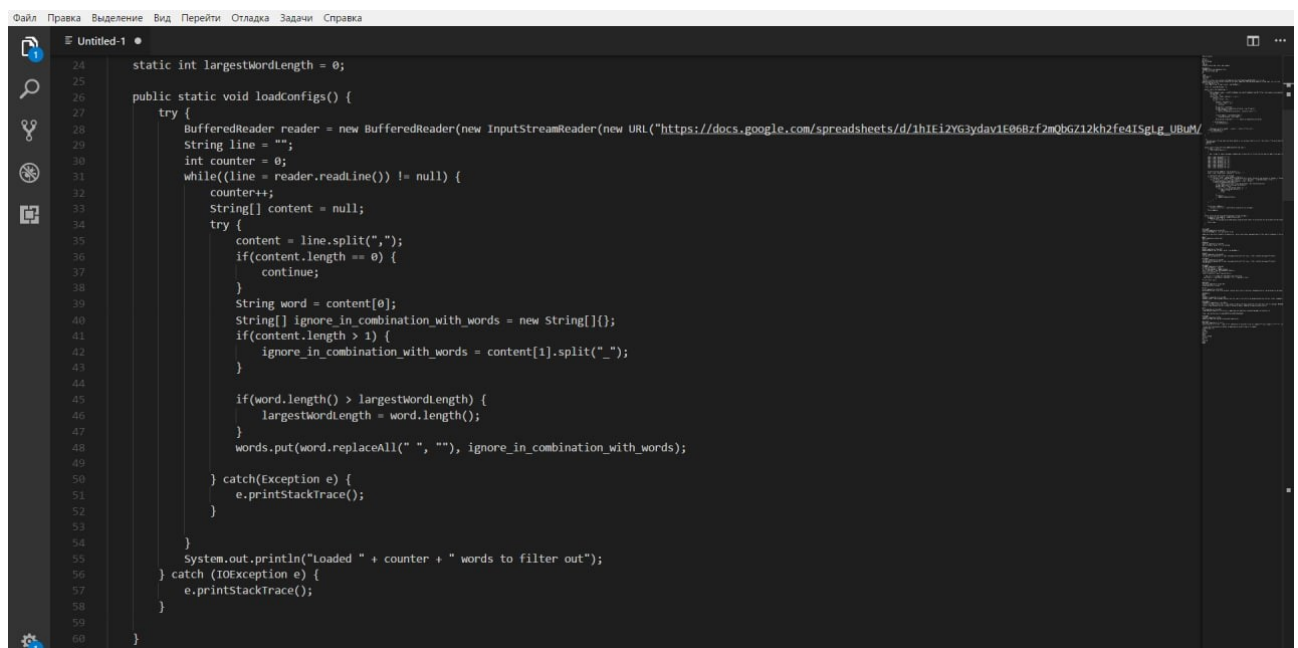


Рисунок 4.1 – Класифікація подій по частоті трапляння

Удосконалений метод реалізовано з використанням платформи ProM (що скорочено від process Mining framework).

У той час як існує безліч систем, що підтримують виконання таких процесів, існуюча практика моніторингу та аналізу цього виконання в організаційній реальності все ще залишає бажати кращого.

Process mining надає революційні засоби для аналізу та моніторингу процесів підприємства. Дослідження інтелектуального аналізу процесів пов'язане з отриманням знань про бізнес-процес з журналів його виконання.

ProM-це розширюваний фреймворк, який підтримує широкий спектр методів інтелектуального аналізу процесів у вигляді плагінів. Він не залежить від платформи [35].

В ході виконання робробленого проекту був створений програмний додаток до найбільш вживаної програми інтелектуального аналізу даних PROM.

Основне меню програми ProM 6.0, показано на рисунку 4.2.



Рисунок 4.2 – Робочий простір ProM 6

Після цього ми загрузили журнал подій, натиснувши кнопку Імпорт, і вибрали збережену копію файлу журналу для нашої задачі.

Як тільки файл був додан у вигляді журналу подій. Ми отримали екран, показаний на рисунку 4.3.

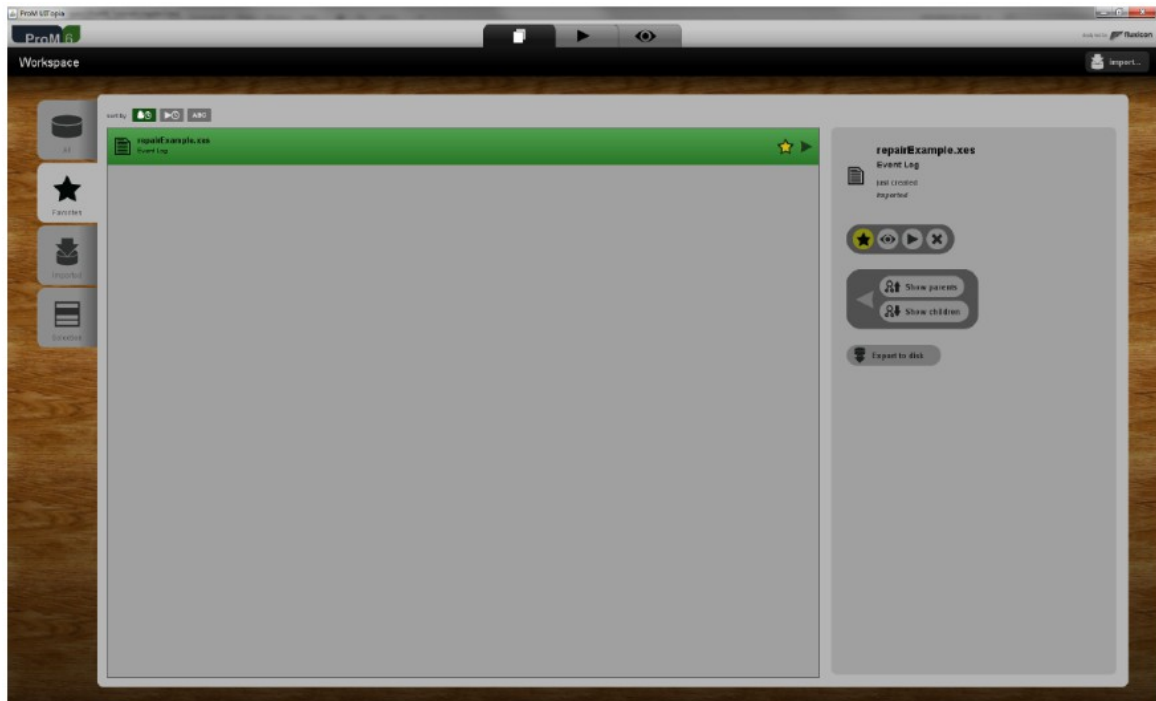


Рисунок 4.3 – Приклад журналу був імпортований

Коли журнал був імпортований, ми приступили до фактичної перевірки журналу. Щоб переглянути ми вибрали запис "repairExample.xes" у робочій області і натиснули кнопку "візуалізувати", це показано на рисунку 4.4.

Log Summary

Total number of events: 11858

MXML Legacy Classifier

Event classes defined by MXML Legacy Classifier

All events

Total number of classes: 12

Class	Occurrences (absolute)	Occurrences (relative)
Test Repair-complete	1518	12.72%
Test Repair-start	1518	12.72%
Register-complete	1114	9.313%
Analyze Defect-complete	1114	9.313%
Analyze Defect-start	1114	9.313%
Inform User-complete	1112	9.296%
Archive Repair-complete	1010	8.435%
Repair (Simple)-complete	781	6.622%
Repair (Simple)-start	781	6.622%
Repair (Complex)-start	721	6.116%
Repair (Complex)-complete	724	6.107%
Restart Repair-complete	481	4.020%

Start events

Total number of classes: 1

Class	Occurrences (absolute)	Occurrences (relative)
Register-complete	1114	100.0%

End events

Total number of classes: 5

Class	Occurrences (absolute)	Occurrences (relative)
Archive Repair-complete	1010	96.75%
Test Repair-complete	15	0.753%
Inform User-complete	27	2.462%
Repair (Complex)-start	1	0.001%
Repair (Complex)-complete	1	0.001%

Рисунок 4.4 – Зведений вигляд на прикладі журналу подій

Потім ми вибрали запис "repairExample.xes" у робочій області та натиснули "Action".

Це відкрило вид дії з попередньо вибраним журналом. Подано тільки дії, що приймають журнал в якості вхідних даних. Вони пофарбовані в зелений колір (Доступні всі вхідні дані), а дії, що вимагають додаткових вхідних даних, пофарбовані в жовтий колір (Доступні деякі, але не всі вхідні дані). Це показано на рисунку 4.5

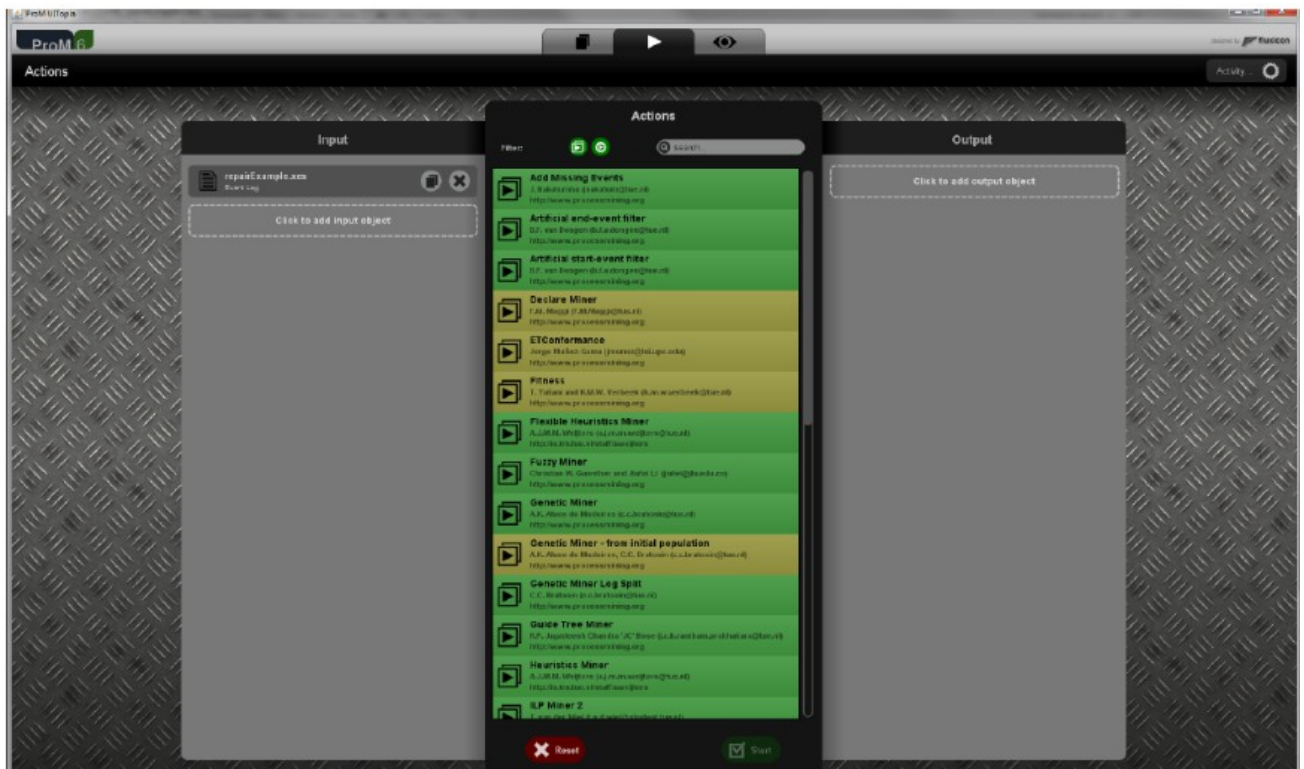


Рисунок 4.5 – Представлення дій з обраним прикладом журналу

Третій фільтр-це фільтр кінцевих подій, який фільтрує показаний на рисунку 4.6 журнал таким чином, що зберігаються тільки ті трасування (або випадки), які закінчуються зазначеними завданнями.

Повзунок внизу дозволяє вибрати найбільш часті кінцеві події.

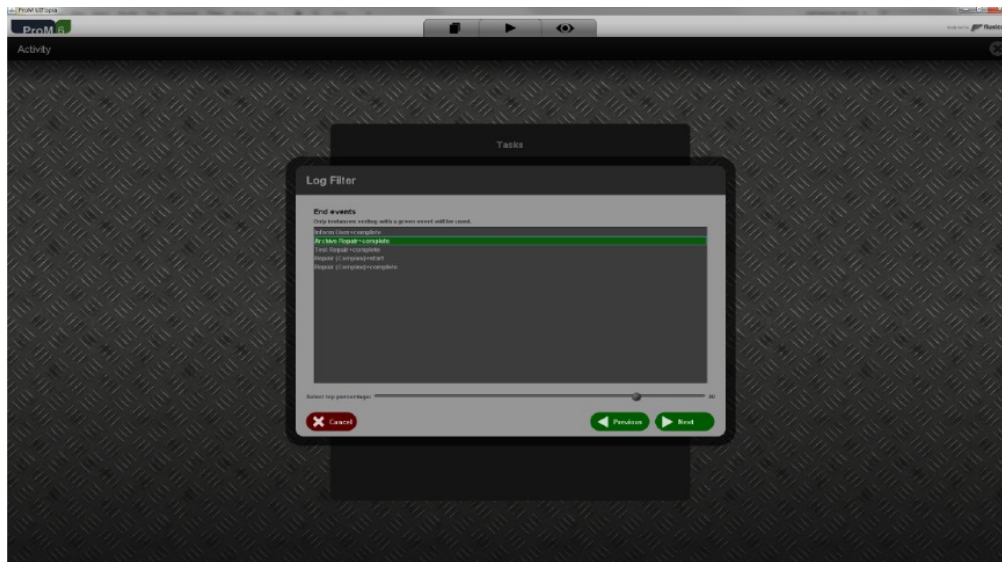


Рисунок 4.6 – Вибір кінцевих подій.

На рисунку 4.7 показано, що більше 80% трас закінчуються словами Archive repair+complete.

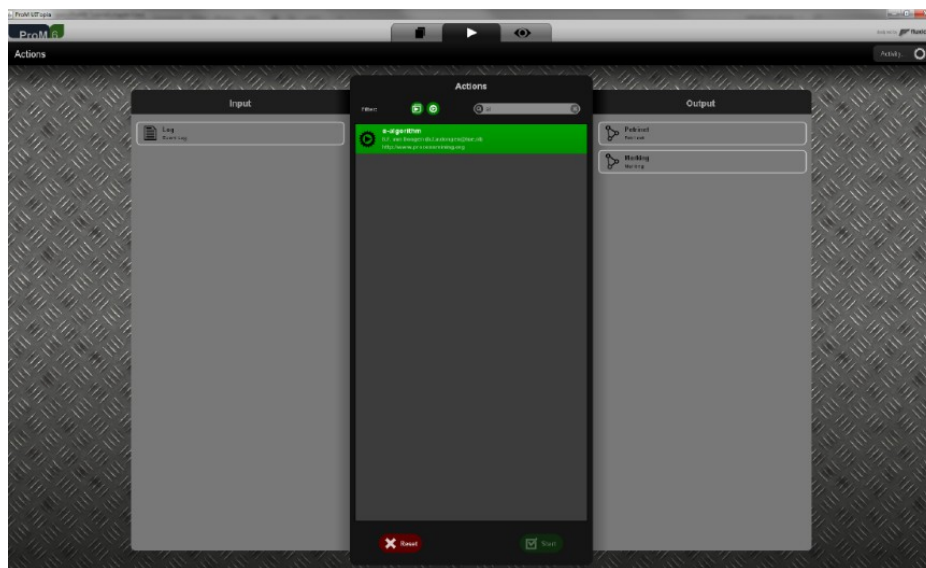


Рисунок 4.7 – Дія алгоритму

Отримана модель без доданого додатку показана на риунку 4.8

Після додавання додатку модель має такий вигляд, який показано на рисунку 4.9

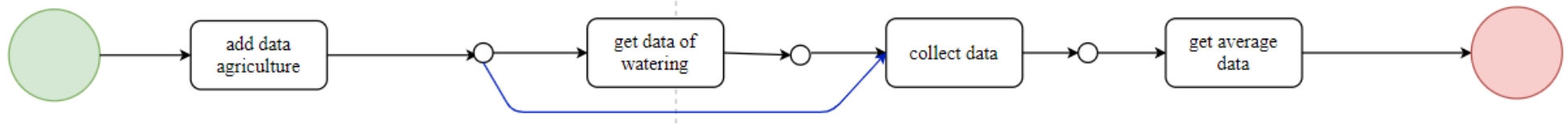


Рисунок 4.8 – Приклад побудови моделі без використання розробленого додатку

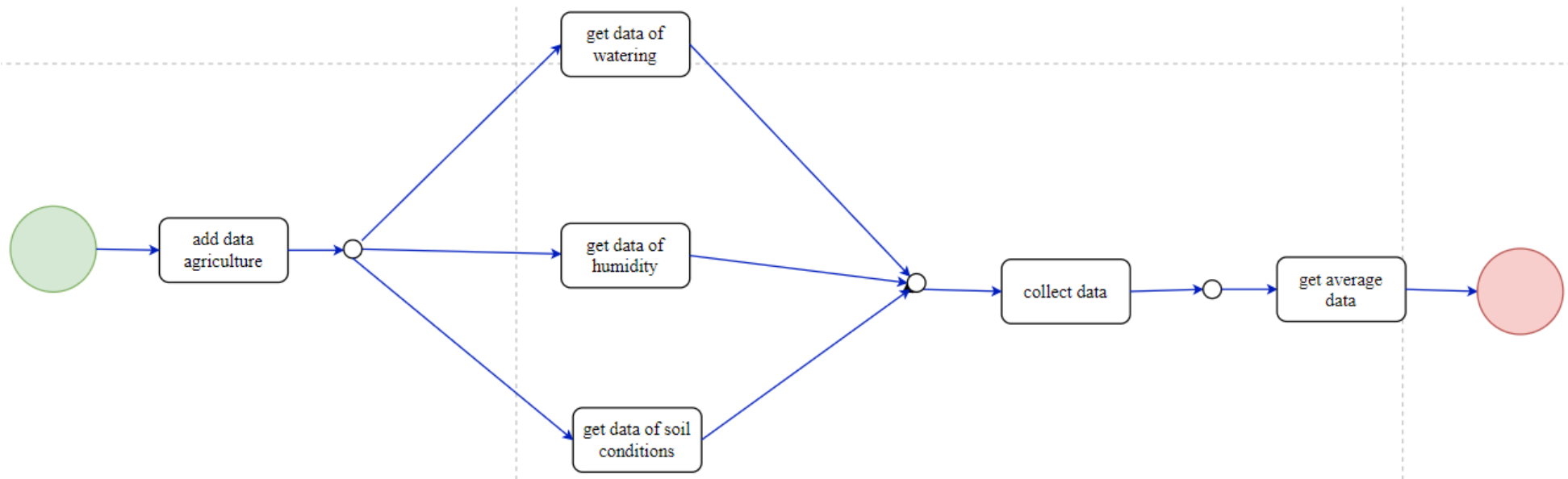


Рисунок 4.9 – Приклад побудови моделі після використання розробленого додатку

4.2 Експериментальна перевірка отриманих результатів

Ми можемо порівняти результати експерименту для різних методів, як показано в таблиці 4.2

Таблиця 1.2. - Порівняння отриманих результатів

	Без використання модифікованого методу, відсоток подій що включений до складу моделі	Із використанням модифікованого методу, відсоток подій що включений до складу моделі
лог 1	70%	97%
лог 2	35%	90%

Після виконання експерименту можна побачити, що із використанням модифікованого методу, відсоток подій, що включений до складу моделі значно виріс.

Виявлено збільшення відсоткової кількості подій у 2.5 рази на 3 лог.

Логи- це процеси у сільськогосподарській інформаційній системі. При виконанні побудови моделі з використанням розробленого додатку були додані:

- get data of watering ;
- get data of humidity;
- get data of soil conditions;

ВИСНОВКИ

В магістерській атестаційній роботі було досліджено проблему реалізації процесного управління в сільськогосподарських інформаційних системах. При переході до процесного управління використовується моделі процесів аграрних робіт. Побудова моделей таких процесів потребує виконання попередньої обробки даних.

В результаті виконання магістерської роботи був виконаний огляд методів попередньої обробки даних у сільськогосподарських інформаційних системах.

Був проведений аналіз особливостей сільськогосподарських інформаційних систем та досліджено методи process mining, які в них застосовуються.

Було досліджено методи структуризації вхідних даних у задачах process mining.

Було проведено аналіз вимог до вхідних даних, методи структуризації вхідних даних, процес конвертації вхідних даних до формату XES, послідовність перетворення вхідних даних.

Було удосконалено метод попередньої обробки неструктурованих даних у сільськогосподарських інформаційних системах .

Метод реалізовано у вигляді програмного додатку, що виконує попереднє упорядкування вхідних даних для побудови процесної моделі сільськогосподарського виробництва засобами process mining. Виконано експериментальну перевірку методу з використанням платформи ProM. Результати показали, що із використанням модифікованого методу, відсоток подій, що включений до складу моделі значно виріс. Практичним результатом стало підвищення точності побудови процесних моделей на неструктурованих даних.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Методичні вказівки щодо розробки та оформлення магістерської атестаційної роботи за спеціальністю 8.05010101 – Інформаційні управляючі системи та технології. Освітньо-кваліфікаційний рівень – магістр / Упоряд.: Левикін В.М., Міхнов Д.К., Саєнко В.І., Євланов М.В., Міхнова А.В., Керносов М.А. – Харків: ХНУРЕ, 2012. – 28 С.
2. Aggarwal CC. Data Mining: The Textbook. Berlin, Germany: Springer; 2015. 2. Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. IEEE Trans Knowl Data Eng. 2014;26(1):97–107.
3. Fernández A, del Río S, López V, Bawakid A, del Jesús MJ, Benítez JM, et al. Big data with cloud computing: an insight on the computing environment, mapreduce, and programming frameworks. Wiley Interdisc Rew Data Min Knowl Discov. 2014;–409 С.
4. Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. In: OSDI 2004. San Francisco, CA; 2004, – 137 С.
5. White T. Hadoop, The Definitive Guide. Sebastopol: O'Reilly Media, Inc; 2012– 28 С.
6. Karau H, Konwinski A, Wendell P, Zaharia M. Learning Spark: Lightning-Fast Big Data Analytics. Sebastopol: O'Reilly Media; 2015;–409 С.
7. García S, Luengo J, Herrera F. Data Preprocessing in Data Mining. Berlin: Springer; 2015;–409 С.
8. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques, 3rd ed. Burlington: Morgan Kaufmann Publishers Inc; 2011;–409 С.
9. Zaki MJ, Meira W. Data Mining and Analysis: Fundamental Concepts and Algorithms. New York: Cambridge University Press; 2014;–121 С.
10. Wang H, Wang S. Mining incomplete survey data through classification. Knowl Inf Syst. 2010;–49 С.

11. Чалый С. Ф. Модель бизнес-процесса с изменяемой на основе правил структурой / С. Ф. Чалый, И.Б. Буцукина// АСУ и приборы автоматики : збірник наукових праць. – Х. : ХНУРЭ, 2012. – № 161. – С. 103–105.
12. Чалый С. Ф. Сервис-ориентированная модель бизнес-процесса с изменяемой структурой /С. Ф. Чалый, Али Джамиль Альшейх // АСУ и приборы автоматики : збірник наукових праць. – Х. : ХНУРЭ, 2012. – № 161. – С. 106–108.
13. Чалый С. Ф. Формирование набора ограничений в задаче предварительной обработки неструктурированных журналов регистрации событий/ С. Ф. Чалый, Е. О. Богатов, А. Ю. Кальнищкая// Вісник Національного технічного університету "Харківський політехнічний інститут" : збірник наукових праць. – Серія : Інформатика та моделювання. – Х. : НТУ "ХПІ", 2013. – № 2 (976). – С. 139–148.
14. Левыкин В. М. Модели представления журналов регистрации событий в задачах интеллектуального анализа процессов / В. М. Левыкин, С. Ф. Чалый, Е. О. Богатов // Вісник Академії митної служби України. Сер. : Технічні науки. - 2013. - № 1. - С. 25–32.
15. Чалый С. Ф. Предварительная структуризация исходных данных в задачах интеллектуального анализа процессов/ С. Ф. Чалый, Е. О. Богатов// Известия Волгоградского государственного технического университета. Вып. №22 (125). - Том 18. – (Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах", 2013. – С. 86–90.
16. Чалый С. Ф. Модель бизнес-процесса с изменяемой на основе правил структурой / С. Ф. Чалый, И.Б. Буцукина// АСУ и приборы автоматики : збірник наукових праць. – Х. : ХНУРЭ, 2012. – № 161. – С. 103–105.

17. Luengo J, García S, Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowl Inf Syst.* 2013. – C. 86–90.
18. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, 1st ed. New York: Wiley; 2013. – C. 86–90.
19. Frénay B, Verleysen M. Classification in the presence of label noise: A survey. *IEEE Trans Neural Netw Learn Syst.* 2012. – 28 C.
20. Zhu X, Wu X. Class Noise vs. Attribute Noise: A Quantitative Study. *Artif Intell Rev.* 2002. – 28 C.
21. Bellman RE. *Adaptive Control Processes - A Guided Tour*. Princeton, NJ: Princeton University Press; 1961.– 28 C.
22. Hall MA. Correlation-based feature selection for machine learning. Waikato University, Department of Computer Science. 1999. – 28 C.
23. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* 2003. – 128 C.
24. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng.* 2014. – 218 C.
25. Kim JO, Mueller CW. *Factor Analysis: Statistical Methods and Practical Issues (Quantitative Applications in the Social Sciences)*. New York: Sage Publications, Inc; 1978. – 228 C.
26. Duntelman GH. *Principal Components Analysis*. A Sage Publications. Thousand Oaks: SAGE Publications; 1989. – 22 C.
27. Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding. *Science.* 2000. – 28 C.
28. Tenenbaum JB, Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science.* 2000. – 218 C.
29. Liu H, Motoda H. On issues of instance selection. *Data Min Knowl Disc.* 2002. – 128 C.
30. Olvera-López JA, Carrasco-Ochoa JA, Martínez-Trinidad JF, Kittler J. A review of instance selection methods. *Artif Intell Rev.* 2010. – 24 C.

31. García S, Derrac J, Cano JR, Herrera F. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Trans Pattern Anal Mach Intell.* 2012. – 224 C.
32. Triguero I, Derrac J, García S, Herrera F. A taxonomy and experimental study on prototype generation for nearest neighbor classification. *IEEE Trans Syst Man Cybern Part C.* 2012. – 243 C.
33. Liu H, Hussain F, Tan CL, Dash M. Discretization: An enabling technique. *Data Min Knowl Discov.* 2002. – 214 C.
34. Wu X, Kumar V, (eds). *The Top Ten Algorithms in Data Mining.* Boca Ratón, Florida: CRC Press; 2009. 38. Quinlan JR. *C4.5: Programs for Machine Learning.* San Francisco, CA: Morgan Kaufmann Publishers Inc.; 1993. – 222 C.
35. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: *Proceedings of the 20th Very Large Data Bases Conference (VLDB)*; 1994. – 27 C.