МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

АДЖІТ ПРАТАП СІНГХ ГАУТАМ

УДК 004.912:007.51

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ЕКСТРАКЦІЇ БІЗНЕС ЗНАНЬ З ТЕКСТОВОГО КОНТЕНТУ ІНТЕГРОВАНОЇ КОРПОРАТИВНОЇ СИСТЕМИ

Спеціальність 05.13.06 – інформаційні технології

Автореферат дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2016

Дисертацією є рукопис.

Робота виконана на кафедрі інтелектуальних комп'ютерних систем Національного технічного університету «Харківський політехнічний інститут» Міністерства освіти і науки України.

Науковий керівник доктор технічних наук, професор
Шаронова Наталія Валеріївна,
Національний технічний університет «Харківський політехнічний інститут», завідувач кафедри інтелектуальних комп'ютерних систем

Офіційні опоненти: доктор технічних наук, професор
Асєєв Георгій Георгійович,
Харківська державна академія культури, завідувач кафедри інформаційних технологій

кандидат технічних наук, доцент
Чала Лариса Ернестівна,
Харківський національний університет радіоелектроніки, доцент кафедри штучного інтелекту

Захист відбудеться 3 листопада 2016 р. о 13.00 годині на засіданні спеціалізованої вченої ради Д 64.050.07 в Національному технічному університеті «Харківський політехнічний інститут» за адресою: 61002, м. Харків, вул. Багалія, 21.

З дисертацією можна ознайомитись у бібліотеці Національного технічного університету «Харківський політехнічний інститут» за адресою: 61002, м. Харків, вул. Багалія, 21.

Автореферат розісланий «_____» 2016 р.

Вчений секретар

спеціалізованої вченої ради  Дорофєєв Ю. І.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Інформаційний простір підприємства включає у себе не тільки явним чином визначені знання й структуровані дані, які представлені у базах даних, але й у більшості неструктуровану інформацію, яку представлено у різноманітних текстових документах корпорації. Дослідження свідчать, що від 80 % до 85 % інформації, критичної для прийняття бізнес рішень, зберігається саме у неструктурованій формі, в основному у вигляді текстів. У зв'язку з цим у додаток до методів, що традиційно використовуються для підтримки прийняття бізнес рішень, таких як аналіз часових рядів цін акцій, методи фундаментального аналізу, що базуються на звітній інформації поточного фінансового й економічного стану компаній, потужним додатковим засобом прийняття бізнес рішень стають знання, які вилучаються з текстів відповідної проблематики.

У наукові дослідження, що спрямовані на ідентифікацію знань у слабоформалізованій текстовій інформації, значний внесок зробили видатні вчені Ю. Апресян, Г. Бєлоногов, М. Бондаренко, М. Вілкс, Т. Віноград, Т. Гаврилова, В. Глушков, А. Жолковський, І. Мельчук, М. Мінський, О. Палагін, Р. Піотровський, Д Поспєлов, Ч. Филлмор, Н. Хомський, Ю. Шабанов-Кушнаренко, Р. Шенк, В. Широков та ін.

Кількість пропонованих підходів до розв'язання задач ідентифікації та екстракції знань з текстів постійно зростає. Однак, у переважній більшості застосовані на сьогоднішній день технології обробки текстової інформації базуються на традиційних статистико-імовірнісних підходах, що не використовують або слабо використовують опрацювання смислу. У зв'язку з цим системи, які автоматизують процес вилучення екстракції знань з великих обсягів текстових даних, що представлені у інтегрованих корпоративних системах, мають достатньо низьку повноту й точність екстракції знань з текстового контенту.

Усе вищезазначене обумовлює актуальність вирішення науково-практичного завдання створення інформаційної технології екстракції бізнес знань інтегрованої корпоративної системи на базі використання моделей та методів смислового опрацювання текстового контенту.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана на кафедрі інтелектуальних комп'ютерних систем НТУ «ХПІ» у межах держбюджетних тем МОН України: «Розробка математичних моделей та методів розв'язання задач інтелектуальної обробки інформації» (ДР № 0108U003926), «Розробка моделей та методів для інформаційно-пошукових, лексикографічних інтелектуальних систем» (ДР № 0111U002258), у яких здобувач брав участь як виконавець.

Мета і задачі дослідження. Метою дисертаційної роботи є створення інформаційної технології екстракції бізнес знань інтегрованої корпоративної системи на основі інформаційно-логічних моделей і методів смислового опрацювання текстового контенту.

Відповідно до зазначеної мети поставлено задачі:

– проаналізувати існуючі інформаційні технології, моделі й методи екстракції та ідентифікації знань з текстів і сформулювати основні вимоги до розробки інформаційного забезпечення підсистеми екстракції бізнес знань з текстового контенту інтегрованої корпоративної системи (ІКС);

– дослідити можливість використання інструментарію алгебри скінченних предикатів у інформаційно-логічних моделях екстракції фактів з текстових потоків;

– побудувати математичну модель генерації фактів з текстових потоків корпорації;

– розробити метод виявлення актуальної множини класифікованих сутностей предметної області;

– реалізувати метод компарації для структурування відношень фактів бізнес знань ІКС;

– удосконалити інформаційну технологію формування єдиного інформаційного простору представлення ресурсів бізнес діяльності корпорації;

– впровадити результати дисертаційного дослідження у практику створення підсистем екстракції знань з текстового контенту ІКС.

Об'єктом дослідження є процес екстракції знань з текстового контенту інтегрованої корпоративної системи.

Предметом дослідження є інформаційна технологія екстракції бізнес знань з текстового контенту інтегрованої корпоративної системи, яка заснована на смисловій обробці тексту.

Методи досліджень базуються на використанні теорії інтелекту, алгебри скінченних предикатів і методу компараторної ідентифікації, які комплексно застосовуються при створенні інформаційно-логічних моделей смислової обробки і методів екстракції бізнес знань з текстового контенту. Використовується математичний апарат логічної алгебри, який розроблено науковою школою теорії інтелекту професорів Шабанова-Кушнаренка Ю. П. і Бондаренка М. Ф. Алгебра предикатів застосована для формалізації знань та опису природномовних відношень; метод компараторної ідентифікації використаний для опису процедури формування інформаційного простору інтегрованої корпоративної системи. Для підтвердження достовірності отриманих результатів використаний метод текстових колекцій та методи математичної теорії вибірки.

Наукова новизна отриманих результатів:

– *уперше* розроблено логіко-лінгвістичну модель генерації фактів з текстових потоків інформаційної корпоративної системи, яка базується на використанні поверхневих граматичних характеристик сутностей, предикатів та атрибутів, що дозволяє ефективно екстрагувати з текстового контенту профільні знання про суб'єкти моніторингу;

- *уперше* розроблено метод створення інформаційного простору фактів інтегрованої корпоративної системи, заснованої на побудові ієрархічної структури кластерів фактів, яка базується на гіпонімічних відношеннях більш високого порядку узагальнення, що дозволяє структурувати вилучені знання про економічну діяльність корпорації за видами продукції, галузями виробництва, геог-

рафічному включенню тощо;

– *отримав подальший розвиток* метод компараторної ідентифікації, який використано для структурування відношень фактів бізнес знань ІКС, реалізація якого дозволяє класифікувати атрибути сутностей за класами відношень за рахунок смислової тотожності триплетів фактів, що об'єктивно визначено компаратором;

– *удосконалено* метод виявлення актуальної множини класифікованих сутностей предметної області, який відрізняється комплексним використанням лінгвістичних, статистичних й смислових характеристик в наївному байєсівському класифікаторі, що дозволяє класифікувати сутності, які екстрагуються з тексту, за апріорно виділеними типами;

– *удосконалено* інформаційну технологію формування єдиного інформаційного простору бізнес діяльності корпорації, яка дозволяє за рахунок використання алгебро-логічних перетворень здійснювати породження складного знання шляхом експліцитного узагальнення інформації, що прихована у сукупності часткових фактів.

Практичне значення одержаних результатів полягає у розробці технології формування єдиного інформаційного простору бізнес діяльності корпорації. Розроблена технологія включає логіко-лінгвістичну модель генерації фактів з текстових потоків ІКС, метод структурування відношень фактів бізнес знань корпорації, метод виявлення актуальної множини класифікованих сутностей предметної області, а також спеціалізовані етапи Web Content Mining лінгвістичного процесора. Запропоновані у дослідженні математичні моделі можуть бути використані у системах автоматичного опрацювання текстів, системах вилучення знань, добування інформації (Information Extraction) і розпізнавання сутностей (Named Entity Recognition).

Результати дослідження імплементовані у веб-додаток, який здійснює моніторинг фактографічної інформації заздалегідь визначених компаній та корпорацій. Практична реалізація результатів дослідження знайшла застосування у підсистемах формування фактографічних баз даних й фактографічного пошуку наукових бібліотек НТУ «ХПІ» і Харківського національного університету радіоелектроніки. Використання розроблених у роботі моделей і методів дозволило підвищити ефективність технологій екстракції бізнес знань з текстового контенту за рахунок підвищення середніх значень коефіцієнтів повноти й точності видачі фактографічної інформації.

Теоретичні результати дисертації використовуються в навчальному процесі на кафедрі інтелектуальних комп'ютерних систем НТУ «ХПІ» при викладанні спеціальних дисциплін «Інформаційно-ресурсне забезпечення лінгвістичної діяльності», «Штучний інтелект: лінгвістичні проблеми», «Автоматизована обробка природної мови» для студентів спеціальності «Прикладна лінгвістика» та при виконанні курсових й дипломних робіт.

Особистий внесок здобувача. Усі основні результати дисертаційної роботи, що виносяться на захист, отримані здобувачем особисто, серед них: технологія формування єдиного інформаційного простору представлення ресурсів бізнес діяльності корпорації; математична модель генерації фактів з текстових

потоків інформаційної корпоративної системи; метод виявлення актуальної множини класифікованих сутностей предметної області; метод компарації для структурування відношень фактів бізнес знань ІКС.

Апробація результатів дисертації. Результати дисертаційної роботи доповідались та обговорювались на: Міжнародній конференції Applicable Information Models: Joint International Scientific Events on Informatics (Варна, Болгарія, 2015); III, IV, V Всеукраїнських науково-практичних конференціях «Інтелектуальні системи і прикладна лінгвістика» (Харків, 2014, 2015, 2016); Міжнародній конференції Managing Innovation in Business & Technology (Варанасі, Індія, 2011); X Міжнародній конференції International scientific and technical conference «Computer science and information technologies CSIT» (Львів, 2015).

Публікації. Основні результати дисертації опубліковані у 14 наукових працях, серед яких 4 статті у фахових наукових виданнях України, 6 статей у іноземних періодичних фахових виданнях, 4 – у матеріалах конференцій.

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Повний обсяг дисертації складає 153 сторінки, з них 14 рисунків по тексту, 5 таблиць по тексту, списку з 137 найменувань використаних джерел на 17 сторінках, додатки на 8 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертації, зазначено зв’язок роботи з науковими темами, сформульовано мету і задачі дослідження, визначено об’єкт, предмет і методи дослідження, показано наукову новизну та практичне значення отриманих результатів, наведено інформацію про практичне використання, апробацію результатів та їх висвітлення у публікаціях.

У **першому розділі** на основі аналітичного огляду інформаційних джерел проаналізовано існуючі задачі і проблеми в області автоматизації корпоративних систем. ІКС управління територіально розподіленою корпорацією засновані на поглибленому аналізі даних, широкому використанні систем інформаційної підтримки прийняття рішень, електронному документообігу і діловодстві. Вказані системи, які закликані поєднати стратегію управління підприємством та провідні інформаційні технології, засновані на єдиній програмно-апаратній платформі й спільній базі знань.

Відзначено, що у явному вигляді бізнес знання корпорації представляють собою доступні знання, які використовуються для підвищення ефективності роботи даної корпорації. Завдання сучасної ІКС – накопичувати структуровані формалізовані знання, які дозволяють повторно вирішувати реальні виробничі та організаційні завдання на рівні всієї корпорації. У той же час інформаційний простір підприємства, структуру якого представлено на рис. 1, включає в себе не тільки явним чином визначені знання, але й структуровані дані, представлені у базах даних, а також неструктуровану текстову інформацію, яка міститься у документах корпорації. Якщо компанія не має відповідних технологій екстрак-

ції та актуалізації знань з текстів, корисні дані втрачаються.

Проведений аналітичний огляд існуючих технологій, підходів і методів екстракції знань з текстового контенту показав, що вилучення інформації є інтелектуальним процесом і доцільно розширити застосування існуючих статистико-позиційних методів. Практичними нагальними завданнями, які стоять перед системами витягу знань з текстів на природній мові є: (1) пошук і витяг елементів знань, явно присутніх у текстовій колекції у вигляді тверджень або фактів; (2) експліціювання узагальненого знання, прихованого у сукупності часткових тверджень і/або фактів.

Сучасні процедури витягу знань, які використовують методи обробки текстів на природній мові, як правило, спрямовані лише на розв'язання дуже вузького класу задач (відбір обмеженого набору тем (питань, проблем), а найчастіше тільки однієї теми). Це призводить до того, що розроблені процедури опрацювання тексту не можуть бути застосовані у якості універсального засобу витягу знань у ІКС.

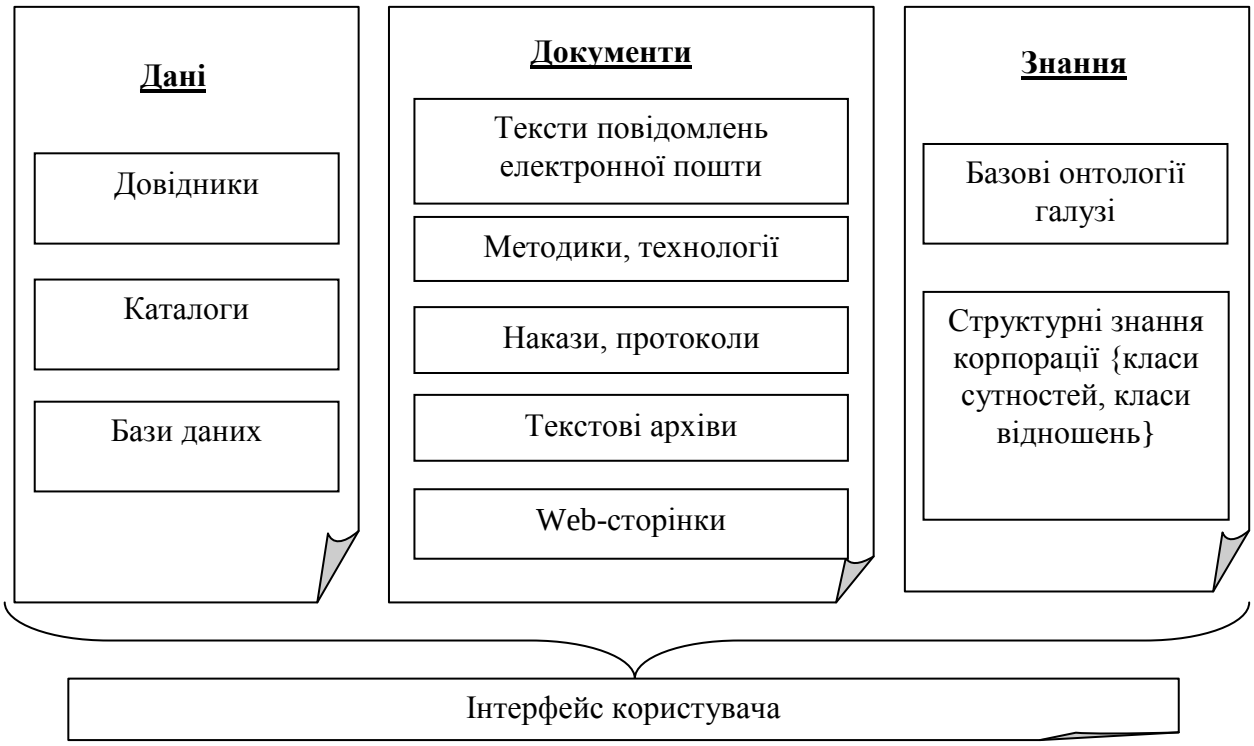

Рисунок 1 – Структура інформаційного простору інтегрованої корпоративної системи

На підставі критичного аналізу існуючих невирішених проблем та задач вилучення знань з різноманітних текстів корпорації сформульовано й обґрунтовано доцільність дослідження моделей та методів інтелектуальної екстракції бізнес знань з текстового контенту для впровадження їх в автоматизовані ІКС.

У другому розділі дисертації обґрунтовано вибір математичних засобів для формалізації процесу витягу знань з текстового контенту ІКС. У якості математичного апарату опису дискретних, детермінованих і скінченних об'єктів системи представлення знань, добутих з текстового контенту ІКС, запропоно-

вано використання алгебри скінченних предикатів (АСП), методу компараторної ідентифікації та засобів теорії логічних мереж.

Для побудови базового апарату інформаційної технології водиться універсум елементів *U*, який включає: (1) об'єкти текстової інформації ІКС (довідки, виписки, текстові звіти, листи електронної пошти, тексти веб-сторінок і т. п.), (2) суб'єкти та об'єкти фактів предметної області, (3) граматичні і семантичні характеристики лінгвістичних одиниць тексту. З елементів універсуму у відповідності до конкретної задачі обробки інформації утворюється m підмножин: M_1, M_2, …, M_m. Декартовий добуток $S = M_1 \times M_2 \times \times \ldots \times M_m$ називається предметним простором S з координатними предметними осями M_1, M_2, …, M_m над універсумом U. Кількість осей m називається розмірністю простору S. Вводиться множина $V = \{x_1, x_2, \ldots, x_m\}$ предметних змінних простору S. Значеннями змінної x_i, $i = \overline{1,m}$ служать елементи множини M_i такі, що $x_1 \in M_1$, …, $x_m \in M_m$, тобто множини M_1, M_2, …, M_m є областями визначення змінних x_1, x_2, …, x_m. На декартовому добутку $M_1 \times M_2 \times \ldots \times M_m$ визначаються предикати, які характеризують роботу моделі. У ролі базисного елемента АСП виступає предикат x_i^a впізнавання предмета a по змінній x_i

$$x_i^a = \begin{cases} 1, якщо\ x_i = a, \\ 0, якщо\ x_i \neq a. \end{cases}$$

При реалізації методу компараторної ідентифікації для роботи з інформаційними об'єктами текстового контенту вводяться: множина токенів $P = \{p_i\}$, $1 \le i \le k$, які визначають суб'єкти бізнес знань {{NN (noun)}, {NP (proper noun)}, {COL (collocation)}}; множина токенів $O = \{o_i\}$, $1 \le i \le n$ іменованих сутностей об'єктів фактів бізнес знань {{NN objective case}, {NP objective case}, {COL objective case}, {regular expression}}, а також чітко окреслена множина іменованих відношень $V = \{v_j\}$, $1 \le j \le m$, які встановлюються існуючими фактами між іменованими сутностями. Три множини *P, O* і *V* є базовими при застосуванні методу компараторної ідентифікації для структурування відношень фактів бізнес знань ІКС. В процесі обробки слабоструктурованих текстів й вилучення з них триплетів фактів, схему ідентифікації яких представлено на рис.2, компаратор реалізує тернарний предикат характеристики факту (Feature Fact) $Q(p,o,v)$, заданий на декартовому добутку $P \times O \times V$ множин іменованих сутностей (суб'єктів і об'єктів) та іменованих відношень. Предикат характеристики факту $Q(p,o,v)$ задає відношення між трьома операндами, які визначають триплет факту *Subject* $\rightarrow$*Predicate* $\rightarrow$*Object*. Компаратор сприймає триплет (p,o,v), створений іменованими сутностями й відношеннями, та встановлює, чи відповідають три параметри даному факту.

Визначено, що два іменованих атрибути (властивості) v та v' відносяться до одного типу $v \sim v'$ ($v, v' \in V'$) тоді і тільки тоді, коли для $\forall o$ і $\forall p$

$$Q(p,o,v) \sim Q(p,o,v'),$$

й показано, що відношення ~ є відношенням еквівалентності, яке розбиває множину V на класи еквівалентності. На декартовому квадраті $V \times V$ вводиться предикат відповідності атрибутів одному підтипу:

$$E_1(v, v') = (\, Q(p, o, v) \sim Q(p, o, v')) \text{ для } \forall o \in O\ ,\ \forall\ p \in P,$$

який однозначно визначений предикатом Q. Предикат $E_1(v, v')$ використовується для об'єктивного визначення відношення будь-яких атрибутів v та v', що належать множині V, до одного підтипу.

Показано, що розподіл атрибутів на підтипи здійснюється за смисловою тотожністю триплетів фактів, яка формально описується предикатом Q, об'єктивно визначеним компаратором. Класу L_a всіх атрибутів $v \in V$, що відносяться до одного підтипу, який містить атрибут $\alpha \in V$, відповідає предикат $L_a(v) = E_1(v, \alpha)$:

$$L_a(v) = (Q(p, o, v) \sim Q(p, o, \alpha)) \text{ для } \forall o \in O,\ \forall p \in P.$$

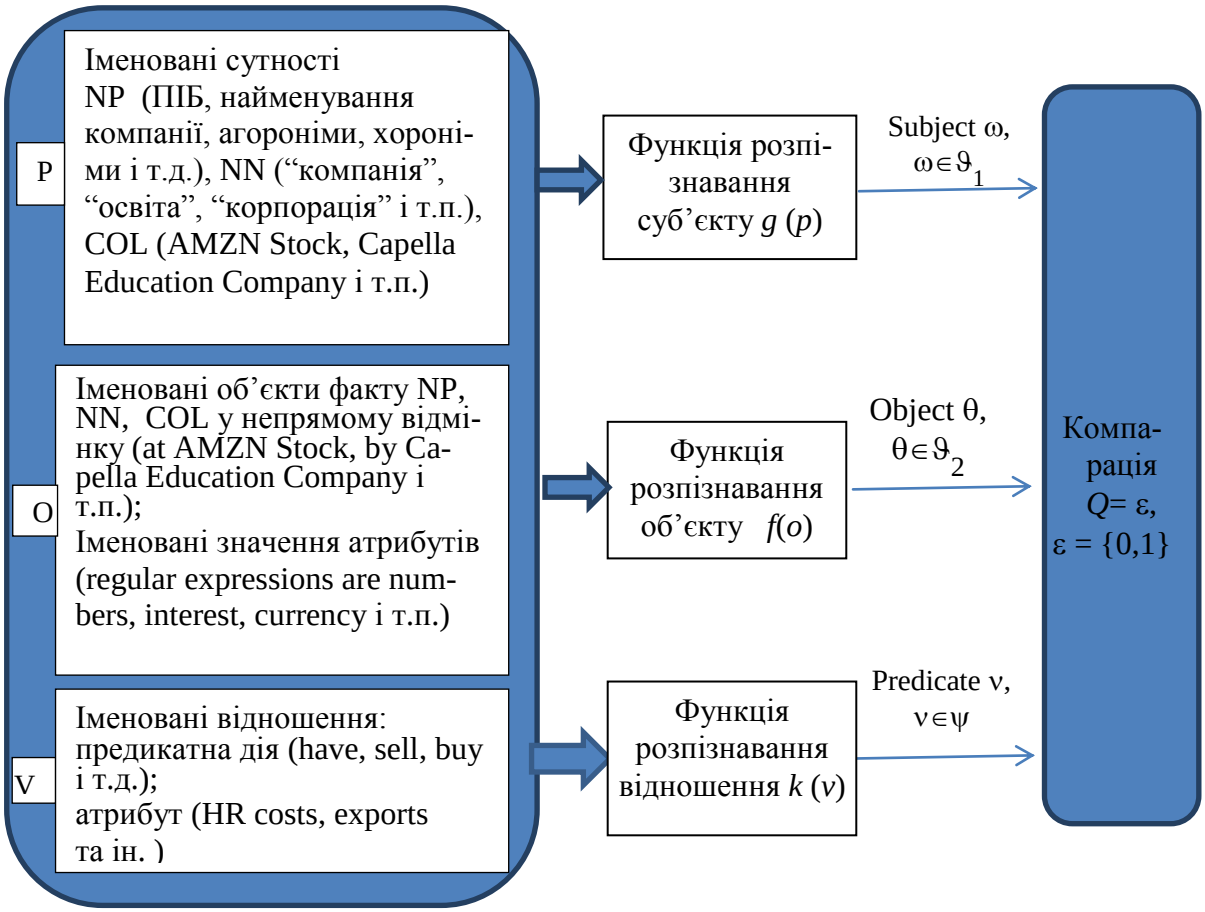

Рисунок 2 – Схема компараторної ідентифікації фактів бізнес аналізу корпорації

На рис. 3 показано приклад роботи компаратора, який оброблює факти бізнес знань корпорації. У даному прикладі множину суб'єктів $P = \{p_1,\ldots,p_5\}$ представлено змінними: p_1= "HMRC", p_2= "company", p_3= "Slater&Gordon", p_4 = "*businessday.com.au*", p_5 ="*Maria Hayek*". Множину атрибутів $V=\{v_1,\ldots, v_6\}$,

що розбиваються на підтипи, представлено: v_1 = "*HR costs*", v_2 = "*Financing costs*", v_3 = "*Dividend yield*", v_4 ="*Founded*", v_5 = "*sale Date*", v_6 = "*Birthday*". Множину розглянутих значень властивостей O= {o_1,…, o_8} подано регулярними виразами:

o_1 = (^\$(\d{1,3},?(\d{3},?)*\d{3}(\.\d{1,3})?|\d{1,3}(\.\d{2})?)$|^\d{1,2}(\. \ d{1,2})? *%$|^ 100%$);

o_2= ^\s*-?((\d{1,3}(\.(\d){3})*)|\d*)(,\d{1,2})?\s?(\u20AC)?\s*$;

o_3 = ^\$?\d{1,3}(,?\d{3})*(\.\d{1,2})?$;

o_4 = ^(( )*\£{0,1}( )*)\d*(.\d{1,2})?$;

o_5 = ^ *(1[0-2]|[1-9]):[0-5][0-9] *(a|p|A|P)(m|M) *$;

o_6= ^((((0[13578])|([13578])|(1[02]))[\/](([1-9])|([0-2][0-9])|(3[01]))) |(((0 [469]) | ([469])|(11))[\/](([1-9])|([0-2][0-9])|(30)))|((2|02)[\/](([1-9])|([0-2][0-9]))))[\/]\d{4} $|^\d{4}$;

o_7= (^\$(\d{1,3},?(\d{3},?)*\d{3}(\.\d{1,3})?|\d{1,3} (\.\d{2})?)$|^\d{1,2} (\.\d {1,2})? *%$|^100%$);

o_8 = ^100$|^[0-9]{1,2}$|^[0-9]{1,2}\,[0-9]{1,3}$.

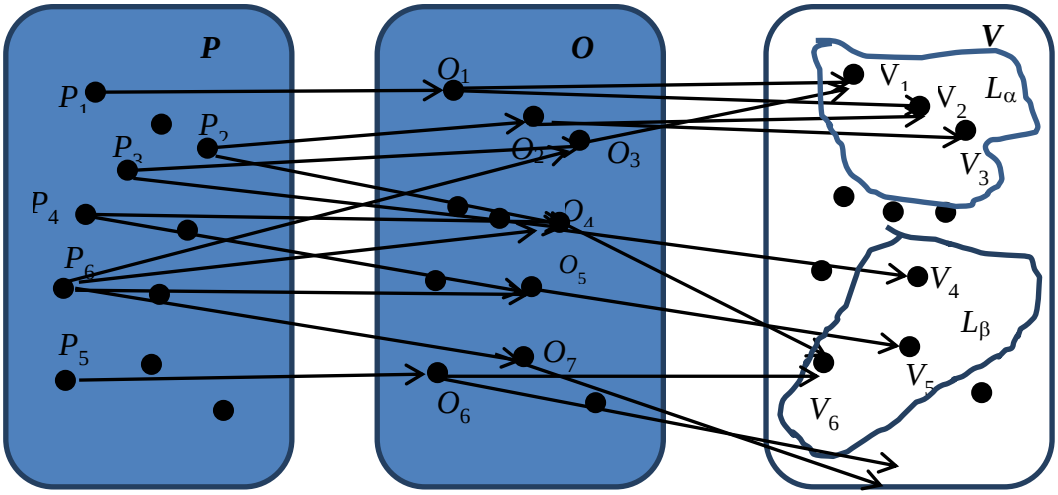

Рисунок 3 – Приклад розбиття на підтипи атрибутів фактів бізнес знань корпорації

У **третьому розділі** представлено процедури формування інформаційного простору ІКС, що включає відбір концептів, які визначають базові сутності знань корпорації; класифікацію базових сутностей через продукування класів еквівалентності концептів; формування атрибутів сутностей; визначення відношень між сутностями у вигляді *Is-A, Part Of , Include* та ін.

Динамічна екстракція знань з текстів потребує ідентифікації сутностей, які включаються у інформаційне поле бізнес інтересів корпорації. Для витягу з тексту імен сутностей, які позначають деякі об'єкти або суб'єкти дискретного світу, і віднесення їх до певного семантичного класу, що визначає тип його поведінки у триплеті факту, удосконалено метод виявлення актуальної множини класифікованих сутностей предметної області. Для зменшення типів класу використовується кодування, що дозволяє об'єднувати характеристики різних то-

кенів на основі аналізу ступеня близькості оформлення до їх графемного виду. Приклади структурного позначення графемних характеристик слів ("word shape sequences features") представлено в табл. 1, що використовує наступне схематичне позначення атрибутів: *X* – великі літери, *x* – малі літери, d – цифри, δ – грецькі символи. При цьому кодуються тільки перші та останні два символи слова, визначаючи множини символів, які проявляються всередині слова. Обраний спосіб кодування дозволяє зменшити не тільки кількість об'єктів класу, але й кількість аналізованих типів фактів, оскільки схожість у написанні сутностей пов'язана зі схожістю у написанні назви, що веде до відповідної схожості зв'язків і проявів в фактах.

Таблиця 1 – Приклади структурного позначення графемних характеристик слів

Boeing	*Xxxx*
Seattle	*Xxxx*
is	*xx*
located	*xxxx*
XP Windows	*XX Xxxx*
Pieter Pen	*Xxxx Xxx*
ASX1	*XXXd*
Colgate-Palmolive	*Xx-Xxx*

Метод виявлення актуальної множини класифікованих сутностей предметної області базується на використанні наївного байєсівського класифікатора і дозволяє розпізнавати згадування персоналій – <*Per*>, компаній і організацій – <*Org*>, дат – <*Date*> фінансових метрик – <*FinM*> і невизначених токенів – <*Other*>. У якості незалежних ознак класифікації використовуються характеристики графемного оформлення токенів і характеристики, визначені контекстом його використання. Модель враховує характеристики поточного токена w_0, двох попередніх токенів w_{-1}, w_{-2} і наступного токена w_1.

Основні характеристики поточного токена w_0 визначаються на етапі передлінгвістичного аналізу: 3 < *length* < 20, наявність апострофа і множину графемних характеристик – "*word shape sequences features*" (*x* – малі літери, *X* – прописні літери, *d* – цифри, δ – грецькі символи). Характеристиками w_{-1}, що визначаються, є належність токена до множини {"*the*", "*mr.*", "*Dr.*", "*mis.*", "*chairman*", "*Dear*" та ін. } і належність двох токенів w_{-1}, w_{-2} до множини послідовностей {<*Org*> *and*, <*Per*> *and*, <*FinM* > *and*, <*Date* > *and*, <*Org*> «,», *Per* «,» , <*FinM* > «,», <*Date* > «,»}. Характеристиками, що визначаються для наступного токена w_1, є його належність до множин $w_1 \in$ {"company", "corporation", та ін.} і $w_1 \in$ *Verbs*.

Використання наївного класифікатора Байєса і текстів корпорації, розмічених мітками апріорно виділених класів сутностей <*Org*>, <*Per*>, <*Date*>, <*FinM*>, <*Other*>, дозволяє віднести кожний токен до класу, що визначає тип

його поведінки, тобто тип фактів, у яких даний токен може представляти суб'єкт або об'єкт триплету.

Запропоновано логіко-лінгвістичну модель генерації фактів з текстових потоків ІКС, яка базується на використанні поверхневих граматичних характеристик ідентифікації сутностей дій та атрибутів.

Для формалізації та явного представлення засобами поверхневої структури суб'єкта і об'єкта триплету факту *Subject→Predicate→Object*, яке називається реченням англійської мови, виділені й описані предметними змінними такі скінченні множини синтаксичних і морфологічних категорій:

$$z^{\text{to}} \vee z^{\text{by}} \vee z^{\text{with}} \vee z^{\text{about}} \vee z^{\text{of}} \vee z^{\text{on}} \vee z^{\text{at}} \vee z^{\text{in}} \vee z^{\text{out}} = 1,$$
$$y^{\text{ap}} \vee y^{\text{aps}} \vee y^{\text{out}} = 1, \; x^{\text{f}} \vee x^{\text{l}} \vee x^{\text{kos}} = 1,$$
$$m^{\text{is}} \vee m^{\text{are}} \vee m^{\text{havb}} \vee m^{\text{hasb}} \vee m^{\text{hadb}} \vee m^{\text{was}} \vee m^{\text{were}} \vee m^{\text{out}} = 1,$$
$$p^{\text{III}} \vee p^{\text{ed}} \vee p^{\text{I}} \vee p^{\text{ing}} \vee p^{\text{II}} = 1,$$

де z – предметна змінна, що визначає синтаксичні характеристики наявності (*to*, *by*, *with*, *about*, *of*, *on*, *at*, *in*) або відсутності (*out*) прийменника у англійській фразі; y – предметна змінна, яка визначає наявність (*ap*, *aps*) або відсутність (*out*) апострофа наприкінці слова; x – предметна змінна, яка визначає позицію іменника перед (*f*) чи після (*l*) особистим дієсловом або після непрямого доповнення (*kos*); m – предметна змінна, яка визначає існування будь-якої форми дієслова "*to be*" (*is*, *are*, *havb*, *hasb*, *hadb*, *was*, *were*, *out*); p – предметна змінна, яка визначає форму основного дієслова (III, *ed*, I, *ing*, II).

Семантичне значення учасників дії, що іменуються словами речення, визначається предикатом

$$P\,(x, y, z, m, p) \rightarrow P(x) \wedge P(y) \wedge P(z) \wedge P\,(m) \wedge P(p).$$

У кон'юнкції предикатів, яка описує взаємозв'язок граматичних характеристик слів, предикат γ_k вилучає частину зв'язків поверхневої структури, не властивих сутностям триплету факту

$$P\,(x, y, z, m, p) = \gamma_k\,(x, y, z, m, p) \wedge P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p),$$

де $k \in [1; h]$, h – кількість розглянутих у системі фактів. Предикат γ_k приймає значення 1, якщо комплекс вибраних характеристик для n-ої фрази формує деяке семантичне значення учасника триплету, і значення 0 у протилежному випадку.

Досліджено декілька профільних типів фактів: 1) твердження про володіння (або приналежності) деякої сутності суб'єкта деякою сутністю об'єкта (рис. 4); 2) твердження про переміщення суб'єктом об'єкта; 3) твердження про втрату (продаж) деякого об'єкта деяким суб'єктом; а також факти-атрибути трьох вищеназваних типів фактів – часу дії, локації дії та ієрархічної належності суб'єкта або об'єкта дії факту іншої сутності.

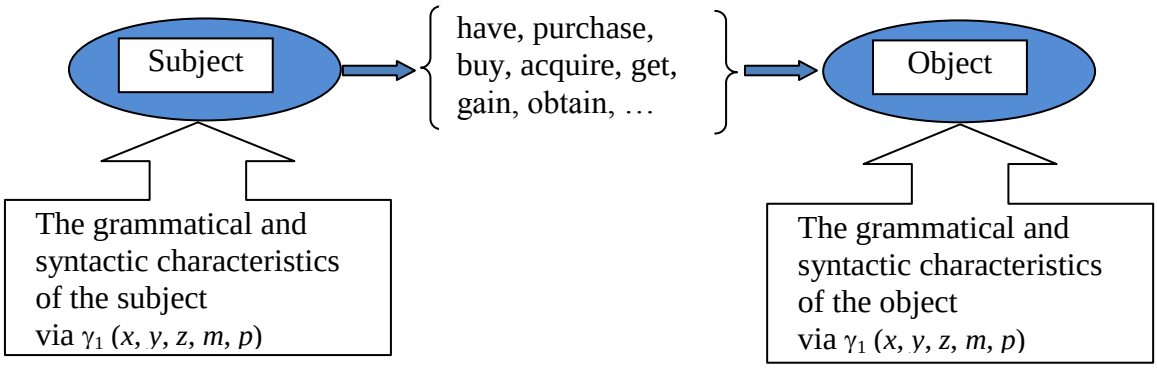

Рисунок 4 – Схема ідентифікації факту належності (власності)

Введені наступні предикати, що визначають:

- граматичні і синтаксичні характеристики *Subject* триплету факту

$$\gamma_1(x, y, z, m, p) = z^{out}\, y^{out}\, x^{f}\, m^{out}\, p^{I} \vee z^{out}\, y^{out}\, x^{f}\, m^{out}\, p^{II} \vee z^{out}\, y^{out}\, x^{f}\, m^{out} p^{ed} \vee z^{by}\, y^{out}\, x^{l}\, p^{ed} (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \vee z^{by}\, y^{out}\, x^{l}\, p^{III} (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}); \quad (1)$$

- граматичні і синтаксичні характеристики об’єкта факту

$$\gamma_2(x, y, z, m, p) = z^{out}\, y^{out}\, x^{l}\, m^{out}\, p^{I} \vee z^{out}\, y^{out}\, x^{l}\, m^{out} p^{ed} \vee z^{out}\, y^{out}\, x^{l}\, m^{out} p^{II} \vee z^{out}\, y^{out}\, x^{f} p^{III} (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \vee z^{out}\, y^{out}\, x^{f} p^{ed} (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}); \quad (2)$$

- граматичні характеристики атрибута часу дії факту

$$\gamma_3(x, y, z, m, p) = (z^{on} x^{kos}\, y^{out} \vee z^{in} x^{kos}\, y^{out} \vee z^{at} x^{kos} (p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II}) (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{out}); \quad (3)$$

- граматичні характеристики атрибута ієрархічної належності суб’єкта або об’єкта дії факту іншої сутності

$$\gamma_4(x, y, z, m, p) = z^{out}\, x^{f} (y^{ap} \vee y^{aps}) (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{out}) (p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II}). \quad (4)$$

Наведено приклад, що висвітлює екстракцію фактів та атрибутів із речення «*The companies' shares were sold by the investor on Tuesday*». За допомогою формули (1) іменник "*investor*" визначається як суб’єкт дії, формули (2) – іменник "*shares*" визначається як об’єкт факту продажу, номінований дієсловом "*sold*"; формули (3) – слово “*Tuesday*” визначається як атрибут часу факту продажу; формули (4) – слово "*companies*"' визначається як атрибут ієрархічної належності об’єкта дії.

Четвертий розділ присвячено удосконаленню інформаційної технології формування єдиного інформаційного простору бізнес діяльності корпорації, яка включає наступні функції:

– витяг даних, добування інформації і знань з контенту Веб-сторінки;
– пошук і витяг елементів знань, явним чином присутніх у тексті у вигляді твердження або факту, які здійснюються на базі логіко-лінгвістичної моделі генерації фактів з текстових потоків ІКС;
– породження складного знання шляхом узагальнення, яке здійснюється за рахунок структурування відношень фактів бізнес знань ІКС.

Етап Web Content Mining запропонованої технології дозволяє використати специфіку структури HTML сторінок при передлінгвістичній обробці текстів та включає наступне:

1) виділення блоків інформаційної навантаженості (основний блок, маловажна інформація, інформаційне сміття),
2) метод виявлення актуальної множини класифікованих сутностей предметної області (ПО),
3) ідентифікацію відношень між концептами, які представляють сутності ПО, на базі регулярних виразів, подібних наступним шаблонам:

<concept>s such as <subconcept>*;
such<concept>s as <subconcept>*;
<subconcept> is a <concept>;
*<subconcept> (and|or) other <concept> s**,
*<concept>s*such as <subconcept>*,
*<concept_r>s**[.,] *e.g.*, (*<concept>s+*)+.

У результаті використання розробленої технології екстракції бізнес знань з текстового контенту база знань ІКС доповнюється системою класифікації економічної і бізнес діяльності корпорації, яка здійснюється через структурування продукції, компаній конкурентів, партнерів і клієнтів, а також галузей виробництва. Зокрема, для 370 організацій моніторингу отримано ієрархічну мережу відношень «Частина-ціле → Адміністративне включення», «Частина-ціле → Географічне включення», а також структуру гіпонімічних відношень атрибутів фактів бізнес знань корпорації, яка розподіляє їх по типах і підтипах (рис. 5).

Оцінювання ефективності інформаційної технології екстракції знань з текстового контенту виконується за допомогою методу текстових колекцій, який полягає у порівнянні результатів роботи імплементованої технології з деяким еталонним результатом для кожного конкретного набору текстів. Визначення обсягу експериментально досліджуваних текстів проводиться методом математичної теорії вибірки, який визначає механізми формування репрезентативної поворотної вибірки. Виконані розрахунки показали, що перевищуючий обсяг тексту (ACCII кодування), який представляє репрезентативну вибірку речень, що оброблюються корпоративною інформаційною системою, з урахуванням вступних і наказових речень, приблизно дорівнює 38 кБ.

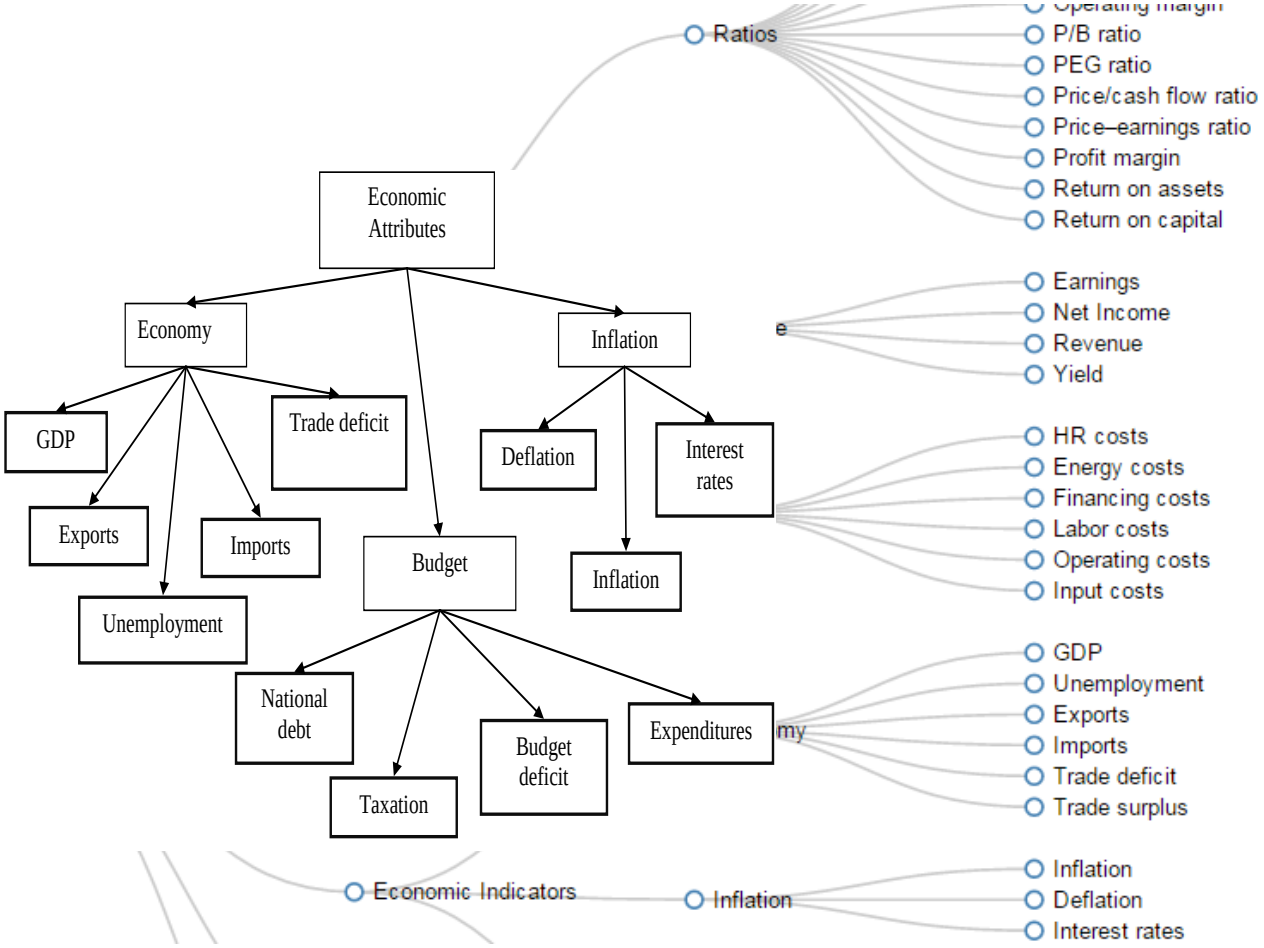

Рисунок 5 – Схематичний фрагмент отриманої структури гіпонімічних відношень атрибутів фактів ІКС

У якості критеріїв оцінки ефективності розробленої технології екстракції бізнес знань з текстового контенту застосовуються інтегральні показники оцінки якості знань, добутих з різнорідних електронних джерел, а саме коефіцієнти повноти і точності, які затверджені міждержавним стандартом з інформації, бібліотечної та видавницької справи:

$$precision = n_{yy} /(n_{yy} + n_{yn}),$$
$$recall = n_{yy} /(n_{yy} + n_{ny}),$$

де n_{yy}. – число правильно ідентифікованих програмним додатком фактів; n_{yn} – число ідентифікованих програмним додатком некоректних фактів (фактів, які визначені експертом, як некоректні); n_{ny} – число фактів, які залишились не ідентифікованими програмним додатком.

Результат досліджень для трьох типів фактів та їх атрибутів представлено в табл. 2.

Середній коефіцієнт повноти, який визначається відношенням числа правильно ідентифікованих програмним додатком коректних фактів до загального числа коректних фактів даного типу, представлених у тексті, дорівнює 0,94.

Середній коефіцієнт точності, який визначається відношенням числа правильно ідентифікованих програмним додатком фактів до загального числа визначених системою фактів (як коректних, так і некоректних), дорівнює 0,84. Дані показники перевищують середні показники ефективності роботи систем фактографічного пошуку необмежених предметних областей.

Таблиця 2 – Результати експериментального оцінювання ефективності технології екстракції знань з текстів

	Facts of the lack of…	Facts of possession of …	Facts of a displacement	Attributes of time	Attributes of location	Attributes of belonging	Attributes of industry
recall	0,92	0,91	0,91	0,97	0,96	0,95	0,94
precision	0,81	0,79	0,76	0,90	0,90	0,89	0,84

Середній коефіцієнт точності, який визначається відношенням числа правильно ідентифікованих програмним додатком фактів до загального числа визначених системою фактів (як коректних, так і некоректних), дорівнює 0,84. Дані показники перевищують середні показники ефективності роботи систем фактографічного пошуку необмежених предметних областей.

ВИСНОВКИ

У дисертаційній роботі вирішено науково-практичну задачу створення інформаційної технології екстракції бізнес знань інтегрованої корпоративної системи на основі інформаційно-логічних моделей і методів смислового опрацювання текстового контенту, що дозволило підвищити ефективність екстракції бізнес знань інтегрованої корпоративної системи. У процесі виконання дисертаційної роботи отримані такі результати.

1. Проаналізовано інформаційні технології, моделі й методи екстракції та ідентифікації знань з текстів, що дозволило сформулювати основні вимоги до розробки інформаційного забезпечення підсистеми екстракції бізнес знань з текстового контенту інтегрованої корпоративної системи.
2. Досліджені можливості використання інструментів алгебри скінченних предикатів у інформаційно-логічних моделях екстракції фактів з текстів, що дозволяє створити моделі вилучення фактографічної інформації з текстових потоків інтегрованої корпоративної системи
3. Побудовано логіко-лінгвістичну модель генерації фактів з текстових потоків інформаційної корпоративної системи, яка базується на використанні поверхневих граматичних характеристик ідентифікації сутностей дій і атрибутів, що дозволяє інтегрувати видобуті факти до єдиного інформаційного простору корпорації.

4. Розроблено метод виявлення актуальної множини класифікованих сутностей предметної області, який дозволяє класифікувати сутності, що екстрагуються з тексту, за апріорно виділеними типами.
5. Реалізовано метод компарації для структурування відношень фактів бізнес знань ІКС, який дозволяє класифікувати атрибути сутностей по класам відношень за рахунок смислової тотожності триплетів фактів, об'єктивно визначеної компаратором.
6. Удосконалено інформаційну технологію формування єдиного інформаційного простору бізнес діяльності корпорації, яка дозволяє здійснювати породження складного знання шляхом експліцитного узагальнення інформації.
7. Виконано впровадження результатів дисертаційного дослідження у підсистеми формування фактографічної бази даних наукових бібліотек НТУ «ХПІ» і Харківського національного університету радіоелектроніки. Результати досліджень впроваджені у навчальний процес кафедри інтелектуальних комп'ютерних систем НТУ «ХПІ» і використовуються у програмах спеціальних дисциплін та при виконанні дипломних робіт студентів спеціальності «Прикладна лінгвістика».

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Аджит Пратап Сингх Гаутам. Информационно-лингвистическая технология извлечения актуальных фактов из потоков специализированных текстов / Н.Ф. Хайрова, Д.Ю. Узлов, Аджит Пратап Сингх Гаутам // Системи управління, навігації та зв'язку. – Полтава : Полтавський національний технічний університет ім. Ю.Кондратюка, 2013. – № 4 (28). – Т. 2. – С. 134-139.

Здобувачем запропоновано основні підходи до побудови інформаційної технології вилучення актуальних бізнес фактів.

2. Аджит Пратап Сингх Гаутам. Особенности экстракции и идентификации знаний WEB-контента / Н.Ф. Хайрова, Аджит Пратап Сингх Гаутам // Системи управління, навігації та зв'язку. – Полтава : Полтавський національний технічний університет ім. Ю.Кондратюка, 2014. – № 1 (29). – Т. 2. – С. 122-125.

Здобувачем наведено основні підходи до опису особливостей вилучення фактографічних знань з WEB-контенту.

3. Аджит Пратап Сингх Гаутам. Интеллектуальные технологии идентификации фактографической информации / А.Ю. Дорошенко, Е.А. Оробинская, Аджит Пратап Сингх Гаутам // Проблеми інформаційних технологій. – Херсон : ХНТУ, 2014. – № 2 (016). – С. 103-106.

Здобувачем запропоновано основні підходи до побудови інформаційної технології ідентифікації фактографічної інформації у ІКС.

4. Аджит Пратап Сингх Гаутам. Информационное пространство фактов интегрированной корпоративной системы / Н.Ф. Хайрова, Аджит Пратап Сингх Гаутам // Вісник Національного технічного університету «Харківський політехнічний інститут». – Харків : НТУ «ХПІ», 2015. – № 11 (1120). – С. 96-102.

Здобувач запропонував новий метод представлення інформаційного прос-

тору бізнес фактів ІКС.

5. Ajit Pratap Singh Gautam. An Analysis of Data Mining Techniques for Customer Relationship Management / Ajit Pratap Singh Gautam, Dr. D. B. Singh // Information Technology. New Delhi : A.P.H. Publishing Corporation, 2010. – № 4. – P. 36–47.

Здобувачем запропоновано аналіз основних засобів інформаційної технології Data Mining стосовно моделі взаємодії корпорації з користувачами.

6. Ajit Pratap Singh Gautam. Face Detection in Still Images using Neural Network and Back Propagation Algorithm/ Ajit Pratap Singh Gautam // Global Journal of Computational Intelligence Research. – Research India Publication, 2011. – Vol. 1. – № 1. – P. 21-30.

7. Ajit Pratap Singh Gautam. A Survey of Association Rule Mining for Customer Relationship Management. / Ajit Pratap Singh Gautam, Dr. Natalia Sharonova // International Journal of Application or Innovation in Engineering and Management (IJAIEM), 2014. – Vol. 3. – № 8. – P. 180-186.

Здобувачем запропоновано аналіз асоціативних правил інформаційної технології Data Mining стосовно моделі управління взаємодії з користувачами.

8. Аджит Пратап Сингх Гаутам. Логико-лингвистическая модель генерации фактов из текстовых потоков информационной корпоративной системы / Нина Хайрова, Наталья Шаронова, Аджит Пратап Сингх Гаутам // International Journal "Information Theories and Applications". Sofia, Bulgaria, 2015. – Vol. 22. – № 2. – P. 142-152.

Здобувачем запропоновано аналіз засобів побудови логіко-лінгвістичної моделі генерації бізнес фактів з контенту ІКС.

9. Ajit Pratap Singh Gautam. The logic and linguistic model for automatic extraction of collocation similarity / N. Khairova, S. Petrasova, Ajit Pratap Singh Gautam // ECONTECHMOD: An International Quarterly Journal on Economics in Technology. – Lublin–Rzeszow, Polish Academy of Science, 2015. – Vol. 4. – № 4. – P. 43-48.

Здобувачем запропоновано метод створення логіко-лінгвістичної моделі для автоматизованої екстракції фактів, які виражено колокаціями.

10. Аджит Пратап Сингх Гаутам. Экстракция фактов из слабоструктурированной текстовой информации /Н. Хайрова, Н. Шаронова, Аджит Пратап Сингх Гаутам // International Journal "Information Model & Analyses". – Bulgaria, 2016 . – Vol. 5. – № 1. – P. 66-77.

Здобувачем запропоновано інформаційну технологію вилучення фактів із слабо структурованої текстової інформації.

11. Ajit Pratap Singh Gautam. Multimodal Biometrics: Need of the Hour/ Ajit Pratap Singh Gautam // Managing Innovation in Business & Technology. – New Delhi, Excel Publishing Services, 2011. – P. 237-244.

12. Аджит Пратап Сингх Гаутам. Применение интеллектуальных технологий экстракции и идентификации знаний в корпоративных информационных системах / Аджит Пратап Сингх Гаутам, С.В. Шкапо // Матеріали ІІІ Всеукраїнської науково-практичної конференції «Інтелектуальні системи та прикладна лінгвістика». – Харків : НТУ «ХПІ», 2014. – С. 15-17.

Здобувач запропонував аналіз застосування інтелектуальних технологій у ІКС.

13. Аджит Пратап Сингх Гаутам. Основные проблемы обработки текстов в интегрированных корпоративных информационных системах / Аджит Пратап Сингх Гаутам, Н.В. Шаронова // Матеріали IV Всеукраїнської науково-практичної конференції «Інтелектуальні системи та прикладна лінгвістика», – Харків : НТУ «ХПІ», 2015. – С. 9-10.

Здобувачем запропоновано аналіз основних проблем опрацювання текстів у інтегрованих корпоративних інформаційних системах.

14. Аджит Пратап Сингх Гаутам. Модели извлечения фактографической информации в системе формирования библиографических описаний полнотекстовых документов научной библиотеки / Аджит Пратап Сингх Гаутам // Матеріали V Всеукраїнської науково-практичної конференції «Інтелектуальні системи та прикладна лінгвістика». – Харків : НТУ «ХПІ», 2016. – С. 37-38.

АНОТАЦІЇ

Аджіт Пратап Сінгх Гаутам. Інформаційна технологія екстракції бізнес знань з текстового контенту інтегрованої корпоративної системи. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Національний технічний університет «Харківський політехнічний інститут», Харків, 2016.

Мета дисертаційного дослідження – створення інформаційної технології екстракції бізнес знань інтегрованої корпоративної системи (ІКС) на основі інформаційно-логічних моделей і методів смислового опрацювання текстового контенту. Основні результати: вперше розроблено логіко-лінгвістичну модель генерації фактів з текстових потоків ІКС, яка базується на використанні поверхневих граматичних характеристик сутностей, предикатів та атрибутів, що дозволяє ефективно екстрагувати з текстового контенту профільні знання про суб'єкти моніторингу. Отримав подальший розвиток метод компараторної ідентифікації, який використано для структурування відношень фактів бізнес знань ІКС. Реалізація методу дозволила класифікувати атрибути сутностей за класами відношень за рахунок смислової тотожності триплетів фактів, які об'єктивно визначені компаратором. Удосконалено метод виявлення актуальної множини класифікованих сутностей предметної області, який відрізняється комплексним використанням лінгвістичних, статистичних й смислових характеристик в наївному байєсівському класифікаторі. Метод дозволяє класифікувати сутності, що екстрагуються, за апріорно виділеними типами. Удосконалено інформаційну технологію формування єдиного інформаційного простору бізнес діяльності корпорації, яка дозволяє за рахунок використання алгебро-логічних перетворень здійснювати породження складного знання шляхом експліцитного узагальнення інформації, що прихована у сукупності часткових фактів.

Ключові слова: інформаційна технологія, інтегрована корпоративна система, екстракція бізнес знань, ідентифікація сутностей, логіко-лінгвістичні мо-

делі.

Аджит Пратап Сингх Гаутам. Информационная технология экстракции бизнес знаний из текстового контента интегрированной корпоративной системы. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 – информационные технологии. – Национальный технический университет «Харьковский политехнический институт», Харьков, 2016.

Цель диссертационного исследования – создание информационной технологии экстракции бизнес знаний интегрированной корпоративной системы на основе информационно-логических моделей и методов смысловой обработки текстового контента. В работе проанализированы существующие информационные технологии, модели и методы экстракции и идентификации знаний из текстов, сформулированы основные требования к разработке информационного обеспечения подсистемы экстракции бизнес знаний из текстового контента интегрированной корпоративной системы. Обосновано использование инструментов алгебры конечных предикатов в информационно-логических моделях экстракции фактов из текстовых потоков; построена математическая модель генерации фактов из текстов корпорации.

В диссертационной работе впервые разработана логико-лингвистическая модель генерации фактов из текстовых потоков ИКС, базирующаяся на использовании поверхностных грамматических характеристик идентификации сущностей действий и атрибутов, что позволяет эффективно экстрагировать из текстового контента профильные знания о субъектах мониторинга. Получил дальнейшее развитие метод компараторной идентификации, который использован для структурирования отношений на множестве фактов бизнес знаний ИКС. Обрабатывая слабоструктурированные тексты и извлекая из них триплеты фактов, компаратор воспринимает триплет, образованный именованными сущностями и отношениями. Применение компаратора позволяет определить соответствие атрибутов рассматриваемых триплетов фактов однотипным отношениям, тем самым относя атрибуты к определенным классам эквивалентности. Применение метода позволило распределить атрибуты сущностей по классам отношений за счет смыслового тождества триплетов фактов, объективно определенных компаратором.

Усовершенствован метод выявления актуального множества классифицированных сущностей предметной области, отличающийся комплексным использованием лингвистических, статистических и смысловых характеристик в наивном байесовском классификаторе. Метод позволяет классифицировать экстрагируемые сущности по априорно выделенным типам. Усовершенствована информационная технология формирования единого информационного пространства бизнес деятельности корпорации, которая позволяет за счет использования алгебро-логических преобразований осуществлять порождение сложного знания путем эксплицитного обобщения информации, скрытой в совокупности частичных фактов.

Результаты диссертационного исследования внедрены в практику разработки и создания подсистем экстракции знаний из текстового контента реальных ИКС. На основе разработанных в диссертационном исследовании методов и моделей интеллектуальной обработки текстового контента предложена информационная технология формирования единого информационного пространства бизнес деятельности корпорации. При этом под информационным пространством интегрированной корпоративной системы понимается совокупность некоторых актуальных сведений, данных, оформленных таким образом, чтобы обеспечивать качество и оперативность принятия решений в области целевой деятельности корпорации. Предложенная информационная технология позволять извлекать знания из всего многообразия информационных ресурсов современного предприятия: Интернет- и интранет- сайтов предприятий и организаций, почтовых сообщений, файловых систем, хранилищ документов различных ведущих производителей, текстовых полей баз данных, репозитариев, различных бизнес-приложенийт.п.. Технология включает логико-лингвистическую модель генерации фактов из текстовых потоков ИКС, метод структурирования отношений фактов бизнес знаний, метод выявления актуального множества классифицированных сущностей предметной области, а также специализированные этапы Web Content Mining лингвистического процессора.

Разработанные в исследовании математические модели могут быть использованы в различных системах автоматической обработки текстов, системах извлечения знаний, экстракции информации (Information Extraction) и распознавания сущностей (Named Entity Recognition).

Ключевые слова: информационная технология, интегрированная корпоративная система, экстракция бизнес знаний, идентификация сущностей, логико-лингвистические модели.

Ajit Pratap Singh Gautam. Information technology of business knowledge extraction from text content of the integrated corporate system – Manuscript.

Thesis for a candidate degree in technical science, speciality 05.13.06 – Information Technologies. – National Technical University «Kharkiv Polytechnic Institute». – Kharkiv, 2016.

The aim of the thesis is to develop information technology of extraction of business knowledge of integrated corporate system (ICS) based on the information logic models and methods of text content sense processing. The main results are as follows: a logic linguistic model of fact generation from ICS text streams has been developed which is based on surface grammar characteristics of identification of entities of actions and attributes which allows to effectively extract industry specific knowledge about the subjects of monitoring from text content. The thesis further develops the method of comparator identification used for structuring of ICS business knowledge facts relationship. The method allows to classify the attributes of entities according to class relationships due to sense identity of fact triplets which are determined by the comparator objectively. The paper improves the method of determination of actual set of classified entities of a subject domain which is distinguished by an integral use of linguistic, statistical and sense characteristics in the naïve Bayes

classifier. The method allows to classify entities extracted according to types determined a priori. The thesis improves the development of information technology of common information space of corporation business activity which allows complicated knowledge generation by means of explicit generalization of information hidden in the collection of partial facts using algebra logic transformations.

Key words: information technology, integrated corporate system, business knowledge extraction, identification of entities, logic linguistic models.