МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

ЧЕРЕНКОВ ІГОР ОЛЕКСАНДРОВИЧ

УДК 004.6

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ФОРМУВАННЯ ЦІНОВОЇ СТРАТЕГІЇ ПІДПРИЄМСТВА НА ОСНОВІ ПОТОКУ ІНТЕРНЕТ НОВИН

Спеціальність 05.13.06 – інформаційні технології

Автореферат дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2014

Дисертацією є рукопис.

Робота виконана на кафедрі автоматизованих систем управління Національного технічного університету «Харківський політехнічний інститут» Міністерства освіти і науки України.

Науковий керівник кандидат технічних наук, доцент
Орєхов Сергій Валерійович,
Національний технічний університет
«Харківський політехнічний інститут»,
доцент кафедри автоматизованих систем
управління

Офіційні опоненти: доктор технічних наук, професор
Федорович Олег Євгенович,
Національний аерокосмічний університет
імені М. Є. Жуковського
«Харківський авіаційний інститут»,
завідувач кафедри інформаційних
управляючих систем

доктор технічних наук, професор
Філатов Валентин Олександрович,
Харківський національний університет
радіоелектроніки,
професор кафедри штучного інтелекту

Захист відбудеться « 20 » березня 2014 р. о 14-30 годині на засіданні спеціалізованої вченої ради Д 64.050.07 в Національному технічному університеті «Харківський політехнічний інститут» за адресою: 61002, Харків, вул. Фрунзе, 21.

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

Автореферат розісланий « 15 » лютого 2014 р.

Вчений секретар
спеціалізованої вченої ради В. П. Северин

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Інформаційні технології формування цінової стратегії підприємства є затребуваними для багатьох ринків України з еластичним попитом та, насамперед, ринків сировини. Результат процесу функціонування таких інформаційних технології на підприємстві зводиться до задачі прогнозування ціни на товар у майбутньому, при цьому найбільша увага приділяється розробкам автоматичних підходів формування ціни.

Серед існуючих методів вирішення задачі цінового прогнозування найбільше поширення одержали методи математичної статистики, зокрема експоненціального згладжування та авторегресійні моделі. Однак, точність загальноприйнятих підходів обмежена. Отже, все більша увага приділяється розробкам альтернативних підходів цінового прогнозування, які дозволять безпосередньо включати вплив зовнішніх факторів у прогноз. В цьому напрямку виникає ряд проблемних питань. По-перше, чи існує модель та алгоритм отримання маркетингових даних про зовнішні фактори та класифікації цих зовнішніх факторів-подій, що відбуваються на ринку. По-друге, з якого інформаційного джерела можна отримувати інформацію про ці фактори. По-третє, чи можливо в автоматичному режимі виділяти ці фактори та зберігати ретроспективний аналіз їх змін у часі. Четверте, чи розроблено інформаційну технологію, яка веде облік факторів та дозволяє аналізувати їх вплив на ціну товару у теперішньому та минулому. І, п’яте, як візуально презентувати ціновий тренд та перелік цих факторів в процесі формування цінової стратегії.

Таким чином, є актуальною науково-практична задача побудови сервіс орієнтованої WEB-базованої інформаційної технології, яка дозволяє дослідити в автоматизованому режимі вплив зовнішніх факторів на ціновий тренд, коли основним джерелом маркетингових даних виступає потік інтернет новин.

Зв’язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана на кафедрі автоматизованих систем управління Національного технічного університету «Харківський політехнічний інститут» (НТУ «ХПІ») в рамках завдань фундаментальних держбюджетних НДР МОН України «Розробка систем підтримки прийняття рішень в складних інформаційно-управляючих комплексах» (ДР № 0109U002424) і «Розробка систем підтримки прийняття рішень з управління розвитком складних розподілених техніко-економічних та соціально-економічних систем» (ДР № 0111U002287), де здобувач був виконавцем окремих розділів.

Мета і задачі дослідження. Метою є підвищення точності прогнозів ціни при формуванні стратегії підприємства на основі автоматичного виявлення впливу зовнішніх факторів на цінові зміни шляхом аналізу текстів з потоку інтернет новин.

Для досягнення зазначеної мети поставлені наступні задачі:

– провести аналіз сучасного стану проблеми формування цінової стратегії підприємства, дослідити математичні моделі та інформаційні технології, необхідні для її розв’язання;

– розробити комплексний метод формування цінової стратегії підприємства шляхом вибору фінального значення ціни на товар;

– розробити комплекс синтаксичних моделей для аналізу текстів інтернет новини, як контейнера маркетингової інформації про існуючі зовнішні фактори;

– створити технологію, яка дозволяє ідентифікувати зовнішні фактори у потоці інтернет новин;

– сформувати метод для визначення фінального значення ціни на товар з урахуванням ринкових подій в потоці інтернет новин;

– спроектувати сервіс-орієнтовану інформаційну технологію, яка дає змогу виявляти та візуально відображати вплив зовнішніх факторів на фінальну ціну товару в заданому ринковому сегменті;

– провести апробацію та впровадження розробленої інформаційної технології, перевірити її працездатність у процесі формування цінової стратегії на підприємстві.

Об'єктом дослідження є процес формування цінової стратегії підприємства.

Предмет дослідження – сервіс орієнтована WEB-базована інформаційна технологія формування цінової стратегії підприємства на основі аналізу гіпертекстів з потоку інтернет новин.

Методи дослідження. Досягнення мети роботи базується на використанні: теорії системного аналізу для дослідження процесу формування цінової стратегії на підприємстві; методів класифікації та кластерізації для обробки текстів; технології видобутку даних для розробки синтаксичних моделей аналізу інтернет новини; методів математичної статистики, комп'ютерного навчання для розробки технології визначення фінальної ціни на основі асоціативних правил; сучасних підходів до побудови WEB-базованого сервіс-орієнтованого програмного забезпечення (ПЗ) для реалізації інформаційної технології.

Наукова новизна отриманих результатів полягає у тому, що:

– вперше розглянуто у якості об'єкту дослідження процес формування цінової стратегії під впливом потоку інтернет новин, котрий відрізняється від існуючих тим, що фінальна ціна формується на основі асоціативних правил, які відображають вплив зовнішніх факторів (ринкових подій), виявлених у профільному потоці інтернет новин, що дозволяє підвищити точність встановлення фінального значення ціни на товар в заданому сегменті ринку;

– удосконалено методи видобутку даних, котрі відрізняються від відомих тим, що сформульовано ключові етапи видобутку даних з текстових інтернет новин за допомогою детально описаних моделей граматик безпосередніх складових, які розроблені для ключових категорій потоку новин для заданого ринку, що дозволяє прискорити процес добутку маркетингових даних;

– удосконалено метод ієрархічної кластеризації, котрий відрізняється від відомих тим, що формалізовано опис інтернет новини та визначені умови для оцінки ступеню близькості двох і більше новин для заданого ринку. Це дозво-

ляє виявляти унікальні події серед дублікатів та сюжетних ланцюжків, що завжди існують у потоці інтернет новин;

– удосконалено метод цінового прогнозування на основі асоціативних правил, що відрізняється від існуючих аналогів тим, що урахування впливу випадкових зовнішніх факторів здійснюється безпосередньо на прогнозне значення, враховуючи параметри підтримки й вірогідності. Таке удосконалення дозволяє підвищити точність прогнозу;

– удосконалена інформаційна технологія формування цінової стратегії підприємства на базі WEB базованого сервіс-орієнтованого програмного забезпечення, яка відрізняється від існуючих тим, що дає змогу виявляти зовнішні фактори (ринкові події) з потоку інтернет новин та досліджувати їх вплив на ціновий тренд для заданого ринку. Такий підхід дозволяє прискорити отримання цінової стратегії та покращити її характеристики, такі як точність.

Практичне значення отриманих результатів для формування цінової стратегії вітчизняних підприємств полягає у розробці WEB базованої сервіс-орієнтованої інформаційної технології для вирішення задач видобутку даних з потоку текстових інтернет новин та автоматичного виявлення впливу зовнішніх факторів на поведінку ціни.

Розроблена WEB базована сервіс-орієнтована інформаційна технологія була впроваджена на підприємствах: ТОВ «Енергетичні технології» (м. Харків), ТОВ «Енергетехінвест» (м. Харків). Результати дисертаційної роботи було використано у навчальному процесі кафедри автоматизованих систем управління НТУ «ХПІ» в дисциплінах: «Алгоритми і структури даних», «Інформаційні технології управління знаннями на підприємстві».

Особистий внесок здобувача. Положення і результати, винесені на захист дисертаційної роботи, отримані здобувачем особисто. Серед них: комплекс синтаксичних моделей аналізу інтернет новини; методи класифікації та кластеризації для визначення унікальних ринкових подій (зовнішніх факторів), що впливають на процес формування цінової стратегії; інформаційна технологія на базі сервіс-орієнтованих складових для визначення та візуалізації впливу текстів з потоку інтернет новин на вибір цінової стратегії.

Апробація результатів дисертації. Основні положення та результати роботи доповідались на: XVIII и XIX Міжнародних науково-практичних конференціях «Інформаційні технології: наука, техніка, технологія, освіта, здоров'я» (м. Харків, 2010 р, 2011 р.); XXV Міжнародній науковій конференції «Математичні методи в техніці та технологіях (ММТТ-25)» (м. Харків, 2012 р.); XVII Міжнароднім форумі «Радіоелектроніка та молодь у XXI столітті» (м. Харків, 2012 р.); Міжнародній науково-практичній конференції «Актуальні питання економіки: проблеми, гіпотези, дослідження» (м. Сімферополь, 2012 р.); на наукових семінарах кафедри автоматизованих систем управління НТУ «ХПІ».

Публікації. Основний зміст дисертації відображено у 9 наукових публікаціях, з них: 5 статей у наукових фахових виданнях МОН України (2 з яких входять до наукометричних баз), 4 у збірниках трудів наукових конференцій.

Структура й обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Загальний обсяг дисертації становить 172 сторінки, з них 73 рисунки по тексту, 15 таблиць по тексту, список використаних джерел зі 168 найменувань на 17 сторінках, 7 додатків на 29 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність дисертації, сформульовано її мету і задачі, визначено об'єкт, предмет і методи дослідження; висвітлено зв'язок роботи з програмами, планами та темами НДР; сформульовано напрямки та наукові задачі дослідження, що забезпечують значущість одержаних результатів; визначено наукову новизну та практичну значущість роботи, а також особистий внесок здобувача в надрукованих роботах.

Перший розділ присвячено системному аналізу інформаційних технологій процесу формування цінової стратегії підприємства на основі автоматичного прогнозування ціни.

Показано, що необхідність визначення цінової політики підприємства обумовлюється її місцем у маркетинговій стратегії. Формування цінової політики підприємства є одною із ключових задач маркетингу. Встановлено, що визначення цінової політики конкурента здійснюється за допомогою прогнозування ціни на продукт.

Питання встановлення ціни є ключовим для будь-якого ринку, тому що ціна є грошове вираження вартості товару, роботи або послуги. На ціну впливає безліч факторів і в першу чергу це попит і пропозиція, простір і час, а також конкуренція, так звані елементи ринку. Тому, питання виявлення інформаційного впливу через фактори, що формують ціну товару або послуги, шляхом застосування сучасних інтернет технологій у зв'язці з математичним забезпеченням і економічним обґрунтуванням потребує подальшого дослідження.

Існуюче ПЗ для дослідження потоку інтернет новин включає три групи підходів: перший – сайти з редакційним або WEB 2.0 наповненням новин; другий – e-mail alert системи; третій – повноцінні автоматичні системи по агрегації і фільтрації новин. Кожний з підходів відповідає вимогам своєї цільової аудиторії. Однак, для розв'язання задачі формування цінової стратегії на основі потоку новин ці підходи можна використати тільки на окремих етапах.

Встановлено, що інформаційна технологія повинна реалізовувати формування цінової стратегії шляхом автоматичного прогнозу цінових значень на основі виявлення зовнішніх факторів (ринкових подій), візуального подання факту їх наявності, а також доводити факт їх впливу на цінову динаміку в заданому сегменті ринку. Тому на основі критичного аналізу існуючих підходів, методів та інформаційних технологій сформульована постановка задачі дослідження та запропонована схема її вирішення на базі побудови асоціативних правил між факторами (ринковими подіями), що виявлено в потоці інтернет новин.

У **другому розділі** систематизовані способи вирішення задачі вияву випадкових зовнішніх факторів шляхом аналізу текстів з потоку інтернет новин.

Запропоновано розглядати інтернет новину як контейнер маркетингових даних, що містить різні ринкові події (зовнішні фактори). На основі морфологічно-синтаксичного аналізу сформульовано підхід до ідентифікації такої ринкової події. Підхід реалізується на основі множини синтаксичних моделей, які отримані за допомогою онтології предметної області. Ці моделі враховують категорії ринкових подій (зовнішніх факторів) таких як споживання та пропозиція, профіль конкурента, інфляція, світові ціни, науково-технічний розвиток, профіль споживача, психологія споживача та інші.

Розробка онтології є обов'язковим етапом і дозволяє сформувати множини семантичних полів, синтаксичних моделей та лексем відповідно до предметної області (заданого ринку). На прикладі онтології подій ринку полімерів (рис. 1), що є типовим представником ринку сировини з еластичним попитом, побудована синтаксична модель, що описує категорію «профіль конкурента».

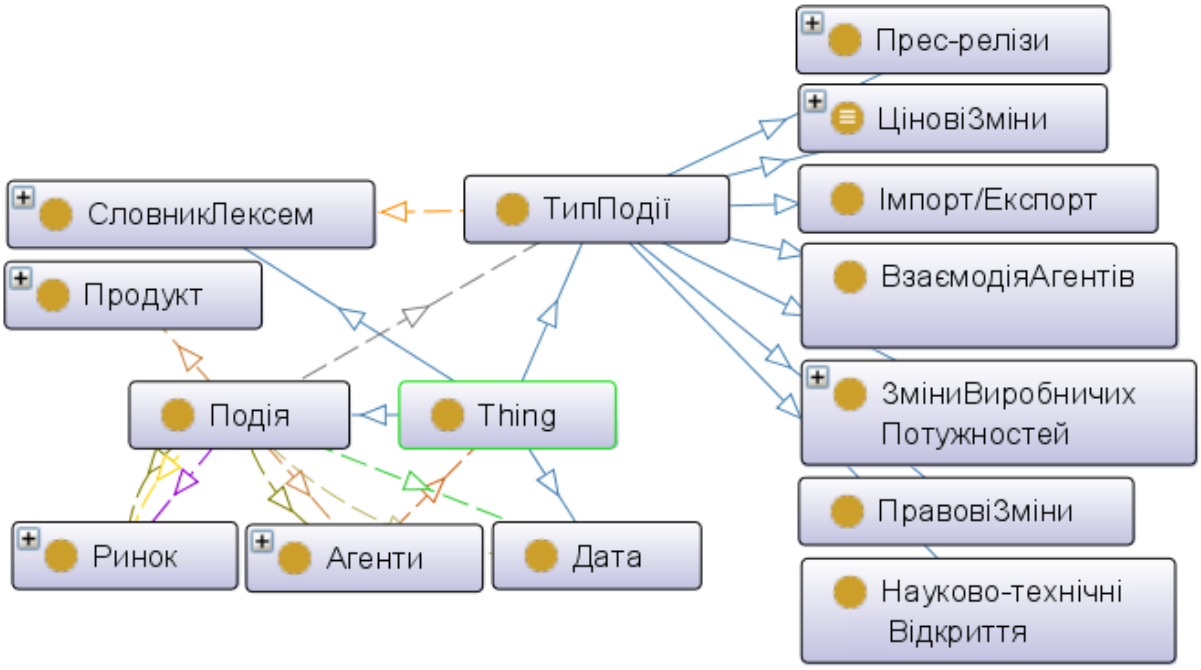

Рисунок 1. – Фрагмент онтології новин полімерного ринку

Загальний опис синтаксичної моделі, користуючись методологією граматик безпосередніх складників (БС), наведено на рис. 2. Модель включає множину синтаксичних моделей фраз, що діляться на дієслівні та іменникові фрази, а також множини лексем на основі морфологічного аналізу.

Робота з елементами тексту, як фразами, здійснюється на основі граматик безпосередніх складників, що оперує відповідними лінгвістичними одиницями. Під безпосередніми складниками варто розуміти кожну із двох конструкцій максимального об'єму, які можна виділити у складі речення і далі у складі кожної безпосередньої складової.

Рисунок 2. – Модель новини як текстового об’єкту

Атомарною одиницею в аналізі є не слово, а морфема або лексема. Графічне подання речення в термінах БС для ринку полімерів зображено на рис. 3. Для видобутку даних із усього потоку новин схожі моделі (правила граматики) повинні бути сформовані для назви та ліда кожної категорії новин.

Кожна інтернет новина, як контейнер маркетингових даних, включає чотири основних блока лексем (рис. 3): «хто», «що», «де», «коли». Перший блок містить опис контрагента або продукту. Другий дає уяву про ринкову дію. Третій визначає в який проміжок часу відбулася подія, а четвертий – де географічно вона відбулася. Для кожного з цих блоків формуються відповідні словники (множини) лексем: M – про продукти ринку, E – про контрагентів, G – про географію ринкової події. Окремо виділяється дата появи інтернет новини – d .

Для виявлення випадкових зовнішніх факторів в роботі запропоновано виконати двоетапну обробку інтернет новини. По-перше, це класифікація. Вона здійснюється шляхом синтаксично-морфологічного аналізу, який виконується на основі множин M , E , G . В результаті формуються значення вектору категорії події $\vec{c}$. Кожна координата c_k $k=\overline{1;K}$ цього вектору приймає значення 1, якщо новина належить до k –ої категорії та нуль у іншому випадку.

На другому етапі виконується виділення кластерів подій однієї природи, що дозволяє виключити появу дублікатів та сюжетних ланцюжків та отримати потік подій високої якості.

Умовою об’єднання двох новин-дублікатів в одну подію є збіг координат векторів $\vec{c}'$ однієї новини та $\vec{c}''$ іншої новини. Дата новин, що відповідає події, повинна розрізнятися в межах граничного значення d_t , достатнього для відображення динаміки ринку, тобто повинно виконуватися нерівність $\left|d'-d''\right|\le d_t$, де d' та d'' - дати однієї та другої новин, що порівнюються.

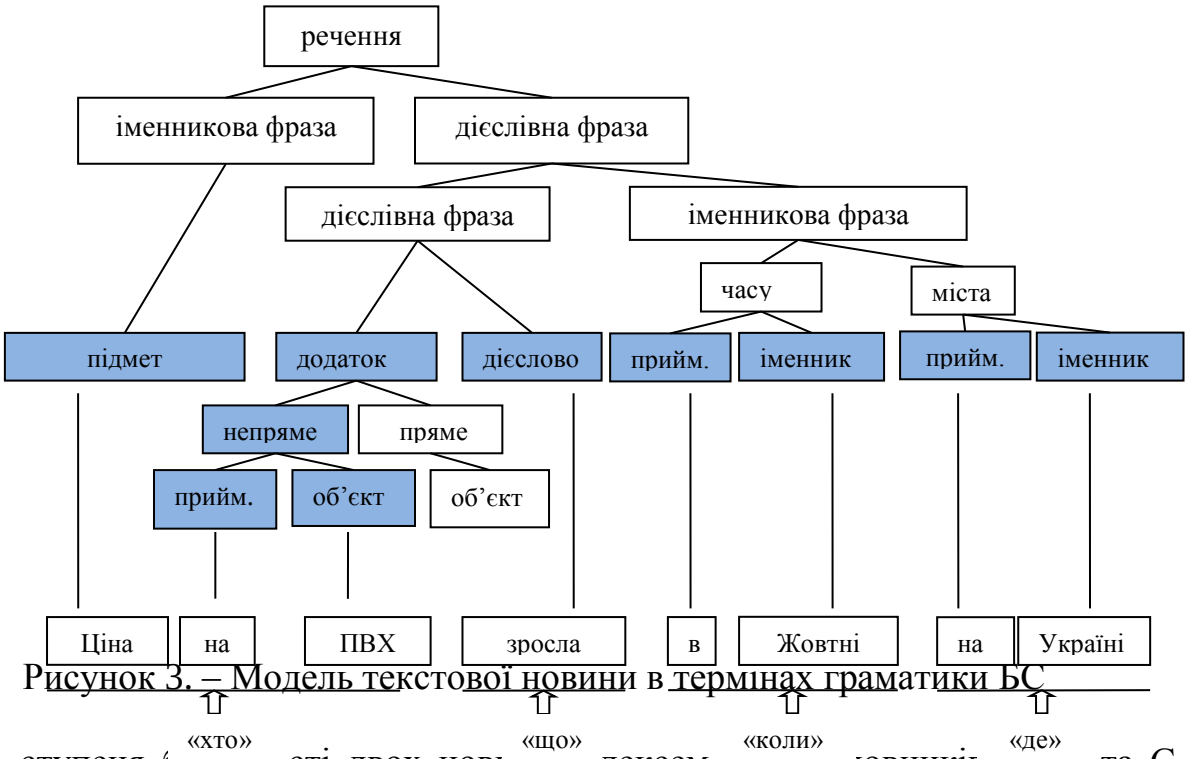

Рисунок 3. – Модель текстової новини в термінах граматики БС

Оцінка ступеня близькості двох новин за лексемами зі словників M, E та G здійснюється за формулою:

$$F_p = \left[1 + \sum_{i=1}^{I} (m'_i - m''_i)^2\right]\left[1 + \sum_{j=1}^{J} (e'_j - e''_j)^2\right]\left[1 + \sum_{h=1}^{H} (g'_h - g''_h)^2\right], \qquad (1)$$

де m'_i, m''_i координати векторів $\vec{m}'$, $\vec{m}''$, що формуються відповідно до лінейно-упорядкованої множини $\tilde{M} = M^1 \cup M^2$ лексем продуктів обох новин та мають однакову розмірність $I = |\tilde{M}|$, $i = \overline{1;I}$, M^1 та M^2 – неупорядковані множини лексем продуктів кожної окремої новини, що порівнюються. Для першої новини координати $\vec{m}'$ приймають значення

$$m'_i = \begin{cases} 1, & l_i^M \in M^1; \\ 0, & l_i^M \notin M^1, \end{cases}$$

де l_i^M – одна з лексем зі словника M усіх можливих ринкових лексем, що сформовано у процесі дослідження предметної області новин, при цьому $|M| > I$. Схожим чином формується також вектор $\vec{m}''$ для другої новини. Аналогічно формуються вектора контрагентів $\vec{e}', \vec{e}''$ з відповідними координатами e'_j, e''_j, що формуються схожим чином до координат m'_i на основі лексем l_j^E зі словника E усіх можливих лексем контрагентів відповідно до лінійно-упорядкованої множини $\tilde{E} = E^1 \cup E^2$, де E^1 та E^2 – неупорядковані множини лексем контрагентів, та мають однакову розмірність $J = |\tilde{E}|$, $j = \overline{1;J}$, при цьому $|E| > J$. Так

саме формуються вектора географії ринків новин $\vec{g}'$, $\vec{g}''$ з відповідними координатами g'_h, g''_h, що формуються схожим чином до координат m'_i на основі лексем l_h^G зі словника G усіх можливих лексем географії подій відповідно до лінійно-упорядкованої множини $\tilde{G} = G^1 \cup G^2$, де G^1 та G^2 – неупорядковані множини лексем географії подій, і мають однакову розмірність $H = \left|\tilde{G}\right|$, $h = \overline{1;H}$, при цьому $\left|G\right| > H$.

В більшості випадків множини словникових лексем N_1, відповідно для однієї новини $N_1^1 = M^1 \cup G^1 \cup E^1$ та іншої $N_1^2 = M^2 \cup G^2 \cup E^2$, будуть відрізнятися, тому повного збігу, коли формула (1) дорівнює одиниці, спостерігатися не буде. Дана проблема вирішується або шляхом розрахунку експертних оцінок порогів об'єднання новин, або за рахунок формування граничного значення аналітичним способом.

Пропонується спосіб аналітично розрахунку допустимої похибки α, що дозволяє оцінювати ступінь близькості новин за первинною умовою $F_p \le \alpha$. Розрахунок α засновано на множині додаткових лексем N_3. До множини N_3 варто віднести ті лексеми, що входять у новину, але при цьому не відображають саму подію безпосередньо, а лише уточнюють характер події, зокрема, вказують на спрямованість події в часі, на характер, силу впливу й т.п. Для подальшого розрахунку α вводиться проміжна ступінь оцінки близькості за N_3

$$F_a = \begin{cases} \left|N_3^1 \cup N_3^2 \setminus N_3^1 \cap N_3^2\right|, \left|N_3^1 \cup N_3^2 \setminus N_3^1 \cap N_3^2\right| \neq 0; \\ 1, \left|N_3^1 \cup N_3^2 \setminus N_3^1 \cap N_3^2\right| = 0, \end{cases} \quad (2)$$

де N_3^1 та N_3^2 – множини N_3 однієї та іншої новини, що порівнюються.

Безпосередньо на підставі F_a з (2) отримається коефіцієнт α

$$\alpha = \left|N_3^1 \cup N_3^2\right| \Big/ F_a \,. \quad (3)$$

Коефіцієнт α показує ступінь близькості новин за сторонніми лексемами, і чим ближчі новини, тим більше α. Значення коефіцієнта α з (3) полягає в тому, що чим він більше, тим більше ймовірність того, що новини описують одну й ту саму подію внаслідок зв'язків між словами у природних мовах, і тим менша схожість потрібна для подібності по первинній та вторинній умовах.

Слід зазначити, що завдяки виду умови $F_p \le \alpha$ неможливо компенсувати розбіжність по F_p за рахунок росту α, тому що швидкість росту F_p набагато більше, ніж α, окрім цього що α є обмеженою величиною і приймає значення $\alpha \in [1, \left|N_3^1 \cup N_3^2\right|]$. Також набагато більша швидкість росту F_p ніж α свідчить про більш питому вагу F_p над α, тобто лексеми з N_1 мають більш питому вагу

ніж лексеми з N_3.

Таблиця 1 показує поведінку частин умови $F_p \le \alpha$ для випадку, коли $\alpha \in [1,10]$, де n_p – число розбіжних словникових лексем, тобто $\left| N_1^1 \cup N_2^2 / N_1^1 \cap N_2^2 \right|$, а n_α – число збігів несловникових лексем, тобто $\left| N_3^1 \cap N_3^2 \right|$.

Таблиця 1. – Значення F_p та α

n_p	0	1	2	3	4
F_p	1	2	4	6	12

n_α	0	3	5	7	8	9	10
α	1	1,2	1,5	2,1	3	5,5	10

Для більш точної оцінки ступеня близькості новин вводяться множини N_2 відновлених лексем. Множини N_2^1 та N_2^2 відповідно для першої та другої новини формуються на основі N_1^1 та N_1^2 з доданням зібраної інформації про предметну область галузі. Наприклад, додаються дані про товари та географічні ринки агентів. Отже, на основі відновлених векторів лексем було введено вторинну умову оцінки ступеня близькості новин $F_s \le \alpha F_p$, ліва частина якої розраховується за формулою аналогічній (1), але на основі множин N_2^1 та N_2^2

$$F_s = \left[1 + \sum_{i=1}^{I} (\tilde{m}_i' - \tilde{m}_i'')^2 \right] \left[1 + \sum_{j=1}^{J} (\tilde{e}_j' - \tilde{e}_j'')^2 \right] \left[1 + \sum_{h=1}^{H} (\tilde{g}_h' - \tilde{g}_h'')^2 \right]. \quad (4)$$

Формули (1), (3), (4) поєднуються в одну комплексну формулу з урахуванням передумов щодо збігу категорії подій та близькості дат новин у межах d_t

$$F = \begin{cases} \vec{c}' = \vec{c}''; \\ \left| d' - d'' \right| \le d_t; \\ F_p \le \alpha; \\ F_s \le \alpha F_p \end{cases} \quad (5)$$

Умова (5) інтерпретується в такий спосіб: $F_p \le \alpha$ означає, що новини повинні бути схожі по множинам лексем зі словників N_1 тим більше, чим менше вони схожі за множинами інших лексем N_3; $F_s \le \alpha F_p$ означає, що різниця за множинами відновлених лексем N_2 неповинна перевищувати різницю за множинами лексем зі словників N_1 в межах погрішності α.

Слід зазначити, оскільки швидкість росту F_s, F_p однакова, тому вид нерівності $F_s \le \alpha F_p$ свідчить про більш питому вагу F_p над F_s, тобто більш питому вагу множини лексем зі словників N_1 ніж N_2 при об'єднанні двох новин у кла-

стер.

Отриманий після етапів класифікації й кластеризації потік новинних об’єктів буде мати майже взаємо-однозначну відповідність із реальними подіями, що породили відповідні їм новини.

У **третьому розділі** детально розглянуті способи вирішення задачі отримання фінальної ціни, як задачі цінового прогнозування на основі асоціативних та послідовних правил.

Для формування цінового прогнозу, а, значить, фінального значення ціни на товар у складі цінової стратегії, необхідно на основі отриманого потоку подій побудувати множину асоціативних правил поведінки ціни в залежності від цих подій. Для цього введемо додаткові позначення (рис. 4). Будемо вважати, що протягом часового відрізку τ на ринку m відбулася ринкова подія Y_i^τ, $i = \overline{1; Z_\tau}$. Вона призвела до зміни ціни p^τ. Позначимо додаткову подію Y_0^τ, що відображає напрямок зміни ціни

$$Y_0^\tau = \begin{cases} +1, & p^{\tau+1} - p^\tau > \tilde{p}_m; \\ -1, & p^\tau - p^{\tau+1} > \tilde{p}_m; \\ 0, & \left| p^{\tau+1} - p^\tau \right| \le \tilde{p}_m, \end{cases} \qquad (6)$$

де $\tilde{p}_m$ – мінімальний поріг цінових коливань для конкретного ринку, $\tilde{p}_m > 0$.

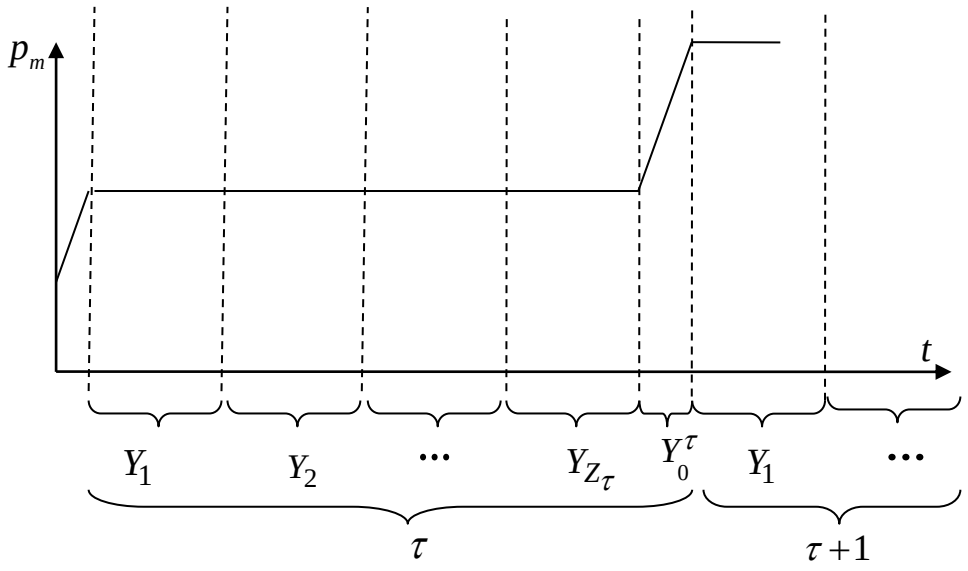

Рисунок 4. – Графічна інтерпретація потоку подій, доповнених змінами ціни

Тоді задача прогнозування цінових змін інтерпретується як задача пошуку для заданого ринка послідовностей подій у вигляді

$$Y_i^{\tau} Y_j^{\tau} \to Y_k^{\tau+1} \to Y_0^{\tau+2} \ . \qquad (7)$$

Послідовність (7) називається правилом. Правило (7) показує, що одночасне виникнення подій Y_i^{τ} та Y_j^{τ}, після яких з'являється подія $Y_k^{\tau+1}$, приводить до події зміни ціни $Y_0^{\tau+2}$ за формулою (6), де $i, j \in Z_{\tau}$, $k \in Z_{\tau+1}$. В роботі задача побудови правил типу (7) вирішується за допомогою алгоритму SPADE.

Тоді отримання фінального значення ціни на основі асоціативних правил, побудованих завдяки потоку інтернет новин, виконується за рахунок ідентифікації поточного ринкового стану та вибору правила, що найбільш відповідає цій поточній ситуації.

Позначимо множину добутих правил типу (7) як $r \in R$, де r – це послідовність (набір) ринкових подій, що виникає перед зміною ціни. Такі правила будуть визначати фінальне значення ціни на товар в заданому сегменті ринку. Тому запропоновано ідентифікувати ринкову ситуацію, що призводить до заданого значення фінальної ціни як відповідне правило r. При цьому кожному видобутому правилу r ставляться у відповідність два атрибути: s – підтримка, що характеризує абсолютну частоту появи правила у вихідній вибірці; c – ймовірність виникнення цінової зміни в зв'язку з появою набору подій, описаних правилом r.

Але при застосуванні даного підходу виникає наступна проблема: при ідентифікації сформованої ситуації і виборі правила існує невизначеність, тому що асоціативні правила r мають різну вірогідність і підтримку. Правило може мати як дуже високу підтримку, тобто бути очевидним правилом, так і навпроти мати дуже низьку та бути неочевидним правилом. Отже, якість прогнозів прямо залежить від методу ідентифікації сформованої ситуації.

Для вирішення цієї проблеми в роботі запропоновано наступний алгоритм рис. 5. Основа ідея алгоритму полягає в тому, що для заданого рівня вірогідності C і підтримки S відібрати всі правила, для яких $c \geq C$, $s \geq S$ відповідно. Тоді для обраного періоду часу τ треба знайти найбільш підходяще правило на основі наступної послідовності дій:

1. Задати $n = 1$.
2. Відібрати множину правил R_n, що задовольняють поточній ситуації, яка існувала на ринку протягом часового відрізку τ.
3. Якщо знайдено тільки одне правило $\left|R_n\right| = 1$, тоді формулюється фінальне значення ціни на товар. Якщо знайдено більше одного правила, зрівнюються прогнозні значення, якщо прогноз спрямований у один бік зміни ціни (нагору/униз), то сформувати сумарне прогнозне значення;

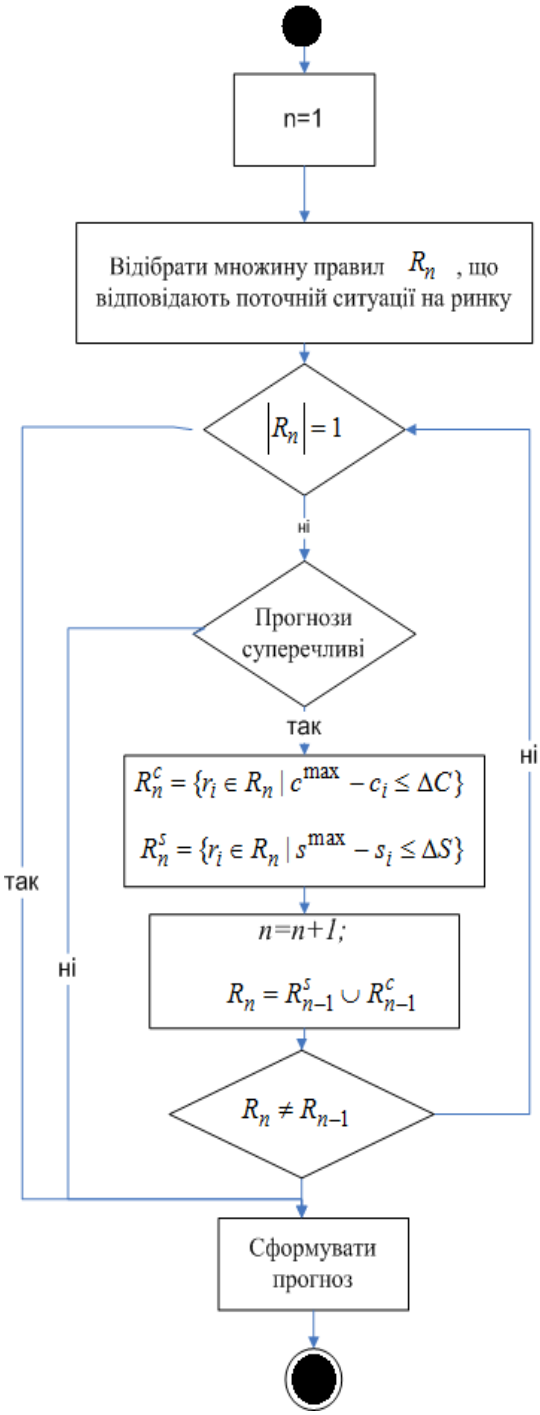

Рис. 5. Алгоритм прогнозування на основі асоціативних правил

4. Якщо знайдено більше одного правила та прогнозні значення суперечливі, то сформувати із множини R_n множину шляхом виключення менш достовірних правил у відповідність із умовою

$$R_n^c = \{r_i \in R_n \mid c^{\max} - c_i \le \Delta C\},$$

де $c^{\max}$ – найбільший параметр вірогідності в поточній множині правил, $\Delta C = \alpha c^{\max}$ – припустима погрішність, α – коефіцієнт погрішності. Сформувати множину правил R_n^s шляхом виключення менш неочевидних правил відповідно до формули

$$R_n^s = \{r_i \in R_n \mid s^{\max} - s_i \le \Delta S\},$$

де $s^{\max}$ – правило з найбільшим параметром підтримки в поточній множині правил, $\Delta S = \beta s^{\max}$ – припустима погрішність, β – коефіцієнт погрішності.

5. Збільшити n на одиницю, сформувати нові множини $R_n = R_{n-1}^s \cup R_{n-1}^c$.

6. Якщо $R_n \neq R_{n-1}$, то перейти до пункту 3. Інакше вибрати найбільш очевидне правило та достовірне правило з R_n та сформувати прогноз.

Слід зазначити, що значення параметрів C й S безпосередньо впливають на якість прогнозів, і визначення їхніх оптимальних значень є окремою дослідницькою задачею.

Розглядаючи ціну, як випадковий процес, доцільно також додаткове підвищення точності прогнозів за рахунок застосування апарата теорії викидів, яка дозволяє вести спостереження за наступними параметрами: середнє число викидів (перетинання деякого рівня ціни H), тривалість викидів, тривалість переходів між викидами.

У результаті, маючи чисельні дані про середній час і дисперсію перебування ціни над рівнем H, розширюються й уточнюються прогнози значення ціни, отримані з використанням асоціативних та послідовних правил, що дозво-

лить корегувати майбутнє значення, з урахуванням часу перебування ціни в області даного значення.

У **четвертому розділі** представлена структура розробленого програмного продукту, надана детальна інструкція користувача, розглянуто по крокам приклад використання розробленої WEB-базованої сервіс-орієнтованої інформаційної технології та наведені результати експериментальних досліджень.

Процес функціонування програмного продукту зображено на рис. 6. На першому етапі відбувається збір новин з потоку інтернет новин. На другому етапі, на основі множин синтаксичних моделей відбувається класифікація та формування словників лексем новин. Прикладом такої множини лексем є

$$N_1 = (Изменение\ производст.\ мощн., Таиланд, Ритас).$$

На третьому етапі вирішується проблема наявності дублікатів і сюжетних ланцюжків у потоках новин за допомогою застосування методу ієрархічної кластеризації до множини векторів новин. Результатом виконання третього етапу є очищений потік подій, на підставі якого за допомогою технологій пошуку асоціативних правил аналізуються закономірності поведінки ціни на четвертому етапі.

На п’ятому етапі визначається поточна ситуація на ринку й обираються найбільш адекватні правила для прогнозу, та виконується розрахунок прогнозу ціни.

Програмний продукт розроблено на основі клієнт-серверної архітектури з тонким клієнтом для російськомовного потоку інтернет новин. Для роботи із розробленим WEB-базованим сервіс-орієнтованим ПЗ користувачу необхідна наявність WEB браузера з підключенням до інтернет.

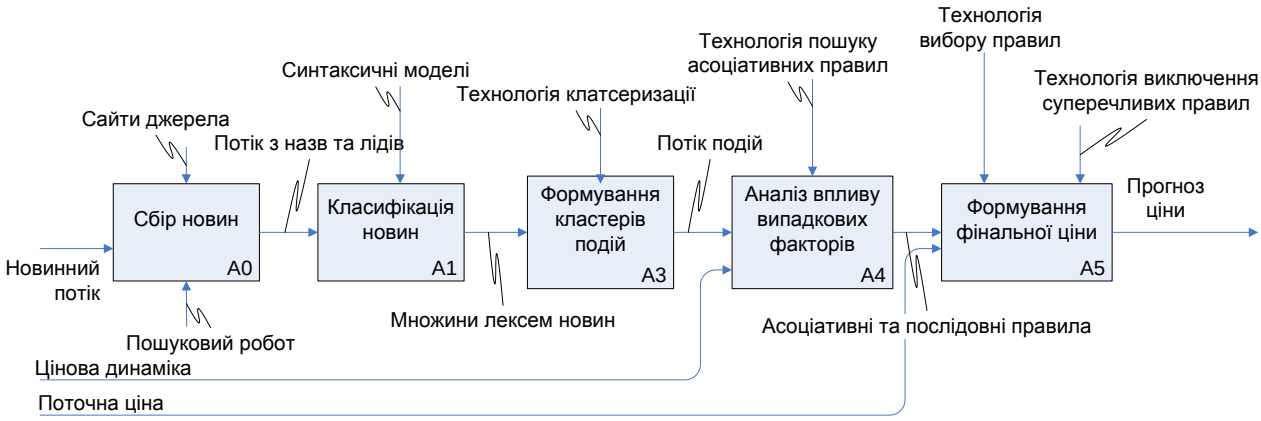

Рисунок 6. – Процес функціонування ПЗ

Серверна частина програмного продукту складається з компонентів Java Server Pages. Компонента Forecasting tool обробляє запити користувача. Компонента News grabber є «роботом», який завантажує інтернет сторінки із заданих джерел. Візуальна компонента, що формує графік прогнозу здійснюється компонентою GoogleGraphs, що розташована на сервері GoogleApps. Взаємодія між

компонентами серверної частини ПЗ і GoogleApps сервером здійснюється за допомогою протоколу TCP/IP. Для зберігання даних, а також виконання алгоритму пошуку асоціативних правил використовується компонента БД. У якості серверу БД виступає PostgreSQL, взаємодія меж серверною частиною ПЗ и сервером БД виконується із використанням драйверу JDBC.

Фрагмент інтерфейсу користувача представлено на рис. 7. На ньому відображена цінова ломана, що включає історичні значення ціни (точки J та K) та прогнозну ціну (точка L).

Буквами на графіку цінової ламаної історичних значень відображені події отримані з потоку інтернет новин. До кожної букви надається детальний опис: які саме події відбулися та їх типи. Також користувач може вибирати довжину перегляду прогнозу, працювати з переглядом правил і новин, працювати з інструментарієм вибору параметрів вірогідності та складності правил прогнозу, параметрів «історичного» періоду для навчання, можливість примусово відновлювати список асоціативних правил.

Формування множин подій, які відбулися у кожній окремій точці, виконується із використанням розроблених технологій класифікації та кластеризації, процес функціонування яких є прихованим відносно користувача. Отримані множини подій не містять дублікатів та сюжетних ланцюжків.

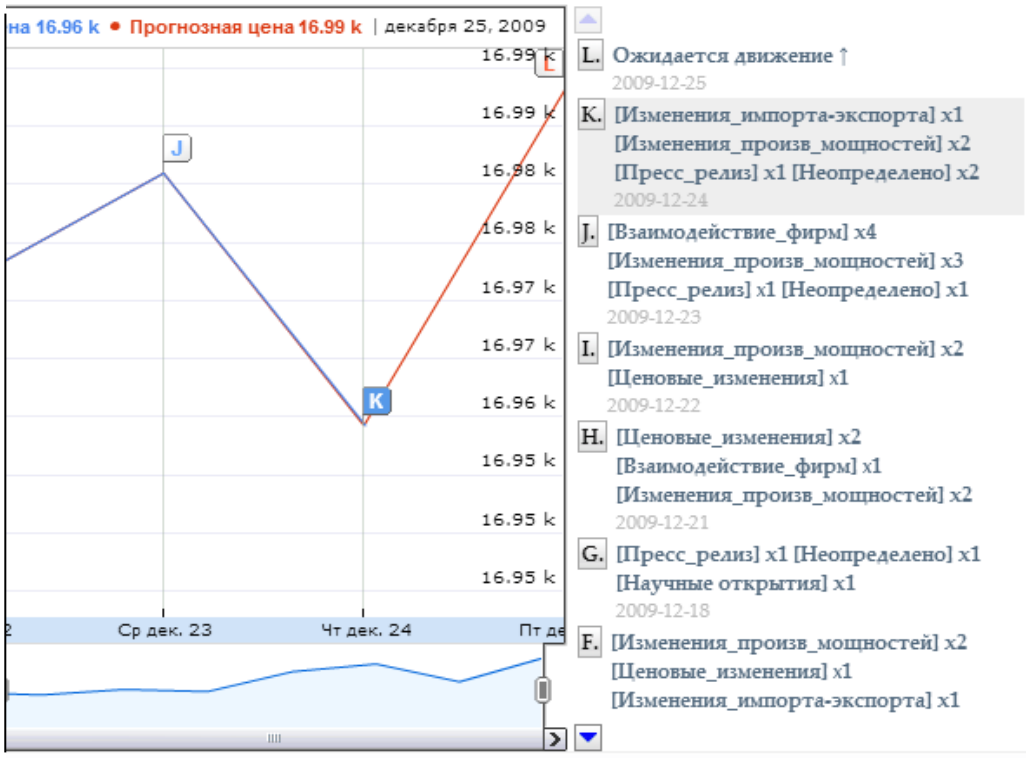

Рисунок 7. – Фрагмент інтерфейсу користувача

Зображеній на рис. 7 точці «К» поточного стану ринка при довжині правил 4 та вірогідності не нижче 0,7 в процесі функціонування WEB-базованого сервіс-орієнтованого ПЗ отримані наступні правила:

[Взаимодействие_фирм] + ↓ ▷ *[Изменения_произв_мощн]* ▷ ↑
C =0,83 S=29,
↓ ▷ *[Изменения_произв_мощн]* + *[Изменения_импорта-экспорта]* ▷ ↑
C=0,72 S=16,
[Пресс_релиз] +↓ ▷ *[Изменения_импорта-экспорта]* ▷↓
C=0,82 S=7,
[Взаимо_фирм] + *[Изм_произв_мощн]* ▷ *[Изм_импорта-экспорта]* ▷↓
C=0,74 S=25.

Кожному правилу ставиться у відповідність два параметри: підтримки S, та достовірності C.

Технологія формування множини правил для прогнозу у точці «К» виконана поетапно, спочатку були відкинуті усі правила, ключ яких не закінчувався однією з подій в точці «К», далі, доки не була досягнута глибина правил 4, на кожному з етапів було відкинуто ті правила, у яких попередня частина не мала жодного типу події з попередньої точки стану ринку.

У відповідності до технології виключення суперечливих правил (рис. 5), виключаються усі правила, що мають недостатні параметри достовірності та підтримки. Фінальний прогноз ціни формується як агрегація несуперечливих правил прогнозу. Множина правил, що сформувала фінальний прогноз ціни в точці «К» стану ринку зображена на рис. 8.

№	
1	[Взаимодействие фирм] + ↓ ▷ [Изменение производственных мощностей] ▷ ↑
2	↓ ▷ [Изменение производственных мощностей] + [Изменения импорта-экспорта] ▷ ↑

Рисунок 8. – Відібрані правила прогнозу

Розроблена сервіс орієнтована WEB-базована інформаційна технологія була застосовано для формування цінової стратегії на прикладі ринку полімерів України. В якості вхідної інформації використано цінові значення по поліетилену високого тиску (ПЕВТ) та низького тиску (ПЕНТ) за 2009 – 2011 рр. і відповідний до цієї цінової динаміки поток інтернет новин. Розмір вибірки цінових значень склав 800 записів, вибірки інтернет новин – 2700 записів, при цьому як навчальна вибірка були взяті перші 600 значень ціни й відповідні їм 2100 новин. Зі значень, що залишилися, сформовані дві контрольні вибірки.

Проведені експерименти дозволили оцінити якість роботи розробленої сервіс орієнтованої WEB-базованої формаційної технології, якість методів прогнозування оцінювалася на основі моделей, побудованих з мінімізованим значенням функції правдоподібності, множина асоціативних правил отримана за допомогою алгоритму SPADE. Для оцінки точності розробленого методу прогнозування розглянуті методи, представлені в табл. 2.

Аналіз отриманих результатів свідчить, що прогнози, одержані на основі асоціативних правил, на 6% точніше, ніж у загальноприйнятих методів.

Таблиця 2. – Експериментальні значення помилки прогнозів MAPE

	Вибірки	
Метод	Контрольна 1	Контрольна 2
Експонентне згладжування	38.2%	39.5%
ARIMA	37.4%	37.1%
Асоціативні правила	30.9%	31.2%

Більша точність досягається завдяки тому, що прогнози на основі новинних потоків за допомогою асоціативних правил дозволяють безпосередньо включати події, що впливають на формування ціни, у прогнозне значення, у той час як регресійні методи включають ці події побічно.

Економічний ефект від впровадження розробленого розробленої сервіс орієнтованої WEB-базованої формаційної технології на ТОВ «Енергетехінвест» (м. Харків) стосовно ринків ПЕВТ і ПЕНТ склав 32 тис. грн.

У **додатках** наведено акти впровадження розробленого програмного продукту на ТОВ «Енергетичні технології» та ТОВ «Енергетехінвест» (м. Харків).

ВИСНОВКИ

У дисертаційній роботі вирішена науково-практична задача розробки моделей та інформаційної технології, що дозволяють провести дослідження впливу потоку інтернет новин на формування цінової стратегії підприємства. Проведені в дисертації дослідження дали можливість отримати наступні наукові та практичні результати:

1. На основі аналізу сучасного стану задачі формування цінової стратегії підприємства на базі сервіс-орієентовинної WEB-технології встановлено, що для вибору цінової стратегії з урахуванням нового джерела маркетингових даних – потоку інтернет новин в форматі RSS-XML слід розробляти оновлену модель вибору цінової стратегії.

2. Для формування фінального значення цін на товар, розроблено оновлену модель формування цінової стратегії, яка ґрунтується на аналізу двох підзадач: виявлення випадкових зовнішніх факторів (ринкових подій) в потоці інтернет новин і формування асоціативних та послідовних правил поведінки ціни.

3. Обґрунтована необхідність використання морфологічно-синтаксичних моделей тексту для первинного аналізу новин. Створений комплекс синтаксичних моделей новин дозволяє аналізувати інтернет новини, як контейнери маркетингової інформації. Використання розроблених 6 груп моделей новин дає змогу класифікувати до 85% новин на прикладі ринку полімерів. Велика точність класифікації досягається за рахунок проведення виключно вибіркового аналізу змісту новини: назви і ліда.

4. Для виключення дублікатів та сюжетних ланцюжків розроблено технологію кластеризації, що дозволяє проводити якісний аналіз потоку подій, що

побудована на основі комплексу умов, що дозволяють оцінити ступінь близькості новин та сформувати кластери унікальних ринкових подій.

5. Досягнення на 6% більшої точності прогнозу ціни ніж у загальноприйнятих методів досягається за рахунок використання розробленого методу формування фінального значення ціни на основі асоціативних та послідовних правил із параметрами підтримки $S = 7$ та реалізації $C = 0{,}8$.

6. Для формування фінального значення ціни на основі випадкових зовнішніх факторів шляхом аналізу впливу зовнішніх факторів на цінову динаміку запропоновано застосовувати розроблену інформаційну технологію на базі сервіс-орієнтованної архітектури на базі WEB сервісів компанії Google.

7. Проведено апробацію розробленої інформаційної технології у процесах формування цінової стратегії на підприємствах Харківщини, зокрема в ТОВ «Енерготехінвест» та ТОВ «Енергетичні технології» (м. Харків). Отримані результати наукового дослідження використовуються в навчальному процесі на кафедрі автоматизованих систем управління НТУ «ХПІ».

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Черенков И.А. Обоснование прогнозирования цен полимеров посредством новостного потока / И.А. Черенков, С.В. Орехов // Східно-Європейський журнал передових технологій. – Харків : Технологічний центр, 2010. – №5/7 (47). – С. 18-21.

Здобувач дослідив проблематику цінового прогнозування і розглянув підходи до прогнозування на базі потоку новин.

2. Черенков И.А. Автоматический поиск данных из новостей на примере рынка полимеров / И.А. Черенков // Системи обробки інформації. – Харків: ХУПС, 2011. – №8 (98). – С. 156-160.

3. Черенков И.А. Прогнозирование на основе новостного потока посредством ассоциативных правил / И.А. Черенков // Енергозбереження. Енергетика. Енергоаудит. – Харків: БЕТ 2012. – №11 (105). – С. 38-43.

4. Черенков И.А. Добыча данных из текстовых новостей на примере рынка полимеров / И.А. Черенков, С.В. Орехов // Системи обробки інформації. – Харків: ХУПС, 2012. – №9 (107). – С. 224-228.

Здобувач дослідив процес видобутку даних з текстових новин на основі морфологічного і синтаксичного аналізу.

5. Черенков И.А. Подход выделения событий в новостном потоке / И.А. Черенков, С.В. Орехов // Східно-Європейський журнал передових технологій. – Харків : Технологічний центр, 2013. – №1/4 (61). – С. 62-64.

Здобувач дослідив процес обробки та виділення кластерів дублікатів та сюжетних ланцюжків у потоці інтернет новин.

6. Черенков И.А. Методы решения задачи краткосрочного ценового прогнозирования / И.А. Черенков // Актуальные вопросы экономики:

проблемы, гипотезы, исследования: Международная научно-практическая конференция. – Симферополь : Economics, 2012. – С. 145-147.

7. Черенков И.А. Автоматическое ценового прогнозирование на основе анализа интренет новотных потоков / И.А. Черенков // Радиоэлектроника и молодёжь в 21 веке: материалы XVII Международного молодёжного форума. – Харьков : ХНУРЭ 2013. – т. 7. – С. 211-212.

8. Черенков И.А. Решение задачи ценового прогнозирования на примере рынка полимеров / И.А. Черенков, С.В. Орехов // Математические методы в технике и технологиях ММТТ-25 : XXV Международная конференция (Харьков, 2012г.). – Саратов : СГТУ. – 2012. – т. 5. – С. 27-30.

Здобувач дослідив процес цінового прогнозування на прикладі ринку полімерів.

9. Черенков І.О. Вирішення задачі прогнозування зміни ціни на полімери за допомогою аналізу потоку новин / С.В. Орехов, І.О. Черенков // Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: XVIII Міжнародна науково-практична конференція (Харків, 12-14 травня 2010р.). – Харків : НТУ «ХПІ». – 2010. – ч. 1. – с. 21.

Здобувач дослідив процес зміни ринкової ціни в залежності від потоку новин на прикладі ринку полімерів.

АНОТАЦІЇ

Черенков І.О. Інформаційна технологія формування цінової стратегії підприємства на основі потоку інтернет новин. – 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, 2014.

Дисертація присвячена розробці WEB-базованої сервіс-орієнтованої інформаційної технології, що дозволяє завдяки виявленню зовнішніх факторів ринкової середи проводити вибір фінальної ціни на товар шляхом автоматичного короткострокового прогнозуванню ціни на основі текстового потоку інтернет новин. На підставі досліджень підходів видобутку даних з текстових об'єктів визначено, що задля успішного видобутку даних з новини необхідне проведення морфологічно-синтаксичного аналізу. Класифікатор новин, заснований на даному підході, дозволяє виділяти категорію події і ключову інформацію про подію. Для достовірного прогнозування на основі пошуку асоціативних правил крім класифікації новинних об'єктів, необхідно виділення кластерів новин, що виключають дублікати і сюжетні ланцюжки. Розроблений підхід кластеризації в сукупності з умовами поєднання новинних об'єктів дозволяють отримувати потік подій високої якості, достатній для формування цінового прогнозу. Цінове прогнозування здійснюється на основі пошуку асоціативних правил у вигляді наборів подій, що призводять до конкретної зміни ціни. Ключовими перевагами підходу є мінімізація суб'єктивного фактора в прогнозах та підвищення їх точ-

ності, завдяки безпосередньому обліку в прогнозі поточних подій, що впливають на ціну.

Ключові слова: машинне навчання, цінове прогнозування, синтаксичний аналіз, класифікація, кластеризація, асоціативний пошук, data minig, цінова стратегія, web технологія.

Черенков И.А. Информационная технология формирования ценовой стратегии предприятия на основе потока интернет новостей. – 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, 2014.

Диссертация посвящена разработке WEB-базированной сервис ориентированной информационной технологии, позволяющей с помощью выявления внешних факторов рыночной среды формировать финальное значение цены на товар путем автоматического краткосрочного прогнозированию цены на основе потока интернет новостей.

Исследования показали, что большинство общепринятых средств прогнозирования, используемых в формировании ценовой стратегии, обладают недостаточной точностью, что обусловлено механизмом формирования прогноза: исключительно на основе предшествующей закономерности поведения цены. Основным недостатком такого подхода является тот факт, что одному и тому же поведению цены могут соответствовать разные наборы событий.

Разработка подхода прогнозирования, учитывающего влияние внешних факторов на поведение цены, требует реализации механизма обработки происходящих событий описываемых новостями, а также реализации непосредственно механизма прогнозирования. Оптимальным источником данных для автоматической обработки событий является новостной поток в интернет, имеющий текстовую природу.

На основании исследований подходов добычи данных из текстовых объектов было определено, что для высокоточной добычи данных из новости необходимо проведение морфологическо-синтаксического анализа, который в свою очередь должен осуществляться применительно к ключевым элементам новости, таким как: название, лид. Классификатор новостей, основанный на данном подходе, позволяет выделять категорию события и ключевую информацию о событии. Формирование множества моделей морфологическо-синтаксического анализа основывается на теории грамматик непосредственных составляющих.

Для достоверного прогнозирования на основе поиска ассоциативных правил помимо классификации новостных объектов и выделения категорий событий, описанных в новостях, необходимо выделение кластеров новостей, исключающих дубликаты и сюжетные цепочки. Сформулированный подход кластеризации в совокупности с условиями объединения новостных объектов позво-

ляют получать поток событий высокого качества, достаточный для формирования ценового прогноза.

Ценовое прогнозирование осуществляется на основе поиска ассоциативных правил в виде наборов событий, приводящих к конкретному изменению цены. В результате, после формирования множества правил, для выбранного момента времени производится поиск правила максимально приближённого к текущей рыночной ситуации.

Дополнительное повышение точности прогнозов осуществляется за счёт применения элементов теории выбросов, позволяющих определить среднее время пребывания цены над некоторым уровнем, среднее количество переходов цены через границу уровня.

Ключевыми преимуществами подхода является минимизация субъективного фактора в прогнозах и повышение их точности, благодаря непосредственному учёту текущих событий в прогнозе цены.

Ключевые слова: машинное обучение, ценовое прогнозирование, синтаксический анализ, классификация, кластеризация, ассоциативный поиск, data minig, ценовая стратегия, web технология.

Cherenkov I.A. Information technology of the enterprise's pricing strategy based on internet news stream. – On the rights of the manuscript.

Thesis for granting the Degree of Candidate of Technical sciences in speciality 05.13.06 – information technologies. – National Technical University «Kharkiv Polytechnic Institute», Kharkiv, 2014.

The thesis is dedicated to the development of service oriented WEB based information technology. It allows us by analysis of external market's factors to support decision making in pricing strategy by an automatic approach for short-term forecasting based on online text news flow. The research of existing approaches of data mining from text objects shows that it is necessary to conduct the morphological-syntactic analysis of news for highly accurate data extraction. News classifier based on this approach allows to extract the event's category and key information about it. Developed approach in combination with a set of syntactic models can accurately classify objects text news flow. For reliable price prediction based on search of associative rules in addition to news classification it is necessary to form according of news clusters, which will exclude duplicates and plot strings. Proposed hierarchical clustering algorithm, which is based on news objects' proximity function allows to form news clusters thus get the news flow of according quality for forecasting. The price forecasting is based on the search of associative rules of news events which lead to specific changes in price. The key advantage of developed approach is a reduction of human factor in the predictions. The accuracy of predictions is based on direct accounting of forecast events that affect on a price.

Keywords: machine learning, price prediction, syntactic analysis, classification, clustering, associative search, data mining, pricing strategy, web solutions.

Підписано до друку 14.02.2014 р. Формат 60х84/16.

Папір офсетн. Друк – різографічний. Умовн. друк. арк. 0,9.

Гарнітура Times New Roman. Наклад 100 прим. Замовлення №

Надруковано у копі-центрі «МОДЕЛІСТ»

ФО-П Миронов М.В., Свідоцтво ВО4№022953
м. Харків, вул. Червонопрапорна, 3 літер Б-1
