УДК 004.89

В.В. Литвин, А.С. Мельник, В.Я. Крайовський
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

ОЦІНКА НОВИЗНИ ЗНАНЬ ПІД ЧАС АВТОМАТИЧНОЇ РОЗБУДОВИ ОНТОЛОГІЙ

Розглядається оцінка новизни знань під час автоматизації побудови онтології на основі аналізу текстових ресурсів. Розвиток онтології починається з деякої базової онтології, яка задає досліджувану предметну область. Розроблено алгоритм оцінки новизни знань.

Ключові слова: онтологія, концепт, новизна знань.

In the paper the evaluation of the novelty of knowledge during the automation of ontology-based text analysis resources. Development of ontology begins with some basic ontology, which specifies the subject area studied. The algorithm estimates the novelty of knowledge.

Key words: ontology, concept, novelty of knowledge, quality assessment.

Постановка проблеми у загальному вигляді

Функціонування інтелектуальних систем підтримки прийняття рішень (ІСППР) – це постійне прийняття рішень на основі аналізу поточних ситуацій для досягнення певної мети. Типова схема функціонування ІСППР складається з таких трьох кроків: 1) планування цілеспрямованих дій та прийняття рішень, тобто аналіз можливих дій і вибір тієї дії, яка якнайкраще узгоджується з метою системи; 2) зворотна інтерпретація прийнятого рішення, тобто формування робочого алгоритму для здійснення реакції системи; 3) реалізація реакції системи, наслідком чого є зміна зовнішньої ситуації та внутрішнього стану системи.

Центральною підсистемою ІСППР є база знань (БЗ), яка займається зберіганням, впорядкуванням та керуванням інформацією про навколишній світ. Найважливіший параметр БЗ – якість та повнота знань про ПО, яку вона задає. Якість БЗ залежить від структури та формату знань, способу їх подання.

Для широкого впровадження будь-якої технології чи методики необхідний чіткий і арґументований стандарт. У галузі розроблення БЗ таким стандартом стали онтології. Онтологією називається експліцитна специфікація концептуалізації. Формально онтологія складається з термінів (понять, концептів), організованих в таксономію, їх визначень і атрибутів, а також пов'язаних з ними аксіом і правил виведення [1].

Сьогодні розрізняють три типи онтологій: предметно-орієнтовані (Domain-oriented), орієнтовані на прикладну задачу (Task-oriented) та загальні онтології (Top-level). Онтологія ПО містить таксономію понять, додаткові відношення, екземпляри класів і різні види обмежень (аксіом). Аксіоми встановлюють семантичні обмеження для системи відношень. Мета онтології задач – зробити знання доступними для повторного використання. Онтології задач визначають ступінь використання знань в процесі логічного виведення. Загальна онтологія описує категорії – поняття верхнього рівня. Ми будуємо єдину онтологію, яка містить відразу три наведені типи онтологій. Ієрархічно це виглядає так: загальна онтологія знаходиться на верньому рівні ієрархії, а онтології ПО та задач до неї під'єднуються. Такий підхід дає можливість цілісно розглядати усі задачі у межах ПО.

Для побудови онтологій використовуються усі вищенаведені чотири моделі подання знань: для задання понять використовують фрейми, для задання відношень – семантичні мережі, для задання аксіом – логіку 2-го порядку, для побудови правил виведення – продукційну систему. Семантичну мережу фреймів (концептів) називають концептуальним графом (КГ).

Для того, щоб вручну побудувати повну зв'язану онтологію для певної ПО, необхідно витратити достатньо багато часу та ресурсів. Причина таких затрат полягає у тому, що такі онтології повинні містити десятки тисяч елементів, щоб бути придатними для розв'язування широкого кола прикладних задач, які виникають у цих ПО. Отже, ручна побудова онтології людиною-оператором – це довгий рутинний процес, який до того ж вимагає ґрунтовних знань ПО та розуміння принципів побудови онтологій [2, 3]. Для цього необхідно розробити методи та алгоритми автоматичної побудови онтології. Ми вважаємо, що базові терміни та відношення між ними повинні бути введені людиною-експертом в онтологію вручну. Таку початкову онтологію називатимемо базовою і позначатимемо

$$O_{base} = \langle C_b, R_b, F_b \rangle,$$

де C_b – скінченна множина понять (концептів, термінів) предметної області, яку задає онтологія O_{base}; R_b – скінченна множина відношень між концептами (поняттями, термінами) заданої предметної області; F_b – скінченна множина функцій інтерпретації (аксіоматизація, обмеження), заданих на концептах чи відношеннях онтології O_{base}. Отже, процес побудови онтології починається з моменту, коли в ній вже є якісь дані. Тому такий процес називатимемо розбудовою базової онтології і позначатиме

$$c : O_{base} \to O. \qquad (2)$$

Онтологія – це мова науки. Мова науки як структуроване наукове знання задає багатошарове ієрархічне утворення, в якому виділяються блоки: терміносистема; номенклатура; засоби та правила формування понятійного апарата і термінів.

Отже, з точки зору процесу побудови онтології необхідно побудувати її терміносистему O_T та номенклатуру O_N. У нашому підході базова онтологія повинна чітко включати в себе частину терміносистеми (див. рис. 3), тобто $O_B \cap O_T \neq \varnothing$.

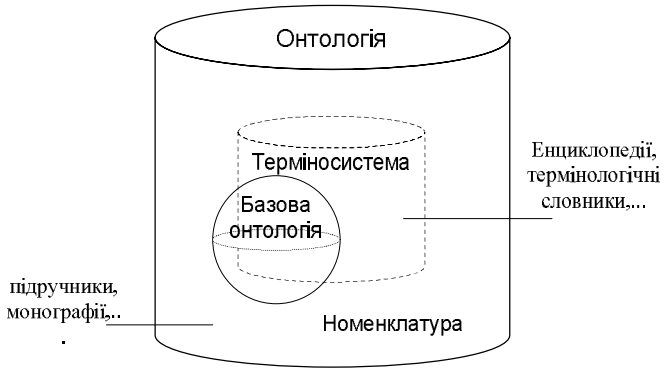

Рис. 1. Архітектура онтології

Енциклопедії, термінологічні та толкові словники, на основі яких будується терміносистема ПО, як правило, мають чітку структуру і складаються із словникових статей. Тому необхідно дослідити можливі їх структури з метою розпізнавання понять і відношень між ними. Побудова номеклатури складніша. Якщо в словниках терміни в деякий спосіб вже виділені, то в наукових текстах (підручники, монографії тощо) їх необхідно виділяти, здійснити пошук властивостей понять і відношень між поняттями. Ми вважаємо, що ефективність онтології напряму залежить від новизни знань, які до неї додаються. Тому ми запропонували під час розбудови онтології оцінювати її новизну знань. Тоді базова онтологія поповнюється знаннями, новизна яких більша. Ще у 1999 році економіст Хал Варіан припустив, що з точки зору економіки, "лише нова інформація має значення" [4, с. 122].

Аналіз останніх досліджень та публікацій

Для побудови онтологій, які адекватно описують семантичні моделі ПО, необхідно насамперед розв'язати задачі видобування знань із різних джерел для виявлення множини концептів і встановлення ієрархії на цій множині. Оскільки значна частина інформації міститься у природно-мовних текстах (ПМТ), перспективним є видобування знань із текстової інформації, а також інтелектуальне опрацювання спеціально підібраних колекцій ПМТ.

Існує багато перспективних лінгвістичних розробок, серед яких доцільно виділити метод лексико-граматичного аналізу (Part-of-Speeach-tagging), який полягає в автоматичному розпізнаванні, до якої частини мови належить кожне слово у тексті. Для підвищення точності такого аналізу використовують два типи алгоритмів: ймовірнісно-статистичні та алгоритми на основі продукційних правил, що оперують словами і кодами. Щодо останніх, то вони можуть використовувати правила, автоматично зібрані з корпусу текстів або ж підготовлені кваліфікованими лінгвістами.

Серед систем, розроблених в Україні, треба відзначити розробку кафедри математичної інформатики Київського національного університету імені Тараса Шевченка – систему опрацювання текстів природною мовою. Система створена для розв'язування таких задач, як аналіз та синтез текстів природною мовою, автоматичне генерування реферату тексту, автоматична індексація (визначення тематики) тексту. Найвагомішим технічним рішенням в системі є можливість "зважувати" вершини семантичної мережі тексту. При цьому найважливішими вершинами мережі вважаються вершини, які мають найбільшу кількість зв'язків з іншими. Ця процедура може застосовуватися під час побудови образу реферату зважуванням вершин і відкиданням найлегших – "маргінальних".

Синтаксично-семантичний аналізатор Link Parser – один із найефективніших підходів до автоматичного аналізу тексту, розроблений та фактично реалізований в Університеті Карнегі-Мелона. Аналізатор використовує апріорну інформацію про типи зв'язків, які може мати кожне слово з іншими словами, розташованими в реченні справа та зліва від нього, а також порівняно невелику кількість загальних граматичних правил. Вихідні коди цього аналізатора опубліковані зі статусом "open source", що дає змогу вільно використовувати їх для аналізу семантичної структури тексту.

Семантичний розбір є лише підготовчим етапом інтелектуального аналізу тексту, метою якого є прийняття інтелектуальною системою рішення про класифікацію певного тексту (розпізнавання) чи про існування деякого нового знання, яке треба внести до БЗ (навчання).

Оцінка новизни знань у найширшому своєму значенні включає визначення будь-яких аномалій, відхилень від норми у певному наборі даних. Вона має широкий спектр застосувань, від пошуку новоутворень (ракових клітин) в організмі людини і до виявлення роботом незнайомого типу ландшафту. Проте у цій роботі ми розглядаємо лише проблему оцінки новизни знань у текстових фрагментах. Навіть за такого обмеження задача пошуку нової інформації широко застосовується в сучасних інформаційних сервісах – під час створення тимчасових оглядів новин, побудови мінімального набору документів відповідно до запиту, асистування експерта під час аналізу великих обсягів текстових документів, що містять дублікати. Наприклад, коли лікар аналізує історію хвороби пацієнта, він проглядає інформацію про багаторазові обстеження зі здебільшого подібними результатами. Зазвичай, головне, що його цікавить – це поява нових симптомів, змін у самопочутті пацієнта, відхилення результатів аналізів від норми чи попереднього стану тощо, тобто "нові" дані.

Дамо визначення поняттю новизни:

Новизна, або нова інформація – це нові відповіді на потенційні питання в контексті користувацького запиту.

Зауважимо, що у визначенні вжито фразу "в контексті користувацького запиту", тобто пошук нових даних проводиться з урахуванням певних меж зацікавленості користувача. Це дуже важливо, оскільки два різні речення дуже часто міститимуть певну нову інформацію одне стосовно іншого. Наведемо приклад:

1. “Конференцію відвідали 200 науковців”.
2. “Конференцію відвідали 200 молодих науковців”.

Якщо не визначено жодного контексту, зрозуміло, що друге речення містить нові дані стосовно першого (той факт, що учасники конференції – молоді). Проте, якщо запит користувача звучить як “Кількість учасників конференції”, стає зрозуміло, що обидва речення повторюють одне одного у цьому контексті.

Пошук нової інформації може відбуватись на рівні окремих подій або окремих речень. У першому випадку під подією розуміємо документ, статтю, новину тощо, – тобто певний фрагмент інформації, визначений у межах системи, що трактується як єдине ціле. У цьому випадку фрагмент даних може розбиватись на частини в процесі аналізу, але висновок щодо новизни приймається повністю для усього фрагмента. Наприклад, агрегатор новин може аналізувати окремі статті на рівні речень, проте в результаті він приймає рішення щодо того, чи показувати повністю статтю користувачу, чи ні. У разі пошуку новизни на рівні окремих речень будь-який фрагмент тексту розбивається на менші неподільні частини, обмежені границями одного речення. Результатом пошуку нових знань є набір речень. Саме такий підхід розглядається у цій роботі.

Формування цілей

Мета роботи – побудувати методи та алгоритми оцінки новизни знань, які містяться у природомовних текстах під час автоматичної розбудови онтології. Враховуючи мету роботи, визначимо *об’єкт дослідження* – процес аналізу великих текстових фрагментів з високим рівнем надлишкової (дублюючої) інформації з метою виділення лише унікального вмісту у певному контексті. *Предмет дослідження* – гнучкі алгоритми виявлення нових знань у текстових фрагментах, здатні до самонавчання на основі відгуку користувача.

Основний матеріал

Розроблена у межах цієї роботи підсистема оцінки знань використовує алгоритми навчання, ґрунтуючись на відгуку від користувача. Самонавчання як підхід до розв’язання задач штучного інтелекту не є сам по собі принципово новим, але в сфері визначення новизни знань майже не використовувався. Це може бути пов’язано з тим, що навчання системи часто вимагає експертів і є доволі дорогим з економічного погляду. З врахуванням цього аспекта ми використали дуже простий спосіб відгуку користувача – підсистема отримує дані лише про те, які речення користувач визначив вартими уваги. У межах зовнішньої інформаційної системи це може бути реалізовано дуже прозоро для користувача – фрагменти інформації, визначені підсистемою аналізу новизни як дублюючі, можуть показуватись у підсумованій формі з можливістю розгортання певного блока тексту. У разі його розгортання цей факт передається назад у підсистему виявлення новизни для зміни її параметрів. У такий спосіб навчання системи може відбуватись у режимі її нормальної роботи, без необхідності залучення експертів. Початкові дані системи отримані на основі аналізу бази знань Acquiant Collection.

Оцінка ефективності системи виявлення новизни – це доволі суб’єктивний показник, що залежить від експерта, який використовує систему. Саме експерт (або користувач системи) визначає, які речення, на його думку, несуть в собі “нову” інформацію. Ця оцінка змінюватиметься від одного користувача до іншого. Тим не менше, з усередненої оцінки розробленого прототипу після початкового навчання можна зробити висновок, що точність виявлення новизни розробленої підсистеми становить близько 70 %. Це доволі високий показник серед результатів сучасних досліджень (для порівняння, алгоритм-переможець змагання у межах TREC Novelty Track у 2004 році досягнув рівня точності 60–65 %). Це вказує на перспективність обраного підходу.

Автоматичне визначення смислової подібності речень – непросте завдання. Коли людина аналізує прочитаний текст, вона використовує асоціативне мислення, здатність абстрагувати та наводити паралелі, попередній досвід, лексичні особливості мови тощо – процеси, які надзвичайно важко змоделювати на комп’ютері. Якщо комп’ютерна програма може зробити певний висновок про подібність слів “тварина” та “тварини” (ґрунтуючись на знанні про лексичні правила утворення

множини понять у певній мові), то оцінити подібність термінів “тварина” та “собака” набагато важче. А для людини це очевидно, що “собака” – підвид “тварини”, тобто терміни, пов’язані між собою зв’язком типу IS-A.

Для вирішення проблеми проведення семантичних асоціацій ми використаємо WordNet – лексичну базу слів англійської мови [5]. У межах цієї бази даних для кожної словоформи визначено набір синонімів (synsets), частоту вживання словоформи та інші лексичні дані. Найбільша цінність цієї системи для нас – це можливість визначити смислові зв’язки між словами, що не пов’язані між собою спільним коренем, але належать до одного класу сутностей. Такі зв’язки характеризуються семантичною відстанню.

Підсистема оцінки новизни знань орієнтована на опрацювання інформації лише англійською мовою. Це пов’язано з тим, що для англійської мови існують потужні лексичні бази, одна з яких – WordNet – була використана у розробленій системі для визначення семантичних відстаней між реченнями.

WordNet – це лексична база даних англійської мови, розроблена у межах Прінстонського університету під керівництвом Джорджа Міллера. Цю базу можна безкоштовно завантажити і після інсталяції на файлову систему здійснювати пошук. Для пошуку засобами Java існує бібліотека JAWS (Java API for WordNet Searching). Вона надає простий та інтуїтивний інтерфейс для оперування лексичними даними.

Відстань між батьківським та дочірнім вузлами в WordNet обчислюватимемо як

$$Dist(c, p) = IC(c) - IC(p), \tag{3}$$

де

$$IC(c) = -\log P(c) = -\log\{\frac{\sum freq(w)}{N}\}; \tag{4}$$

$w \in c^*$, c^* – набір усіх гіпонімів (підвидів цього слова у смисловому значенні) c.

Семантична відстань між двома словоформами в WordNet:

$$Dist(c_1, c_2) = IC(c_1) + IC(c_2) - 2 * IC(LSuper(c_1, c_2)), \tag{5}$$

де $LSuper(c_1, c_2)$ – найближчий гіпернім двох словоформ c_1 та c_2.

Слово може мати різні словоформи і відповідно різні набори синонімів, гіпонімів та гіпернімів. У такому випадку ми обираємо найчастіше вживану словоформу.

Ґрунтуючись на цьому визначенні семантичної відстані між словами, сформулюємо спосіб обчислення семантичної відстані між реченнями. Для цього введемо поняття мінімальної відстані між конкретним словом w і усіма іншими словами у певному реченні S: WSSD (Word-Sentence Semantic Distance):

$$WSSD(w, S) = \min\{ Dist(w, w_i) \mid w_i \in S\}. \tag{6}$$

На основі мінімальної відстані певного слова та усіх інших слів у реченні дамо визначення семантичної відстані між двома реченнями (SSD – Sentence Semantic Distance):

$$SSD(S_1, S_2) = \frac{\sum_{w_i \in S_1} WSSD(w_i, S_2) + \sum_{w_j \in S_2} WSSD(w_j, S_1)}{|S_1| + |S_2|}, \tag{7}$$

де $|S_1|$ та $|S_2|$ – кількості слів у першому та другому реченнях відповідно.

Схему обчислення семантичної відстані між реченнями показано на рис. 2. Тут a_n позначає n-не слово першого речення, b_n – n-не слово другого речення.

Мета роботи – розробити гнучку підсистему оцінки новизни знань для застосування до великих обсягів текстової інформації. Ця підсистема повинна інтегруватись у зовнішню інформаційну систему певної предметної області. Оскільки ця предметна область не є наперед відомою, то важливо забезпечити механізм навчання, який уможливив би підсистемі вивчити особливості предметної області і відповідно пристосуватись. З погляду економічної ефективності залучення експертів для навчання системи є небажаним. Тому ще однією вимогою до

розроблюваної підсистеми виступає забезпечення простого механізму отримання відгуку від користувача.

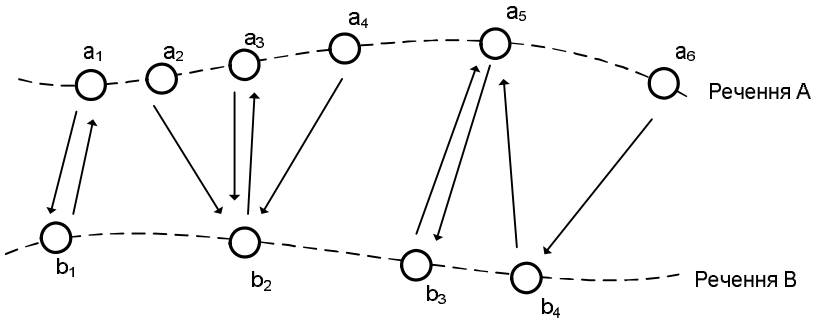

Рис. 2. Схема обчислення семантичної відстані між реченнями

Вимоги до системи можна ієрархічно зобразити у вигляді дерева цілей (рис. 3):

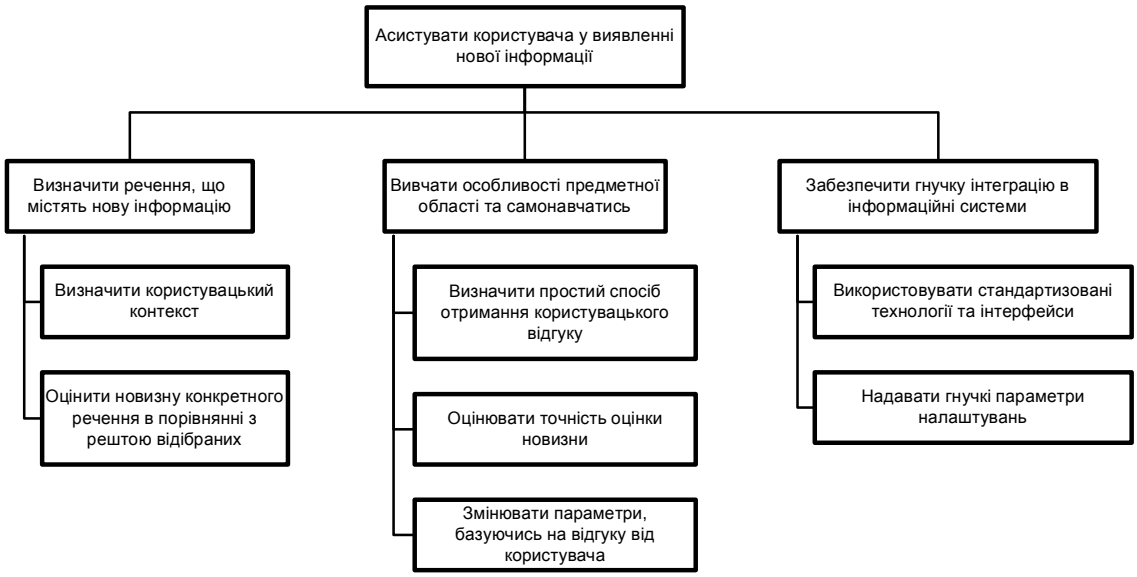

Рис. 3. Дерево цілей підсистеми

Розглянемо їх детальніше:

1. *Визначити користувацький контекст.* Як було зазначено вище, під час визначення поняття новизни інформації пошук нових даних відбувається у певному контексті – в іншому разі кожне відмінне від раніше відібраних речення можна вважати новим (навіть якщо воно не надає користувачу відповідей на його потенційні запитання). Контекстом може виступати запит користувача (якщо в інформаційній системі це підтримується). В іншому випадку контекст формується автоматично, ґрунтуючись на найвживаніших іменованих сутностях та термінах у наборі текстових фрагментів.

2. *Оцінити новизну конкретного речення порівняно з рештою відібраних.* Це центральна мета системи, що полягає у визначенні рівня новизни окремого речення та формуванні мінімального набору речень, з виключенням дублювань.

3. *Вивчати особливості предметної області та самонавчатись.* Оскільки розроблювана підсистема буде призначена для інтеграції в різні інформаційні системи, необхідно забезпечити механізм пристосування цієї підсистеми до вимог конкретної сфери використання.

4. *Визначити простий спосіб отримання користувацького відгуку.* Навчання підсистеми оцінки новизни знань з залученням експертів є економічно невигідним, а тому однією з цілей

системи є забезпечення простого механізму надання відгуку, який би уможливив проводити навчання системи прозоро у процесі її використання.

5. *Оцінювати точність оцінки новизни.* Для того, щоб проводити корегування своїх параметрів в процесі навчання, підсистема повинна мати механізм оцінки її ефективності. Це дає змогу визначати коефіцієнти зміни параметрів під час самонавчання.

6. *Використовувати стандартизовані технології та інтерфейси.* Для забезпечення гнучкої інтеграції підсистема повинна надавати чітко документований та зрозумілий API (Application Programming Interface), а той, своєю чергою, повинен використовувати стандартизовані канали та формати комунікації. Це дасть змогу звести до мінімуму зусилля під час інтеграції підсистеми.

На основі описаних вище цілей визначимо основні модулі підсистеми:

1. Модуль оцінки новизни знань.
2. Модуль навчання та корегування параметрів системи.
3. Модуль визначення контексту.
4. Модуль взаємодії з користувачем.

На рис. 4 показано DFD-діаграми розроблюваної підсистеми.

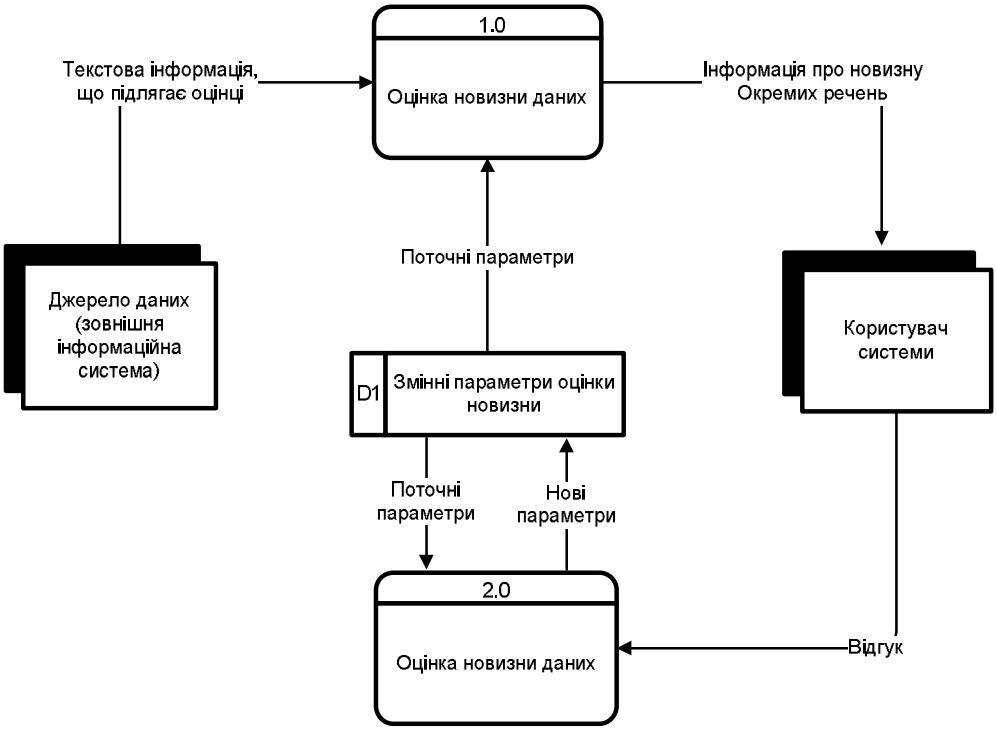

Рис. 4. DFD-діаграма першого рівня

Для визначення новизни інформації та для налаштування параметрів системи необхідно провести аналіз іменованих сутностей. Іменовані сутності (англ. "*named entity*") – це відповіді на питання про ДАТУ ("коли?"), МІСЦЕ ("де?"), ПЕРСОНУ ("хто?"), ОРГАНІЗАЦІЮ ("хто/що?") і КІЛЬКІСТЬ ("скільки?"). У межах розроблюваного алгоритму ми виділяємо п'ять типів іменованих сутностей – термінів, що позначають одне з п'яти понять: особу, місце, організацію, час або кількість.

Виявлення таких слів потрібне для двох цілей:

1. Розмітка дублюючих речень з метою полегшення їх аналізу для користувача. Залежно від конкретної інформаційної системи, в яку вбудовується підсистема оцінки новизни знань, інформація, що вважається дублюючою, може показуватись користувачу у вигляді короткого підсумку, щоб користувач міг окинути цей фрагмент оком та оцінити, чи зацікавлений він в його прочитанні. Оскільки іменовані сутності, зазвичай, несуть набагато більше смислового

навантаження, ніж решта слів у реченні, їх зручно використовувати в автореферуванні дублюючих фрагментів інформації.

2. Кількість іменованих сутностей, що належать до поточного контексту, в реченні виступають як один з параметрів визначення новизни цього речення. Статистичним шляхом показано, що речення з новими (такими, що раніше не зустрічались) іменованими сутностям в середньому у шість разів частіше [6, с. 247] несуть в собі нову інформацію порівняно з реченнями без іменованих сутностей.

Оскільки аналіз іменованих сутностей є лише допоміжним засобом у розроблюваній системі, ми не розглядаємо детально алгоритм виявлення цих термінів.

В реалізації системи використано розробку BBN, Identifier [7], яка здатна вхідний текст

Jim bought 300 shares of Acme Corp. in 2006.

перетворити у таку розмітку:

<enamex type="person">*Jim</* ***enamex*** *> bought <****numex type="quantity"****>300</****numex****> shares of <* ***enamex type="organization"****>Acme Corp.</* ***Enamex*** *> in <****timex type="date"****>2006</****timex****>*

Цей формат визначений організацією MUC (Message Understanding Conferences) у 1990 році [8].

Як було визначено в меті цієї роботи, значна увага під час розроблення підсистеми виявлення нових даних приділялась гнучкості інтеграції та здатності підсистеми до самонавчання у процесі роботи.

Різні предметні області потребують різних підходів до визначення нової інформації. Наприклад, у сфері медицини поріг новизни даних може бути набагато нижчим, ніж, скажімо, в сфері новин. Це пов'язано з тим, що для лікаря може бути важливою навіть незначна зміна у симптомах пацієнта, в той час, як в процесі аналізу новин незначна різниця в тоні висвітлення подій є не дуже важливою.

В такий спосіб були виділені основні параметри системи, що підлягають зміні на основі користувацького відгуку:

1. Вплив кількості іменованих сутностей в реченні на рівень його новизни.
2. Вагомість кожного типу іменованої сутності на оцінку новизни речення.
3. Поріг новизни (мінімальна семантична відстань між реченнями).
4. Вплив контексту на оцінку новизни даних.
5. Вплив хронологічної послідовності на оцінку новизни даних.

Як вже згадувалось, розроблювана підсистема має дуже простий спосіб отримання відгуку від користувача, – власне, єдина інформація, яка подається у підсистему щодо коректності її роботи, – це дані про те, які речення користувач насправді переглянув. Отримавши інформацію про те, що система помилково визначила речення як дублююче, незрозуміло, які параметри потрібно змінити для точнішої роботи програми. Для правильного вибору параметра для зміни ми використовуємо оцінку точності роботи програми до поточного моменту (за умови, що користувач уже прочитав певну кількість інформації). Отже, для кожного відгуку, що надійшов від користувача, ми ізолюємо зміну кожного параметра і робимо переоцінку точності. Та зміна, що дає найкращий приріст точності оцінки новизни, застосовується. Інші параметри залишаються незмінними. Алгоритм корегування параметрів системи показаний на рис. 5.

У межах цієї роботи була розроблена гнучка підсистема оцінки новизни знань з метою її подальшої інтеграції у зовнішні інформаційні системи. Ця підсистема розроблялась з використанням модульного підходу та має гнучкий API (Application Programming Interface). Ці особливості забезпечують зручну інтеграцію, проте зрозуміло, що підсистема сама по собі не має користувацького інтерфейсу. З метою оцінки ефективності обраного підходу було також розроблено прототип інформаційної системи, в яку вбудовано підсистему оцінки новизни знань.

Для забезпечення гнучкої інтеграції підсистеми необхідною вимогою є використання стандартизованих форматів даних за взаємодії з навколишньою інформаційною системою. Тому увесь обмін даних розробленої підсистеми з навколишнім середовищем відбувається у форматі XML.

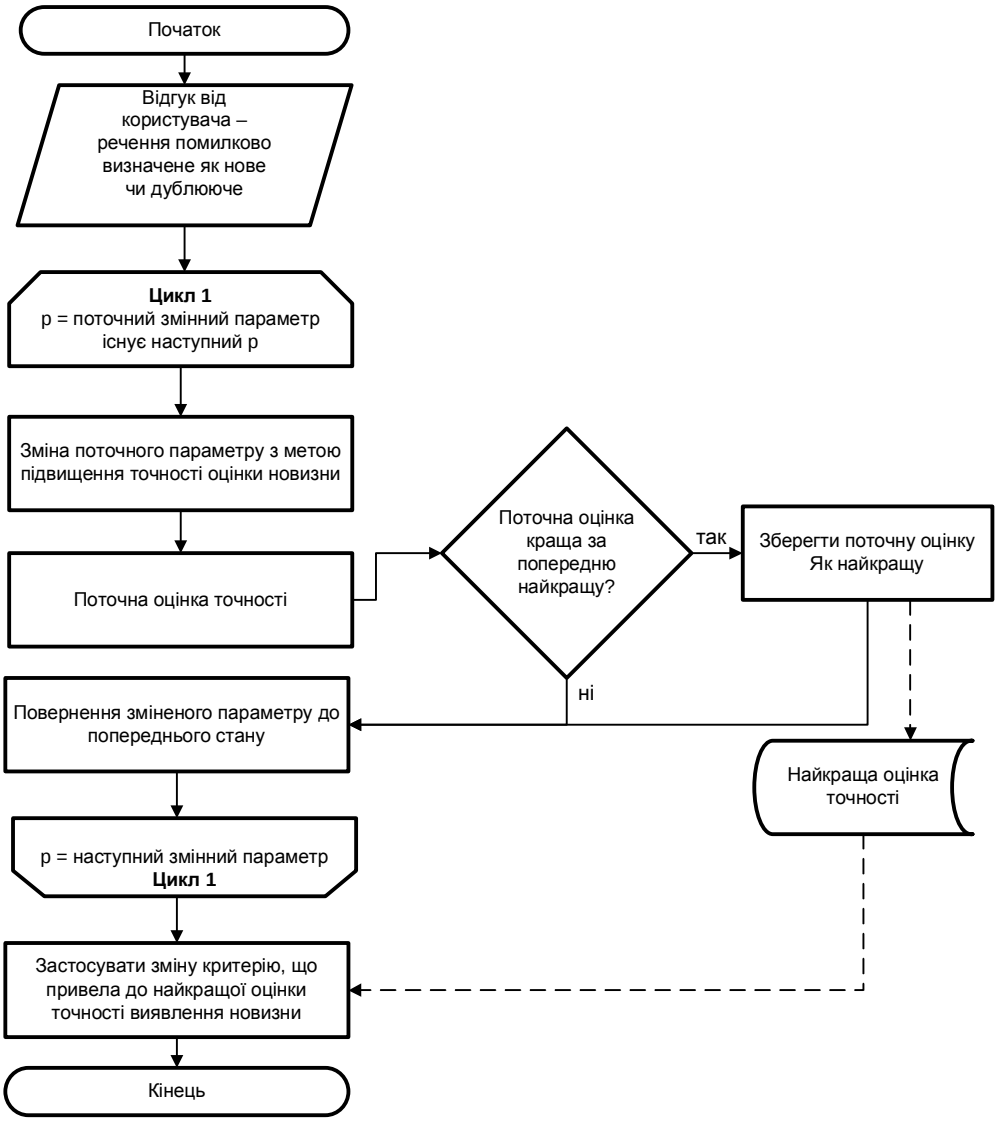

Рис. 5. Алгоритм корегування параметрів системи

Висновки

Оцінка ефективності системи виявлення новизни – це доволі суб'єктивний показник, що залежить від експерта, який використовує систему. Саме експерт (або користувач системи) визначає, які речення, на його думку, несуть в собі "нову" інформацію. Ця оцінка змінюватиметься від одного користувача до іншого. Тим не менше, з усередненої оцінки розробленого прототипу після початкового навчання можна зробити висновок, що точність виявлення новизни розробленої підсистеми становить близько 70 %. Це доволі високий показник серед результатів сучасних досліджень (для порівняння, алгоритм-переможець змагання в межах TREC Novelty Track у 2004 році досягнув рівня точності 60–65 %). Це вказує на перспективність обраного підходу. У межах цього дослідження була проаналізована можливість запровадження поняття контексту окремого речення під час визначення новизни. На нашу думку, застосування цього підходу в поєднанні з використовуваними алгоритмами допоможе збільшити точність виявлення нових знань і саме у цьому напрямі варто продовжувати дослідження.

1. Інтелектуальні системи, базовані на онтологіях // Д. Г. Досин, В. В. Литвин, Ю. В. Нікольський, В. В. Пасічник. – Львів: "Цивілізація", 2009. – 414 с. 2. Литвин В. В. Автоматизація процесу розвитку базової онтології на основі аналізу текстових ресурсів /

В. В. Литвин // Комп'ютерна та математична лінгвістика: Вісник НУ "Львівська політехніка". – 2010. – № 673,. – С. 319–325. 3. Литвин В. В. Математичне забезпечення розвитку базової онтології предметної області / В. В. Литвин, Д. Г. Досин, Н. В. Шкутяк // 2-га Міжнар. наук.-практ. конф. "Сучасні інформаційні та інноваційні технології на транспорті". – MINTT-2010. – Херсон. – 2010. – С. 345–348. 4. Varian H. R. Economics and Search / Hal Varian – Berkeley, California: ACM Press, 1999. – 421 с. 5. Local source annotation with WordNet [Електронний ресурс]. Режим доступу: www.cogsci.princeton.edu/~wn. 6. Xiaoyan L. Novelty Detection Based on Sentence Level Patterns / Li Xiaoyan, Croft W. – Bremen, 2005 – 751 с. 7. Daniel M. An Algorithm that Learns What's in a Name / Daniel M. Bikel, Richard L. Schwartz, Ralph M. – N.Y., Columbia University Press, 1999. – 231 с. 8. Message Understanding Conference [Електронний ресурс]. Режим доступу: http://en.wikipedia.org/wiki/Message_Understanding_Conference.

УДК 004.773.2

О.Ю. Тимовчак-Максимець, А.М. Пелещишин, К.О. Слобода
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

АНАЛІЗ КОМУНІКАТИВНОЇ ВЗАЄМОДІЇ НА ВЕБ-ФОРУМАХ: ІНФОРМАЦІЙНА ПОВЕДІНКА ТА УЧАСНИКИ

Рзглядаються особливості комунікативної взаємодії у віртуальних спільнотах на основі веб-форумів. Проаналізовано типи інформаційних продуктів та інформаційної поведінки, розроблено класифікацію інформаційної поведінки на основі впливу на комунікаційний процес. Запропоновано класифікацію веб-форумів на основі розподілу інформаційних ролей.

Ключові слова: віртуальні спільноти, інформаційний продукт, інформаційна поведінка, інформаційна роль.

This paper deals with peculiar features of communicative interaction in virtual communities on the basis of Web-forums. Types of information products and information behaviour have been analyzed, the classification of information behaviour on the basis of influence on communication process has been developed. The classification of Web-forums on the basis of information roles distribution has been suggested.

Key ords: irtual communities, information product, information behaviour, information roles.

Вступ

Віртуальні соціальні спільноти формуються довкола певного інтернет-середовища комунікації, наприклад, веб-форуму, і забезпечують задоволення комунікаційних та інформаційних потреб їх учасників. Кожен із учасників задовольняє власну потребу взаємодією іншими учасниками спільноти. Взаємодія учасників реалізується у тривалій в часі комунікації, тобто у формі полілога учасників на певну задану тематику. Життєдіяльність віртуальної спільноти визначається бажанням учасників долучитися до обговорення. Отже, обов'язковою умовою існування віртуальної спільноти є наявність у достатньої кількості її учасників бажання вступати у комунікативну взаємодію.