Список використаних джерел:

1. Якимова О. Ю. Методы оценки эффективности корпоративных информационных систем управления / О. Ю. Якимова // Современные наукоемкие технологии. – 2006. – Т. 1. – № 3. – С. 95–98.

2. Каленская Н. В. Методология формирования инфраструктурного обеспечения инновационного развития промышленных предприятий : автореф. дисс. на соискание науч. степени д-ра экон. наук : спец. 08.00.05 / Н. В. Каленская. – Казанский гос. фин.-экон. ин-т. – Казань, 2010. – 52 с.

3. Юденко М. Н. Теория и методология формирования институциональной инфраструктуры предпринимательской деятельности в строительстве : автореф. дисс. на соискание науч. степени д-ра экон. наук : спец. 08.00.05 / М. Н. Юденко ; С.-Петерб. гос. инж.-экон. ун-т.– СПб., 2010. – 38 с.

4. Денисова А. Л. Теория и практика экспертной оценки товаров и услуг : учеб. пособие / А. Л. Денисова, Е. В. Зайцев. – Тамбов : Изд-во Тамб. гос. техн. ун-та, 2002. – 72 с.

✧ ✧ ✧ ✧ ✧ ✧ ✧

УДК 004.912

Б. І. Мороз, доктор технічних наук, декан факультету інформаційних систем та технологій Університету митної справи та фінансів
Д. Є. Костенко, старший викладач кафедри інформаційних систем та технологій Університету митної справи та фінансів
В. В. Костенко, старший викладач кафедри інформаційних систем та технологій Університету митної справи та фінансів
І. В. Лавренюк, старший викладач кафедри інформаційних систем та технологій Університету митної справи та фінансів

ЗАСТОСУВАННЯ НОВИХ ПІДХОДІВ ДО ІНТЕЛЕКТУАЛЬНОГО ПОШУКУ ТА АНАЛІЗУ ЕЛЕКТРОННИХ ДОКУМЕНТІВ В ІНФОРМАЦІЙНИХ СИСТЕМАХ

Розглянуто проблеми та можливості застосування регулярних виразів. Досліджено деякі підходи застосування регулярних виразів (з використанням літералів, метасимволів і квантифікаторів) на базі онтологій та перспективи використання вказаного підходу для застосування в експертних системах.

Пропонується розробка методу й алгоритму семантичного пошуку з урахуванням особливостей обробки масивів неструктурованих текстів в електронних документах. Розроблено систему семантичного пошуку та розглянуто питання про можливе використання онтологій.

Ключові слова: *інформація; літерали; метасимволи; неструктуровані документи, онтології; регулярні вирази.*

Using the ontologies is one of research issues of the day. Every year the volumes of information increase and traditional information context query facilities in documents do not provide the desired result. There are examine the problems and management possibilities by knowledge on the basis of the unstructured texts arrays in documents. Examines some approaches for using the regular expressions(with using metacharacters and quantifiers) based on ontologies and perspectives of this approach for using in expert systems.

Regular expressions is a powerful tool. They can be useful everyone who works with text. With regular expressions, we can automate many search and replace task.

The development of method and algorithm of semantic search are assumed by the authors of this article taking into account the features of treatment of the unstructured texts arrays in official documents. Development of the system of semantic search is planned. A question about the possible use of ontologies is affected.

Key words: *information; quantifiers; metacharacters; unstructured documents; ontologie; regular expressions.*

Постановка проблеми. З кожним роком збільшується обсяг доступних користувачу масивів текстової інформації, що сприяє більшій актуалізації завдання пошуку необхідних документів у таких масивах. Для виконання подібних завдань дуже часто використовуються різноманітні елементарні програмні засоби – тематичні класифікатори, рубрикатори тощо, які дозволяють шукати (автоматично або вручну) документи в невеличкій підмножині документної бази, що задовольняє інтереси користувача [1]. Але не завжди результат пошуку може відповідати поставленим завданням.

З'ясованим фактом є те, що сучасні моделі інформаційного пошуку не використовують знань, описаних у тезаурусах та онтологіях, а базуються на моделях тексту як набору слів, пропонуючи методи врахування частоти появи слів у реченні, тексті, наборі документів. Нерідко враховується кількість спільної появи слів.

Це означає, що традиційні механізми роботи з електронними документами вже застаріли й не задовольняють потреби сучасного користувача [2].

Тому необхідні нові підходи до інтелектуального пошуку та аналізу електронних документів, їх інтеграції в інформаційні системи [2].

За допомогою регулярних виразів розв'язується безліч задач обробки текстової інформації. Наприклад [3]:

– пошук фрагментів тексту, що відповідають шаблону;

– перевірка тексту на відповідність шаблону;

– заміна тексту за шаблоном.

Найактуальніші на даний момент такі проблеми [2]:

1) експонентне зростання кількості документів ускладнює пошук необхідних документів та їхню організацію у вигляді структурованих за змістом сховищ [4]. Зі збільшенням простору пошуку пропорційно зростає і кількість документів у відгуку пошукової системи;

2) відсутність стандартизованих механізмів семантичного індексування також згубно впливає на ефективність роботи з електронними документами. Більшість сучасних технологій підготовки і роботи з документами (текстові редактори, HTML) орієнтовані на організацію зручної роботи з інформацією для людини;

3) неструктурований характер інформації більшості електронних документів не дозволяє застосувати традиційні механізми її обробки й аналізу. Неструктурована інформація становить значну частину сучасних електронних документів, основні знання розташовуються саме в таких документах. Для розв'язання таких проблем необхідно розширити поняття традиційного документа: з документом потрібно пов'язати метадані, що дозволяють інтерпретувати й обробляти інформацію, яка зберігається в цьому документі, тобто включити в документ інформацію, яка описує структуру і семантику його змісту [2].

Мета роботи полягає в аналізі, дослідженні, покращанні деяких можливостей ефективного пошуку інформації у сукупності неструктурованих документів та у формулюванні вимог до моделі предметно-орієнтованої системи для ефективного семантичного пошуку. Проблема полягає в тому, щоб зробити пошук динамічним і зручним для користувача. Для будь-якого типу запиту, що виникає в практичній діяльності, потрібно знайти адекватні знання в інформаційному просторі. При цьому мова для формулювання пошукової вимоги не має бути занадто складною.

Аналіз останніх досліджень і публікацій. Основу онтології становлять представлені в ній терміни. Втім, не тільки терміни. В онтологічну сукупність входять також відомості про предметні області, про області визначень тощо.

Словосполучення "регулярні вирази" чув (або бачив) кожен, чия діяльність так чи інакше пов'язана з використанням комп'ютерів. Багато користувачів застосовують найпростіші варіанти регулярних виразів мало не щодня, навіть не підозрюючи це. Зазвичай їм приділяється не надто багато уваги і, як правило, в контексті конкретної мови програмування (Perl, Python і т. д.).

Питання "А що таке регулярний вираз взагалі?" доволі складне. Можна сказати, що це спеціалізована мова опису символьного шаблона (послідовності символів) пошуку в рядках тексту. Тут важливо те, що під час пошуку збігів виконується саме посимвольне порівняння.

Джеффрі Фрідл радить розвивати звичку буквально інтерпретувати регулярні вирази [5]. Наприклад, дивлячись на шаблон "cat", що означає "рядок повинен починатися зі слова cat", слід міркувати так: збіг буде знайдено, якщо ми перебуваємо на початку рядка і виявляємо символ "c", безпосередньо за яким розташовується символ "a", відразу після якого стоїть символ "t". Це дозволяє максимально точно оцінити зміст і сутність регулярного виразу [5].

Більшість користувачів знають, що для пошуку досить задати слово-зразок. Наприклад, у web-браузері в полі "Пошук" після введення "Linux" буде отримано

довгий список посилань на сторінки, в тексті яких знайдено збіг із заданим шаблоном “Linux”.

Не всі, але деякі користувачі вміють застосовувати метасимволи (*.?) в шаблонах пошуку. Ще менша кількість людей знає про можливість застосування модифікаторів та інших витончених засобів для конструювання регулярних виразів, тобто в багатьох випадках потужність механізму регулярних виразів використовується ледве на третину.

Важливість поняття регулярних виразів, яке пов’язане з онтологіями, обумовлена також тим, що знання, яке не описано і не тиражовано, зрештою стає застарілим і непотрібним. Навпаки, знання, яке поширюється, є генератором нових знань [6].

Д. Мерзляков вважає, що використання регулярних виразів дозволяє гнучко враховувати різні структурні перестановки всередині тексту, різні варіанти написання одних і тих самих понять, а так само відношення синонімії. Однак у класичних регулярних виразах немає знань про онтології, до якої належить текст, що перевіряється. Як наслідок, ручне складання класичного регулярного виразу для перевірки відкритих тестів практично неможливе, оскільки воно, зрештою, зводиться до повного перебирання всіх можливих варіантів відповіді, що, навіть з урахуванням використання регулярних конструкцій для скорочення виразу і перевірки лише найімовірніших варіантів відповіді, під час роботи з текстом на природній мові потребує величезних затрат [3].

У разі зміни знань про предметну область може знадобитися перегляд усіх регулярних виразів, порушених цими змінами [3]. Один з можливих розв’язків зазначеної проблеми – привнесення семантичної складової в регулярні вирази, що дозволить автоматично та гнучко перебудовувати їх зі зміною знань про предметну область [3].

Мета статті – аналіз, дослідження та покращання можливостей ефективного пошуку інформації у сукупності неструктурованих документів. Розглядається питання використання онтологій та регулярних виразів.

Актуальність теми дослідження визначається тим, що багато завдань пошуку і заміни можна автоматизувати, створивши регулярні вирази, які складаються з текстових літералів і мета символів узагальнення.

Виклад основного матеріалу. У величезних масивах інформації мають працювати пошукові системи, які забезпечували б користувачу результат – швидкий і точний. Вважається, що системи, які базуються на онтологічному підході, досконаліші та відповідають потребам користувачів.

Процес аналізу текстової інформації передбачає такі етапи:

1) графематичний аналіз (токенізація, визначення іменованих сутностей);

2) морфологічний аналіз (нормалізація, стемінг);

3) синтаксичний аналіз (побудова дерева синтаксичного розбору);

4) семантичний аналіз (побудова семантичного графа).

Онтологічну систему можна побудувати за допомогою діаграми варіантів використання (рис. 1).

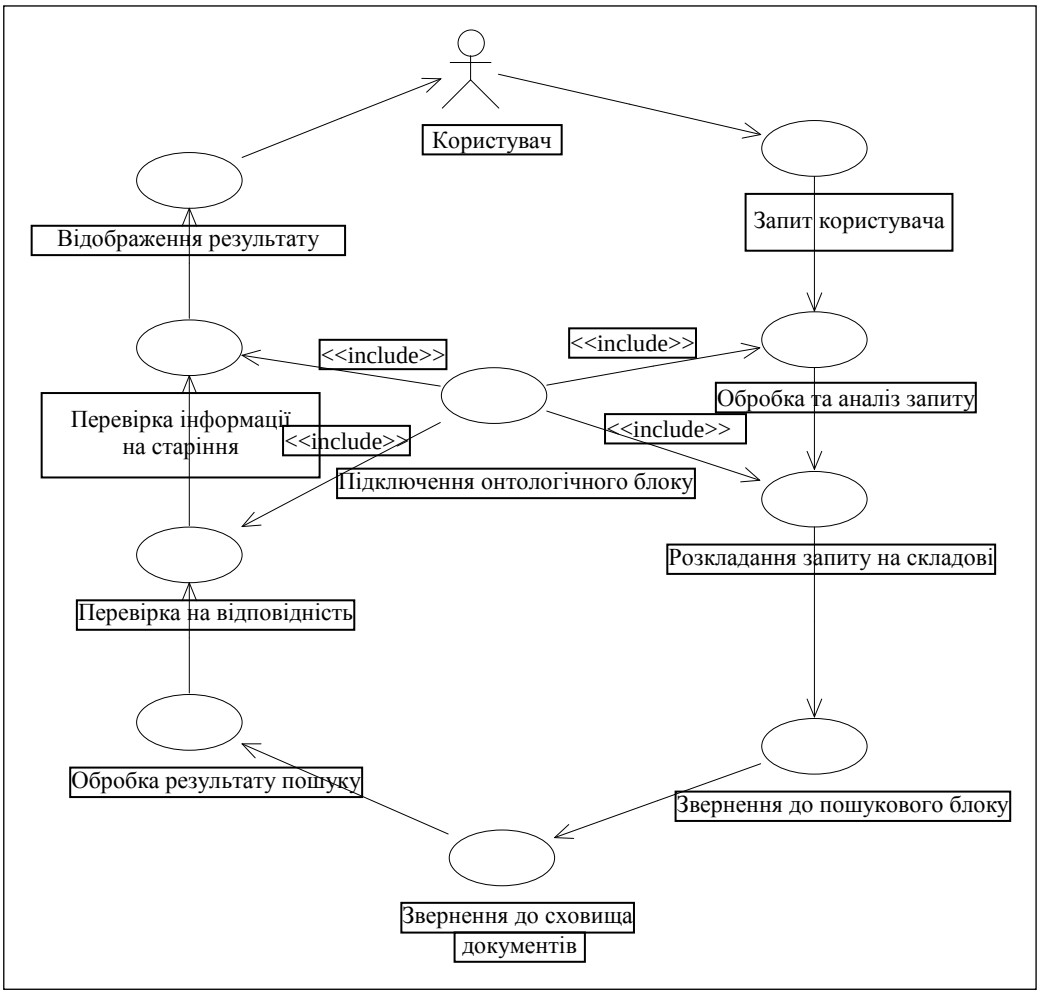

Рис. 1. Діаграма варіантів використання системи онтологічного пошуку

Онтології необхідно застосувати у ролі посередника між користувачем і процесом пошуку, між процесом пошуку і пошуковою системою. Для побудови онтології потрібно формальне декларативне подання чітко організованих конструкцій, які містять словник термінів тематичної області, опис визначень цих термінів, наявні взаємозв'язки між ними і взагалі теоретично можливі й неможливі взаємозв'язки.

Взаємодія з онтологією відбувається на таких етапах:

1) обробка та аналіз запиту;
2) розкладання запиту на складові;
3) перевірка інформації на старіння та відповідність;
4) перевірка на старіння інформації, яка була обрана зі сховища даних.

Робота з регулярними виразами (далі – РВ) має проводитися в ті моменти, коли запит аналізується та розкладається на складові.

Регулярний вираз є послідовністю, що описує множину рядків. Ці послідовності використовуються для того, щоб дати точне описання множини, не перелічуючи всі її елементи. Наприклад, множина, що складається зі слів “грати” та “ґрати”, може бути описана регулярним виразом “[гґ]рати”. В більшості формалізмів, якщо існує регулярний вираз, що описує задану множину, тоді існує нескінченна кількість варіантів, які описують цю множину.

У регулярних виразах використовуються звичайні (літерали) та спеціальні символи (метасимволи).

За допомогою РВ виконують такі основні завдання:

1) перевірка на відповідність (перевірка тексту, що вводиться на відповідність деякому шаблону);

2) пошук та аналіз (пошук у тексті фрагментів, відповідних заданим шаблоном з метою подальшого статистичного аналізу: аналізу кількості та частоти входжень, аналізу оточення входження тощо);

3) пошук і заміна (пошук у тексті фрагментів, відповідних заданим шаблоном з метою подальшої їх заміни на інші фрагменти; саме в цьому аспекті сервіси підтримки регулярних виразів у різних мовах програмування можуть дещо відрізнятися один від одного набором наданих можливостей);

4) синтаксичний аналіз виразів (рішення полягає в складанні регулярного виразу, що описує граматику вихідного виразу; з його допомогою спочатку здійснюємо перевірку на відповідність складеним шаблонам і потім розбиваємо вираз на лексеми).

Синтаксис регулярних виразів залежить від інтерпретатора, що використовується для їхньої обробки. Однак, із незначними відхиленнями, майже всі поширені інтерпретатори регулярних виразів мають спільні правила.

Реалізація компонентів експертної системи передбачає виконання кількох завдань:

1) використання РВ;

2) групування метасимволів за функціональним критерієм;

3) ефективного використання РВ у рамках експертної системи;

4) розробка редактора для побудови регулярних виразів та збереження їх у спецформатах файлів для подальшого використання.

Наведемо деякі фрагменти моделі експертної системи, що розробляється. Створення, редагування і збереження РВ можна подати за допомогою діаграми варіантів використання (рис. 2).

Діаграма є безпосереднім концептуальним відображенням роботи компонента експертної системи, на який покладено функції роботи з РВ.

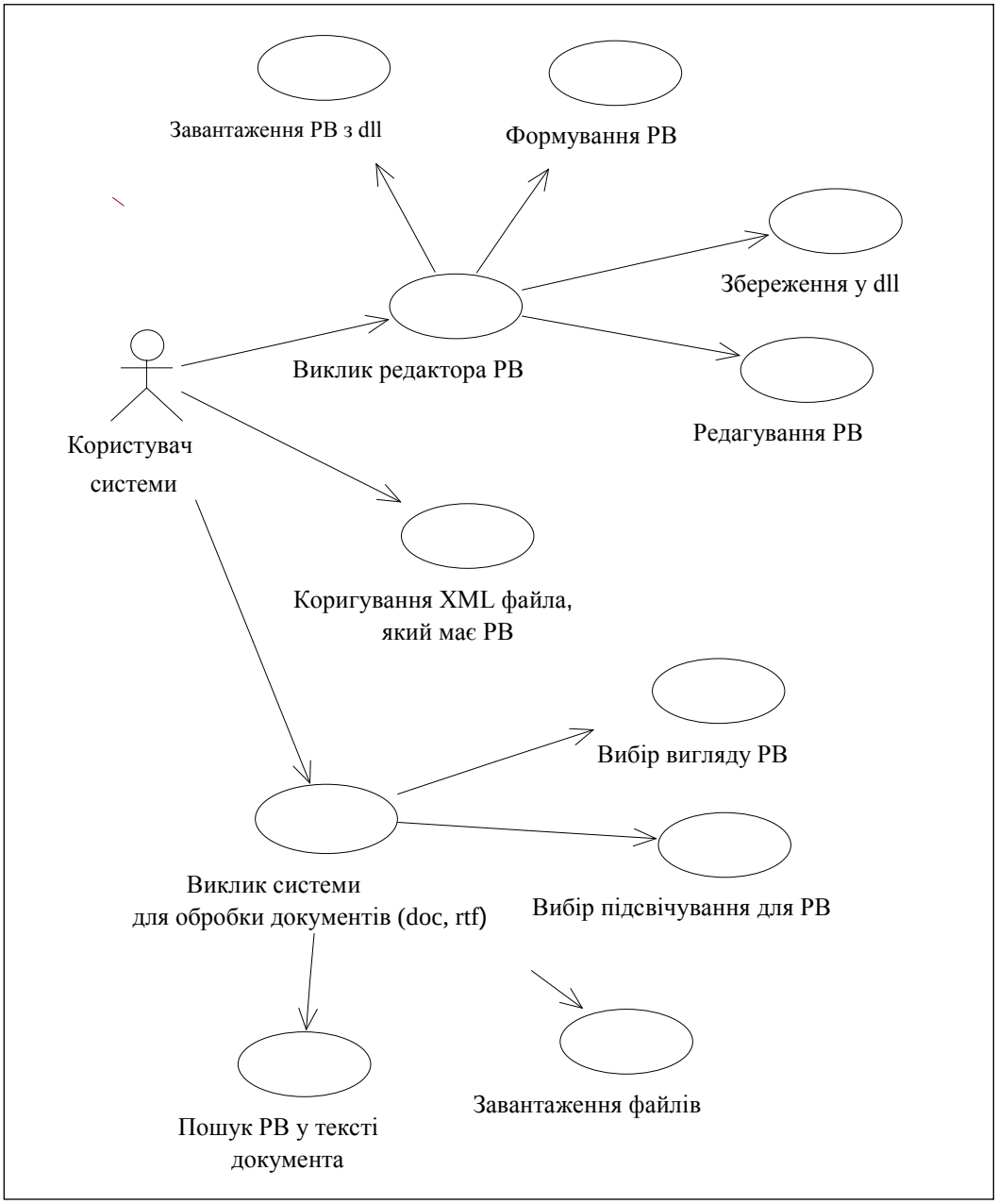

Рис. 2. Діаграма варіантів використання регулярних виразів

На рис. 3 та 4 більш детально показано процес власне обробки та переформатування (за потреби) РВ та їх використання безпосередньо у текстових документах.

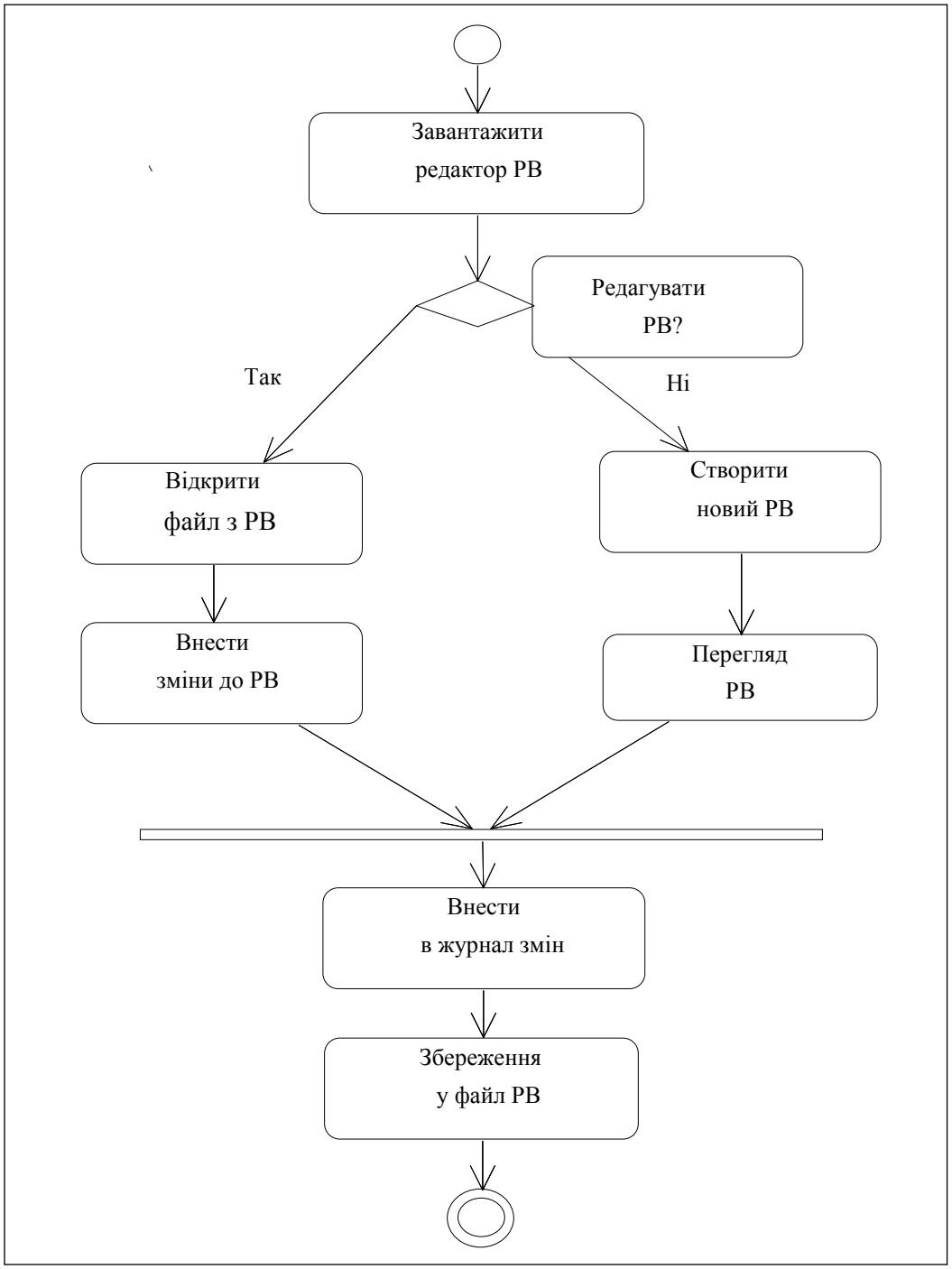

Рис. 3. Діаграма діяльності користувача системи у створенні РВ

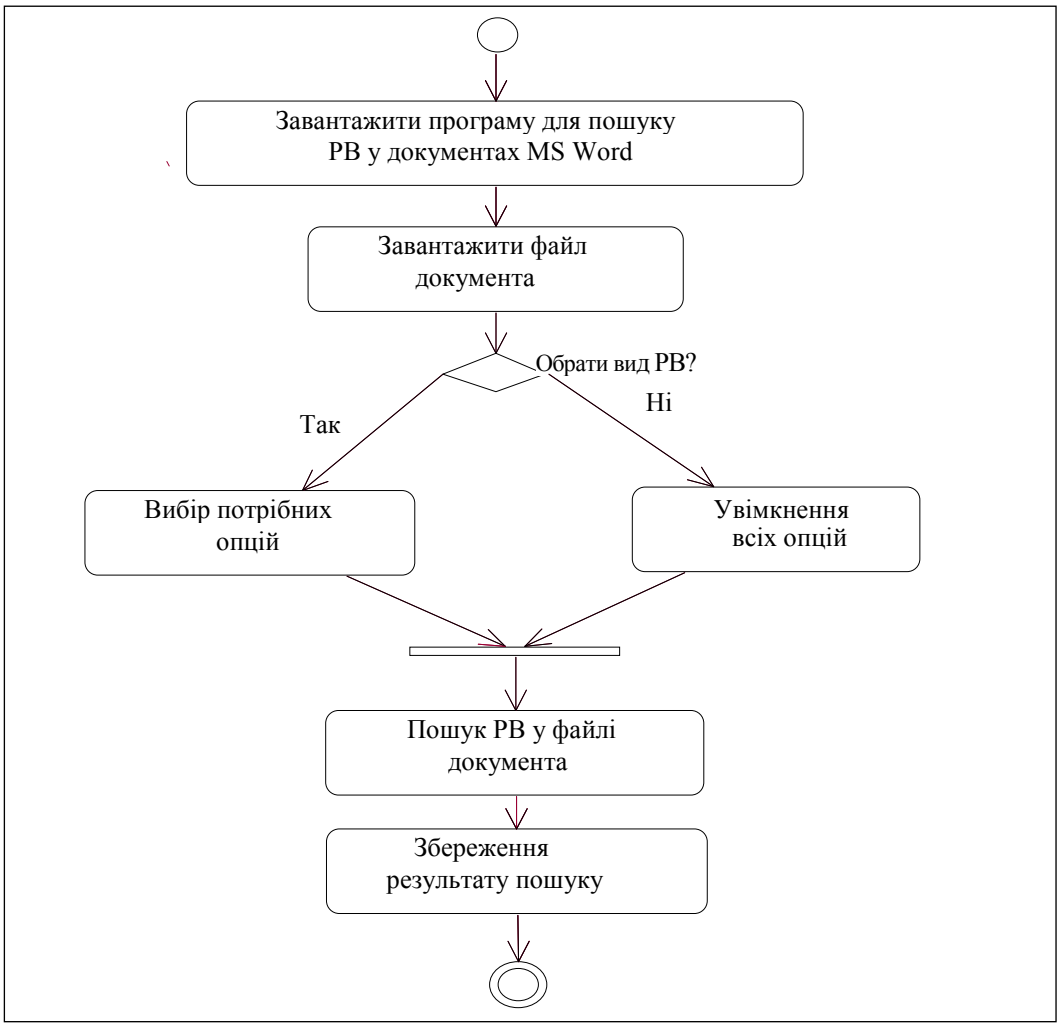

Рис. 4. Діаграма діяльності користувача системи у використанні РВ
під час процесу пошуку в документах

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямку. Модель системи пошуку повинна мати такі властивості:

– цілісність (система розглядатиметься як єдине ціле, що складається із взаємодіючих модулів, можливо, неоднорідних, але одночасно та "точково" сумісних між собою; при цьому не виключена можливість того, що деякі блоки існують окремо і підключаються у необхідний момент, але в даному випадку вважаємо, що про порушення цілісності системи не йдеться);

– зв'язність (наявність істотних стійких зв'язків між елементами та їх властивостями, причому із системних позицій значення мають не будь-які, а лише істотні зв'язки, які визначають інтегративні властивості системи);

– організованість (наявність певної структурної та функціональної організації, сюди ж можна додати один із процесів роботи системи – обробка регулярних виразів);

– інтегративність (наявність якостей, властивих системі в цілому, але не властивих жодному з її елементів окремо, тобто властивості системи хоча й залежать від властивостей елементів, але не визначаються ними повністю);

– мобільність (у даному випадку було складно дібрати термін, поки що під цим розумітимемо можливість швидкої перебудови моделі та системи під нові обставини; обов'язковою умовою необхідно додати процес "самонавчання" системи).

Використання регулярних виразів у експертній системі дозволяє гнучко враховувати різні структурні перестановки всередині неструктурованого документа. Звісно, що це має відбуватися з урахуванням онтологічної складової.

Завдяки цьому можна використовувати і складати прості й складні регулярні вирази в рамках конкретної предметної області.

Перспектива використання регулярних виразів на базі онтологій полягає не тільки в пошуку окремих слів або словосполучень, але й у можливості здійснення пошуку точних фраз, фраз зі списку, в пошуку слова в різних варіантах написання або зі спеціальними нестандартними символами, у пошуку слова зі змінними символами. Такий підхід змінить швидкість і точність пошуку потрібних документів.

Список використаних джерел:

1. Шатовская Т. Интегрированный подход текстовой кластеризации для неструктурированных документов / Т. Шатовская, И. Каменева // INTERNET – EDUCATION – SCIENCE : материалы 6-й Международной конференции (Винница 7–11 октября 2008 г.). – Винница, 2008. – С. 504–506.

2. Ланин В. Онтологии как основа функционирования систем обработки электронных документов / Ланин В. // "Знания–Онтологии–Теории (ЗОНТ-09)": материалы конференции с международным участием. – Новосибирск, 2009. – Т. 2. – С. 173–177.

3. Мерзляков Д. А. Генерация регулярных выражений для автоматизации проверки тестов открытого характера [Электронный ресурс] / Мерзляков Д. А. "Студенческий научный форум" (15.02.2013–31.03.2013) : материалы 5-й Международной студенческой электронной научной конференции. – М., 2013. – Режим доступа : http://www.scienceforum.ru/2013/147/2470

4. Ефремов В. Search 2.0: огонь по "хвостам" / В. Ефремов // Открытые системы. СУБД. – 2007. – № 8. – С. 72–74.

5. Фридл Дж. Регулярные выражения : пер. с англ. / Фридл Дж. –3-е изд. – СПб. : Символ-Плюс, 2008. – 597 с.

6. Гаврилова Т. А. Онтологический подход к управлению знаниями при разработке корпоративных информационных систем / Т. А. Гаврилова // Новости искусственного интеллекта. – 2003. – № 2. – С. 24–30.