УДК 004.652.4+004.827

В.Я.Крайовський, В.В.Литвин, Н.Б.Шаховська

ОСНОВНІ ПІДХОДИ ДО РОЗРОБЛЕННЯ ПРОГРАМНОГО КОМПЛЕКСУ АВТОМАТИЧНОГО РЕФЕРУВАННЯ ТЕКСТОВИХ ДОКУМЕНТІВ

The architecture and refer algorithm are described. There are described mining of refer and his main characteristics and functions.

Вступ

Значні обсяги текстової інформації, яку необхідно опрацювати, приводить до побудови програмних засобів, які б дозволяли здійснювати автоматизоване реферування цієї інформації. Анотування й реферування є невід'ємною частиною сучасного видавничого процесу. Будь-яке видання, чи це монографія, підручник, аналітичний огляд тощо, завжди випереджуються вторинним документом (рефератом або анотацією). Реферування використовується не тільки для економії часу при ознайомленні з великою кількістю джерел, але й з метою пришвидшення повнотекстового пошуку за множиною документів, оскільки обсяг реферату у декілька разів менший, ніж обсяг вхідного документу чи їх множини.

Стаття присвячена розробленню системи реферування текстів з декількох джерел.

Огляд літературних джерел та постановка задачі

Реферування – це процес видобування найважливішої інформації з одного або декількох джерел для складання їхньої скороченої версії для потреб певних користувачів або задач [1]. **Реферат** – це семантично адекватний виклад основного змісту первинного документа, що відрізняється ощадливим знаковим оформленням, сталістю лінгвістичних і структурних характеристик і призначений для виконання різноманітних інформаційно-комунікативних функцій у системі наукової комунікації. Рефератом називається текст, що передає основну інформацію джерела у згорнутому вигляді й складений у результаті його змістовного перетворення.

Існує два підходи до рішення задачі реферування. У першому підході видобувається невелика кількість фрагментів, у яких якнайповніше поданий вміст документа. Це можуть бути пропозиції, що містять терми запиту; фрагменти пропозицій з оточенням термів декількома словами та ін. Під час другого підходу реферування є синтезованим документом у вигляді короткого вмісту. Реферат, сформований відповідно до першого підходу, якісно поступається отриманому під час другого підходу. Однією з проблем, що виникає під час синтезу, є відсутність засобів семантичного аналізу і синтезу тексту природною мовою, тому сервіси реферування орієнтовані або

на вузьку предметну область, або вимагають участі людини.

Є такі функції рефератів: розповідні, інформативні, критичні. Розповідні реферати формуються за класичним принципом видобування інформації: вони надають достатній об'єм інформації, щоб створити у користувача уявлення про відповідні джерела, з тим щоб їх можна було вибрати для уважнішого перегляду.

Інформативні реферати замінюють собою текст, в основному вони містять основну або нову фактичну інформацію в скороченій формі.

Критичні реферати (або огляди) повідомляють не тільки про суть інформації, але й пропонують певну думку про неї. Критичні реферати володіють додатковою цінністю в порівнянні з оригіналом, оскільки пропонують логічне виведення, якого немає в самому тексті.

Результати порівняння комерційних систем реферування подано у табл.1 [3].

Табл. 1. Порівняльна характеристика систем автоматичного реферування

Системи автоматичного реферування\Властивості	Розміщення у тексті	Ключові фрази	Довжина речення	Машинне навчання	Дискурсивна модель	Один документ	Багато документів	Фрагментарний текст	Зв'язний текст	Загальний реферат	Спеціальний реферат	Налаштування стиснення	Багатомовний	Можливість налаштування реферування
Auto Sumarizer (MS Word)						+		+				+		
CONTEXT						+		+		+		+		
Data Hammer						+		+				+	+	+
DimSum				+		+						+	+	+
Extractor				+		+		+		+			+	+
GE Summarizer	+		+		+	+		+	+	+	+	+		
Intelligent Miner	+					+		+		+		+		
IntellScope	+					+		+		+	+	+	+	
InText						+		+			+	+		
InXihtSummarizerPlus	+	+	+	+		+		+		+		+	+	+
ProSum	+		+			+				+				
Search'2005 Developer Kit	+	+	+			+		+		+	+	+		
SMART						+			+		+			
SUMMARIST						+		+		+	+		+	
TexNet32						+		+		+	+	+		
TextAnalyst2.0				+		+	+				+			

Як бачимо, лише система TextAnalyst2.0 формує реферат з декількох джерел. Тому **метою** статті є розроблення архітектури системи реферування з декількох джерел. Джерелами реферату виступатимуть текстові файли довільного формату (.txt, .doc, .rtf), веб-сторінки, xml-файли, які зберігаються у локальній чи глобальній мережі.

Для підвищення ефективності автоматизованого реферування текстових документів нами пропонується використовувати простори даних, як засіб інтеграції інформаційних ресурсів та онтології предметних областей (ПО) для визначення важливості понять та термінів ПО [4]. Очевидно, що у склад реферату повинна входити інформація про ті поняття для яких коефіцієнт важливості є вищим. Виходячи з цієї точки зору, нами дано такі поняття простору даних та адаптивної онтології.

Основний текст

Оскільки за мету ми поставили складання реферату з декількох джерел, то однією із задач, яку необхідно вирішити для розроблення системи автоматичного реферування, є забезпечення доступу та опрацювання різноформатних даних.

Одним із засобів опрацювання різноформатних джерел інформації є простір даних. **Простір даних реферування** – це множина даних, поданих у різних моделях (статичних Web-сторінок **Wb**, неструктурованих текстових даних **Nd**, графічних та мультимедійних даних **Gr**), локальних сховищ джерел **ODW**, а також засобів інтеграції **Int**, пошуку **Se** та опрацювання інформації **Wo**, об'єднаних середовищем керування моделями **EM**.

$$\mathbf{DS=<Wb, Nd, Gr, ODW, Int, Se, Wo, EM>.}$$

Необхідною умовою для здійснення доступу та опрацювання джерел є наявність повної інформації про них: місця розташування, тип та формат даних, методи пошуку та опрацювання, притаманні цьому формату даних, частота оновлення, власник даних тощо). Інформація такого типу зберігається у каталозі даних.

Каталог (**CG**) – це реєстр ресурсів даних, що містить базову інформацію про кожного з них: джерело, ім'я, місцезнаходження в джерелі, розмір, дату створення і власника та ін., а також взаємозв'язок між ними та опис онтології ПО простору даних (**Dic**). Каталог є інфраструктурою для більшості інших сервісів простору даних, але він також може підтримувати базовий, призначений для користувача, інтерфейс переглядання простору даних.

$$\text{Metadata}(\mathbf{Wb}, \mathbf{Nd}, \mathbf{Gr, Dic}) \Rightarrow \mathbf{Cg}.$$

Наявність онтології ПО є необхідною умовою якісного реферування, оскільки дозволяє виявляти зв'язки між семантичними одиницями та їх вагами [5].

Виходячи з цього, онтологію визначимо як п'ятірку:

$$O=<C, R, F, W, L>,$$

де C – скінченна множина концептів (понять, термінів) ПО, яку задає

онтологія O; R – скінченна множина відношення між концептами (поняттями, термінами) ПО; F – скінченна множина функцій інтерпретації (аксіоматизація), заданих на концептах або відношеннях онтології O, W – важливість понять C, L – важливість відношень R. Цю онтологію будемо називати адаптивною, оскільки вона адаптується до ПО за рахунок модифікації понять та коефіцієнтів важливості цих понять та зв'язків між ними.

Визначення інформаційної ваги елементів онтології виконується за наступними правилами [6]:

1. Перерахунок ваги вершин відбувається по вертикалі знизу вгору.

2. Розрахунок вагових коефіцієнтів поняття є рекурсивною процедурою. Повна вага W_j^i класу онтології дорівнює сумі власної ваги Wo_j^i, ваги підкласів Ws_j^i та ваги суміжних класів Wn_j^i (класів, зв'язаних з даним класом не is-a зв'язком):

$$W_j^i = Wo_j^i + Ws_j^i + Wn_j^i, \qquad (1)$$

де: $Ws_j^i = \sum_k Wc_k^{i+1} \cdot L_{j,k}$ – вага k підкласів j-го класу i-го рівня;

$Wc_k^{i+1} = Wo_k^{i+1} + Ws_k^{i+1}$ – вага класу C_k^{i+1};

$L_{j,k}$ – вага зв'язку між класами C_j^i та C_k^{i+1}.

Схема перерахунку окремих компонент повної ваги класу показана на рис. 1.

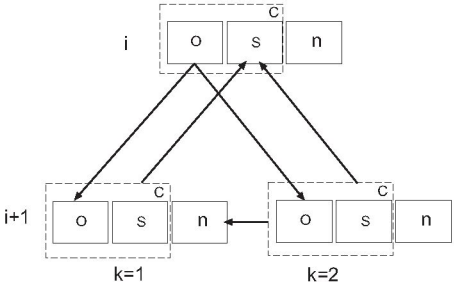


Рис.1. Схема перерахунку окремих компонент повної ваги класу

3. У момент внесення на i+1-й рівень нового підкласу йому присвоюється власна вага Wo_j^{i+1}, рівна половині власної ваги класу, вищого i-го рівня.

4. Під час встановлення зв'язку між поняттями k_1 та k_2 між відповідними вершинами графа з'являється ребро, а до ваги суміжних класів Wn_1 додається вага Wc_2 і навпаки – до Wn_2 додається вага нового суміжного до нього класу Wc_1.

5. Вага екземпляра дорівнює повній вазі його класу.

Даний набір правил покладено в основу автоматичного перерахунку ваги класів онтології та екземплярів бази знань в процесі її експлуатації.

Для пошуку елементів онтології у джерелах даних розроблено інтелектуальний агент (ІА), здатний в процесі самонавчання адаптуватися до конкретних інформаційних потреб користувача та виявляти, зберігати і використовувати релевантні до відповідних задач знання. Основою ІА є його база знань (каталог даних **Cg**), а ядром бази знань є онтологія **Dic** (рис. 2). Загальні властивості ІА визначаються його онтологією, яка задає спосіб подання знань, механізми міркувань та прийняття рішень [7].

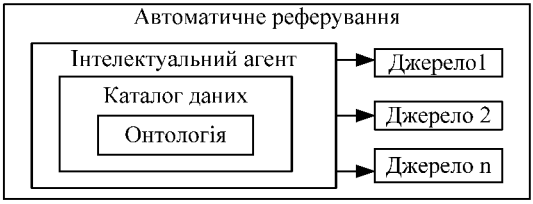


Рис. 2. Місце інтелектуального агента у загальній структурі системи автоматичного реферування

Особливості функціонування спеціалізованого інтелектуального агента визначаються його інтересом – вектором оцінок бажаності можливих станів агента. Для опису інтересу агента, за допомогою якого він розрізняє стани довколишнього світу та позиціонує себе у ньому, застосовується функція корисності, котра є числовою оцінкою його бажаності для агента. Корисності об'єднуються з ймовірностями дій для визначення очікуваної корисності кожної дії. Отже, вхідними даними для процесу автоматичного реферування є: n - кількість документів, D - масив з n документів, $D=\{d_1, d_2, d_3, \ldots\ldots, d_n\}$.

Для $d_i \in D$ формується множина тем $T=\{t_1, t_2, t_3, \ldots. tm\}$ і множина ваг кожної теми $P=\{p_1, p_2, p_3, .. p_m\}$. $t_j \in T$ описується множинами слів і словосполучень і частотами їх появи у тексті $W=\{w(1)_j, w(2)j, w(3)j, \ldots w(lj)j\}$ і $F=\{f(1)j, f(2)j, f(3)j, \ldots f(lj)j\}$, $w(lj)j$ – слово або словосполучення з документа d_i , що визначає тему t_j; $f(lj)j$ – частота появи у документі d_i слова чи словосполучення $w(lj)j$; lj – кількість слів або словосполучень, що описують тему t_j.

Наступні функції виконуються у автоматичному режимі:

- визначення тематичних рубрик документа;
- визначення об'єктів на основі онтологічного опису;
- формування пошукового образа документа;
- формування частотного словника ключових слів і словосполучень;

У автоматизованому режимі забезпечується:

- складання словників формалізованих описів рубрик за тестовими вибірками документів;
- ведення словників, необхідних для роботи програми.

Загальну схему автоматичного складання реферату подано на рис. 3.

Алгоритм реферування множини документів складається з двох укрупнених кроків:

1. реферування кожного документа (див. рис. 4);
2. реферування отриманих проміжних рефератів.

Алгоритм реферування проміжних рефератів:

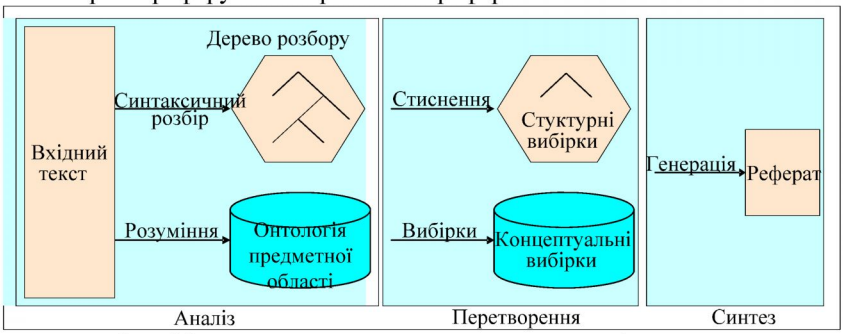


Рис. 3 Загальна схема процесу автоматичного реферування

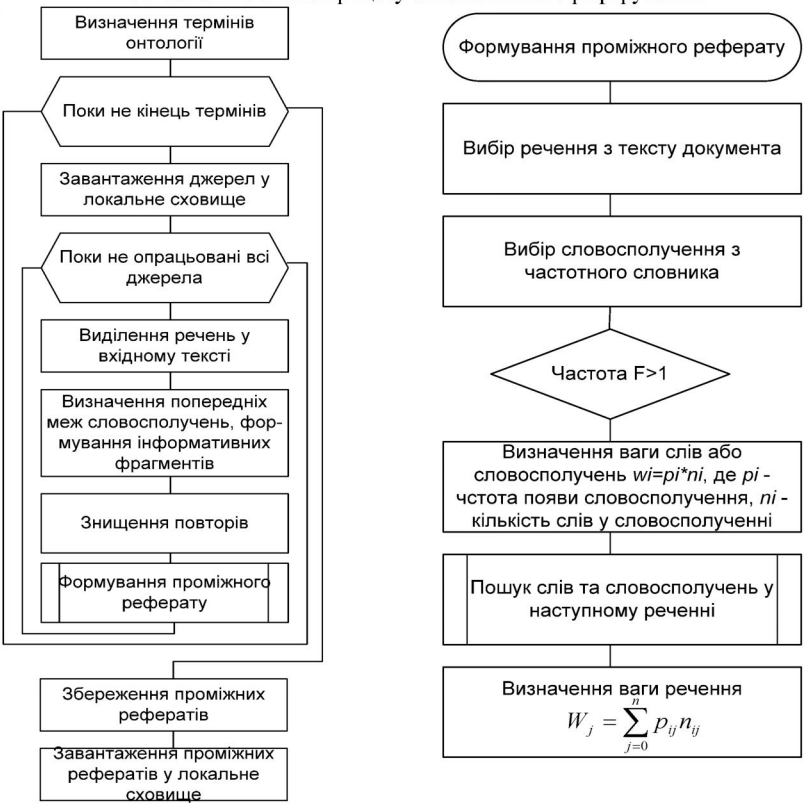


Рис. 4. Алгоритм формування проміжних рефератів для множини вхідних джерел

1. Формування списку речень, у які входять слова й словосполучення, що характеризують теми.
2. Видалення з тексту кожного речення неінформативної лексики.
3. Обчислення ваги кожного речення.
4. Перевірка речень на тотожність.
5. Обчислення коефіцієнта стиску реферату.
6. Видалення речення із найменшою вагою, якщо був отриманий реферат з коефіцієнтом стиснення більше заданої величини.
7. Повторення п. 5-6 доти, поки не буде отриманий реферат, що задовольняє критерію стиснення.

Опишемо програмну реалізацію розроблених алгоритмів. Як предметну область нами обрано веб-сервіси. Підмножина ключових слів, які входять у онтологію цієї ПО наведена на рис. 5. Додавання джерел на базі яких буде сформовано реферат та визначення їх структури здійснюється у формі, поданій на рис. 6.

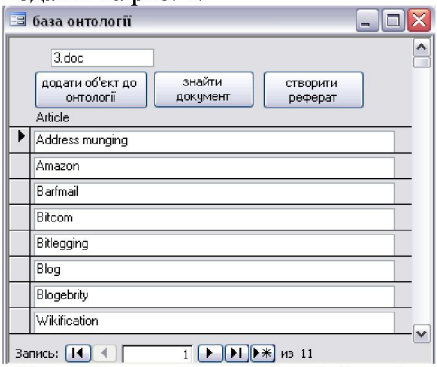


Рис. 5. Онтологія предметної області веб-сервісів

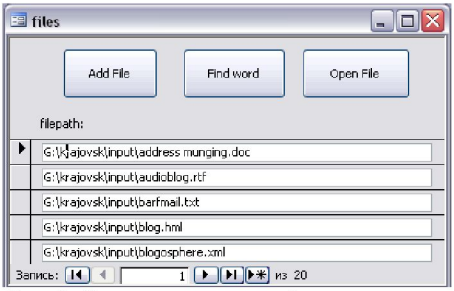


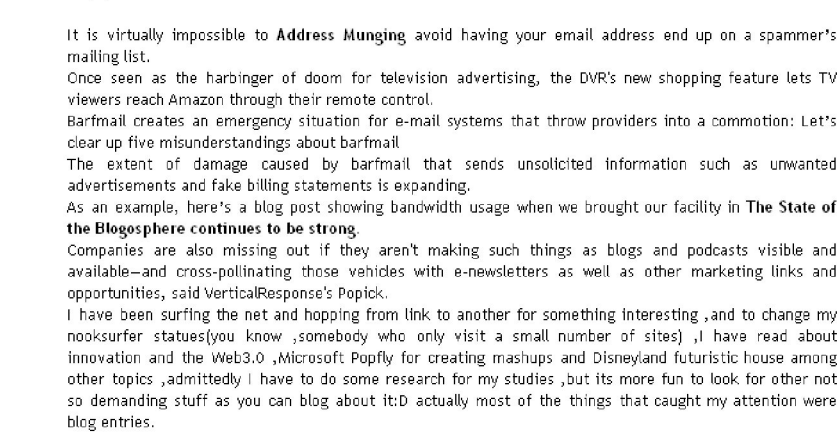
Рис. 6. Опис джерел для реферування

Використовуючи алгоритм формування проміжних рефератів (див. рис. 4), будуємо множину проміжних рефератів. Їх формування полягає у створенні реферату кожного джерела та його перегляді за необхідності. Далі на основі проміжних рефератів знову запускається процес реферування та формується кінцевий реферат. Приклад такого реферату наведено на рис. 7.

Висновки

1. Реферування множини документів є актуальною та важливою задачею, розв'язання якої практично відсутнє у сучасних засобах реферування.
2. Ресурси, що підлягають реферуванню, мають різноманітну природу, традиційно по різному обліковуються і потребують попереднього опрацювання. Для цього запропоновано використовувати простори даних. Визначено формальне поняття простору даних з точки зору математичного забезпечення.

3. На якість отриманого реферату впливає важливість термінів та понять відповідної предметної області. Тому запропоновано використовувати адаптивні онтології для врахування взаємозв'язків та ієрархій між поняттями, на основі чого здійснювати перерахунок їх ваг.


Реферат

It is virtually impossible to **Address Munging** avoid having your email address end up on a spammer's mailing list.
Once seen as the harbinger of doom for television advertising, the DVR's new shopping feature lets TV viewers reach Amazon through their remote control.
Barfmail creates an emergency situation for e-mail systems that throw providers into a commotion: Let's clear up five misunderstandings about barfmail
The extent of damage caused by barfmail that sends unsolicited information such as unwanted advertisements and fake billing statements is expanding.
As an example, here's a blog post showing bandwidth usage when we brought our facility in **The State of the Blogosphere continues to be strong**.
Companies are also missing out if they aren't making such things as blogs and podcasts visible and available—and cross-pollinating those vehicles with e-newsletters as well as other marketing links and opportunities, said VerticalResponse's Popick.
I have been surfing the net and hopping from link to another for something interesting ,and to change my nooksurfer statues(you know ,somebody who only visit a small number of sites) ,I have read about innovation and the Web3.0 ,Microsoft Popfly for creating mashups and Disneyland futuristic house among other topics ,admittedly I have to do some research for my studies ,but its more fun to look for other not so demanding stuff as you can blog about it:D actually most of the things that caught my attention were blog entries.

Рис. 7. Кінцевий реферат

4. Розвиток розроблюваної системи передбачається у напрямі отримання реферату на українській мові з можливостями перекладу вихідних джерел на нашу мову. У цьому напрямі вже розпочато побудову онтології української мови. Вона служитиме для генерування кінцевого реферату з метою максимального наближення до природної мови.

1. *Соловьев В.И.* Составление и редактирование рефератов: Вопросы теории и практики // *А.А.Гречихин, И.Г.Здоров, В.И.Соловьев.* Жанры информационной литературы. Обзор. Реферат. – М., 1983. – С. 217.
2. *Хан У.* Системы автоматического реферирования / *У.Хан, И.Мани.* – Открытые системы. – Режим доступу до журналу: http://www.osp.ru/os/2000/12/178370
3. *Barzilay R.* Sentence Ordering in Multidocu-ment Summarization. Computer Science at Columbia University, Web seit, 2007,
http://www.cs.columbia.edu/nlp/papers/2001/barzilay_al_01.pdf
4. *Даревич Р.Р.* Оцінка подібності текстових документів на основі визначення інформаційної ваги елементів бази знань / *Р.Р.Даревич, Д.Г.Досин, В.В.Литвин, З.Т.Назарчук* // Исскуственный интеллект. – Донецк. – № 3. – 2006. – С. 500-509.
5. *Даревич Р.Р.* Метод автоматичного визначення інформаційної ваги понять в онтології бази знань / *Р.Р.Даревич, Д.Г.Досин, В.В.Литвин.* // Відбір та обробка інформації. – Львів. – Вип. 22(98). – 2005. –С.105-111.
6. *Литвин В.В.* Спосіб введення метрики для визначення відстані між текстовимми документами / *В.В.Литвин* // Інформаційні системи та мережі. – Львів. – №621. – 2008. – С.162-170.
7. *Шаховська Н.Б.* Про задачу автоматизованого анотування події на основі простору даних / *Н.Б.Шаховська, В.В.Литвин* // Фізика. Електроніка. (Комп'ютерні

системи та компоненти). – Чернівці. – №426. – 2008. – С. 58-62.
8. *Шаховська Н.Б.* Простір даних області наукових досліджень //Моделювання та інформаційні технології. - ІПМЕ ім. Г.Є.Пухова «Моделювання та інформаційні технології». – Вип. 45, К.: 2008, с.132-140.

Поступила 19.01.2009р.

УДК 621.3

П.Й. Омеляновський, г.п. «Західенерго», Львів, Л.С. Сікора, д.т.н., НУ «ЛП», Львів, І.О. Малець, н.с. ЦСД «ЕБТЕС», Львів, Ю.Г. Міюшкович, асп., НУ «ЛП», Львів.

СТРАТИФІКАЦІЯ І ІНТЕГРАЦІЯ ІЄРАРХІЧНИХ СИСТЕМ УПРАВЛІННЯ

Анотація. На основі концепції Месаровича, в статті розглянуто та введено поняття термінального ситуаційного простору, процедури інтеграції і стратифікації ієрархічних систем управління на інтервалі термінального часу в застосуванні до енергетики та виробництва.

Аннотация. На основе концепции Месаровича, в статье рассмотрено понятие терминального ситуационного пространства, процедуры интеграции и стратификации иерархических систем управления на интервале терминального времени применительно к энергетике и производству.

Annotation. The concept of terminal situational space, procedures of integration and stratification of hierarchical control system on the interval of terminal time, is considered the article as it applies to energy and production.

Ключові слова: ієрархія, система, інтеграція. страта, термінальне управління.

Актуальність. Сучасний етап розвитку промислових структур характеризується їх виробничою, інформаційною, ресурсною інтеграцією на основі вироблення корпоративних стратегій, як основи реалізації глобальних цілей.

Прийняття рішень в таких структурах як в нормальних штатних, так і в надзвичайних ситуаціях є складною проблемною задачею, і тому для свого розв'язання вимагає необхідних інформаційних і системних технологій та логіко-математичних методів для побудови процедур і каналів рішень та моделей об'єктів і образів ситуацій в них [1,2].

Проблема опису динаміки об'єктів і систем.

Моделі динаміки систем можна описати і представити (рис. 1а) в фазовому просторі через координати (Rx_1,Rx_2) в базисі $(\vec{l}_1, \vec{l}_2, \varphi)$, де ($x_{10}$,$x_{20}$) –