Препринти Інституту фізики конденсованих систем НАН України розповсюджуються серед наукових та інформаційних установ. Вони також доступні по електронній комп'ютерній мережі на WWW-сервері інституту за адресою http://www.icmp.lviv.ua/

The preprints of the Institute for Condensed Matter Physics of the National Academy of Sciences of Ukraine are distributed to scientific and informational institutions. They also are available by computer network from Institute's WWW server (http://www.icmp.lviv.ua/)

Олеся Ігорівна Мриглод

Автоматизований алгоритм пошуку термінів у наукових публікаціях

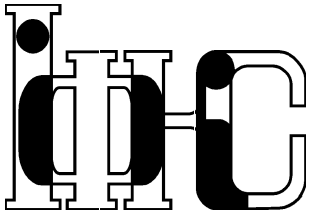
Роботу отримано 30 жовтня 2015 р.

Затверджено до друку Вченою радою ІФКС НАН України

Рекомендовано до друку Лабораторією статистичної фізики складних систем

Виготовлено при ІФКС НАН України

Національна академія наук України



ІНСТИТУТ ФІЗИКИ КОНДЕНСОВАНИХ СИСТЕМ

ICMP–15–04U

О. Мриглод

АВТОМАТИЗОВАНИЙ АЛГОРИТМ ПОШУКУ ТЕРМІНІВ У НАУКОВИХ ПУБЛІКАЦІЯХ

ЛЬВІВ

УДК: 001.893; 519.248; 53.08; 311.21

PACS: 01.30.-y, 02.50.-r, 07.05.Kf

Автоматизований алгоритм пошуку термінів у наукових публікаціях

О. Мриглод

Анотація. У роботі описано послідовність застосування одного з алгоритмів автоматизованого пошуку наукових термінів, модифікованого з огляду на специфіку поставленої задачі. Було проаналізовано сукупність наукових документів з обраної тематики, погрупованих за кількома дисциплінарними напрямками. В результаті комбінації лінгвістичного та статистичного підходів до аналізу текстів було знайдено перелік найбільш важливих термінів, що дають змогу оцінити спектр дрібніших тематичних напрямків у публікаціях з кожної дисципліни.

Semi-automatic algorithm for terms identification in scientific publications

O. Mryglod

Abstract. The application of partially modified semi-automatic algorithm for scientific terms identification is described in this paper. The set of research papers of a given topic within several disciplines were analyzed. The combination of linguistic and statistical approach to the analysis of texts gave a possibility to get the list of the most important terms. These terms can be used to reveal the spectra of subtopics in the set of selected publications within each discipline.

Подається в Вісник НУ "Львівська політехніка": Комп'ютерні науки та інформаційні технології
Submitted to Bull. Lviv Polytechnic Nat. Uni.

Вступ

Серед наукометричних досліджень важливе місце посідають проблеми вивчення структури науки та її еволюції. Виявлення так званих "гарячих напрямків" та спостереження за розвитком окремих тематик – це задачі, розв'язок яких може бути вельми корисним для практичного використання. Адже інформація про те, які напрямки у науці на даний момент є особливо затребуваними та гіпотетично перспективними, є необхідною для прийняття рішень, починаючи від постановки задачі для майбутніх дослідників і завершуючи розподіленням державних або ґрантових коштів. Завдання моніторингу наукових напрямків, на перший погляд, може видатись достатньо простим, проте з огляду на складність системи науки та процесів, що відбуваються у ній, однозначного вирішення досі немає [1,2]. Для того, щоб погрупувати публікації або, скажімо, видання за тематичною ознакою – і тим самим визначити певну структуру наукових досліджень – використовуються різні методи. Окрім експертного аналізу, тобто класифікації чи сортування "вручну", використовуються алгоритми автоматизованої кластеризації на основі даних про співавторство або ж співцитування (про одні із перших спроб див. у [3,4]). Також проблема визначення тематичного забарвлення наукових текстів є суміжною до проблеми виділення основних тематичних концепцій та побудови тезаурусів, що часто використовує лінвістичні підходи для аналізу власне змісту текстів – так званого контент-аналізу. А вже підзадача виявлення тематики публікації виявляється потенційно корисною для цілого спектру практичних застосувань: організації релевантного інформаційного пошуку, каталогізації та рубрикації електронних ресурсів або ж публікацій у виданні, автоматичного пошуку рецензентів та багатьох інших. Загалом, завдання зводиться до розроблення методів автоматичного чи хоча б автоматизованого (за часткової участі людини-експерта) аналізу текстів з виділенням ключових тематичних концепцій, представлених у вигляді ключових слів. У випадку наукових текстів результатом може бути перелік значущих наукових термінів (більш формальне визначення *терміну* обговорюється далі), що, власне, відображають ці концепції.

1. Коротко про підходи до виявлення ключових слів у текстах

Намагання розробити автоматичний чи хоча б напівавтоматичний спосіб виділення ключових термінів, що описували б основні конце-

пції документів здійснюються вже не перше десятиліття (див., наприклад, [5–8]). Адже така задача потенційно має ряд практичних застосувань: тематичне маркування виявлених груп документів, завдання інформаційного пошуку, каталогізації, та інші, згадані вище.

Перш ніж перейти власне до обговорення способів аналізу текстів, необхідно визначити, що ж ми розуміємо під *терміном*. Насправді не існує чітко формалізованого визначення, проте зазвичай термінами називають так звані "змістовні" (чи "сигнальні" [5]) слова або словосполучення, що передають основні змістові ідеї тексту, тобто відображають ту чи іншу тематичну концепцію, висвітлену у документі. На відміну від таких "змістовних" слів, "функціональні" використовуються для зв'язки речень та передачі додаткової інформації [9]. Можна собі уявити, що у "функціональні" слова є середовищем, яке забезпечує розташування "змістовних" слів – наче риб у воді. До "функціональних", зокрема, відносять усі види сполучників та службових слів.

Важливо розуміти, що "змістовність" кожного конкретного слова є невід'ємною від контексту. Так, слово "шум" може бути вторинним у тексті з біології, проте стати терміном для фізичної публікації. Релевантність слова може змінюватись навіть в рамках однієї дисципліни, тому множина термінів завжди є індивідуальною для конкретно обраного набору документів.

Окремою проблемою є включення в аналіз не лише одиничних, але й складених термінів (з кількох слів). Їх автоматичне виділення технічно є більш проблематичним, проте часто вони допомагають уточнити більш загальні за значенням одиничні терміни, детальніше описати ту чи іншу концепцію документа. Скажімо, коли іменник "пухлина" може означати досить широкий медичний спектр тематик, то словосполучення "тироїдна пухлина" вже значно звужує коло пошуку. У цьому випадку знову не знімається питання про те, що в подальшій роботі все-таки вважати терміном: слово "пухлина" чи словосполучення "тироїдна пулина". Тут, як правило, потрібно приймати рішення знову ж таки для кожного конкретного набору документів.

Вважається, що найбільш надійним методом визначення множини ключових термінів для корпусу документів чи певної галузі є залучення експертів – фахівців у відповідній ділянці. Автоматизувати цей процес поки що не вдається власне через відсутність абсолютних критеріїв, багатозначність мови, її контекстність тощо. Проте вже тривалий час пропонуються та вивчаються напівавтоматичні методи аналізу текстів та визначення ключових слів, термінів або ж

концептів. Для знаходження множини слів чи словосполучень, що потенційно можуть виступати такими ключовими словами, можна використовувати різні принципові підходи: на основі статистичного, синтаксичного чи змішаного аналізу слів у наборі документів [6, 9]. У першому випадку текст розглядається лише як випадковий набір або ж впорядкована послідовність елементів – слів. Тоді можна робити частотний аналіз, знаходити типові послідовності елементів та застосовувати інші статистичні підходи. У другому випадку враховується синтаксис, частини мови та структура слів у реченнях. А, зрештою, на практиці найчастіше використовується комбінація цих двох методів [9, 11].

Якщо знехтувати структурою документа та вважати його "мішком зі словами", то можна проаналізувати частоту вживання слів k, побудувавши її розподіл. Ще у середині минулого століття було доведено, що в результаті такого аналізу буде одержано степеневий закон розподілу слів, відомий як закон Зіпфа [10]. Останній говорить про те, що у тексті типово є велика кількість різних слів, що зустрілися один раз або кілька разів, і лише декілька таких, що вживаються дуже часто (див. далі Рис. 3). Певний парадокс полягає у тому, що хоч частота вживання слів є одним із базовим понять, проте на основі лише частотного розподілу неможливо визначити, які ж зі слів можна вважати ключовими або такими, сукупність яких описує власне основні концепції документа. Найчастіше вживані слова, як правило, є дуже загальними за змістом, а рідко вживані – дуже конкретними, проте не можуть вважатися статистично значущими. Вважається, що найкращі кандидати у "змістовні" слова знаходяться десь посередині згаданого частотного розподілу (див. Рис. 3) – такі, що вживається не найчастіше, проте і не надто рідко [5, 9, 11]. В залежності від зростання довжини текстів, частота "функціональних" та "змістовних" слів змінюється по-різному: для перших вона пропорційно зростає, тоді як для других просто відбувається розширення словника (більший текст – більша імовірність обговорення нової концепції – нові ключові слова/терміни) [9, 11].

Окрім частоти вживання, додаткову інформацію можна отримати при врахуванні структури корпусу та окремих його документів. Відомо, що "функціональні" та "змістовні" слова неоднаково розподілені серед документів чи структурних частин одного документа. Тоді як перші характеризуються скоріше рівномірною розкиданістю по корпусу та документах, другі вживаються нерівномірно – сконцентровано у групі документів (чи в певному місці окремого документа) та рідко в інших.

Вже побіжний огляд показує велику кількість неоднозначностей, пов'язаних із намаганням автоматично виявити значущі слова, що б відображали тематичний спектр набору документів. При цьому досі мова йшла лише про формування списку кандидатів на терміни – слів/словосполучень, які в подальшому потрібно оцінити на предмет їх "значущості", тобто релевантності, специфічності та важливості для конкретного набору текстів. За відсутності визначення такої значущості, можна використовувати різні методи для "зважування" потенційних термінів і порівняння їх ваг між собою. Наприклад, можна використати вже згадану властивість неоднорідного розподілу "змістовних" слів. Про міру "специфічності" слова/словосполучення може говорити добуток частоти його вживання k_i на так звану обернену частоту документів idf_i, в яких воно зустрічається. Остання величина рівна відношенню загальної кількості документів у корпусі N до кількості документів, в яких трапилось дане слово/словосполучення n_i: з огляду на типово велике значення N, можна розглядати логарифм відношення, тобто $idf_i = \log(N/n_i)$. З одного боку, величина $k_i \times idf_i$ буде пропорційною до загальної вживаності кандидату у терміни, а з іншого боку – обернено пропорційною до кількості *різних* документів, де він зустрічається [9]. Існує ціла низка методів зважування та нормалізації, що можуть враховувати довжини документів чи їх структуру (тобто місця локалізації слів у певних частинах документа), факти співпояв слів, їх контекст, тощо.

Якщо мова йде виключно про наукові статті та виявлення у них наукових термінів, то задача дещо полегшується з огляду на чітку структурованість таких документів. Вважається, що значущі терміни найбільше сконцентровані у певних структурних частинах, таких як заголовок, анотація, перший абзац статті чи висновки. Деякі дослідники вважають, що комбінація заголовка та анотації є достатньою основою для аналізу (наприклад, [6, 12]), інші вважають найбільш релевантними окремо взяті заголовки (див. [13]), проте завжди потрібно враховувати невеликий розмір цих фрагментів з точку зору статистичного підрахунку частот. Крім того, значущість тих чи інших структурних лементів статей є різною для природничих та гуманітарних наук.

Можна виділити типові риси наукових термінів загалом. Так вважається, що найчастіше це одноосібні іменники або ж словосполучення, що формуються навколо головного іменника, доповнюючись іншими іменниками, прикметниками, тощо. Потрібно зауважити, що такі висновки наразі найбільш обгрунтовані для англомовних текстів як найбільш досліджених [9, 11].

Далі у роботі описано покрокову напівавтоматичну процедуру виявлення наукових термінів у публікаціях обраної тематики. Таке завдання було поставлене в рамках ширшої задачі дослідження реакції наукової спільноти – що відображається власне в опублікованих роботах – на визначену подію.

2. Постановка задачі: приклад тематичного аналізу наукових публікацій

Нещодавно в рамках вивчення реакції наукової спільноти на Чорнобильську аварію [14, 15] було зібрано бібліографічні дані про усі релевантні до проблеми наукові публікації в базі даних Scopus (`www.scopus.com`) станом на початок 2015 року. Загалом кінцевий перелік включав більше 9,5 тис. бібліографічних записів про публікації, що містили різні написання слова "Chornobyl" у заголовках, анотаціях чи ключових словах. Таким чином, було зібрано тематичну колекцію наукових документів із вузької тематики. Окрім дослідження розподілу публікацій за галузями науки та по роках, аналізу відповідної мережі співпраці на рівні країн, виявлення зміни зацікавленості в рамках різних дисциплін та інших завдань, цікаво було дослідити тематичний спектр всередині зібраного корпусу документів. Адже поруч із загальнодисциплінарними тенденціями до підвищення чи загасання інтересу в рамках тієї чи іншої галузі науки, можна очікувати зміни тематичного спектру на більш тонкому масштабі, в рамках однієї дисципліни – адже з часом актуальність одних проблем втрачається, в той час як інші починають активно досліджуватись. Такий детальніший аналіз тим більше цікавий з огляду на те, що наразі для домінуючих дисциплін (за числом чорнобильських публікацій у Scopus) спостерігається більш-менш стала картина, тобто щорічна кількість публікацій коливається навколо певного значення. З іншого боку, для низки інших дисциплін було спосетрежено тенденції до загасання (скажімо, для ветеринарії) чи зростання (наприклад, для економіки та фінансів) інтересу до чорнобильської тематики [14, 15]. Таким чином, було поставлене завдання виявлення внутрішніх тематик для набору статей хоча б для п'яти найбільше представлених у базі Scopus дисциплін: медицини (3 635 статей); наук про нвколишнє середовище (3 156); енергетики (1 470); фізики та астрономії (1 437); біохімії, генетики та молекулярної біології (1 198).

Виявлення ключових термінів дає змогу судити про найбільш актуальні завдання в розрізі часу. Такого роду аналіз також допомагає побудувати карту наукових дисциплін – тобто візуально предста-

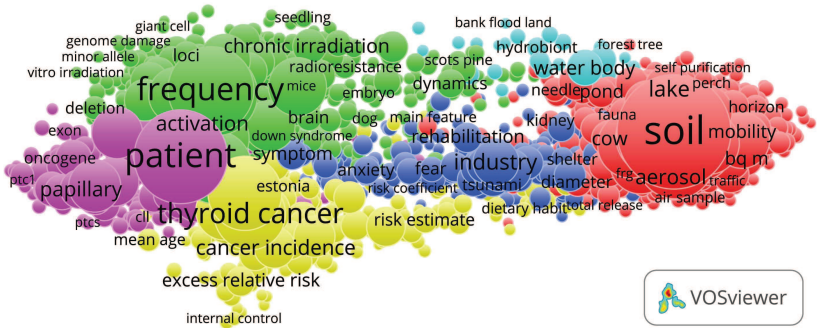


Рис. 1. Карта термінів для чорнобильських публікацій з медицини; наук про нвколишнє середовище; енергетики; фізики та астрономії; біохімії, генетики та молекулярної біології, знайдених у базі Scopus станом на початок 2015 року. Показано 60% найбільш релевантних термінів.

вити, як знайдені терміни (а отже і тематичні піднапрямки) взаємопов'язані між собою. Наприклад, на Рис. 1 показано карту термінів для нашого набору статей з домінуючих дисциплін, згенеровану за допомогою спеціальної програми VOSviewer [16, 17]. У цьому випадку для автоматичного виділення термінів використовувалась інформація про їх співпояви у заголовку та анотації (що в даному випаду трактуються як один документ) кожної статті. Вбудований алгоритм кластерування дає змогу розрізнити декілька тематичних груп: чотири більші та дві менші. І хоча ці групи не відповідають взаємно однозначно п'яти домінуючим дисциплінам, можна чітко розрізнити дві найбільші ділянки досліджень, що стосують аварії на Чорнобильській АЕС : вплив на здоров'я людини (три великі кластери зліва) та наслідки для навколишнього середовища (великий кластер справа). Також достатньо добре візуально розрізняються піднапрямки, що пов'язані із онкологічними захворюваннями, генетичними ефектами та аналізом шляхів забруднень у різних середовищах.

3. Покрокова процедура виявлення термінів

Для того, щоб отримати ключові слова, набір яких найбільш точно описує загальний тематичний зміст чорнобильських наукових публікацій з п'яти домінуючих дисциплін, перелічених вище та двох додаткових – мистецтва (лише 59 публікацій у Scopus) та соціології (310)

як яскравих представників гуманітарних наук, – було модифіковано процедуру, запропоновану в [11]. Основні кроки описані нижче.

Крок 1. В першу чергу потрібно визначити, що буде основою для аналізу: окрім інформації про авторів та видання, у нашій базі зібрані анотації, заголовки та ключові слова. Оскільки ключові слова за замовчуванням можна вважати авторськими термінами, за основу об'єктивного аналізу беремо анотації та/чи заголовки. Заголовок за своїм призначенням мав би найбільш точно та коротко вказувати на суть статті, проте все частіше заголовки покликані скоріше привабити читача, аніж вказати на зміст. Тому не не комбінувалися заголовок+анотація, а порівнювалися два окремі випадки, коли в якості *документів* трактувалися окремо заголовки та окремо анотації.

Крок 2. Для того, щоб одержати інформацію про лінгвістичні властивості слів у документах, можна використати одну із доступних у вільному доступі програм – а саме TreeTagger [18]. Ця програма призначена для маркування слів у тексті за частинами мови та знаходження базової форми для кожного слова – що пізніше дає змогу нехтувати різними закінченнями, що залежать від числа або роду. Приклад результатів обробки фрагмента тексту програмою наведений на Рис. 2.

```
colony-stimulating|NN|colony-stimulating
factors|NNS|factor
for|IN|for
the|DT|the
treatment|NN|treatment
of|IN|of
the|DT|the
hematopoietic|JJ|hematopoietic
component|NN|component
```

Рис. 2. Результат обробки програмою TreeTagger фрагменту тексту: "colony-stimulating factors for the treatment of the hematopoietic component".

Крок 3. Із одержаного переліку легко відібрати потрібні конструкції – які ми далі називатимемо *семантичними одиницями*, слі-

дуючи означенню, введеному в роботах [11, 15, 19] (semantic units) – одиничні іменники або ж словосполучення, що складаються лише з іменників та прикметників. Подібно, як у [11, 19], було використано загальне правило для відбору складених конструкцій: `*прикметник *іменник`. Ця загальна форма означає, що фраза може розпочинатись із довільної кількості прикметників та завершуватись довільною кількістю іменників, наприклад: “post-chernobyl radioactive contamination”, “radionuclide contamination source”. Крім того, певні найпростіші правила перетворення дають змогу не відсіювати ті конструкції, які існують неявно:

- Фраза, що містить на початку декілька прикметників, розділених комами, та один іменник вкінці, перетворюється у декілька фраз, що складаються з одного прикметника та іменника; наприклад, конструкція “personal, political, linguistic, historical complexity” буде врахована у вигляді чотирьох семантичних одиниць: “personal complexity”, “political complexity”, “linguistic complexity”, “historical complexity”;

- Фраза, у якій один прикметник відноситься до декількох іменників, розділених сполучником “and” (“і” чи “та” з англ.), перетворюється у декілька фраз, що містять прикметник та один із іменників; наприклад із конструкції “individual responses and responsibilities” отримаємо “individual responses” та “individual responsibilities”;

Звичайно, неможливо передбачити та забезпечити перетворення усіх можливих варіантів конструкцій коректно. Поруч із технічними помилками чи описками існують такі, які за змістом мали б бути включені до розгляду, проте не відповідають заданому шаблону, містячи сполучники. Скажімо, класичним прикладом є словосполучення “degrees of freedom”.Тому потрібно пам’ятати про неминучі похибки автоматичних процедур, особливо тих, що стосуються обробки природньої мови.

При відборі семантичних одиниць відразу ж варто відсіяти такі, що занесені у так званий стоп-список. Цей перелік, як правило, складається вручну і конкретно для кожного набору документів. Так, наприклад, у нашому випадку стоп-словами стали семантичні одиниці “article” (“стаття”), “Elsevier” (назва видавництва), різноманітні одиниці вимірювання величин, тощо.

В результаті виконання перших трьох кроків, для нашого корпусу було отримано список із 80 094 семантичних одиниць для анотацій та 15 020 – для заголовків.

Крок 4. Наступним кроком є підрахунок семантичних одиниць для проведення частотного аналізу. Проте проста, на перший погляд, процедура не є однозначною.

- Перш за все, всупереч поширеній практиці, що передбачає відсіювання маловживаних семантичних одиниць від самого початку [11, 19], ми їх початково не виключаємо. Оскільки розміри наших вибірок даних, особливо для гуманітарних дисциплін, є невеликими, таке відсіювання завадить побачити загальну форму частотного розподілу. На Рис. 3 зображено частотно-рангові розподіли семантичних одиниць: для їх побудови спочатку треба посортувати усі семантичні одиниці за спаданням частоти поява k (по вертикальній осі на рисунку), а тоді присвоїти їм відповідні ранги r в порядку зростання (горизонтальна вісь, відповідно). Таким чином, на Рис. 3 бачимо відповідні графіки для анотацій (а) та заголовків (б), що демонструють близькість до степеневого закону. Така форма розподілу є типовою для текстів, написаних природною мовою [10,20]. Більше того, не лише форма, але й нахил одержаної кривої (зі значенням експоненти, що близька до -1) свідчить про універсальні характеристики таких коротких текстів, як анотації (Рис. 3 а). Такий результат не є цілком очевидним, адже в даному випадку ми оперуємо не частотою слів, а частотою лише певних семантичних конструкцій.

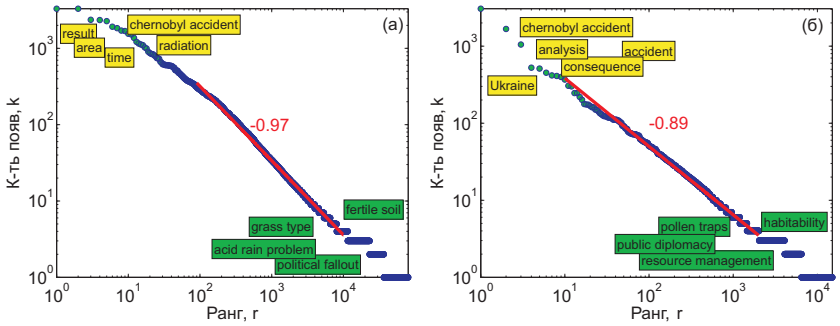


Рис. 3. Частотно-рангові розподіли семантичних одиниць для (а) анотацій та (б) заголовків чорнобильських публікацій з семи обраних дисциплін у базі даних Scopus. Розподіли добре апроксимуються степеневими залежностями $k \sim r^{-\alpha}$ із показниками $\alpha \approx 0{,}97$ та $\alpha \approx 0{,}89$, відповідно.

- Перед тим, як рахувати частоту появ складених конструкцій – тобто тих, що складаються з двох чи більше слів – можна спочатку оцінити їх зв'язаність, статистично підтвердивши, що ціла конструкція зустрічається достатню кількість разів у порівнянні з частотою її окремих елементів. Тобто можна йти шляхом первинної перевірки того, чи знайдені складені семантичні одиниці повинні фігурувати як одне ціле, чи доречно їх розбити на менші елементи (див. [11, 19]). Проте у нашому дослідженні такої перевірки не робиться, оскільки розраховується частота не лише всієї складеної конструкції, але й її складових: окремо останнього слова (іменник, що трактується як головний у семантичній одиниці), а далі останнього слова разом із тими, що йому передують, додаючи їх один за одним. Наприклад, для семантичної одиниці "low-dose radiation exposure" підраховується частота вживань "exposure", "radiation exposure" та "low-dose radiation exposure". Таким чином ми беремо до уваги так звані вкладені терміни (nested terms): терміни, що можуть бути самостійними або входити у склад інших [11].

- Для підрахунку кількості вживань семантичної одиниці було використано так званий бінарний спосіб, тобто якщо в одному і тому ж документі вона зустрілася більше одного разу, все він одно "зараховується" лише одноразово. Таким чином, загальна кількість появ буде дорівнювати кількості документів, у яких зустрілася дана конструкція.

Крок 5. Накінець, із загального списку зібраних та підрахованих *семантичних одиниць* потрібно виділити власне *терміни*. Як вже було вище згадано, важливість семантичної одиниці як терміну не є прямо пропорційною до її частоти вживання. Найчастіше вживані слова, що потрапляють на початок частотно-рангового розподілу типу Зіпфа, є, як правило, найбільш загальними за змістом для даного набору документів. Скажімо, у нашому випадку це такі семантичні одиниці як "Chornobyl accident", "radiation", "Ukraine", тощо (Рис. 3). З іншого боку, слова із найменшою кількістю вживань, що потрапляють у "хвіст" розподілу, дійсно є вельми специфічними, проте не розглядаються як статистично значущі: наприклад, "grass type", "public diplomacy", тощо. Найбільш вірогідними кандидатами у терміни є ті семантичні одиниці, що знаходяться посередині розподілу (Рис. 3) [9]. Тому на цьому кроці важливо застосувати певні частотні (чи інші) фільтри для відокремлення термінів. Спочатку введемо

нижнє критичне значення частоти $k_c = 4$, тобто відсіємо усі семантичні одиниці, що зустрілися у менше, ніж 4 документах. Це значення було обрано шляхом емпіричного спостереження: частота появ семантичних одиниці при $k \leqslant k_c$ спадає повільно, тоді як вище цього значення відбуваться зміна режиму на значно швидший.

Черговий фільтр пов'язаний із нерівномірним розповсюдженням термінів між дисциплінами. Тоді як семантичні одиниці можуть однаково активно використовуватися у статтях, що відносяться до різних галузей знання, термінами будуть називатися ті, що є характерними для певної дисципліни чи кількох дисциплін. Щоб виразити концепцію такого нерівномірного розподілу у числовому вигляді, використаємо ідею так званої "термінності" (*termhood*), запропоновану у [11, 19]. Мова йде про міру специфічності семантичної одиниці, тобто її характерності для певної/певних дисциплін. Для розрахунку термінності спочатку потрібно побудувати ймовірнісний розподіл частоти вживання усіх кандидатів у терміни $P(d)$ по дисциплінах ($d = 1 \dots 7$). Відповідний графік для бази даних анотацій (майже співпадає із аналогічним графіком для заголовків) показано на Рис. 4 (лінія з кружечками).

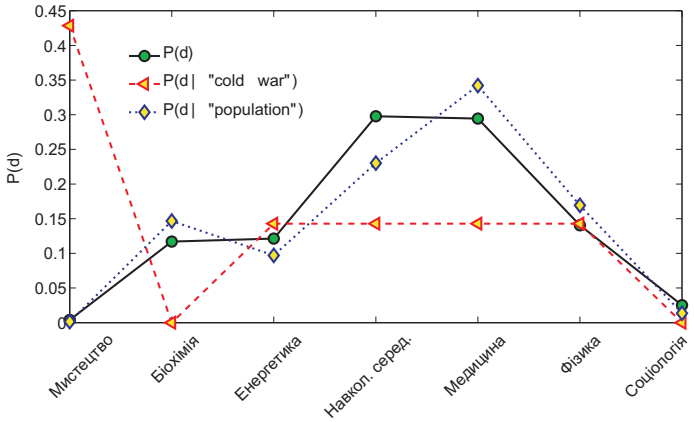


Рис. 4. Ймовірнісні розподіли частоти появ семантичних одиниці у дисциплінах на основі бази даних анотацій чорнобильських публікацій: загальний розподіл $P(d)$ та два індивідуальні розподіли для семантичних одиниць "cold war" та "population".

Із рисунку видно, що найбільш кількість семантичних одиниць відноситься до наук про навколишнє середовище та медицини. Очі-

кувано, найменші значення відповідають обидвом гуманітарним дисциплінам. Далі аналогічним чином для кожної семантичної одиниці s (s змінюється від 1 до загального числа семантичних одиниць у сформованому списку) будується її власний, індивідуальний розподіл $P(d|s)$. Іншими словами, $P(d)$ буде показувати ймовірність *будь-якої* семантчинох одиниці "потрапити" у певну дисципліну d, а $P(d|s)$ – цю ймовірність для конкретної семантичної одиниці s, див. Рис. 4 (лінії з трикутниками та ромбами). Різниця між загальним $P(d)$ та конкретним $P(d|s)$ розподілом, виражена у числовому вигляді і буде шуканою величиною. Існують різні математичні способи порівняння розподілів між собою, нами було використано запропонований у [11, 19]. При цьому для розрахунку рівня "неподібності" між кожною парою розподілів $P(d)$ та $P(d|s)$ вживається поняття так званої від'ємної ентропії. Так, міра termhood для обраної семантичної одиниці j розраховується як:

$$t_s = \sum_{d=1}^{7} \log p_d \,, \qquad p_d = \frac{P(d|s)/P(d)}{\sum_{d'=1}^{7} P(d'|s)/P(d')} \,,$$

приймаючи, що $0 \log 0 = 0$. Чим вище значення величини t_s, тим більш специфічним (більш характерним для певної дисципліни чи кількох дисциплін) вважається семантична одиниця – тим більше підстав її вважати терміном. На Рис. 5 продемонстровано, як для кожної семантичної одиниці змінюються її загальна частота k_s та специфічність t_s.

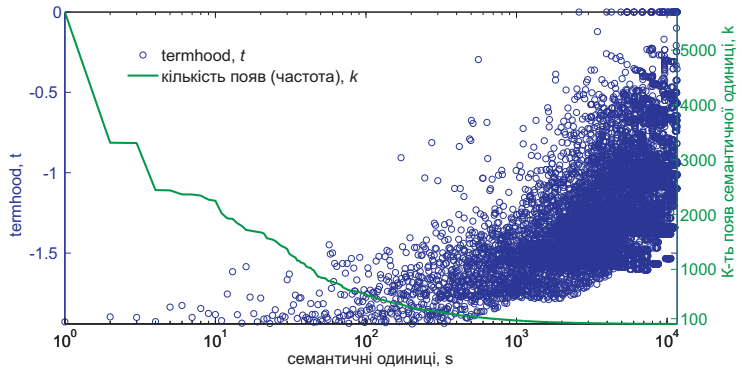


Рис. 5. Специфічність (termhood) t_s відносно частоти появ $k_s > k_c$ для семантичних одиниць на основі бази анотацій.

Очевидно, що необхідно знайти компроміс між цими двома величинами, проте будь-яке рішення буде до певної міри умовним або суб'єктивним. Така ситуація є досить типовою для задач, що вимагають участі експертів та не можуть бути повністю автоматизованими. У даному випадку до термінів було віднесено ті семантичні одиниці, що відповідали таким критеріям:

- $k_s > k_c$, де $k_c = 4$ (див. вище);
- $t_s > t_c$, де t_c рівне медіані;
- Семантична одиниця належить до перших 50-ти у переліку, відсортованому за значенням величини $t'_s \times k'_s$ (добутком t_s та k_s, нормованими так, щоб належати інтервалу $[0 \ldots 1]$) – така вага дає додаткову перевагу специфічності термінів у порівнянні з їх частотою.

Деякі семантичні одиниці були виключені зі списку термінів вручну на фінальній стадії – як пояснювалося вище, обійтися без втручання людини-експерта наразі неможливо.

В результаті було одержано переліки термінів, що характеризують публікації на тему Чорнобильської аварії для кожної із семи обраних дисциплін. На їх основі можна судити про піднапрямки, які були актуальними в рамках ширших областей досліджень. У Табл. 1 та 2 наведено по двадцять найбільш специфічних термінів, що були отримані на основі анотацій та заголовків, відповідно. Можна побачити, що у часовий період, найбільш близький до аварії, виділялися терміни із наук про навколишнє середовище. Більшість термінів, характерних для біохімії генетики та молекулярної біології, вперше з'являються у публікаціях на початку 90-х років. Гуманітарні терміни починають виникати ще дещо пізніше (2002–2006). Це підтверджує думку про те, що чорнобильська тематика досліджувалась в рамках різних дисциплін не синхронно [14,15]. Природньо, що безпосередньо після катастрофи акцентувалася увага на найшвидших її наслідках для навколишнього середовища, здоров'я людей; з ходом часу все актуальнішими стали більш віддалені наслідки, наприклад, генетичного та онокологічного характеру; натомість після кількох десяріч обговорюються також економічні, соціальні та культурні проблеми, пов'язані з аварією на ЧАЕС.

Табл. 1. Перша двадцятка термінів, найбільш специфічних для чорнобильських публікацій в рамках семи досліджуваних дисциплін, відібраних на основі анотацій статей у Scopus.

Терміни (мовою оригіналу)	Терміни (український переклад)	Характерні для:	Рік першої появи у базі даних
1. carcinoma	1. карцинома	біохімія	1992
2. thyroid carcinoma	2. карцинома щитовидної залози	біохімія	1992
3. tumor	3. пухлина	біохімія	1994
4. gene	4. ген	біохімія	1987
5. rearrangement	5. перебудова	біохімія	1993
6. papillary thyroid carcinoma	6. папілярний рак щитовидної залози	біохімія	1995
7. †ptc	7. ptc	біохімія	1992
8. papillary carcinoma	8. папілярна карцинома	гуманітарні	2002
9. science	9. наука	біохімія	1994
10. carcinogenesis	10. канцерогенез	біохімія	1992
11. malignancy	11. злоякісність	навкол. сер.	1987
12. activity ratio	12. коефіцієнт активності	гуманітарні	2006
13. threat	13. загроза	біохімія	1992
14. metastasis	14. метастази	біохімія	1994
15. surgery	15. хірургічна операція, хірургія	соціологія	1989
16. policy	16. політика	біохімія	1999
17. cleanup worker	17. працівник по очищенню (ліквідатор)	біохімія	1997
18. high frequency	18. висока частота	біохімія	1990
19. nuclear disaster	19. ядерна катастрофа	соціологія	1990
20. discharge	20. розряд	навкол. сер.	1988

†ptc – абревіатура від “papillary thyroid carcinoma”

Табл. 2. Перша двадцятка термінів, найбільш специфічних для чорнобильських публікацій в рамках семи досліджуваних дисциплін, відібраних на основі заголовків статей у Scopus.

Терміни (мовою оригіналу)	Терміни (український переклад)	Характерні для:	Рік першої появи у базі даних
1. carcinoma	1. карцинома	біохімія	1993
2. thyroid carcinoma	2. карцинома щитовидної залози	біохімія	1993
3. patient	3. пацієнт	біохімія	1991
4. rearrangement	4. перебудова	біохімія	1991
5. sediment	5. осад	навкол. сер.	1987
6. papillary thyroid carcinoma	6. папілярний рак щитовидної залози	біохімія	1995
7. transport	7. перенесення, транспорт	навкол. сер.	1987
8. mutation	8. мутація	біохімія	1989
9. tumor	9. пухлина	біохімія	1994
10. cleanup worker	10. працівник по очищенню (ліквідатор)	біохімія	1993
11. cleanup	11. очищення (ліквідація)	медицина	1992
12. unit	12. модуль	енергетика	1982
13. history	13. історія	гуманітарні	2009
14. pond	14. ставок	навкол. сер.	1987
15. policy	15. політика	соціологія	1988
16. prevalence	16. поширеність, розповсюдження	біохімія	1995
17. Black sea	17. Чорне море	навкол. сер.	1987
18. thyroid disease	18. хвороба щитовидної залози	біохімія	1991
19. radiation protection	19. захист від радіації	гуманітарні	2006
20. forest ecosystem	20. екосистема лісу	навкол. сер.	1991

4. Висновки

В результаті виконаної роботи можна зробити дві групи висновків. Перша стосується самої процедури виокремлення термінів (ключових/значущих слів) у наукових текстах. Наразі є усі підстави вважати, що її повна автоматизація не є можливою – на тому чи іншому етапі необхідно залучати експертів у відповідній галузі знань для кожного конкретного набору наукових публікацій. Така експертна участь необхідна як на проміжних стадіях, скажімо, для формування списку слів, які завідомо не є змістовними – стоп-списку, так і на кінцевій стадії для верифікації результатів. Проте численні напрацювання у цьому напрямку забезпечили чималий арсенал підходів та методів для автоматизації окремих етапів процедури пошуку термінів. Звичайно, можна повністю покластися на розроблені програми (такі продукти вже існують, наприклад, VOSviewer [16]), проте необхідно допускати відповідну похибку у результатах.

У роботі було реалізовано алгоритм пошуку наукових термінів до бібліометричної бази статей, що стосуються аварії на Чорнобильській АЕС. Цей алгоритм базуєтьсяна комбінації лінгвістичних та статистичних методів опрацювання наукових текстів. З врахуванням усіх нюансів та долею суб'єктивності, вдалося сформувати перелік термінів характерних для 5 дисциплін із найбільшою кількістю публікацій від 1986 до початку 2015 року та 2 гуманітарних дисциплін. В результаті вдалося вирізнити не лише набільш актуальні піднапрямки в рамках кожної галузі, що дає змогу більш детально описати тематичний спектр чорнобильських досліджень, але й відслідкувати їх в часі, спостерігаючи, як одна тематика змінює іншу.

Дослідження проведено в рамках проектів: "Статистична фізика у різноманітних реалізаціях" (7-ма Рамкова угода, FP7–PEOPLE, IRSES project N295302) та "Структура та еволюція складних систем із застосуванням у фізиці та природничих науках" (7-ма Рамкова угода, FP7–PEOPLE, IRSES project N612669). Особлива подяка колегам по проекту, у який увійшла дана задача: Юрію Головачу, Ральфу Кенні та Бертрану Бершу, а також Нілу ван Еку за плідні дискусії та роз'яснення певних моментів роботи програми VOSviewer.

Література

1. Tseng, Y. H., Lin, Y. I., Lee, Y. Y., Hung, W. C., Lee, C. H. (2009). A comparison of methods for detecting hot topics. Scientometrics, 81(1), 73–90.
2. Akritidis, L., Katsaros, D., Bozanis, P. (2012). Identifying attractive research fields for new scientists. Scientometrics, 91(3), 869–894.
3. Griffith, B. C., Small, H., Stonehill, J. A., Dey, S. (1974). The structure of scientific literatures II: Toward a macro- and microstructure for science. Science Studies, 4(4), 339-365.
4. White, H. D., Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. Journal of the American Society for Information Science, 32(3), 163-171.
5. Rip, A., Courtial, J. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. Scientometrics, 6(6), 381-400.
6. Tseng, Y. H. (1998, August). Multilingual keyword extraction for term suggestion. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 377-378). ACM.
7. Jones, L. P., Gassie Jr, E. W., Radhakrishnan, S. (1990). INDEX: The statistical basis for an automatic conceptual phrase-indexing system. Journal of the American Society for Information Science (1986-1998), 41(2), 87.
8. Kageura, Kyo, and Bin Umino. "Methods of automatic term recognition: A review." Terminology 3.2 (1996): 259-289.
9. Schneider, J. W. (2005, June). Verification of bibliometric methods' applicability for thesaurus construction. In ACM SIGIR Forum (Vol. 39, No. 1, pp. 63-64). ACM.
10. Zipf, G. K. (1949). Human behavior and the principle of least effort.
11. van Eck, N. J. (2011). Methodological advances in bibliometric mapping of science (No. EPS-2011-247-LIS). Erasmus Research Institute of Management (ERIM).
12. Resnick, A. (1961). Relative effectiveness of document titles and abstracts for determining relevance of documents. Science, 134(3484), 1004-1006.
13. Zuccala, A., Van Eck, N. J. (2011). Poverty research in a development policy context. Development Policy Review, 29(3), 311-330.
14. Мриглод, О. І., Головач, Ю. В. (2012). Реакція наукової спільноти на Чорнобильську аварію: аналіз розвитку тематики публікацій. Вісник НАН України.

15. Mryglod O., Holovatch Yu., Kenna R., Berche B. Quantifying the evolution of a scientific topic: reaction of the academic community to the Chornobyl disaster. Scientometrics (подано до друку).
16. Van Eck, N. J., Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523-538.
17. Van Eck, N. J., Waltman, L. (2011). Text mining and visualization using VOSviewer. arXiv preprint arXiv:1109.2058.
18. TreeTagger (2015): a language independent part-of-speech tagger http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/. Перевірено доступність 15 вересня 2015 р.
19. van Eck, N., Waltman, L., Noyons, E., Buter, R. (2010). Automatic term identification for bibliometric mapping. Scientometrics, 82(3), 581-596.
20. Simon, H. A. (1955). On a class of skew distribution functions. Biometrika, 425-440.

CONDENSED MATTER PHYSICS

The journal **Condensed Matter Physics** is founded in 1993 and published by Institute for Condensed Matter Physics of the National Academy of Sciences of Ukraine.

AIMS AND SCOPE: The journal **Condensed Matter Physics** contains research and review articles in the field of statistical mechanics and condensed matter theory. The main attention is paid to physics of solid, liquid and amorphous systems, phase equilibria and phase transitions, thermal, structural, electric, magnetic and optical properties of condensed matter. Condensed Matter Physics is published quarterly.

ABSTRACTED/INDEXED IN: Chemical Abstract Service, Current Contents/Physical, Chemical&Earth Sciences; ISI Science Citation Index-Expanded, ISI Alerting Services; INSPEC; "Referatyvnyj Zhurnal"; "Dzherelo".

CONTACT INFORMATION:
Institute for Condensed Matter Physics
of the National Academy of Sciences of Ukraine
1 Svientsitskii Str., 79011 Lviv, Ukraine
Tel: +38(032)2761978; Fax: +38(032)2761158
E-mail: cmp@icmp.lviv.ua http://www.icmp.lviv.ua