УДК 004.912

Д. В. Ланде, В. Б. Андрущенко, І. В. Балагура
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Вікі-індекс популярності авторів наукових публікацій

Запропоновано новий індекс оцінки популярності авторів наукових статей, який розраховується на базі аналізу інтернет-енциклопедії Wikipedia (Wikipedia Index — WI), що містить мільйони статей і репрезентативно представляє практично всі галузі знань. На відміну від інших наукометричних індексів цей індекс дозволяє оцінити саме популярність автора, його вплив у найбільшому «ареалі знань» в Інтернеті — енциклопедії Wikipedia. Запропоновано алгоритми та технологію розрахунку Вікі-індексу шляхом зондування цієї мережевої енциклопедії, наведено приклади розрахунку Вікі-індексу для відомих учених, а також методику побудови інформаційних мереж — моделей предметних областей на базі автоматичного моніторингу та аналізу мережевих інформаційних ресурсів довідкового характеру. Розглянуто мережі понять, що відповідають заголовкам статей із Wikipedia.

Ключові слова: *Wikipedia, оцінка популярності авторів, Вікі-індекс, інформаційні мережі, предметні області.*

Вступ

На сьогодні в наукометрії загальноприйнятими є кілька індексів, відповідно до яких розраховується рівень учених, їхній вплив на науку та суспільство. Так, найбільш простим індексом є кількість публікацій автора. Зрозуміло, що такий індекс не містить якісних параметрів, які краще відображені в іншому індексі — кількості цитувань. Цей індекс, у свою чергу, не ілюструє загальної продуктивності автора, тому що автор всього однієї, але дуже вагомої роботи, може перевершувати за цим показником учених, які регулярно публікують свої результати.

У 2005 році фізиком Хорхе Гіршем з Каліфорнійського університету у Сан-Дієго, було запропоновано самий розповсюджений на сьогодні індекс — індекс Гірша. Принцип його обчислення достатньо простий, при цьому він об'єднує переваги першого та другого підходів. Індекс обчислюється на основі розподілу цитувань робіт даного дослідника. Згідно Гірша вчений має індекс h, якщо h з його N_p статей процитовано як мінімум h разів кожна, в той час, як статті, що залишилися $(N_p - h)$ цитуються не більше ніж h разів кожна. Цей індекс отримав широку

підтримку та використовується в таких наукометричних системах як Scopus, Web Of Science, Google Scholar Citations.

Разом з цим, даний показник, що є орієнтованим на наукову важливість і вагомість автора, не зовсім повно відображає загальну важливість результатів, що ним отримані. Для такої оцінки доцільно застосовувати науково популярні, спеціально значущі джерела, а також ресурси відкритого доступу. Як один з підходів до вирішення зазначеної проблеми авторами запропоновано методологію розрахунку нового індексу — Вікі-індексу популярності автора.

Даний індекс може виступати вагомим інструментом, що у поєднанні з іншими індексами може надавати повну картину впливовості наукових здобутків автора не тільки в науковому середовищі, але й загальний вплив на формування погляду та розуміння досліджень усіма користувачами інформації.

При цьому до розгляду береться мережевий сервіс Wikipedia — найбільшої і найбільш демократичної енциклопедії у мережі Інтернет, доступ до якої не передбачає передплати, крім того система доступна для завантаження в повному обсязі.

На сьогодні Wikipedia (https://www.wikipedia.org/) є найбільш відвідуваним сайтом у мережі Інтернет і одним із найбільш популярних енциклопедичних ресурсів, що охоплюють усі галузі знань, містить відповіді на більшість запитів пошукових систем. На цей час лише англійська версія Wikipedia містить понад 5 млн статей (німецька — понад 2 млн, китайська, російська — понад 1 млн, українська — 680 тис. статей).

Для первинного доступу до системи було застосовано спеціальні терміни — імена вчених і терміни з цільової проблематики, за якими представлені відповідні статті на ресурсі, які створюються та редагуються авторами-експертами. Поряд з індексами вчених за запропонованою методикою будуються мережі предметних областей, що їм відповідають. Цей аспект, на нашу думку, додає вагомості запропонованому підходу.

Дослідженню предметних областей, так само, як і сервісу Wikipedia, присвячено достатню кількість робіт, які підтверджують актуальність досліджень, що були проведені [4]. Серед них, зокрема, методи побудови мереж співавторів, визначення значущих вузлів, структури мережі, дослідження цитувань, а також відповідних корпусів [5].

Крім того, авторами було досліджено масив публікацій, які стосуються підходів до оцінки цитувань та інших аспектів оновлення, існування, наповнення, редагування даних енциклопедичного ресурсу Wikipedia.

Усі дослідження щодо побудови предметних областей та оцінок посилань на ті чи інші публікації або наявність посилань на статті у Wikipedia в наукових статтях (що розміщені в журналах, які індексуються певною наукометричною системою) є достатньо вузько спрямованими і стосуються тільки обмеженого кола наукових напрямків, а також визначення «Wikipedia ризиків» від помилок у наукових публікаціях.

Виходячи з результатів опрацьованої інформації, можна припустити унікальність запропонованих індексів і цінності інформації, яку буде отримано в результаті обчислень, що дозволить оцінити рівень тих чи інших даних у системі популяризації науки та доступності наукової інформації з певних питань.

Застосування індексів можливе в різних напрямках оцінки та аналізу наукової діяльності, воно може виступати додатковим інструментом для прийняття управлінських рішень, формування освітніх програм тощо. Також використання отриманих індексів сприятиме розвитку енциклопедичного ресурсу.

Правило обчислення Вікі-індексу популярності автора

Авторами запропоновано наступні правила обрахунку Вікі-індексу популярності автора. Припустимо, що бібліографічні посилання на автора зустрічаються у ***N*** статтях Wikipedia.

Відсортований за зменшенням ряд із показників, що визначають скільки разів повторюється ім'я автора в бібліографічній частині цих статей позначимо: $R_1,\ R_2,...,\ R_N$.

Вікі-індекс популярності автора (*WI*) відповідає максимальній кількості статей (*WH*) з Wikipedia, в яких кількість бібліографічних посилань на цього автора не менше значення *WH*, що є помножене на певну невід'ємну функцію, яка не зменшується (зокрема, нижче розглядається корінь квадратний) від *N*, тобто:

$$WI = WH \times \sqrt{N} = \max\left(i : R_i > i\right) \times \sqrt{N}\,.$$

Вікі-індекс популярності автора ідейно є близький до індексу Гірша, однак, він враховує не кількість статей, які посилаються на статті автора, а кількість бібліографічних посилань на роботи автора та кількість статей із Wikipedia, які містять дані посилання. Ще одна відмінність від індексу Гірша, множення на певну функцію від *N*, що відображає врахування саме популярності та забезпечує більший розкид значень індексу для різних авторів.

Необхідно відмітити, що індекс популярності автора має бути прив'язаний до його предметної області, з одного боку для того, щоб уникнути помилкового підрахунку для однофамільців, а з іншого — для забезпечення повноти обсягу по свій предметній області.

Приклад.

Припустимо, стаття з Wikipedia з максимальною кількістю бібліографічних посилань на автора Дж. Сміта (у заданій предметній області) містить 100 посилань. Другий за посиланнями — 20 документів, третій — 10, четвертий — 5, п'ятий — 5, ще 4 — по одному посиланню. Таким чином, маємо ряд значень:

$$R_1 = 100;\ R_2 = 20;\ R_3 = 10;\ R_4 = 5;\ R_5 = 5;\ R_6 = 1;\ R_7 = 1;\ R_8 = 1;\ R_9 = 1.$$

1 стаття містить кількість посилань не менше ніж $R_1 = 100$;
2 статті містять кількість посилань не менше ніж $R_2 = 20$;
3 статті містять кількість посилань не менше ніж $R_3 = 10$;
4 статті містять кількість посилань не менше ніж $R_4 = 4$;
5 статей містять кількість посилань не менше ніж $R_5 = 5$.
Не існує 6 статей, що містять кількість посилань не менше ніж 6.
У даному випадку: $N = 9,\ \ WH = 5.$
Таким чином, $WI = 5 \times \sqrt{9} = 15.$

Алгоритми.

При розрахунку Вікі-індексу популярності автора необхідно забезпечити сканування ресурсів сервісу Wikipedia, що відповідають предметній області, в якій працює автор. Відповідно, як «побічний продукт» обчислення Вікі-індексу популярності будується і модель предметної області, що представляє собою мережу, вузлами якої є поняття, яким відповідають статті з Wikipedia, а зв'язками — гіперпосилання між статтями.

Побудова моделі предметної області, в якій працює автор, можливе двома шляхами:

— використанням дампа бази даних Wikipedia (не зовсім актуальний, але доступний за посиланням), за допомогою якого можливий повний обсяг усіх понять і зв'язків. Перевага такого підходу — повнота інформації, недолік — можлива втрата точності через урахування однофамільців, вихід за межі предметної області, значний час розрахунку;

— використанням принципу зондування мережевого сервісу (під зондуванням інформаційних мереж необхідно розуміти вибірку невеличкого об'єму важливого змісту з великих інформаційних мереж, які через технологічні причини не підлягають повному скануванню). Перевага цього підходу — точність отримання інформації суворо в рамках однієї предметної області, вирішення проблеми однофамільців, швидкий час розрахунку. Основний недолік — можлива незначна повнота, яка може бути оцінена додатковими експериментами.

Авторами для розрахунку Вікі-індексу популярності автора та побудови відповідній йому моделі предметної області обраний другий підхід, який було реалізовано у вигляді сервісної програми.

Побудова моделі предметної області шляхом зондування Wikipedia

Для реалізації розрахунку Вікі-індексу популярності авторів розглядався наступний алгоритм побудови предметних областей за даними сервісу Wikipedia, який передбачає уникнення ефекту «зсуву тематик» (TopicDrift).

1. На веб-сторінці https://www.wikipedia.org/ в рядку пошуку задається вихідне слово — ім'я вченого, наприклад **«Albert Einstein»**.

2. Відкривається вікно пошуку, що містить інформацію про поняття, відповідно до заданого на кроці 1. Вихідне слово/словосполучення є вершиною графа, який буде побудований за результатами сканування.

3. До побудованого графа додаються всі терміни-поняття, що відповідають гіперпосиланням на обраній сторінці. Всі обрані слова/словосполучення — гіперпосилання — є вузлами графа. Формуються ребра-зв'язки до цих вузлів від вихідного вузла.

4. Наступний перехід здійснюється за першим ще не задіяним гіперпосиланням, які були вибрані на сторінках, що досліджувалися.

5. На сторінці, на яку було виконано перехід за посиланням, у тексті здійснюється пошук скороченого імені вченого (наприклад, **Einstein**) або перевірочних слів (наприклад, **physics**, **relativity**).

6. У випадку, якщо в тексті присутнє скорочене ім'я вченого або перевірочні слова, відбувається перехід до кроку 4 алгоритму та відповідно від вузла — слова/словосполучення поточного пошуку — будуються нові вузли.

7. Якщо слово/словосполучення в тексті відсутнє — дана гілка графа вважається побудованою.

8. Якщо при переході до наступного слова/словосполучення відбувається перехід на сторінку, що була вже просканованою — слово не додається як вузол графа, а формується зворотній зв'язок із вузлом, який було створено.

9. Дії за пунктами 4–9 повторюються до тих пір, доки залишаються ще не задіяні гіперпосилання, які були вибрані на сторінках, що досліджувалися. В іншому випадку граф вважається побудованим.

Відповідно до наведеного алгоритму процес збору інформації у Wikipedia, починаючи з певного вузла-поняття, припиняється, коли згідно з алгоритмом уже неможливий перехід до нового вузла (базових вузлів для переходу вже не лишається). Таким чином, «зациклювання» є неможливим.

Розрахунок Вікі-індексу популярності автора

Для розрахунку Вікі-індексу популярності до наведеного вище алгоритму вносяться незначні зміни, а саме: на сторінці, перехід на яку було здійснено за гіперпосиланням (крок 5 алгоритму), відбувається пошук згадувань автора в розділах **Publications**, **References**, **Further Reading**. При цьому підраховується кількість цих згадувань, яке відповідає значенням R_i. Якщо $R_i = 0$, то стаття не вважається значущою, поняття визначається як кінцевий вузол, і відбувається перехід до кроку 4. Звісно, дане правило звужує перелік сторінок із Wikipedia, що сканується; це призводить до втрати повноти, однак, як доводять реальні розрахунки, має незначний вплив на загальні результати. Сторінки, що присвячені науковим поняттям та ті, що не містять релевантні публікації, можна не враховувати — пропускати. Разом з тим, час зондування цільового сегмента Wikipedia суттєво зменшується.

У результаті повного зондування мережі формується послідовність $R_1, R_2, ..., R_N$, яка й використовується для підрахунку Вікі-індексу популярності автора, за правилами наведеними вище.

Розрахунок.

Наведені алгоритми були реалізовані у вигляді програмного комплексу, за допомогою якого будуються моделі предметних областей та Вікі-індекси популярності авторів. Наведемо приклади розрахунку Вікі-індексів трьох авторів: Альберта Ейнштейна, Енріко Фермі, Бенуа Мандельброта.

На рис. 1 наведено фрагменти трас виконання програми зондування Wikipedia, на яких відображаються поняття, до яких відбувається перехід від вихідних понять до понять, які містять ім'я автора або перевірочне слово.

На рис. 2–4 наведено візуалізацію за допомогою програми Gephi фрагментів моделей предметних областей, що були отримані шляхом зондування Wikipedia за наведеним вище алгоритмом. Параметри отриманих мереж (моделей предметної області), вузлами яких виступають поняття з Wikipedia, наступні.

Albert_Einstein

1: Albert_Einstein

SCI Links (1): 174
0 Rd +: Ulm
1 Rd -: German_Empire
2 Rd -: Statelessness
3 Rd -: Switzerland
4 Rd -: Kingdom_of_Prussia
5 Rd -: Free_State_of_Prussia
6 Rd -: Weimar_Republic
7 Rd -: Physics
8 Rd -: Philosophy
9 Rd -: Swiss_Patent_Office
10 Rd -: Bern
11 Rd -: University_of_Bern
12 Rd -: University_of_Zurich
13 Rd +: ETH_Zurich
14 Rd -: Kaiser_Wilhelm_Institute
15 Rd -: German_Physical_Societ
16 Rd +: Leiden_University

Enrico_Fermi

1: Enrico_Fermi

SCI Links (1): 28
0 Rd -: Rome
1 Rd -: Chicago
2 Rd -: Physics
3 Rd +: Leiden_University
4 Rd -: University_of_Florence
5 Rd -: Columbia_University
6 Rd +: University_of_Chicago
7 Rd -: Alma_mater
8 Rd -: Doctoral_advisor
9 Rd +: Luigi_Puccianti
10 Rd +: Max_Born
11 Rd -: Paul_Ehrenfest
12 Rd +: Harold_Agnew
13 Rd -: Edoardo_Amaldi
14 Rd +: Owen_Chamberlain
15 Rd +: Geoffrey_Chew
16 Rd +: Jerome_Isaac_Friedman

Benoit_Mandelbrot

1: Benoit_Mandelbrot

SCI Links (1): 47
0 Rd +: Mandelbrot_set
1 Rd -: Warsaw
2 Rd -: Second_Polish_Republic
3 Rd -: Mathematics
4 Rd -: Aerodynamics
5 Rd -: Yale_University
6 Rd -: IBM
7 Rd -: Alma_mater
8 Rd -: University_of_Paris
9 Rd -: Eugene_Fama
10 Rd +: Ken_Musgrave
11 Rd -: Murad_Taqqu
12 Rd +: Mandelbrot_set
13 Rd +: Chaos_theory
14 Rd +: Fractal
15 Rd -: Johannes_Kepler
16 Rd +: Szolem_Mandelbrojt

Рис. 1. Фрагменти трас програми зондування Wikipedia

Для мережі, що відповідає моделі предметної області авторів розраховані наступні параметри.

Альберт Ейнштейн:

1) вузлів — 718;
2) ребер — 22111;
3) середній ступінь вузла — 62;
4) діаметр графа — 4;
5) середній коефіцієнт кластеризації — 0,26;
6) найбільші вузли:

Поняття	Ступінь вузла
Quantum_nonlocality	188
Alain_Aspect	181
Hermann_Weyl	177
Paul_Dirac	174
Electromagnetic_radiation	174
Isaac_Newton	169
Galileo_Galilei	169
Wolfgang_Pauli	169
General_relativity	167
Antimatter	167

$WI = 12 \times 11{,}5 = 138$ (128 статей з бібліографією, $WH = 12$).

Енріко Фермі:
1) вузлів — 605;
2) ребер — 22079;
3) середній ступінь вузла — 73;
4) діаметр графа — 4;
5) середній коефіцієнт кластеризації — 0,47;
6) найбільші вузли:

Поняття	Ступінь вузла
Enrico_Fermi	440
Nobelium	206
Transuranic_element	206
Particle_physics	204
Mendelevium	204
Einsteinium	204
Berkelium	203
Radioactive_decay	195
Radioactive	190
Particle_accelerators	188

$WI = 7 \times 9{,}6 = 67$ (92 статті з бібліографією, $WH = 7$).

Бенуа Мандельброт:
1) вузлів — 34;
2) ребер — 259;
3) середній ступінь вузла — 15,2;
4) діаметр графа — 3,
5) середній коефіцієнт кластеризації — 0,52;
6) найбільші вузли:

Поняття	Ступінь вузла
Benoit_Mandelbrot	22
Pattern	20
Chaos_theory	18
Patterns_in_nature	18
Hausdorff_dimension	17
Patterns	16
Fractal	15
Fractal_dimension	15
Fractal_geometry	15
Fractals	15

$WI = 6 \times \sqrt{11} \approx 20$, (11 статей з бібліографією, $WH = 6$).

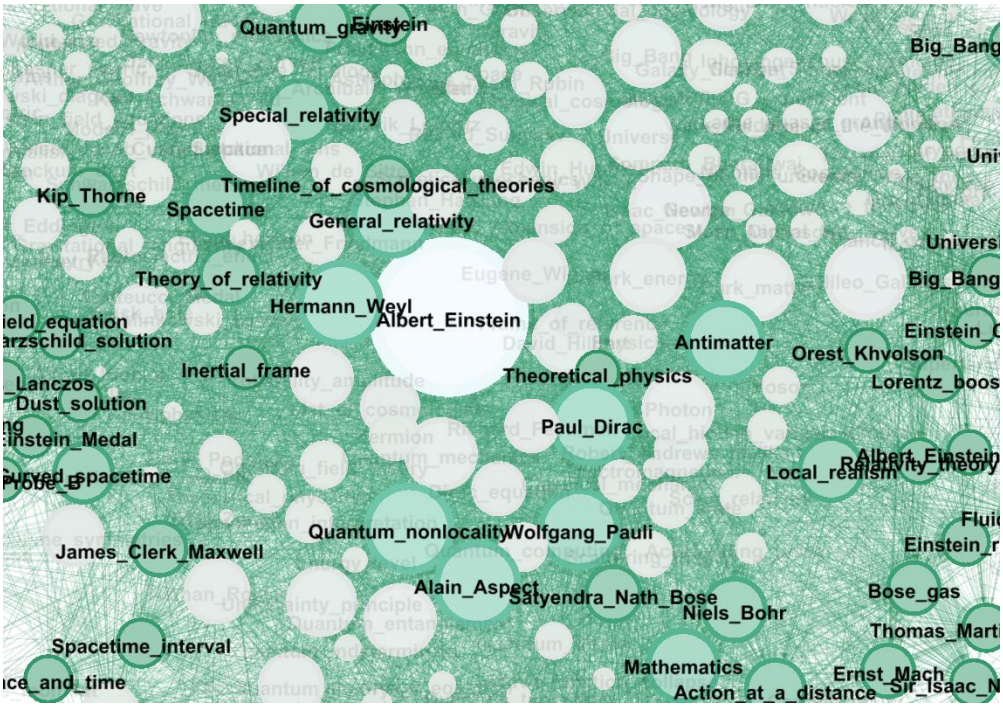

Рис. 2. Фрагменти моделей предметної області: Альберт Ейнштейн

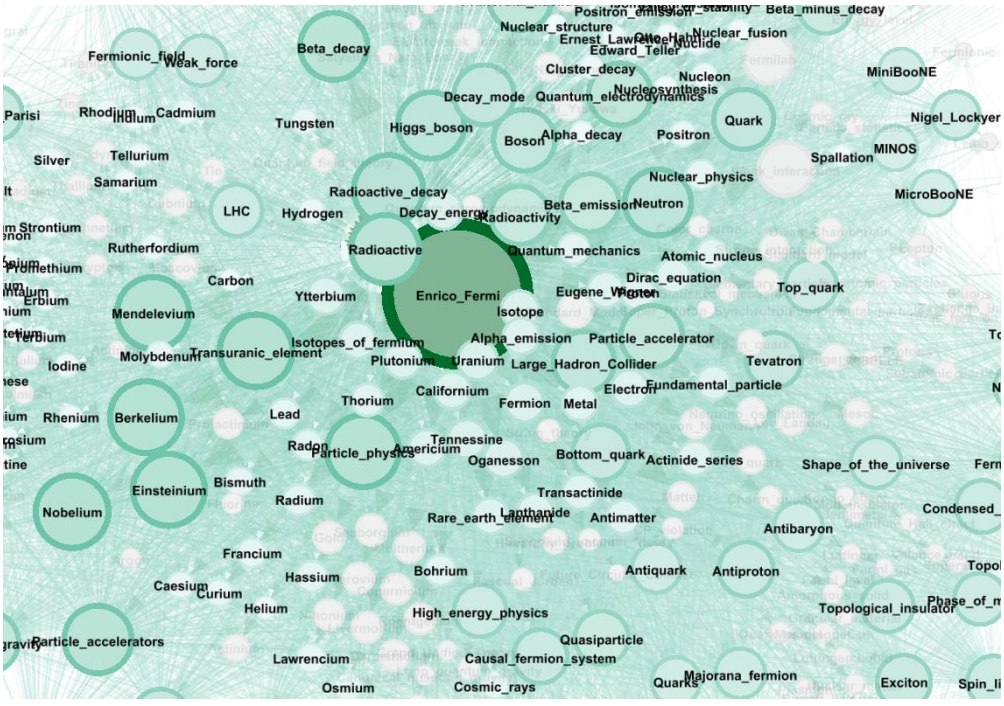

Рис. 3. Фрагменти моделей предметної області: Енріко Фермі

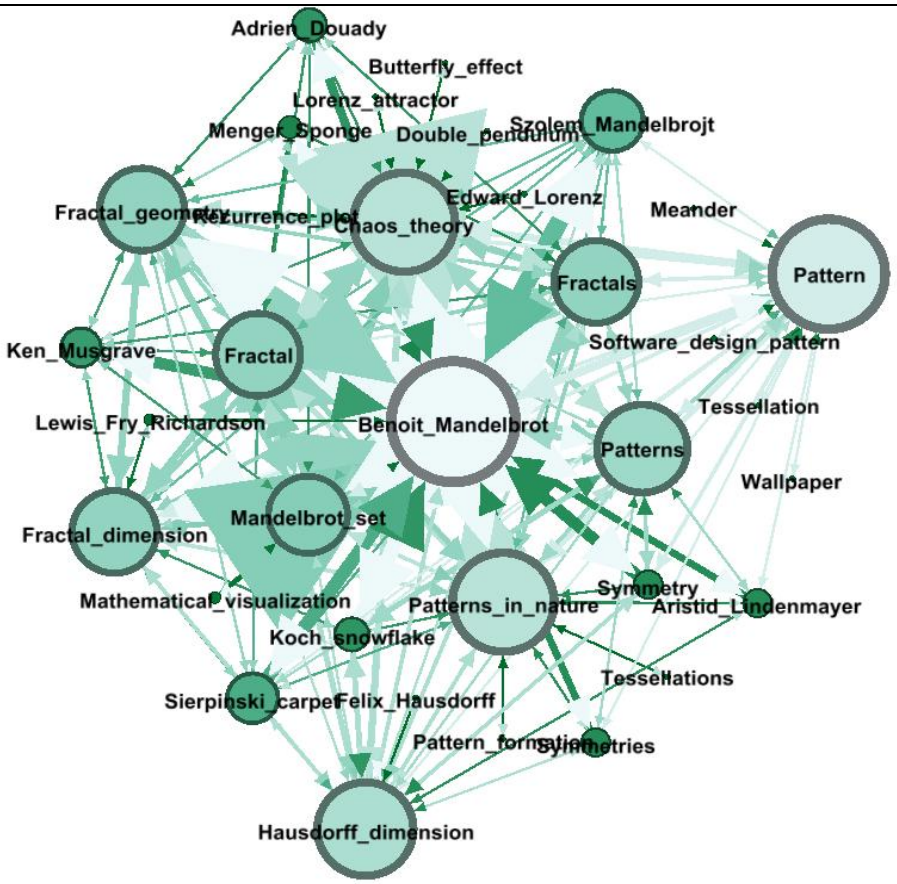

Рис. 4. Фрагменти моделей предметної області: Бенуа Мандельброт

Було проведено порівняння отриманих результатів — Вікі-індекс, розрахований за результатами досліджень із показниками індексу Гірша, що надають провідні світові наукометричні ресурси Scopus, Web Of Science та Google Scholar Citations. Результати відображено в таблиці.

Порівняння показників Вікі-індексу із показниками індексу Гірша
за даними Scopus, Web Of Science та Google Scholar Citations

№ за/п	Учений	Вікі-індекс	*h*-індекс за даними Scopus	*h*-індекс за даними Web Of Science	*h*-індекс за даними Google Scholar Citations
1.	Альберт Ейнштейн	141	36	6	110
2.	Енріко Фермі	67	26	1	49*
3.	Бенуа Мандельброт	20	31	36	90

*Профіль відсутній, розраховувалося за запитом: «e.fermi» до сервісу Google Scholar (Google Scholar Calculator)

За рахунок порівняння можна побачити і оцінити участь інформації щодо досліджень і публікацій у ресурсі загального доступу порівняно з даними ресурсів, що враховують суто наукову інформацію, з певним набором обмежень.

Висновки

У результаті проведених розрахунків і випробовувань запропонованих підходів до формування індексу популярності автора з огляду на присутність посилань на його роботи та згадування у найбільшому в глобальній мережі енциклопедичному ресурсі Wikipedia, можна зробити наступні висновки.

1. Принцип побудови Вікі-індексу відрізняється від тих, що наразі застосовуються в наукометрії, перш за все врахуванням цитувань не з наукових статей, а з популярних сторінок сервісу Wikipedia. Таким чином, у принципі, можна отримати індекс популярності автора в межах даного сервісу. І це суттєво, враховуючи той факт, що Wikipedia є на сьогодні найбільшим і найбільш популярним енциклопедичним ресурсом.

2. У роботі запропоновано технологію «швидкого» розрахунку Вікі-індексу автора, що дозволяє реалізувати розрахунок у вигляді окремого сервісу, а також автоматично формувати модель предметної області.

3. За рахунок застосування та популяризації запропонованих індексів можливе значне розширення ресурсів відкритого доступу (доступних для редагування користувачами мережі Інтернет).

4. Проведена робота може мати продовження шляхом аналізу інших ресурсів і формування показників для оцінки та аналізу впливовості в тому чи іншому оточенні.

Можна відмітити, що система Wikipedia, як і система Google Scholar Citations, що розглядалася раніше [6, 7], є зручною з точки зору доступу до інформації, не передбачає створення профілю користувача для доступу до інформації, доступ — необмежений.

Також необхідно відмітити принципову різницю запропонованої моделі автоматичного формування моделей предметних областей від тих, що вже існують, які базуються на безпосередній участі експертів при виборі конкретних вузлів і зв'язків. У прикладах, які наведено в роботі, дослідник для побудови відповідної мережі використовує тільки невеличку частку знань, що представлена у вигляді ім'я вченого, його скороченого написання, назви кількох ключових термінів-понять. Після цього програма використовує знання, що закладені авторами (редакторами) статей у Wikipedia, теги, що визначаються внутрішніми гіперпосиланнями. В такому випадку експертна середа значно розширюється.

1. Добров Б.В., Соловьев В.Д., Лукашевич Н.В., Иванов В.В. Онтологии и тезаурусы. Модели, инструменты, приложения. Москва: Бином, 2009. 173 с.

2. Ландэ Д.В., Снарский А.А. Подход к созданию терминологических онтологій. *Онтология проектирования*. 2014. № 2(12). С. 83–91.

3. Чанышев О.Г. Автоматическое построение терминологической базы знаний: сб. трудов 10-й Всероссийской научной конференции «Электронные библиотеки: перспективне методы и технологии, электронные коллекции» — RCDL'2008» (2008, г. Дубна, Россия). Дубна, 2008. С. 85–92.

4. Zareen Saba Syed, Tim Finin, Anupam Joshi. Wikipedia as an Ontology for Describing Documents. Proc. 2nd Int. Conf. on Weblogs and Social Media, AAAI Press, March 2008. P. 136–144.

5. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: навигация в сложных сетях: модели и алгоритмы. Москва: Либроком (Editorial URSS), 2009. 264 с.

6. Ланде Д.В., Андрущенко В.Б. Побудова мереж співавторства фахівців з юриспруденції за даними сервісу Google Scholar Citations. *Правова інформатика*. 2016. № 1(46). С. 146–150.

7. Ландэ Д.В. Построение модели предметной области путем зондирования сервиса Google Scholar Citations. *Онтология проектирования*. 2015. Т. 5. № 3 (17). С. 328–335.

Надійшла до редакції 20.12.2016