УДК 004.421

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

Побудова мережі предметних областей на базі ресурсу arXiv

Запропоновано новий спосіб опрацювання інформації системи бібліотеки Корнуельского університету — ресурсу препринтів arXiv. Авторами розроблено та реалізовано алгоритм пошуку публікацій за заданим концептом з урахуванням наукового напрямку, до якого відноситься публікація. Основну увагу зосереджено на розподілі публікацій за визначеними науковими напрямками та відповідними підгрупами, що передбачені ресурсом. Основні методи, які було використано для реалізації роботи, — це робота з текстовими масивами та подальша обробка отриманих результатів, параметри оцінки роботи пошуку та результатів пошуку. Сформульовано визначення мережі предметних областей. Для кожної предметної області складено словник як довідковий інструмент для реалізації поставленої задачі. Також відображено основні етапи побудови мережі предметних областей на базі ресурсу препринтів arXiv. Результатом роботи стало візуальне представлення мережі предметних областей для концепту — «cavitation» та тлумачення отриманих результатів. Дані, що представлені у матеріалі, опрацьовано в лютому-березні 2018 року.

Ключові слова: *концепт, предметна область, мережа предметних областей, архів препринтів, реферативна інформація.*

Вступ

На сьогодні одним із актуальних джерел отримання інформації є глобальні мережі, що містять широкий спектр ресурсів наукової інформації. Залежно від свого наповнення, ресурси можуть містити як повнотекстові документи, так і посилання на них (доступ до яких є безкоштовний чи передплачений з огляду на політику видання чи ресурсу), реферати статей, ресурси, що представлять собою глобальні інформаційні масиви, які задовольнять запити пошуку інформації.

Кожного дня кількість зазначених ресурсів зростає, з'являються нові ресурси для пошуку та зберігання наукової інформації як локальні, що представляють науковий доробок окремої організації, так і глобальні, які передбачають розміщен-

ня інформації з різних джерел різних авторів. Актуальним на сьогодні залишається питання систематизації цих ресурсів. Наразі можна впевнено говорити про наступні різновиди ресурсів наукової інформації [1]:

— репозитарії наукових текстів;

— наукометричні ресурси;

— енциклопедичні ресурси;

— соціальні мережі науковців, які містять повнотекстові документи як опублікованих, так і підготовлених до друку матеріалів.

Не кожен ресурс має набір функціоналу для отримання інформації за конкретним запитом, або може надавати інформацію за обмеженими варіантами запитів і фільтрації інформації за уточненням.

Необхідно зауважити, що на сьогодні однією із актуальних задач є отримання нової наукової інформації, її подальша обробка та інтерпретація. Одним із способів реалізації цієї задачі є робота з різноманітними існуючими системами доступу до наукових даних. Отримані дані та їхня інтерпретація можуть виступати допоміжним інструментом для виокремлення міждисциплінарної складової досліджень, що в подальшому може визначати необхідність залучення до наукових колективів представників різних наукових напрямків і наукових шкіл. Також отримані дані можуть бути використані для пошуку партнерів, що в подальшому сприятиме утворенню колаборацій для спільного вирішення наукових задач у рамках міжнародних проектів.

Архів препринтів arXiv

Для реалізації поставленої задачі було обрано репозитарій наукових текстів — архів препринтів arXiv, який є найбільшим архівом електронних публікацій і їхніх препринтів відкритого доступу.

Архів було створено в 1991 році. Спочатку на ресурсі було розміщено публікації, які було підготовлено до друку з напрямку «Фізика», але на сьогодні ресурс розширюється за рахунок додавання нових розділів і відповідних підрозділів з інших наукових напрямків.

Кожна публікація проходить процедуру схвалення перед публікацією і обов’язково буде переглянута відповідним підрозділом, який відповідає за публікацію статей на ресурсі.

arXiv — щоденний інформаційний інструмент для сотень і тисяч науковців. Серед користувачів більше 50 Нобелівських лауреатів, лауреати престижних наукових премій. Ресурс є актуальним інструментом для користувачів з країн з обмеженим доступом до наукової інформації.

На сьогодні архів містить 8 розділів, за якими можна розмістити свої матеріали:

1) Computer Science (42 напрямки);
2) Economics (1 напрямок);
3) Electrical Engineering and System Science (3 напрямки);
4) Mathematics (32 напрямки);
5) Physics (13 підрозділів);
6) Quantitative Biology (10 напрямків);
7) Quantitative Finance (9 напрямків);

8) Statistics (6 напрямків).

Аналіз останніх досліджень і публікацій

Для складання картини існуючих робіт і відповідно досліджень щодо питання обробки інформації ресурсу препринтів arXiv було оброблено масив публікацій за останні 10 років, що відображає цікавість до архіву та динаміку розвитку за останні 10 років (з 27) його існування. Серед напрямків робіт, які присвячені роботі з архівом і розробки відповідних надбудов можна виокремити наступні:

— запровадження протоколів відкритого доступу в рамках ресурсу arXiv, відповідність існуючої структури архіву протоколу «OAI protocol v1.0» [2]. Зокрема у цій роботі ґрунтовно зображено динаміку розвитку ресурсу та структуру відповідно до поставленої задачі;

— виявлення плагіату в публікаціях, які розміщені на ресурсі arXiv [3]. В роботі зображено широкий спектр методів, що можуть бути застосовані для виявлення запозичень в архіві. Методи було апробовано на 284,834 публікаціях за останні 14 років на момент проведення дослідження, а також відображено можливість перевірки під час розміщення публікації на ресурсі в реальному часі;

— аналіз використання даних ресурсу [4] — аналіз, обґрунтування та пропозиції щодо кількості звернень до архіву, завантажень відповідно до розділів за науковими напрямками на піднапрямками передбаченими на ресурсі;

— аналіз репрезентативності публікацій з астрофізики на архіві препринтів arXiv, ресурсі Mendeley та реферативній базі даних Scopus [5].

Аналіз публікацій дозволив зробити висновок, що на сьогодні розробок щодо аналізу безпосередньо кореляції наукових напрямків і публікацій не проводилося.

Постановка завдання

Основною задачею запропонованого дослідження було визначити класифікацію наукових напрямків ресурсу — за якими підгрупами науковець може розмістити свої матеріали на ресурсі відповідно до заданих концептів.

Також розробити способи оцінки інформації, яка надається ресурсом за результатами пошуку відповідно до приналежності заданого для пошуку концепту з науковим напрямком і відповідними піднапрямками, що передбачені ресурсом arXiv.

За рахунок обробки інформації за результатом пошуку — побудувати мережу предметних областей для заданого концепту. Також важливим фактором є оцінка результатів пошуку.

Поняття концепту та мережі предметних областей

Для запровадження пошуку щодо отримання масиву інформації для подальшої обробки необхідним є задання концепту для пошуку.

Під концептом мається на увазі змістовна словесна одиниця, або сполучення таких одиниць, що визначає рамки наукового сприйняття сенсу конкретного поняття, яке притаманне одній чи кільком предметним областям.

Мережа предметних областей — це спосіб представлення моделі предметних областей за рахунок визначення узагальнених описів предметної області, представлених власне їхньою назвою та назвами підпорядкованих їй структурних одиниць — наукових напрямків, які більш конкретно описують предметну область, що визначені інформаційною системою, на базі якої будується задана мережа, або на базі запропонованої систематизації предметних областей.

Алгоритм побудови мережі предметних областей для заданого концепту

Для заданого концепту, що може бути представлений одним словом чи словосполученням, відбувається пошук масиву публікацій, в яких він відображений.

Алгоритм побудови мережі предметних областей для заданого концепту передбачає визначення предметних галузей та наукових напрямків, для яких заданий концепт є притаманний. Реалізація алгоритму здійснюється за рахунок обробки інформації, що є результатом пошуку.

Для реалізації поставленої задачі було складено 8 словників, що відповідають науковим напрямкам — предметним областям, які представлені на ресурсі arXiv. Словники та їхній зміст містяться у Додатку.

Уся робота здійснювалася на базі реферативної інформації — результатів пошуку для заданого концепту.

Алгоритм побудови мережі, що охоплює предметні галузі, до яких застосовується задане поняття, передбачає виконання наступних кроків [6].

1. Вершиною мережі є вузол, який є тотожнім із концептом, що було задано для пошуку.

2. Для заданого концепту було отримано перелік реферативної інформації, що містить наступні дані (рис. 1), і буде підлягати подальшій обробці:

1) номер за порядком, ідентифікатор у системі, що має наступний вигляд:

arXiv: XXXX.XXXXX [***], де XXXX.XXXXX — номер публікації у системі; *** — перелік доступних форматів файлів для завантаження;

2) назва публікації;

3) автор(и);

4) **Comments** — поле, яке, як правило, містить інформацію про кількість, сторінок публікації, кількість рисунків та інших елементів (для деяких публікацій дане поле може бути відсутнє);

5) **Journal-ref** — поле, що містить інформацію про видання, в якому розміщена дана публікація (є в наявності для публікацій, що вже є опубліковані);

6) **Subject** — поле, що містить назву предметної області або конкретизованої інформації щодо наукового напрямку в рамках предметної області (відповідно до того, в який спосіб автор публікації зазначив при поданні публікації для розміщення її на ресурсі).

3. Для кожної публікації виокремлюється назва наукового напрямку, що зазначена в отриманій реферативній інформації (рис. 2).

4. Назва наукового напрямку, зазначена в реферативній інформації для публікації, порівнюється з усіма словниками предметних областей. Назва предметної області, в словнику якої було знайдено назву наукового напрямку — є наступний вузол графа, що з'єднаний з вершиною, яка відповідає заданому концепту.

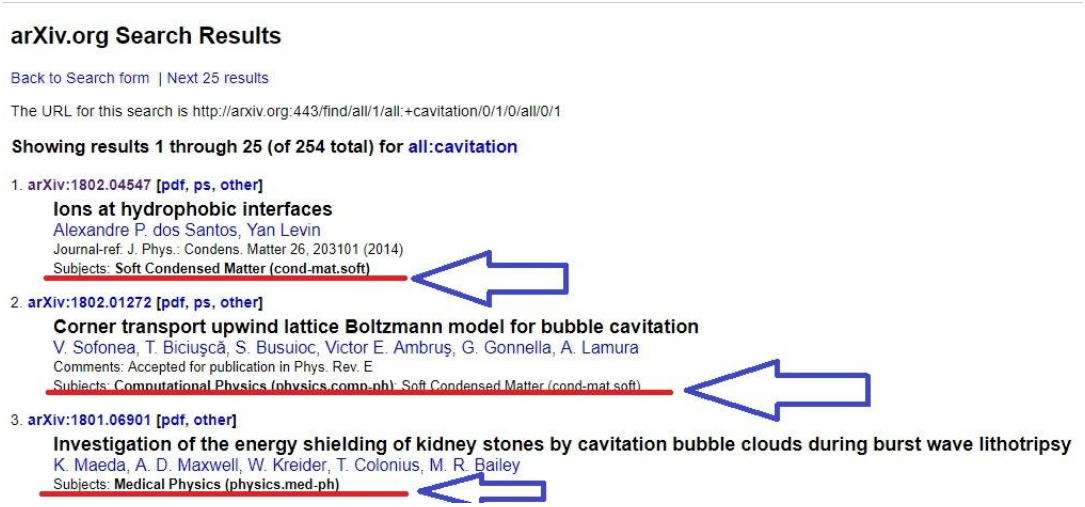

Рис. 1. Перелік результатів за пошуком із позначенням рядка, що містить інформацію про приналежність до наукового напрямку

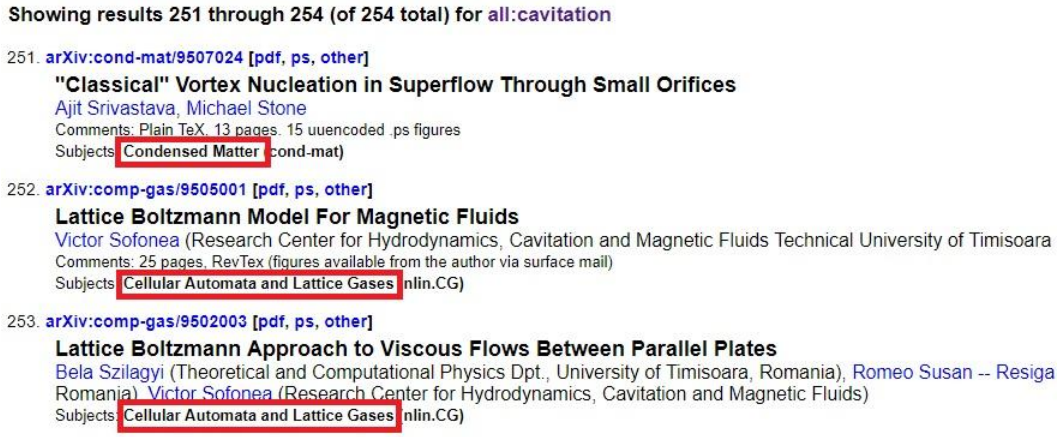

Рис. 2. Виокремленні назви наукових напрямків, які будуть брати участь у подальшій роботі

5. Наступний вузол — назва наукового напрямку, що було виокремлено в результаті роботи з реферативною інформацією. Цей вузол з'єднаний з вузлом, що позначає предметну область, яку деталізує отриманий науковий напрямок (рис. 3).

6. Відбувається перехід до наступного результату пошуку.

7. Якщо для наукового напрямку вже було побудовано вузол — назву предметної області, то будується тільки вузол — назва наукового напрямку, що з'єднується із вузлом — відповідною предметною областю.

8. Якщо для назви наукового напрямку вже було побудовано відповідний вузол, то відбувається перехід до п. 6. Якщо для назви наукового напрямку побудову відповідних вузлів ще не було здійснено, відбувається перехід до п. 4.

9. Якщо назву наукового напрямку, зазначену в полі «Subject», за результатами пошуку було знайдено в двох словниках, то будується два вузли — назви

предметних областей, або той вузол, який ще не було побудовано, і відповідно з'єднаний з ними обома вузол — назва наукового напрямку.

10. Мережа вважається побудованою по завершенні сканування всіх результатів пошуку.

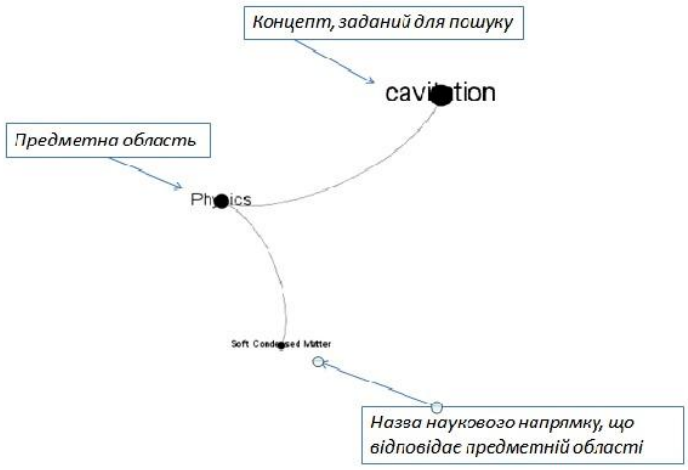

Рис. 3. Початковий етап побудови мережі предметних областей

Для візуалізації отриманих результатів отримані дані імпортуються до середовища Gephi.

Оцінка отриманих результатів за визначеними параметрами

З огляду на поставлену задачу постає необхідність здійснення оцінки результатів, що отримані в рамках пошуку за заданим концептом, та обрахунку його параметрів наступним чином.

1. Вага частоти набору документів (collection frequency weight — CFW) [7], для обрахунку якого до уваги беруться такі показники: n — кількість документів, в яких було знайдено заданий концепт; N — загальна кількість документів, за якими відбувався пошук. Відповідно: $\mathrm{CFW} = \log N - \log n$.

2. Також для заданого поняття обраховується науковий напрямок, який найбільшу кількість разів зустрічається в результатах пошуку і є притаманним для заданого пошуку концепту.

Показник обраховується за наступною формулою:

$$TF = \frac{k_i}{n},$$

де k_i — це кількість разів, коли назва наукового напрямку зустрічається в переліку результатів пошуку, а n — загальна кількість предметних областей, які було виокремлено для побудови мережі для даного поняття.

Для оцінки здійсненого текстового пошуку було обраховано повноту представлення результатів, яка може бути порівняна з експертною думкою для визначення коректності здійснення пошуку [8]. Повнота (recall, r) характеризує здат-

ність системи знаходити необхідні користувачеві результати, але не враховує кількість нерелевантних документів, які йому видаються, і обчислюється як співвідношення знайдених релевантних документів до загальної кількості релевантних документів:

$$r = \frac{a}{a+c},$$

де a — релевантні документи, які знайдені системою; c — релевантні документи, що не знайдені системою.

Реалізація роботи алгоритму та оцінка отриманих результатів

Алгоритм було апробовано для концепту «cavitation». У ході виконання роботи було отримано наступні результати.

1. За результатами пошуку для заданого концепту було системою знайдено 254 публікації.

2. За рахунок реалізації алгоритму було виокремлено 5 наукових напрямків, за якими було розміщено публікації на ресурсі, а саме:

1) Physics — 25 підгруп за напрямком;

2) Computer Science — 5 підгруп за напрямком;

3) Mathematics — 4 підгрупи за напрямком;

4) Quantitative Biology — 2 підгрупи;

6) Statistics — тільки 1 підгрупа.

Отже, за отриманими результатами можна впевнено сказати, що задане поняття є притаманним для 5 предметних областей і серед публікацій, що розміщені на ресурсі, найбільше — з напрямку «фізика», зокрема «фізика рідин» (Fluid Physics) — 82 публікації.

За результатами дослідження було побудовано мережу предметних областей для заданого концепту і в середовищі Gephi виконано візуалізацію. На рис. 4 показано візуалізацію отриманих результатів.

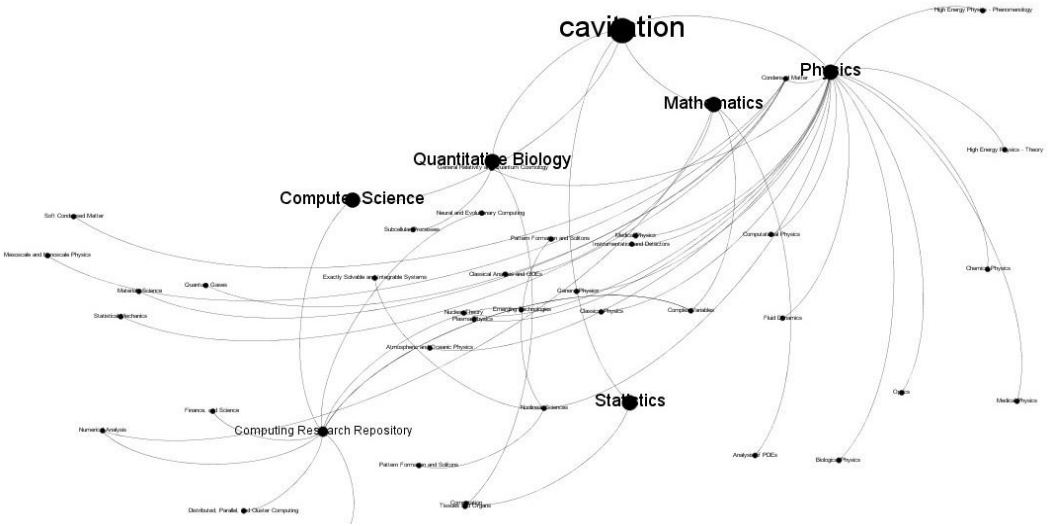

Рис. 4. Мережа предметних областей для поняття «cavitation»

Вага частоти набору документів (collection frequency weight — CFW) для заданого концепту буде дорівнювати:

$$\text{CFW} = \log N - \log n = 3{,}734,$$

де загальна кількість документів ресурсу становить: N — загальна кількість публікацій на ресурсі, цей показник становить 1,377,332 (дані на 23.04.2018 р.); n — кількість публікацій, що містять заданий для пошуку концепт, і дорівнює 254.

Для даної предметної області показник Term frequency буде обрахований наступним чином:

$$TF = \frac{25}{5} = 5.$$

Повнота (для кожного виокремленного наукового напрямку, визначеного системою) виглядатиме наступним чином:

Physics — $r = \frac{25}{25+229} = 0{,}09$;

Computer Science — $r = \frac{5}{254} = 0{,}01$;

Mathematics — $r = \frac{4}{254} = 0{,}02$;

Quantitative Biology — $r = \frac{2}{254} = 0{,}008$;

Statistics — $r = \frac{1}{254} = 0{,}004$.

Відповідно до отриманих результатів можна стверджувати, що робота системи, яка виконує пошук результатів, надає відносно повну картину щодо представлення публікацій у відповідних наукових напрямках, визначених ресурсом.

Висновки

Результати дослідження дозволяють розширити рамки сприйняття понять і наукових термінів. За рахунок сканування найбільшого в глобальній мережі ресурсу препринтів, що містить великий обсяг публікацій як підготовлених до друку, так і розміщених у провідних наукових виданнях, визначається коло наукових напрямків, які пов’язані із заданим концептом.

Застосування додатків розроблених за запропонованим алгоритмом дозволить використовувати мережу предметних областей як додатковий інструмент для пошуку співавторів, розширення застосування поняття в рамках різних наукових напрямків і таким чином отримати можливість для розширення колаборацій і залучення фахівців з різних наукових напрямків.

За результатами роботи може бути побудована модель «Концепт – система наукових напрямків», що може бути застосована та апробована на інших ресурсах наукової інформації. Також робота може бути розвинена за рахунок використання більшого набору даних, що пропонує система за результатами пошуку за заданим концептом, а саме: автори публікації, ключові слова, реферати. Може бути прове-

дений аналіз текстового масиву для визначення найбільш уживаних слів і проведений відповідний розрахунок, аналіз і обґрунтування отриманих результатів.

Запропоновано та реалізовано оцінку якості роботи пошукової системи за рахунок використання параметру повноти.

Для розвитку запропонованих підходів для пошуку, обробки та інтерпретації наукової інформації за рахунок реалізації зазначених алгоритмів, можлива побудова більш розвиненої мережі за рахунок групування назв наукових напрямків у межах словників, а також обрахунку параметрів мережі.

Додаток

Словники предметних областей для ресурсу препринтів arXiv, які містять переліки наукових напрямків, що конкретизують відповідну предметну область

1. Computer Science

Computing Research Repository
Artificial Intelligence
Computation and Language
Computational Complexity
Computational Engineering Finance, and Science
Computational Geometry
Computer Science and Game Theory
Computer Vision and Pattern Recognition
Computers and Society
Cryptography and Security
Data Structures and Algorithms
Databases
Digital Libraries
Discrete Mathematics
Distributed, Parallel, and Cluster Computing
Emerging Technologies
Formal Languages and Automata Theory
General Literature
Graphics
Hardware Architecture
Human-Computer Interaction
Information Retrieval
Information Theory
Learning
Logic in Computer Science
Mathematical Software
Multiagent Systems
Multimedia
Networking and Internet Architecture
Neural and Evolutionary Computing
Numerical Analysis
Operating Systems
Other Computer Science
Performance
Programming Languages
Robotics
Social and Information Networks
Software Engineering
Sound
Symbolic Computation
Systems and Control

2. Economics

Econometrics

3. Electrical Engineering and System Science

Audio and Speech Processing
Image and Video Processing
Signal Processing

4. Mathematics

Algebraic Geometry
Algebraic Topology
Analysis of PDEs
Category Theory
Classical Analysis and ODEs

Combinatorics
Commutative Algebra
Complex Variables
Differential Geometry
Dynamical Systems
Functional Analysis
General Mathematics
General Topology
Geometric Topology
Group Theory
History and Overview
Information Theory
K-Theory and Homology
Logic
Mathematical Physics
Metric Geometry
Number Theory
Numerical Analysis
Operator Algebras
Optimization and Control
Probability
Quantum Algebra
Representation Theory
Rings and Algebras
Spectral Theory
Statistics Theory
Symplectic Geometry

5. Physics

Astrophysics
Astrophysics of Galaxies
Cosmology and Nongalactic Astrophysics
Earth and Planetary Astrophysics
High Energy Astrophysical Phenomena Instrumentation and Methods for Astrophysics
Solar and Stellar Astrophysics
Condensed Matter
Disordered Systems and Neural Networks
Materials Science
Mesoscale and Nanoscale Physics
Other Condensed Matter
Quantum Gases
Soft Condensed Matter
Statistical Mechanics
Strongly Correlated Electrons
Superconductivity
General Relativity and Quantum Cosmology
High Energy Physics — Experiment
High Energy Physics — Lattice
High Energy Physics — Phenomenology
High Energy Physics — Theory
Mathematical Physics
Nonlinear Sciences
Adaptation and Self-Organizing Systems
Cellular Automata and Lattice Gases
Chaotic Dynamics
Exactly Solvable and Integrable Systems
Pattern Formation and Solitons
Nuclear Experiment
Nuclear Theory
Physics
Accelerator Physics
Applied Physics
Atmospheric and Oceanic Physics
Atomic Physics
Atomic and Molecular Clusters
Biological Physics
Chemical Physics
Data Analysis
Statistics and Probability
Fluid Dynamics
General Physics
Geophysics
History and Philosophy of Physics
Instrumentation and Detectors
Medical Physics
Optics
Physics Education
Physics and Society
Plasma Physics

Popular Physics
Space Physics
Quantum Physics

6. Quantitative biology

Biomolecules
Cell Behavior
Genomics
Molecular Networks
Neurons and Cognition
Other Quantitative Biology
Populations and Evolution
Quantitative Methods
Subcellular Processes
Tissues and Organs

7. Quantitative Finance

Computational Finance
Economics
General Finance
Mathematical Finance
Portfolio Management
Pricing and Securities
Risk Management
Staticstical Finance
Trading and Market Microsrtructure

8. Statistics

Machine learning
Methodology
Other Statistics

1. Андрущенко В.Б., Балагура І.В., Ланде Д.В. Інформаційні ресурси доступу та обміну науковою інформацією, системи ідентифікації науковців — можливості, недоліки, переваги. Матеріали міжнародної науково-технічної конференції «Інформаційні технології та безпека» (2 груд. 2016, м. Київ). Київ: ІПРІ, 2017. С. 180–191.

2. Warner Simeon. Open Archives Initiative protocol development and implementation at arXiv. URL: https://arxiv.org/pdf/cs/0101027.pdf (Last accessed: 10.03.2018).

3. Sorokina Daria, Gehrke Johannes, Warner Simeon, Ginsparg Paul. Plagiarism Detection in arXiv. URL: https://ieeexplore.ieee.org/abstract/document/4053155/ (Last accessed: 10.03.2018)

4. Asif-ul Haque, Paul Ginsparg. Positional effects on citation and readership in arXiv. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/asi.21166 (Last accessed: 10.03.2018).

5. Judit Bar-Ilan. Astrophysics publications on arXiv, Scopus and Mendeley: a case study. URL: https://link.springer.com/article/10.1007/s11192-013-1215-1 (Last accessed: 10.03.2018).

6. Андрущенко В.Б. Побудова дерева предметних областей для заданого поняття на базі ресурсу препринтів ArXiv. Матеріали XI Міжнародної науково-технічної конференції «Інтелектуальні технології лінгвістичного аналізу» (24–25 жовт. 2017, м. Київ). Київ: Національний авіаційний університет, 2017. С. 20.

7. TREC. Common evaluation measures. URL: https://trec.nist.gov/pubs/trec16/appendices/measures.pdf (Last accessed: 27.03.2018).

8. Офіційні метрики «РОМІП-2010». URL: http://romip.ru/romip2010/20_appendix_a_metrics.pdf (дата звернення: 27.03.2018).

Надійшла до редакції 22.05.2018