УДК 004.67

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

Лінгвістичні дослідження взаємозв'язків науковців на основі аналізу реферативної бази даних «Україніка наукова»

Запропоновано використовувати методи комп'ютерної лінгвістики для наукометричних досліджень реферативної бази даних «Україніка наукова». Представлено методику наукометричних досліджень наукового напрямку на прикладі комп'ютерних наук реферативної бази даних «Україніка наукова» на основі методів комп'ютерної лінгвістики, що дозволяє детально дослідити тенденції наукової співпраці, виділити найбільш комунікативних науковців, наукові групи та визначити ключові слова для певних наукових груп.

Ключові слова: *комп'ютерна лінгвістика, реферативна БД «Україніка наукова», мережа співавторства, мережа термінів, наукова взаємодія.*

Вступ

Складність визначення кількісного показника ефективності наукових досліджень та необхідність аналізу структури науки та перспектив розвитку науки викликає необхідність дослідження методів наукометрії та розробки відповідних інформаційних систем. Найбільш об'єктивні та визнані інформаційні системи наукометричного аналізу такі як, реферативні бази даних (БД) «Scopus», «Web of knowledge» формуються на основі реферативної інформації та наукометричного інструментарію [1, 2]. Тому актуальною задачею є формування та розвиток української реферативної БД. На сьогодні український реферативний журнал «Джерело» та реферативна база даних (БД) «Україніка наукова» є найповнішим джерелом рефератів українських монографій, статей із збірників наукових праць та серійних (періодичних) видань, матеріалів конференцій, посібників для вузів, авторефератів дисертацій, препринтів [3]. Реферативна БД "Україніка наукова" містить значну кількість української наукової інформації, що є якісною основою для проведення наукометричних досліджень. Наукометричні дослідження реферативної БД «Україніка наукова» проводять науковці з Інституту проблем реєстрації інформації, Національної бібліотеки ім. В.І. Вернадського, Центру досліджень науково-технічного потенціалу та історії науки ім. Г.М. Доброва Національної академії наук України та інших наукових установ [4-7].

Більшість наукометричних інформаційних систем включають методи та технології на основі цитування. Неможливість застосування подібних технологій пояснюється відсутністю обробки посилань в українській реферативній БД. Натомість велика ретроспектива та об'єм

даних надають можливість застосування іншого широко розповсюдженого напрямку – методів комп'ютерної лінгвістики.

В попередніх дослідженнях розглянуто оцінку повноти реферативного ресурсу, доцільність використання індексу рубрикатора НБУВ та проблему співпадіння прізвищ та ініціалів авторів [8]. Визначено ряд публікацій авторів, в яких був вказаний помилковий або пропущений індекс рубрикатора, що негативно впливає на результати наукометричних досліджень. Визначено прізвища, що є найбільш поширеними в реферативній БД. Проблема ідентифікації авторів також вносить значний вплив на достовірність результатів аналізу. Певні недоліки та особливості реферативної бази даних можна врахувати за допомогою застосування методів комп'ютерної лінгвістики для оцінки наукової галузі.

Комп'ютерна лінгвістика – напрямок в прикладній лінгвістиці, що орієнтується на використанні комп'ютерних інструментів – програм, комп'ютерних технологій організації та обробки даних – для моделювання функціонування мови в тих чи інших умовах, ситуаціях, проблемних сферах та інше, а також вся галузь застосування комп'ютерних моделей мови в лінгвістиці та суміжних дисциплінах [9]. Необхідно також зазначити розширення можливостей комп'ютерної лінгвістики, що відбуваються завдяки розвитку комп'ютерних наук вцілому: головним об'єктом дослідження стали інформаційні об'єкти, комп'ютерна лінгвістика стала максимально міждисциплінарною, входить до експериментальної парадигми, найчастіше використовує методи математичного моделювання, теорії складних систем та психофізіології (обробки інформації у людини), в комп'ютерній лінгвістиці з'явились нові об'єкти вивчення (колекції, комп'ютерні кластери і т. ін.) та нові експериментальні можливості (можливості сучасних інформаційних технологій) [9].

Метою досліджень є розробка методики наукометричних досліджень наукового напрямку реферативної бази даних «Україніка наукова» за допомогою методів комп'ютерної лінгвістики.

Методика досліджень

З початку 2000х років в комп'ютерній лінгвістиці почали застосовувати концепцію складних мереж у якості побудови мереж мови, тобто мережевої моделі текстових документів [10].

Аналіз мереж – один із потужних методів в інформатиці, що використовується для аналізу знань [11]. Такий метод використовують майже в усіх галузях знань, і кожне застосування містить окремі особливості та потребує розробки власних методик. Мережі співавторів містять зв'язки спільних робіт окремих науковців та характеризують наукову взаємодію. Мережі термінів основуються на статистиці взаємної зустрічаємості ключових термінів в одному документі. Мережі термінів застосовують для опису тематики різних

наукових галузей та визначення трендів розвитку наукових напрямів [11]. Також дані методики використовують у вебометричних дослідженнях для аналізу веб-сторінок. Зарубіжні автори у бібліометричних дослідженнях використовують аналіз термінів, мереж співавторів та цитування комплексно для всебічної оцінки галузей [12].

Методика наукометричного аналізу реферативної бази даних «Україніка наукова» в своїй основі містить 4 етапи, що включають методи складних мереж, методи фільтрації тексту, візуалізацію даних за допомогою програмних засобів Gephi.

На першому етапі визначається галузь та наукові напрямки, за якими буде проведено аналіз, завантажується та фільтрується файл з реферативною інформацією. Результатом першого етапу маємо відфільтровані за певними індексами рубрикатора дані про авторів та зв'язки між ними, тобто матриця мережі.

Другим етапом є створення мережі співавторів досліджуваної галузі, а також визначення основних характеристик мережі за допомогою програмних засобів Gephi, а також розрахунок додаткових параметрів за допомогою власних програмних засобів. В результаті другого етапу визначаються основні властивості співробітництва науковців, наукові групи та найбільш комунікативні науковці за певним науковим напрямком.

Третій етап присвячений відбору в повнотекстових базах даних наукових публікацій найбільш комунікативних науковців та створення текстового корпусу для виділення основних термінів за науковими напрямками.

На четвертому етапі проводиться візуалізація мереж термінів науковців та галузі вцілому, розрахунок основних параметрів та виділення так званих опорних слів та відповідних словосполучень. Проводиться узагальнення результатів, опис основних характеристик, тенденцій в галузі.

Розглянемо методику наукометричних досліджень за допомогою методів комп'ютерної лінгвістики на прикладі аналізу комп'ютерних наук в реферативній БД «Україніка наукова».

Перший етап. Для проведення наукометричного аналізу галузі «Комп'ютерних наук», на початковому етапі необхідно визначити відповідні індекси рубрикатора НБУВ, після чого буде сформовано таблиці вузлів та ребер мережі співавторів. В дослідженні використано рубрики розділів «Інформаційна та обчислювальна техніка» та «Кібернетика»: «Основи інформатики та обчислювальної техніки», «Аналогова і гібридна обчислювальна техніка», «Цифрова обчислювальна техніка», «Комп'ютери і програмування», «Кібернетичні моделі», «Теорія інформації», «Системний аналіз», «Теорія автоматів» та «Біоніка». Обробка даних рубрик та створення мережі співавторів з комп'ютерних наук детально представлено в роботі [13].

Дані з реферативної БД «Україніка наукова» представлені у форматі УКРМАРК, що є цифровим форматом представлення бібліографічних даних та українською версією міжнародного комунікативного формату UNIMARK. На мові високого рівня Python було створено програмні засоби, що виконували фільтрацію співавторів за індексами рубрикатора НБУВ «З97. Інформаційна та обчислювальна техніка», «З81 Кібернетика» та формували мережу співавторів. Реферативна БД «Україніка наукова» містить записи на українській, англійській та російській мові, тому інформація про публікації авторів може містити похибки. Для корекції результатів було проведено незначну лінгвістичну обробку прізвищ. Ваги зв'язків між авторами для кожної публікації визначалися обернено пропорційно до кількості її авторів [14].

Другий етап. На наступному етапі формується мережа співавторів та обчислюються її основні характеристики та параметри. При необхідності детального визначення причинно-наслідкових особливостей, існує можливість вивчення та візуалізації окремих зв'язків та підмереж. Візуалізація мережі та обчислення основних характеристик, визначення наукових груп було виконано за допомогою засобів програмного продукту Gephi [15]. У результаті отримано мережу співавторів, що містить 1189 автори та 18049 зв'язки. Діаметр графа, тобто найбільша відстань між двома співавторами через інших авторів, дорівнює 23. При чому середня довжина шляху між вузлами близька до 8. Середня кількість людей, з якими співпрацює один співавтор, або середня степінь вузла, приблизно становить 4. Середня зважена степінь приблизно дорівнює 2, тобто у співавторстві один автор має 2 публікації.

Мережа була поділена на групи на основі розрахунку модулярності вузлів. Модулярність вузла – це величина, що оцінює щільність зв'язків у зв'язній компоненті в порівнянні із зв'язками між компонентами. Наявність наукових груп, що можуть мати ознаки наукових шкіл, в галузі «Комп'ютерних наук» можна простежити наочно за особливістю більшості клік, що в своїй основі містять потужного автора з великою кількістю статей у співавторстві та значну кількість маленьких вузлів — авторів-учнів. На рис. 1 наведено фрагмент побудованої мережі співавторів, що містить дані по академіку Палагіну О.В., його співавторам та співавторам його співавторів. Розмір відповідного вузла автора відповідає його кількості зв'язків (табл. 1). Концентрація зв'язків співавторства навколо одного або декількох лідерів можуть характеризувати створення окремих наукових шкіл. У даному випадку їхніми потенційними центрами є такі автори як Палагін О., Сергієнко І., Бодянський Е., Баркалов А. та інші [14].

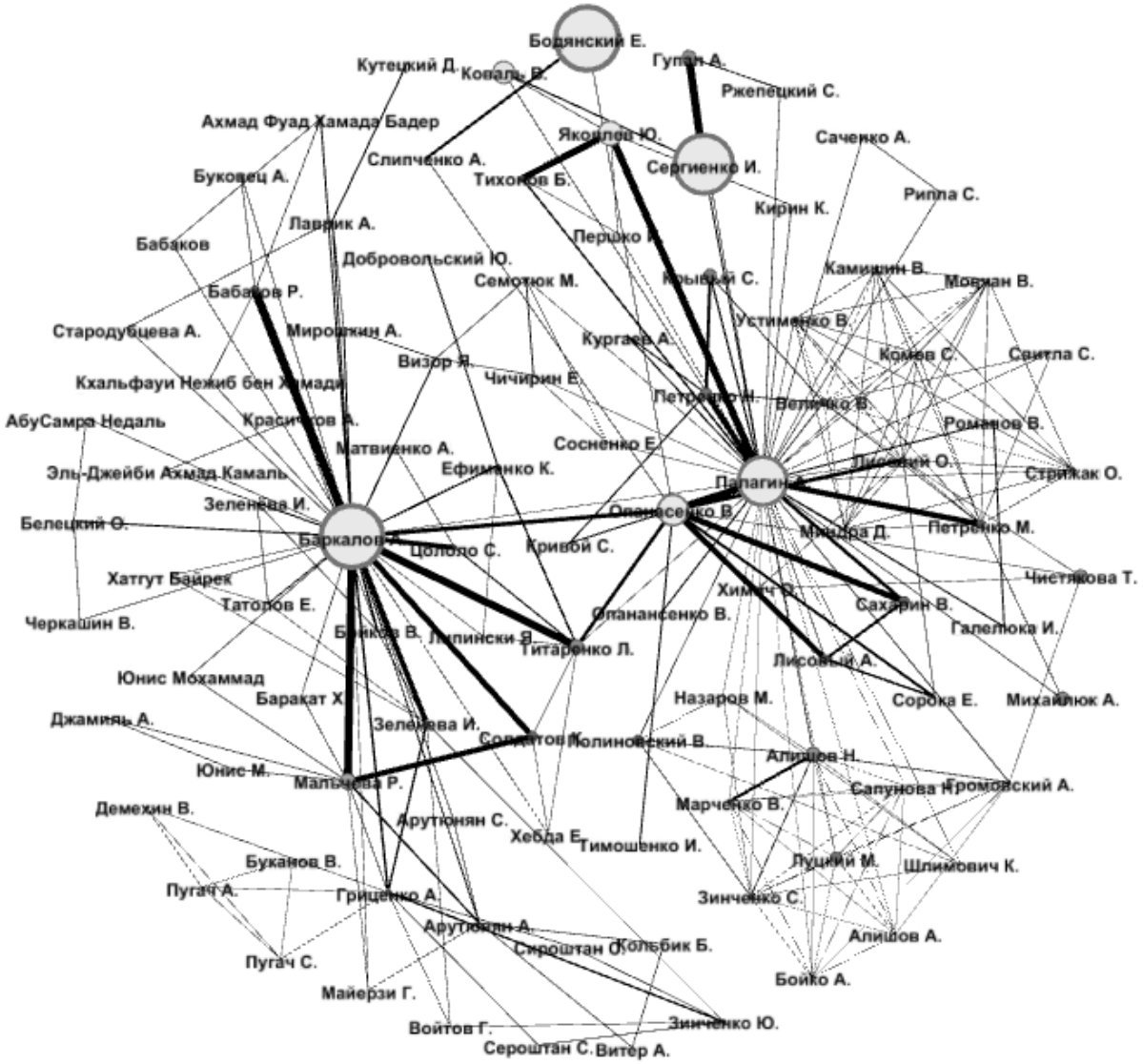

Рис. 1. Співавтори автора Палагіна О. до другого рівня

Визначення важливих вузлів у мережі є актуальною задачею та потребує детального вивчення предмету дослідження, адже існує багато коефіцієнтів, що надають різнобічні характеритики вершин, та доцільність їх застосування визначається тільки відповідністю цілям експериментів. У результаті дослідження було визначено основні показники центральності мережі з комп'ютерних наук [16]. Зважений рівень центральності, що фактично є показником кількості статей у співпраці, відображає об'єм напрацювань автора, а кількість зв'язків автора характеризують коло його співавторів. На рис. 1 рівню центральності вузлів відповідає їх величина. Автори з високим рівнем показника центральності пов'язані в окремих групах — наукових групах та не пов'язані між собою. В роботі [16] також представлено один із модифікованих показників центральності, що можна застосовувати для визначення потенційних центрів в мережі співавторів.

Третій етап. В дослідженні використовувались публікації п'яти авторів з комп'ютерних наук на російській мові за останні 10 років, що представлені в повнотекстовій науковій БД «Наукова періодика України» [17]. Для кожного автора було сформовано

текстовий корпус, що містив від 15 до 50 статей у фахових журналах, тез доповідей на конференціях та патентів. При попередній обробці тексту було вилучено нетекстові символи та виконано стеммінг.

Для виділення термінів використовувався класичний статистичний метод TFIDF (Term Frequency-Invert Document Frequency, «дискримінантна сила» – статистична міра, що використовується для оцінки важливості слів у контексті документа, що є частиною колекції документів або корпуса) [10]. За даним методом значну вагу отримують слова з великою частотою у межах конкретного документу і низькою частотою вживання в інших документах. TFIDF дорівнює добутку частоти слова у фрагменті тексту та двійкового логарифму величини, оберненої до кількості фрагментів тексту, в яких це слово присутнє. Для словосполучень із двох (біграм) та трьох слів (триграм) також розраховується TFIDF.

В наступній обробці для послідовності термінів та їх вагових значень за TFIDF будуються компактифіковані графи горизонтальної видимості та виконується перевизначення значень вагових коефіцієнтів [18]. Далі експертним методом та за допомогою стоп-словника обираються основні терміни, що використовує автор. Основні терміни складаються із одного, двох та трьох слів та описуються за допомогою матриці інцидентності A, а матриці A^TA та AA^T містять асоціативні зв'язки першого (якщо два терміни утворюють третій) та другого (якщо термін утворений двома іншими) роду [19].

Четвертий етап. Відібрані терміни використовуються для створення мережі термінів, де вузлами є терміни та словосполучення, а зв'язки відповідають входженню термінів до словосполучень. Такі мережі мови ще називають L – простором [10]. На рис.2 представлено приклад мережі термінів для автора академіка Палагіна О.В., і яка створена за допомогою програмних засобів Gephi.

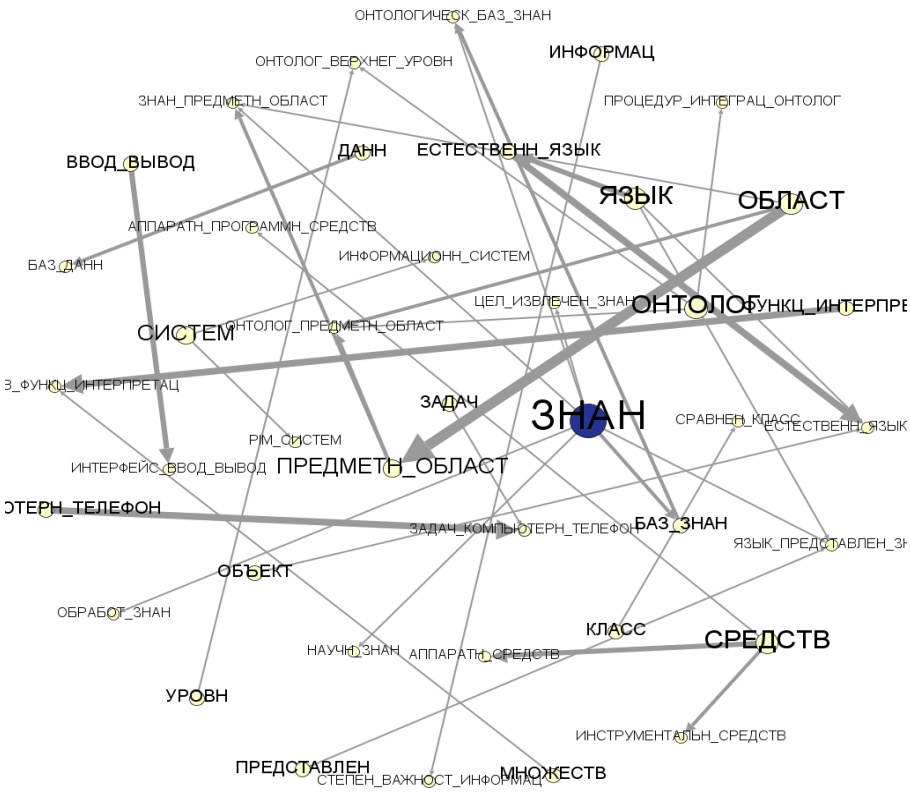

Рис. 2 Мережа термінів на основі аналізу публікацій Палагіна О.В.

Найбільш значимі слова та словосполучення в мережі термінів можна виділяти за допомогою алгоритму HITS (Hyperlink Induced Topic Search – індукований гіперлінковий тематичний пошук). Алгоритм HITS основується на визначенні «авторитетних вузлів» (вузлів, на які багато посилань) та «вузлів-посередників» (вузли, що мають найбільшу кількість посилань). Для автора Палагіна О.В. найбільш значимими термінами є: 1) за значенням «авторитетність» триграмми – «онтология предметной области», «знания о предметной области», «язык представления знаний», «естественно-языковый обьект», «онтология верхнегнего уровня», «онтологичкая база знаний», «задачи компьютперной телефонии», «множество функций интерпретации»; 2) за значенням «посередництво» – «система», «онтология», «знание», «база знаний», «класс», «данные», «задача», «множество», «область», «предметная область», «представление», «информация», «объект», «язык» та ін. Таким чином, можна виділити серед основних інтересів автора методи комп'ютерної лінгвістики, бази знань.

Створення мереж термінів допомогає виділити основні терміни, що характеризують основну тематику робіт автора, а також ефективний інструмент для ідентифікації науковців. В реферативній БД «Україніка наукова» міститься велика кількість записів, але для ідентифікації авторів використовується тільки прізвище та ініціали, що спричинює проблему ідентифікації авторів. При порівнянні термінів нових наукових робіт зі словником, що сформований на основі попередніх робіт автора, в залежності від перетину таких мовних мереж можна стверджувати про приналежність наукової роботи автору [20].

За допомогою перетину L-просторів окремих авторів можна визначити основні терміни наукового напрямку, в якому вони працюють (Рис.3).

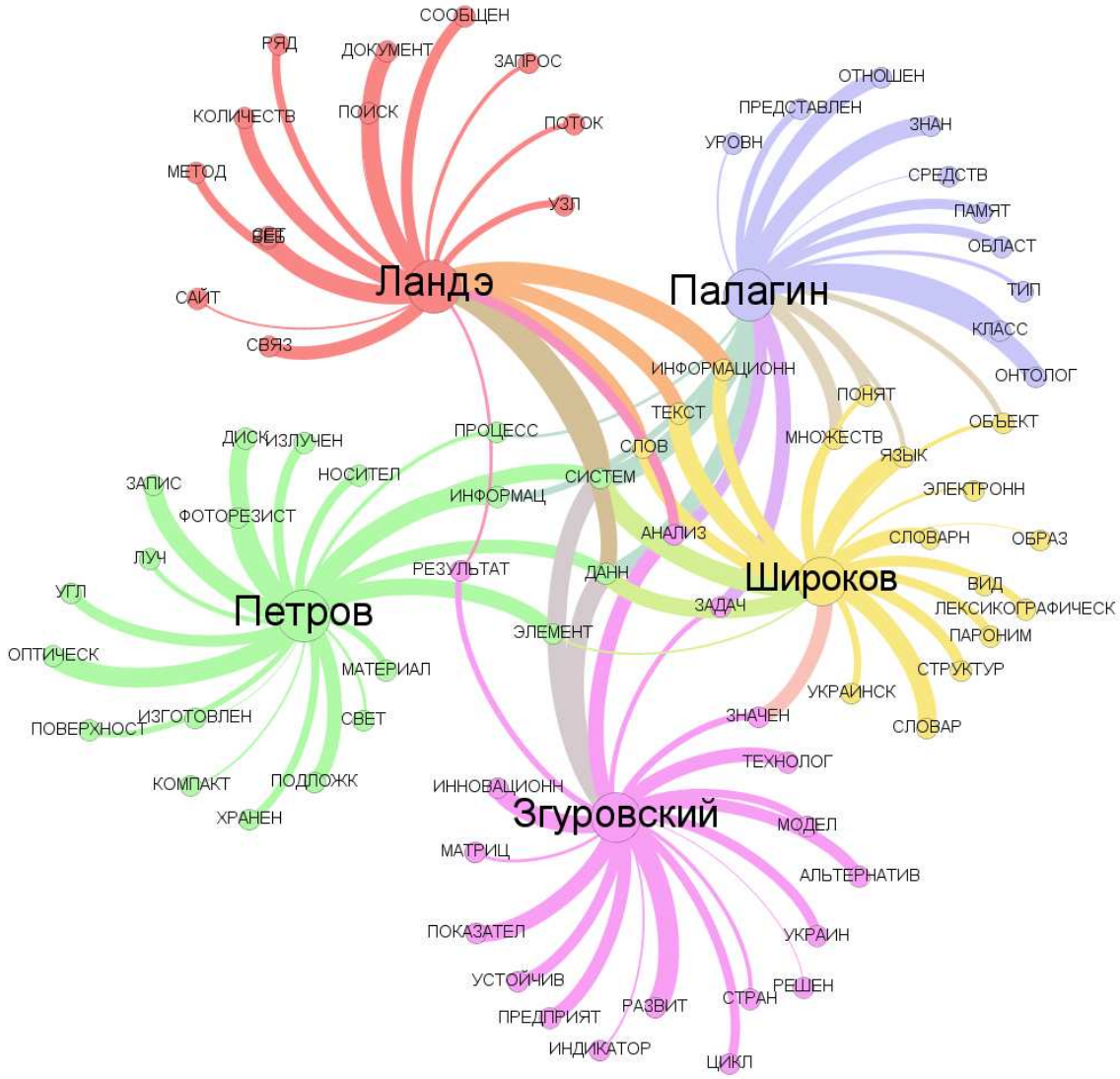

Рис. 3 Мережа термінів одних із найбільш комунікативних науковців з комп’ютерних наук

В результаті роботи було побудовано спільну мережу термінів та отримано загальний словник: «данные», «система», «анализ», «информация», «элемент», «информационный», «слово», «задача», «язык», «понятие», «множество», «текст», «значение», «процесс», «объект», «результат». Такі терміни характеризують перетин досліджень авторів з комп’ютерної лінгвістики, систем зберігання даних, аналізу даних, складних систем, баз даних і знань, досліджень сталого розвитку.

Висновки

Запропоновано використовувати методи комп’ютерної лінгвістики для наукометричних досліджень взаємозв’яків науковців на основі аналізу реферативної бази даних «Україніка наукова». Запропоновано методику наукометричних досліджень наукового

напрямку реферативної бази даних «Україніка наукова» на основі методів комп'ютерної лінгвістики, що дозволяє детально дослідити тенденції наукової співпраці, виділити найбільш комунікативних науковців, наукові групи та визначити ключові слова для певних наукових груп. Показано можливість застосування методики наукометричного аналізу на прикладі аналізу комп'ютерних наук в реферативній базі даних «Україніка наукова». Отримано опис основних характеристик, тенденцій в галузі комп'ютерних наук, виділено основні наукові групи та найбільш комунікативних науковців, основні терміни, створено візуалізацію мереж термінів.

1. Scopus [електронний ресурс] – Режим доступу: http://www.scopus.com. – Назва з екрану.
2. Web of Knowledge [електронний ресурс]: – Thomson Reuters, 2012 – Режим доступу: http://wokinfo.com. – Назва з екрану.
3. Реферативна база даних «Україніка наукова» [електронний ресурс]. – Режим доступу: http://nbuv.gov.ua/db/ref.html. – Назва з екрану.
4. Крючин А.А. Значення видання українського реферативного журналу «Джерело» для розвитку наукових комунікацій в Україні / А.А. Крючин, Л.Й. Костенко, Н.М. Мініна, І.В. Балагура, Л.М. Овсієнко// Наука України у світовому інформаційному просторі. – 2012. – Вип.6. – с.20-23.
5. Костенко Л.Й. Наукова періодика України та бібліометричні дослідження : [монографія] / Л. Й. Костенко, О. І. Жабін, Є. О. Копанєва, Т. В. Симоненко ; НАН України, Нац. б-ка України ім. В. І. Вернадського. – К., 2014. – 173 с.
6. Рибачук В.П. Методологічні проблеми оцінювання продуктивності наукової діяльності / В. П. Рибачук // Наука та наукознавство. — Київ, 2013. — № 2 (80). — С. 46-52
7. Букшина Т.Ф. Бібліометричний аналіз наукових публікацій з питань дошкільної освіти і виховання в загальнодержавній реферативній БД "Україніка наукова" (1998–2012 рр.) [Електронний ресурс] / Т. Ф. Букшина // Інформаційні технології і засоби навчання . - 2014. - Т. 40, вип. 2. - С. 135-150. - Режим доступу: http://nbuv.gov.ua/j-pdf/ITZN_2014_40_2_15.pdf
8. Балагура І.В. Дослідження мережі співавторів з бібліотечної справи та наукознавства [електронний ресурс]. / І.В Балагура// Міжнародна наукова конференція "Місце і роль бібліотек у формуванні національного інформаційного простору", жовтень, 2014. – м. Київ– Режим доступу: http://conference.nbuv.gov.ua/report/view/id/363 – Назва з екрану.
9. Большакова Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков та ін.//. — М.: МИЭМ, 2011. — 272 с.

10. Ланде Д.В. Елементи комп’ютерної лінгвістики в правовій інформатиці. – К.: НДІІП НАПрН України, 2014. – 168 с.

11. Tseng Y.-H. Mining term networks from text collections for crime investigation / Yuen-Hsien Tseng , Zih-Ping Ho, Kai-Sheng Yang, Chun-Cheng Chen // Expert Systems with Applications. —2012. —Vol.39, Issue 11. —P. 10082–10090.

12. Chappin E.J.L. Transition and transformation: A bibliometric analysis of two scientific networks researching socio-technical change / Emile J.L. Chappin, Andreas Ligtvoet // Renewable and Sustainable Energy Reviews. —2014. — Vol. 30, —P. 715–723.

13. Ланде Д.В. Наукометричні дослідження мереж співавторства по базі даних «Україніка наукова» / Д.В. Ланде, І.В. Балагура // Реєстрація, зберігання і обробка даних. – 2012, - Т.14, №4 –с.41-51.

14. Горбов І.В. Визначення потенційних експертних груп науковців в мережі співавторства з використанням методів підтримки прийняття рішень / І.В. Горбов, С.В. Каденко, І.В. Балагура, Д.Ю. Манько, О.В. Андрійчук // Реєстрація, зберігання і обробка даних. – 2013, - Т.15, №4 –С.87-97.

15. The Open Graph Viz Platform Gephi [електронний ресурс]. – Режим доступу: https://gephi.github.io/. – Назва з екрану.

16. Балагура И.В. Характеристики сети соавторов медицинских наук /И.В. Балагура, Д.В. Ландэ, И.В. Горбов // Клин. информат. и Телемед. – 2013. – Т.9, №10 – С.141-144.

17. Наукова періодика України [електронний ресурс]. – Режим доступу: http://www.irbis-nbuv.gov.ua/cgibin/irbis_nbuv/cgiirbis_64.exe?C21COM=F&I21DBN=UJRN&P21DBN=UJRN&S21CNR=20– Назва з екрану.

18. Ландэ Д.В. Использование графов горизонтальной видимости для выявления слов, определяющих информационную структуру текстов на различных язиках [електронний ресурс] /Д.В. Ландэ, А.А. Снарский, Е.В. Ягунова // Тези доповідей міжнародної наукової конференції «Горизонти прикладної лінгвістики і лінгвістичних технологій»(«Megaling-2013») – Київ, 2013. – Режим доступу: http://megaling.ulif.org.ua/prezentatsiyi-dopovidey/– Назва з екрану.

19. Ландэ Д.В. Подход к созданию терминологических онтологий /Д.В. Ландэ, А.А. Снарский // Онтология проектирования – 2014. – Т.12, №2 – С.83-92

20. Ягунова Е.В. Эксперимент и вычисления в анализе ключевых слов художественного текста // Сборник научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика. – Пермь, 2010. – Вып. 1. – С. 85-91.