## Інформатизація, інформаційні технології

УДК 004.67

**ЛАНДЕ Д.В.,** *доктор технічних наук,*
Інститут проблем реєстрації інформації НАН України

# СТВОРЕННЯ ТЕРМІНОЛОГІЧНОЇ МОДЕЛІ ПРЕДМЕТНОЇ ОБЛАСТІ ШЛЯХОМ ЗОНДУВАННЯ GOOGLE SCHOLAR CITATIONS

***Анотація.*** *Пропонується методика побудови мереж – моделей предметних областей на основі зондування контентних мереж. Як така мережа розглядається мережа понять, що відповідають тегам сервісу Google Scholar Citations. Модель застосовано для правової науки, але запропонований підхід можна застосовувати також для інших областей науки.*

***Ключові слова:*** *предметна область, модель предметної області, правова наука, зондування мережі, інформаційна мережа.*

***Аннотация.*** *Предлагается методика построения сетей – моделей предметных областей на основе зондирования контентных сетей. Как такая сеть рассматривается сеть понятий, соответствующих тегам сервиса Google Scholar Citations. Модель использована для правовой науки, однако предложенный подход можно применять и для других областей науки.*

***Ключевые слова:*** *предметная область, модель предметной области, правовая наука, зондирование сети, информационная сеть.*

***Summary.*** *The technique of constructing networks – domain models based on content-based networks sensing is proposed. A network of concepts corresponding to Google Scholar Citations service tags is considered as such network. The model is used for legal science, but the proposed approach can be applied to other areas of science.*

***Keywords:*** *subject area, domain model, legal science, network sensing, information network.*

***Постановка проблеми.*** На цей час залишається актуальною задача створення моделей предметних областей. Під моделлю предметної області на цей час, зокрема, розглядають спеціальним чином сформовану мережу понять (галузеву онтологію). Побудова великих галузевих онтологій – це складна проблема, яка потребує великих ресурсних витрат. Першими етапами цього процесу є побудова термінологічної основи онтології і визначення певних семантичних зв’язків [1].

У цій роботі надається і обґрунтовується підхід до створення моделі предметної області (правової науки) на основі зондування великої інформаційної мережі. Як така мережа у цій статті розглядається мережа понять, що відбиваються у тегах наукометричного сервісу Google Scholar Citations (http://scholar.google.com/citations). На рис. 1 наведено фрагмент інтерфейсу сторінки сервісу Google Scholar Citations, що відповідає заздалегідь заданому тегу comparative law (порівняльне правознавство). На інтерфейсі, що відповідає повному тегу (label:comparative_law) посторінково у ранжированому форматі відображуються імена науковців, що позначили свою діяльність цим тегом, а також інші теги, що ними приписані (наприклад, Alec Stone Sweet визначив для себе ще comparative politics, international law, international relations, European integration).

***Метою роботи*** є опис теоретичних засад і методології автоматизованого формування моделі предметної області, зокрема, правової науки у цілому шляхом зондування великої інформаційної мережі.

***Виклад основного матеріалу.*** Для досягнення поставленої мети застосовується спеціальний алгоритм сканування ресурсів сервісу Google Scholar Citations який надає можливість отримання репрезентативного набору тегів (позначень понять) як основи майбутньої онтології. Під зондуванням інформаційних мереж розуміється витяг невеликого обсягу найважливішого змісту з великих інформаційних мереж, які з технологічних причин не підлягають повному скануванню.

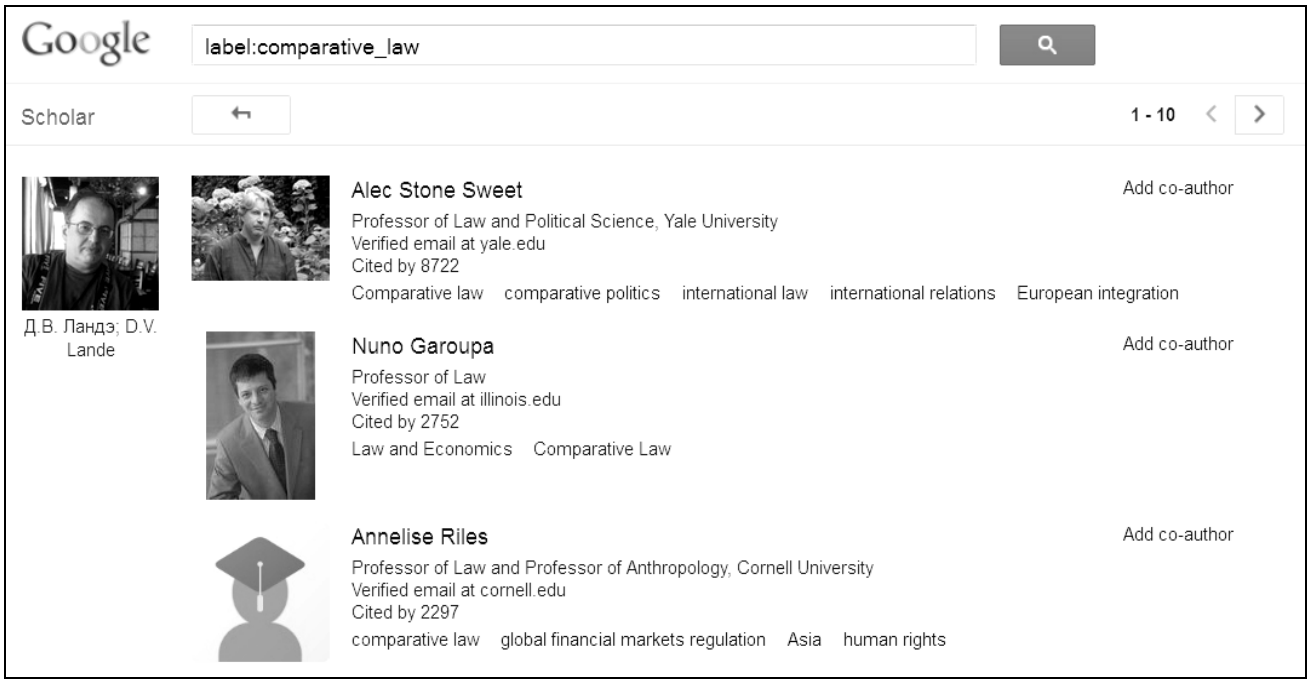


Рис. 1. Інтерфейс сторінки сервісу Google Scholar Citations

*Опис моделі.*

При побудові мереж термінів доцільно застосовувати моделі, що вже випробувані на піринговых мережах (Peer-to-peer, P2P – рівний з рівним), які засновані на рівноправ’ї учасників. У таких мережах відсутні виділені сервери, а кожен вузол (peer) є як клієнтом, так і сервером. У багатьох випадках P2P є накладеними мережами, що використовують існуючі транспортні протоколи мережі Інтернет. Пірингові мережі складаються з вузлів, кожен з яких взаємодіє лише з деякою підмножиною інших вузлів мережі (через обмеженість ресурсів).

Розглянемо модель інформаційної мережі, яку будемо вважати піринговою мережею, до того ж сполученою з глобальною мережею Інтернет, яка розглядається як зовнішнє середовище. Для пошуку необхідних даних (у нашому випадку – тегів) для таких мереж застосовується декілька моделей. Розглянемо, наприклад, модель, що відповідає методу “широкого первинного пошуку” (Breadth First Search, BFS) для мережі розмірності $N$. Нехай на вході є запит, який з вузла q адресується до всіх сусідів (найближчих за деякими критеріями вузлів). Коли вузол p отримує запит, виконується пошук в його локальному індексі. Якщо деякий вузол r приймає запит (Query) і обробляє його, то він генерує повідомлення-відгук (QueryHit), щоб повернути результат. Повідомлення QueryHit включає інформацію про релевантні теги, яка доставляється по мережі вузлу, що запрошує. Інший, так званий “інтелектуальний пошуковий механізм” (Intelligent Search Mechanism, ISM) забезпечує поліпшення швидкості і ефективності пошуку інформації за рахунок мінімізації витрат на кількість

повідомлень, що передаються між вузлами, та мінімізації кількості вузлів, які опитуються для кожного пошукового запиту [4]. Щоб досягти цього, для кожного запиту оцінюються лише ті вузли, які найбільш відповідають даному запиту.

Саме модель, близьку до ISM будемо розглядати у цій роботі.

Зондування опорної модельної мережі здійснюється за таким алгоритмом:

1. Обирається визначена кількість вузлів опорної мережі (мережі, що зондується), що визначаються як базові для нової мережі, що відповідає результатам зондування.

2. Для кожного з базових вузлів визначаються суміжні з ним вузли (“сусіди”), які додаються до мережі, що створюється.

3. Здійснюється перехід до сусіднього вузла опорної мережі, що має найбільшу ступень.

4. Якщо має місце “зациклювання” (вибирається вузол, до якого вже було здійснено перехід за цим алгоритмом), здійснюється перехід до наступного базового вузла з початкового переліку і здійснюється перехід до пункту 2.

5. Якщо перелік базових вузлів завершено, будемо вважати, що мережу, що відповідає результатам зондування, побудовано.

Наведений алгоритм застосовувався для двох найпоширеніших модельних мереж Erdős-Rényi (ER) і Barabási-Albert (BA) (Рис. 2). Відомо, що модель ER – випадкова мережа – будується наступним чином: $N$ спочатку не з’єднаних вузлів попарно поєднуються з ймовірністю $p$. В результаті створюється мережа приблизно з $pN(N-1)/2$ випадково вибраними зв’язками.

Модель BA – одна з декількох моделей мереж із степеневим розподілом ступенів вузлів (так званих, безмасштабних мереж). Ця модель враховує як зростання мережі (динаміку), так і принцип переважного приєднання, який полягає в тому, що чим більше зв'язків має вузол, тим переважніше для нього створення нових зв’язків знову утворюваними вузлами. Вузли з більшим ступенем мають більшу вірогідність приєднання (створення нових зв’язків) до нових вузлів.

Слід відмітити, що безмасштабними є найбільш популярні на цей час мережі, такі як веб-простір із гіперпосиланнями, соціальні мережі, мережі слів у літературних творах, протеїнів, тощо. Автором передбачалося, що мережі понять, які природно формуються учасниками мережевих сервісів вірогідно теж мають властивість безмасштабності, але не завжди можна це перевірити, маючи всеосяжну інформацію. Якщо мережа така складна і велика, як, наприклад, Google Scholar Citations, на допомогу має прийти зондування. Зазначимо, що результати будь-якого зондування не завжди будуть вірно відбивати природу великої мережі, що досліджується. Багато чого тут залежить саме від алгоритму цієї процедури.

Візуально якісні результати зондування мереж ER і BA з близькими параметрами (1000 вузлів, понад 2000 зв’язків) наведені на Рис. 2. Порівняння показує, що зв’язані області (гілки), що відповідають окремим поняттям, у першому випадку досить довгі, а вузлів, за якими йде маршрут зондування, більше, ніж у другому, більш цікавому, випадку. В рамках цього дослідження не є важливими саме чисельні результати і параметри мереж, важливо оцінити лише вигляд зв’язаних ланцюжків, що моделюють гілки понять. Слід зазначити, що реальним мережам притаманний ще й феномен “клуба багатіїв” (Rich Clube), який обумовлює щільнішу пов’язаність найбільших вузлів. Тому передбачалося, що наведений алгоритм буде мати таку особливість, як швидке зациклювання, що приведе до ще більшого скорочення гілок понять.

*а* *б*

Рис. 2. Приклад мережі, побудованої зондуванням модельних мереж, де: *а* – Erdős-Rényi; *б* – Barabási-Albert

Саме виходячи з цих результатів якісного моделювання було зроблено висновок щодо можливості формування невеликих зв’язаних гілок, які відповідають поняттям, що цікавлять користувачів сервісу Google Scholar Citations.

*Зондування мережі Google Scholar Citations.*

Імплементацію наведеного вище алгоритму, який застосовувався до модельних мереж, було адаптовано до реальної мережі понять наступним чином:

1. Експертним шляхом визначається перелік базових тегів (ключових слів, що відповідають поняттям).
2. Обирається тег з визначеного експертами переліку.
3. Відкриваються сторінки веб-сервісу, що відповідають цьому тегу (максимальна кількість таких сторінок обмежується заздалегідь заданим параметром).
4. До мережі, що створюється, додаються усі теги, що містяться на вибраних сторінках.
5. Здійснюється перехід до сторінок, що відповідають тегу, що найбільше повторювався на сторінках, що розглядаються.
6. Якщо має місце “зациклювання” (вибирається тег, до якого вже було здійснено перехід за цим алгоритмом) або “відхід від теми” (виявляється за результатом змістовного аналізу).
7. Якщо перелік базових тегів завершено, будемо вважати, що мережу побудовано. Інакше здійснюється перехід до наступного базового тегу з початкового переліку, тобто перехід до пункту 2.

Для побудови моделі предметної області експертним шляхом було визначено базові теги, що відносяться до неї, наприклад, для області правознавства відомі такі теги англійською мовою: law, public_law, criminal_law, private_law, civil_rights, information_law, legal_theory, criminology та ін.

На Рис. 3. наведено приклади зондування мережі, починаючи з визначених вузлів, на якій для першого ланцюжка вертикальною лінією позначено припинення механізму сканування через відхилення від основної тематики (Topic Shift) (що визначається з урахуванням лексичного складу тегів). Також наведено приклад “зациклювання” цієї процедури на ланцюжку, що відповідає тегу “criminology”.

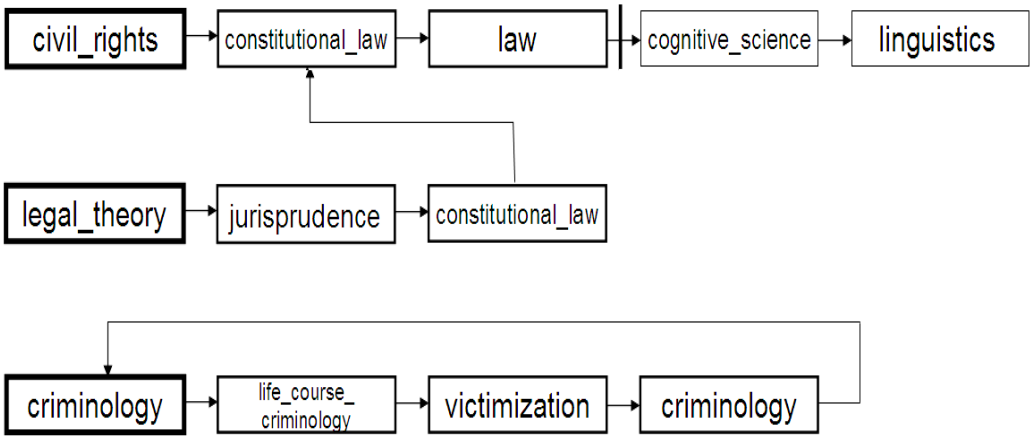


Рис. 3. Приклади зондування мережі, починаючи з визначених вузлів

Формування базового стартового переліку вузлів-понять і правил відбору “кінцевих” вузлів.

На Рис. 4 наведено приклад мережі понять (моделі предметної області), яку було побудовано за наведеним алгоритмом і вказаними базовими тегами.

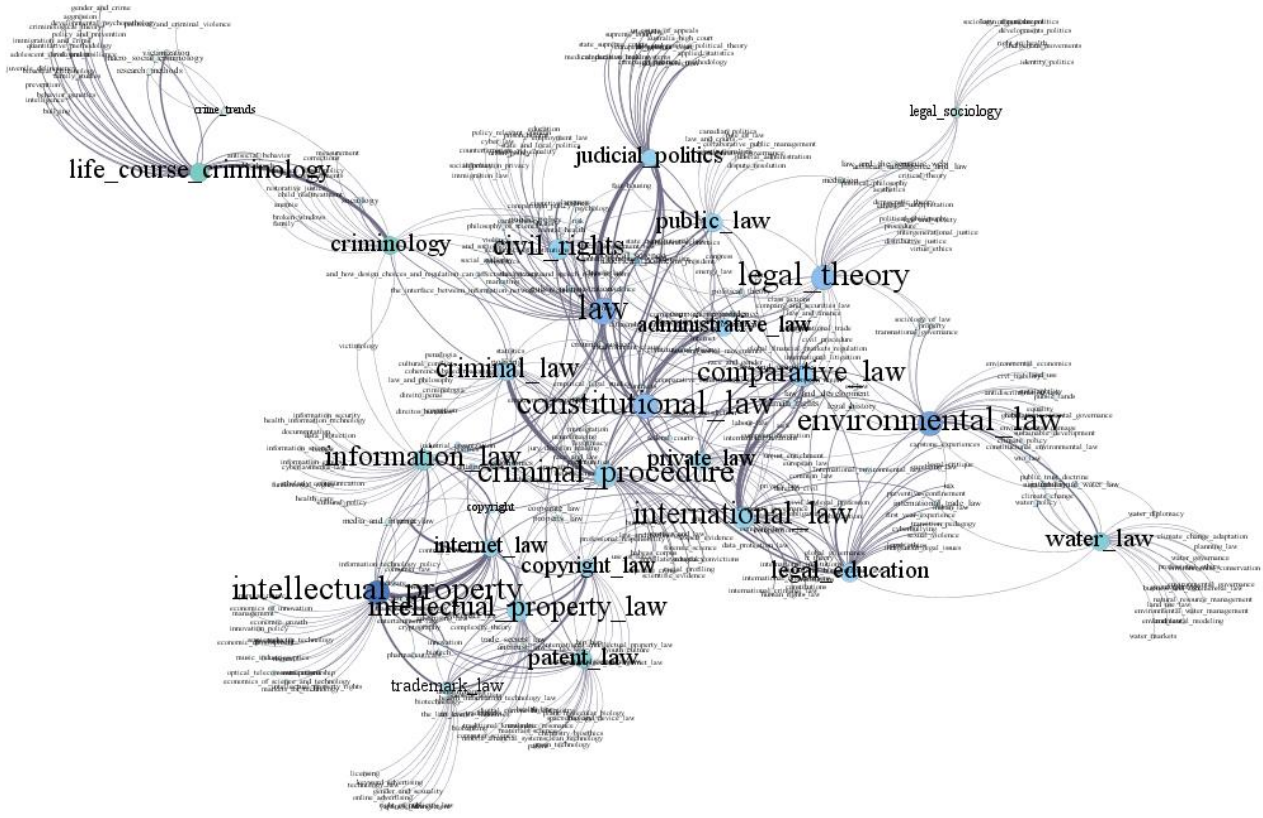


Рис. 4. Фрагмент мережі понять

Побудована мережа понять виявилася зв’язною. При кількості базових тегів 20, загальна кількість вузлів-тегів, що були охоплені алгоритмом, склала 730, а кількість нетермінальних вузлів – лише 54. Розподіл степенів цих вузлів, наведений на Рис. 5, свідчить щодо відсутності степеневого розподілу, тобто наведений алгоритм зондування

не зберіг степеневого розподілу базової мережі тегів, що передбачалася. Середня довжина ланцюжка понять складає 3,5, що відповідає класифікаторам трьох-чотирьох рівнів, прийнятим, зокрема, у бібліографічних дослідженнях.

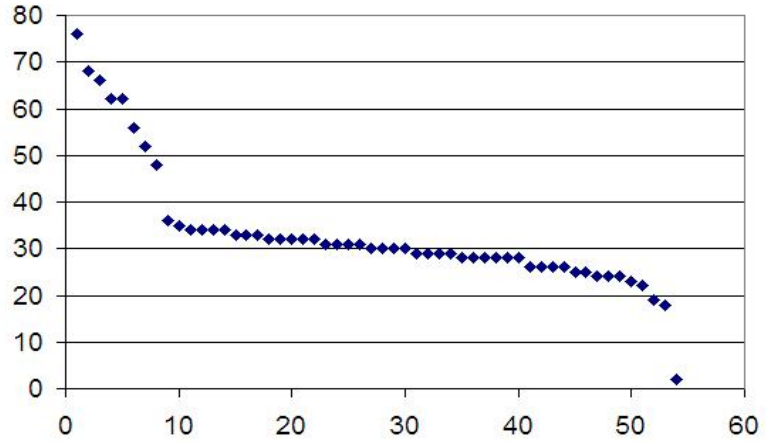


Рис. 5. Розподіл степенів вузлів-тегів мережі понять

***Висновки.***

1. У запропонованій моделі предметної області як онтологічні зв’язки застосовуються зв’язки між областями інтересів окремих вчених. Фактично розглядається компактифікація біографа “вчений – галузі науки, що його цікавлять”. Ці зв’язки дозволили припустити наявність загального наукового апарата, семантичний зв’язок.

2. Запропоновано та реалізовано підхід формування моделі предметної області, при формуванні якої застосовуються знання, заздалегідь вкладені вченими, що є учасниками проекту Google Scholar Citations.

3. Модель застосовано для правової науки, але запропонований підхід можна застосовувати для інших галузей науки. Автором, зокрема, побудовано подібні мережі для напрямку глибинного аналізу текстів (Text Mining) і складних мереж (Complex Networks).

**Використана література**

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

2. Широков В.А. Корпусна лінгвістика / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна. – К. : Довіра, 2005. – 471 с.

3. Добров Б.В. Онтологии и тезаурусы. Модели, инструменты, приложения / [Б.В. Добров, В.Д. Соловьев, Н.В. Лукашевич, В.В. Иванов]. – М., Бином, 2009. – 173 с.

4. Ландэ Д.В., Снарский А.А. Подход к созданию терминологических онтологий // Онтология проектирования. – 2014. – № 2(12). – С. 83-91.

5. Ландэ Д.В. Моделирование контентных сетей // Проблеми інформатизаціх та управління : зб. наук. праць. – К. : НАУ, 2012. – Вип. 1(37). – С. 78-84.

6. Ландэ Д.В. Интернетика : Навигация в сложных сетях : модели и алгоритмы / Д.В. Ландэ, А.А. Снарский, И.В. Безсуднов. – М. : Либроком (Editorial URSS), 2009. – 264 с.

7. Kalogeraki V., Gunopulos D., Zeinalipour-Yazti D. A Local Search Mechanism for Peer-to-Peer Networks // Proc. of CIKM'02, McLean VA, USA, 2002.

~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~
~~~~~~~~~~~~~