Правова інформатика

УДК 004.7:001.8

ЛАНДЕ Д.В., *доктор технічних наук,*
Інститут проблем реєстрації інформації НАН України,
Науково-дослідний інститут інформатики і права НАПрН України

ПОБУДОВА МОДЕЛЕЙ ПРЕДМЕТНИХ ОБЛАСТЕЙ З ЮРИСПРУДЕНЦІЇ ЗА ДАНИМИ СЕРВІСУ WIKIPEDIA

Анотація. *У роботі наводиться алгоритм побудови моделей різних предметних областей з юриспруденції на базі автоматичного аналізу даних сервісу Wikipedia. Показано, як виявляється термінологічна база, що динамічно змінюється при розвитку сервісу-першоджерела, формується мережева структура, розраховується вага різних термінів-понять за двома різними критеріями – ступенями вузлів і PageRank. На прикладах показана адекватність підходів, що пропонуються, а також, що кластери в термінологічних мережах можуть розглядатися як основа для виявлення окремих наукових напрямків.*

Ключові слова: *модель предметної області, термінологічна база, Wikipedia, зв'язкі понять, юриспруденція, сканування мережевого сервісу.*

Аннотация. *В работе приводится алгоритм построения моделей предметных областей по юриспруденции на основе автоматического анализа данных сервиса Wikipedia. Показано, как определяется терминологическая база, динамически изменяемая при развитии сервиса-первоисточника, формируется сетевая структура, рассчитывается вес разных терминов-понятий по двум разным критериям – степеням узлов и PageRank. На примерах показана адекватность предлагаемых подходов, а также то, что кластеры в терминологических сетях могут рассматриваться как основа для выявления отдельных научных направлений.*

Ключевые слова: *модель предметной области, терминологическая база, Wikipedia, связи понятий, юриспруденция, сканирование сетевого сервиса.*

Summary. *The algorithm of creation of models of subject domains on jurisprudence on the basis of automatic data analysis Wikipedia service is offered in the article. It is shown how the terminology database is defined, the network structure is created, weight of different terms-concepts is calculated by two different criteria – node degree and PageRank. Adequacy of the approaches offered is shown by example, and also the fact that clusters of terminological networks can be considered as a basis for detection of certain scientific directions.*

Keywords: *subject domain model, terminology-oriented database, Wikipedia, links of concepts, jurisprudence, scanning of network service.*

Постановка проблеми. Сьогодні під моделлю предметної області, зокрема, розуміють спеціальним чином сформовану мережу понять, онтологію. Побудова великої галузевої онтології – складна науково-практична проблема [1; 2]. Перший етап цього процесу – побудова термінологічної основи онтології і визначення семантичних зв’язків [3].

Аналіз останніх публікацій. Вивченню моделей предметних областей, так само як і сервісу Wikipedia (http://wikipedia.com), присвячена велика кількість робіт, що підтверджує актуальність проведених досліджень [4]. Серед них, зокрема, методи побудови мереж співавторів, визначення значущих вузлів, структури мережі, дослідження цитування, а також відповідних корпусів [5].

Пропонується методика побудови інформаційних мереж – моделей предметних областей на основі автоматичного моніторингу і аналізу мережевих інформаційних ресурсів довідкового характеру. Як така мережа в роботі розглядається мережа понять, що відповідають термінам-заголовкам статей мережевої енциклопедії Wikipedia.

Метою роботи є опис теоретичних принципів і методології та оцінка алгоритмічних засад побудови моделей предметних областей, зокрема, галузі юриспруденції шляхом моніторингу і аналізу мережевих інформаційних ресурсів довідкового характеру. Для досягнення цієї мети розроблено спеціальний алгоритм сканування ресурсів сервісу Wikipedia з метою отримання репрезентативного набору термінів-понять як основи (вузлів) майбутньої мережі.

Виклад основного матеріалу. Очевидно, мережа понять може мати досить великі розміри, якщо її не обмежувати певною тематикою, що відповідає предметній області. Ця властивість значно ускладнює сприйняття сформованої мережі і призводить до такого ефекту, як зсув тематики. Для подолання цього ефекту застосовується елементарна тематична фільтрація – для аналізу використовуються лише ті статті з Wikipedia, які містять базовий термін, що визначається експертом-аналітиком. Відповідність до цих дескрипторів і визначає розмір сформованих мереж – моделей предметних областей, а також динаміку їх формування. Крім того, розпізнавання кластерів в таких мережах може розглядатися як основа для виявлення окремих наукових напрямків [1].

Методика досліджень.

До розгляду було взято систему Wikipedia, що є доступною в глобальній мережі і не передбачає передплати, крім того, доступна для завантаження у повному обсязі. Для первинного доступу до системи було застосовано спеціальні терміни з юридичної проблематики, за якими існують відповідні статті, що створюються і редагуються експертами-авторами (Рис. 1).

З огляду на ці базові терміни (теги), що відповідають певній предметній області, визначено представлення інформації в цій системі. Також було визначено, що вільний перехід за гіперпосиланнями веде до ефекту так званого “зсуву тематик” (Topic Drift).

Розглядався наступний алгоритм побудови моделей предметних областей за даними сервісу Wikipedia, який передбачає уникнення цього ефекту:

1. Обирається перший термін-поняття, з якого починається зондування.
2. Відкривається сторінка веб-сервісу (стаття Wikipedia), що відповідає обраному терміну-поняттю. До створюваної мережі додаються всі терміни-поняття, що відповідають гіперпосиланням на обраній сторінці. Формуються ребра-зв’язки до цих вузлів з вихідного вузла.
3. Статті, що відповідають гіперпосиланням на попередній сторінці, визначаються як базові, якщо на них міститься гіперпосилання на статтю, що відповідає першому терміну-поняттю, з якого починалось зондування.
4. Із списку вузлів мережі, що формується, визначається той, за яким ще не здійснювалося переходу, на сторінку якого планується перейти для подальшого аналізу. Цей вузол має відповідати вимозі, наведеній у попередньому пункті, та не входить до складу тих вузлів, до сторінок яких вже був здійснений перехід.
5. Якщо такий вузол-автор обрано, то здійснюється перехід до пункту 2.
6. Якщо такого вузла не існує, то вважається, що мережу, що відповідає моделі предметної області, побудовано.

Відповідно до наведеного алгоритму процес збирання інформації з Wikipedia, починаючи з певного вузла-поняття, припиняється, коли відповідно до алгоритму вже неможливий перехід до нового вузла (базових вузлів для переходу вже не лишається), тобто “зациклювання” неможливе.

Рис. 1 – Інтерфейс користувача системи Wikipedia, розглядається стаття за терміном-поняттям ***Criminal Law***

Відповідно до наведеного алгоритму процес збирання інформації з Wikipedia, починаючи з певного вузла-поняття, припиняється, коли відповідно до алгоритму вже неможливий перехід до нового вузла (базових вузлів для переходу вже не лишається), тобто “зациклювання” неможливо.

Фрагмент траси виконання програми визначення термінологічної основи предметної області, що відповідає наведеному алгоритму і базовому терміну ***Family Law***, наведено на Рис. 2.

Отримані результати.

Побудовано відповідно до наведеного алгоритму мережі співавторів за базовими термінами-поняттями ***Constitutional law, Family Law, Criminal Law*** без обмежень на кількість сканованих вузлів. За допомогою програмного засобу Gephi отримана візуалізація мереж, що відповідають вибраним предметним областям (Рис. 3 – 5).

Отримані такі характеристики [5] побудованих мереж: ***Constitutional law*** – вузлів: 209, зв’язків: 1535, щільність: 0,035; ***Family Law*** – вузлів: 250, зв’язків: 3871, щільність: 0,062; ***Criminal Law: в***узлів: 1050, зв’язків: 64739, щільність: 0,059. Найбільш вагомі за двома критеріями (*ступенями вузлів і PageRank*) терміни-поняття, що відповідають вибраним предметним областям, наведено у Додатку.

Використання методів кластерного аналізу дозволяє виявляти найбільш тісно пов’язані між собою групи термінів-понять, що можуть застосовуватися для визначення нових наукових областей. На Рис. 3 показано приклад процесу виявлення кластерів шляхом застосування спеціального алгоритму, що застосовується в системі Gephi.

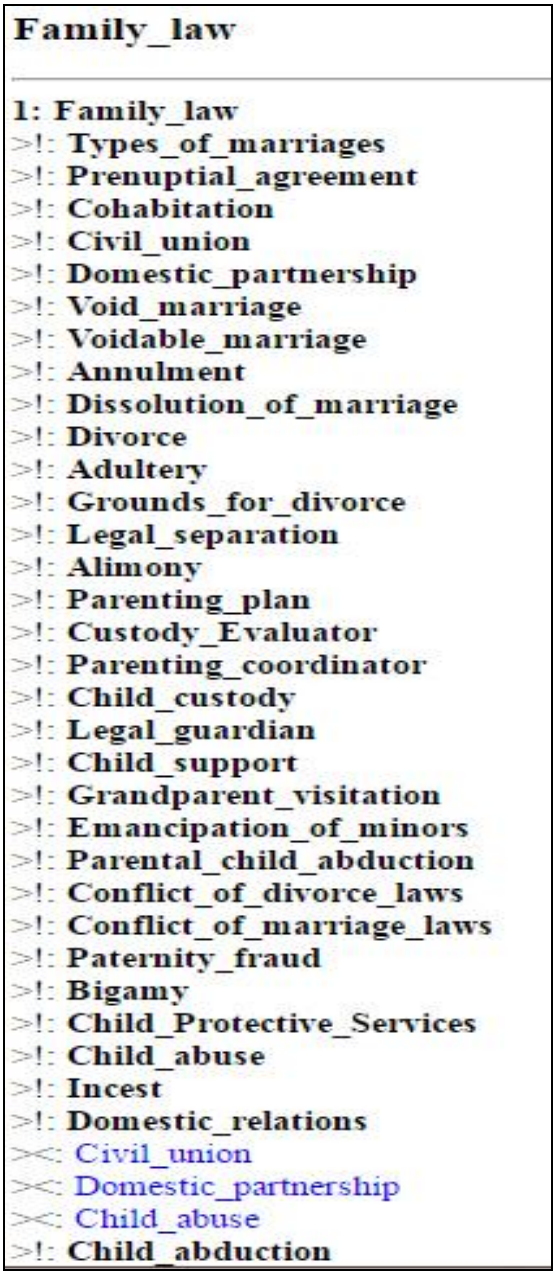

Рис. 2 – Фрагмент траси виконання програми

Висновки.

У роботі запропоновано і реалізовано алгоритм формування моделей предметних областей шляхом автоматичного аналізу мережевого сервісу Wikipedia. Від статичних моделей предметних областей таких підхід відрізняється урахуванням динамічної зміни контенту бази даних цього сервісу, урахуванням нових понять, феноменів, що з’являються, зокрема в юридичній області, що розглядається. При цьому підході визначальними елементами є назви нових статей як маркери знань (теги), що доповнюються авторами матеріалів – учасниками проекту Wikipedia.

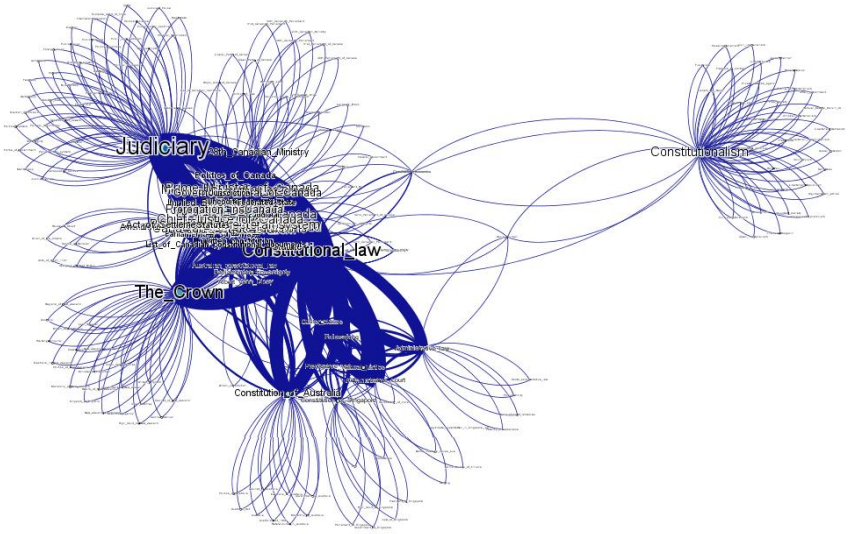

Рис. 3 – Мережа, що відповідає предметній області за темою ***Constitutional law***, окрема підмережа, кластер визначається терміном ***Constitutionalism***

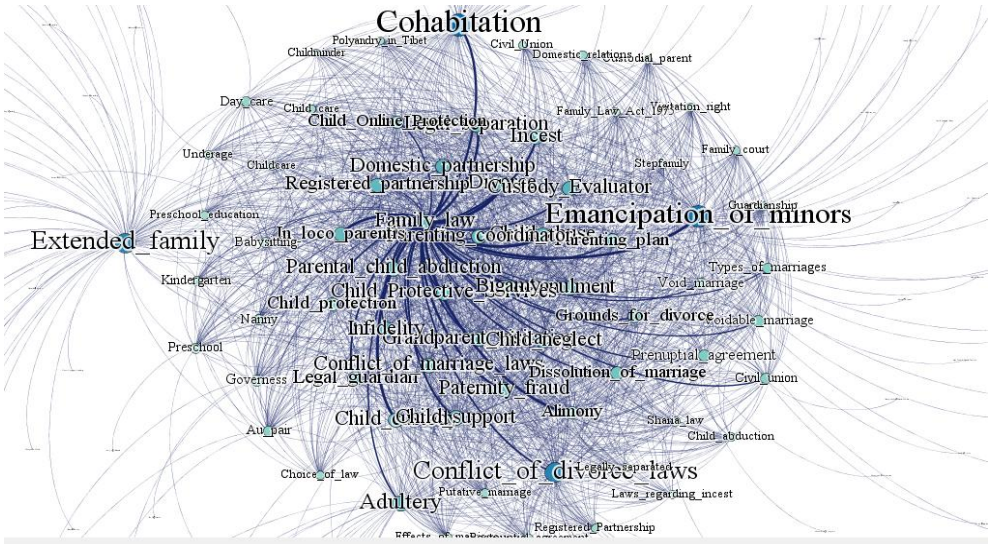

Рис. 4 – Мережа, що відповідає предметній області за темою ***Family Law***

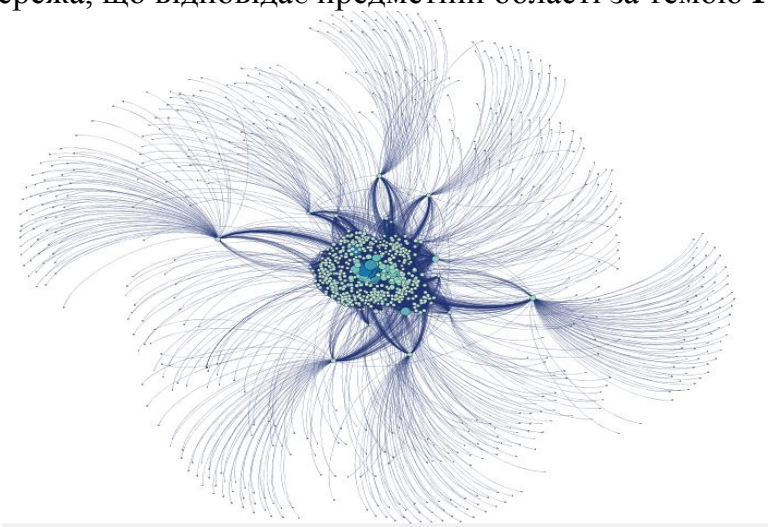

Рис. 5 – Мережа, що відповідає предметній області за темою ***Criminal Law***

Можна зауважити, що система Wikipedia, як і система Google Scholar Citations, що розглядалася раніше [6; 7], є зручною щодо доступу до інформації, не передбачає створення власного профілю користувача для доступу до інформації, доступ є необмеженим.

Слід відзначити принципову відмінність запропонованої моделі автоматичного формування термінологічних мереж від існуючих, що базуються на особистій участі експертів при виборі конкретних вузлів і зв’язків. У випадку, що описується в роботі, дослідник для побудови мережі використовує лише крупицю знань, представлену у вигляді назви першого, ключового терміну-поняття. Після цього програма використовує знання, закладені авторами (редакторами) статей в Wikipedia, теги, що визначаються внутрішніми гіперпосиланнями. У цьому випадку експертне середовище істотно розширюється.

Модель застосовувалася для юридичної науки в рамках сервісу Wikipedia, але запропонований підхід можна використовувати і для інших наукових областей, або для інших текстових масивів, зокрема, баз даних нормативно-правової інформації. Враховуючи дослідження та розробку алгоритмів для системи Wikipedia постає питання застосування цього алгоритму для інших сервісів, зокрема у галузі права, що потребує проведення порівняльного аналізу ресурсів.

Додаток.

Найбільш вагомі терміни-поняття, що відповідають вибраним предметним областям

Constitutional_law

№	Сортування за вагою	Переклад	Сортування за PageRank	Переклад
1	Judiciary	Судова влада	Constitutionalism	Конституціоналізм
2	The_Crown	Корона	Judiciary	Судова влада
3	Constitutional_law	Конституційне право	The_Crown	Корона
4	Constitution_of_Canada	Конституція Канади	Constitutional_law	Конституційне право
5	Parliament_of_Canada	Парламент Канади	Constitution_of_Australia	Конституція Австралії
6	Monarchy_of_Canada	Монархія Канади	Constitution_of_Canada	Конституція Канади
7	Prime_Minister_of_Canada	Прем’єр-міністр Канади	Parliament_of_Canada	Парламент Канади
8	Court_system_of_Canada	Судова система Канади	Monarchy_of_Canada	Монархія Канади
9	Elections_in_Canada	Вибори в Канаді	Prime_Minister_of_Canada	Прем’єр-міністр Канади
10	Canadian_electoral_system	Електоральна система Канади	Court_system_of_Canada	Судова система Канади
11	Chief_Justice_of_Canada	Головний суддя Канади	Elections_in_Canada	Вибори в Канаді
12	Constitutionalism	Конституціоналізм	Canadian_electoral_system	Електоральна система Канади
13	Canadian_federalism	Канадський парламентарізм	Chief_Justice_of_Canada	Головний суддя Канади
14	Canadian_Senate_divisions	Підрозділи Сенату Канади	Canadian_federalism	Канадський парламентарізм
15	Public_Service_of_Canada	Державна служба Канади	Canadian_Senate_divisions	Підрозділи Сенату Канади
16	Prorogation_in_Canada	Перерва у роботі парламенту в Канаді	29th_Canadian_Ministry	29-а Рада міністрів Канади

17	29th_Canadian_Ministry	29-а Рада міністрів Канади	28th_Canadian_Ministry	28-а Рада міністрів Канади
18	Cabinet_of_Canada	Кабінет міністрів Канади	Act_of_Settlement_1701	Акт про спадкування престолу
19	Governor_General_of Canada	Генерал-губернатор Канади	Prorogation_in_Canada	Перерва між парламентськими сесіями в Канаді
20	Patriation	Патріація	Statute	Статут

Family Law

№	Сортування за вагою	Переклад	Сортування за PageRank	Переклад
1	Cohabitation	Сожительство	Extended_family	Розширена сім'я
2	Emancipation_of_minors	Емансипація неповнолітніх осіб	Cohabitation	Сожительство
3	Conflict_of_divorce_laws	Конфлікт законів про розлучення	Emancipation_of_minors	Емансипація неповнолітніх осіб
4	Extended_family	Розширена сім’я	Conflict_of_divorce_laws	Конфлікт законів про розлучення
5	Adultery	Подружня зрада	Adultery	Подружня зрада
6	Incest	Інцест	Incest	Інцест
7	Infidelity	Невірність	Infidelity	Невірність
8	Domestic_partnership	Домашнє партнерство	Child_custody	Опіка над дітьми
9	Child_custody	Опіка над дітьми	Domestic_partnership	Домашнє партнерство
10	Child_support	Аліменти	Child_support	Аліменти
11	Family_law	Сімейне право	Child_neglect	Невиконання обов’язків щодо дитини
12	Child_abuse	Жорстоке поводження з дитиною	Family_law	Сімейне право
13	Divorce	Розлучення	Child_abuse	Жорстоке поводження з дитиною
14	Annulment	Анулювання	Divorce	Ррозлучення
15	Legal_separation	Роздільне проживання за рішенням суду	Annulment	Анулювання
16	Child_Protective_Services	Державне агентство в США CPS	Legal_separation	Роздільне проживання за рішенням суду
17	Conflict_of_marriage_laws	Конфлікт законів про розлучення	Child_neglect	Невиконання обов’язків щодо дитини
18	Paternity_fraud	Шахрайство з батьківством	Family_law	Сімейне право
19	Parental_child_abduction	Батьківське викрадення дітей	Child_abuse	Жорстоке поводження з дитиною
20	Custody_Evaluator	Оцінка, необхідна для опіки	Divorce	Розлучення

Criminal Law

№	Сортування за вагою	Переклад	Сортування за PageRank	Переклад
1	Contract	Контракт	War_crimes	Військові злочини
2	Property_law	Право власності	Forensic_psychology	Судова психологія
3	Criminal_law	Кримінальне право	Crime	Злочин

4	Tort	Делікт	Law	Закон
5	Damages	Відшкодування збитків	Manslaughter	Вбивство
6	Trust_law	Довірча власність	Cohabitation	Співжиття
7	Product_liability	Відповідальність виробника	Contract	Контракт
8	Vicarious_liability	Субсидіарна відповідальність	Emancipation_of_minors	Емансипація неповнолітніх осіб
9	Trespasser	Правопорушник	Tort	Делікт
10	Criminal_conversation	Перелюбство	Property_law	Право власності
11	Malicious_prosecution	Зловмисне судове переслідування	Trust_law	Довірча власність
12	Eggshell_skull	Правило підвищеної вразливості	Criminal_law	Кримінальне право
13	Invasion_of_privacy	Порушення приватності	Damages	Відшкодування збитків
14	Malpractice	Недобросовісна практика	Product_liability	Відповідальність виробника
15	Negligent_entrustment	Недбале ввірення	Vicarious_liability	Субсидіарна відповідальність
16	Medical_malpractice	Відповідальність медичних працівників	Trespasser	Правопорушник
17	Breach_of_promise	Порушення обіцянки	Criminal_conversation	Перелюбство
18	Legal_malpractice	Незаконна судова практика	Malicious_prosecution	Зловмисне судове переслідування
19	Invitee	Запрошений	Eggshell_skull	Правило підвищеної вразливості
20	Duty_to_rescue	Обов’язок порятунку	Invasion_of_privacy	Порушення приватності

Використана література

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

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

3. Чанышев О.Г. Автоматическое построение терминологической базы знаний : зб. трудов 10-й Всероссийской научной конференции [“Электронные библиотеки : перспективные методы и технологии, электронные коллекции” – RCDL’2008”]. – Дубна (Россия), 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. – Seattle (USA) : AAAI Press, March 2008. – Pp. 136-144.

5. 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.

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

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