**Дмитро Володимирович**
**ЛАНДЕ**

# ЕЛЕМЕНТИ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ В ПРАВОВІЙ ІНФОРМАТИЦІ

**Київ - 2014**

НАЦІОНАЛЬНА АКАДЕМІЯ ПРАВОВИХ НАУК УКРАЇНИ

НАУКОВО-ДОСЛІДНИЙ ІНСТИТУТ
ІНФОРМАТИКИ І ПРАВА

**Д.В. ЛАНДЕ**

# ЕЛЕМЕНТИ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ В ПРАВОВІЙ ІНФОРМАТИЦІ

Київ – 2014

УДК 340.13+681.3+519.8
ББК 22.18, 32.81, 60.54
Л95

*Рекомендовано до видання*
*Вченою радою Науково-дослідного інституту*
*інформатики і права Національної Академії*
*правових наук України*
*(протокол № 2 від 18 грудня 2013 року)*

Л95 **Ланде Д.В. Елементи комп'ютерної лінгвістики в правовій інформатиці. – К.: НДІІП НАПрН України, 2014. – 168 с.**

ISBN 978-966-2344-33-2

У монографії наведено засади застосування деяких елементів комп'ютерної лінгвістики в правовій інформатиці, опис окремих підходів, моделей, алгоритмів. Наведено необхідні відомості та окремі процедури виявлення значимих термінів із текстів, зокрема, із нормативно-правових актів, розглянуто приклади побудови мовних мереж, подано алгоритми виявлення подібних текстів, розглянуто принципи організації тезаурусів і онтологій. Також розглянуто питання формування електронних лінгвістичних ресурсів, корпусів, застосування сучасних мережевих інструментальних засобів, зокрема, технології MediaWiki для побудови електронних словників і енциклопедій з правової тематики.

Книгу призначено для широкого кола фахівців в галузях правознавства, прикладної лінгвістики, соціальних комунікацій, студентів і аспірантів з відповідних спеціальностей.

**Рецензенти:**
**Широков В.А.** – доктор технічних наук, академік НАН України,
**Бєляков К.І.** – доктор юридичних наук, професор,
**Мохор В.В.** – доктор технічних наук, професор

УДК 340.13+681.3+519.8
ББК 22.18, 32.81, 60.54

ISBN 978-966-2344-33-2 

# ЗМІСТ

# Вступ

Правова інформатика – це науковий напрямок, в рамках якого досліджуються інформація, інформаційні процеси і системи, що застосовуються в правовій сфері, на базі дослідження правових особливостей об'єктів, явищ і процесів, що вивчаються.

Правова інформатика є частиною більш загального напрямку – інформатики – «інтегральної науки про інформацію взагалі, що складається з трьох частин – теорії інформаційних елементів, теорії інформаційних процесів і теорії інформаційних систем» [*Темников, 1963*]. В Україні, як і в інших країнах колишнього СРСР, трактування терміну «інформатика» офіційно увійшло у науковий обіг у 1983 р. на сесії щорічних зборів Академії наук СРСР з моменту прийняття рішення щодо організації нового відділення інформатики, обчислювальної техніки і автоматизації. Інформатика трактувалася як «комплексна наукова та інженерна дисципліна, що вивчає всі аспекти розробки, проектування, створення, оцінки, функціонування систем обробки інформації, що базуються на ЕОМ, їх застосування і впливу на різні області соціальної практики». Таким чином, інформатика – це галузь людської діяльності, що пов'язана з процесами перетворення інформації за допомогою комп'ютерів і їх взаємодії із зовнішнім середовищем.

Основним об'єктом саме правової інформатики є правова інформація. Термін «правова інформація» у Законі України «Про інформацію» від 02.10.1992 № 2657-XII визначається як сукупність документованих або публічно оголошених відомостей про право, його систему, джерела, реалізацію, юридичні факти, правовідносини, правопорядок, правопорушення і боротьбу з ними та їх профілактику тощо.

Правова інформація досліджується за такими основними напрямками:

- визначення специфічних властивостей цієї інформації, як частини соціальної інформації;

– класифікація інформації, що обертається в правовій сфері;

– якісна і кількісна оцінка правової інформації;

– аналіз ролі інформації у прийнятті юридичних рішень.

На цей час необхідність взаємодії фахівців, урахування особливостей із різних професійних галузей, вимагає від юристів знання всіх технічних і інформаційних особливостей об'єктів, що розглядаються. Правова інформатика є джерелом знань і інструментарієм, необхідними для вирішення проблем правового регулювання суспільних відносин.

Відомо, що правова інформатика є багатопрофільним науковим напрямком, який найбільш тісно пов'язаний з кібернетикою – наукою, що займається вивченням закономірностей керування складними динамічними системами.

Окремий розділ кібернетики – правова кібернетика – займається вивченням особливостей процесів управління в правовій сфері. Правова кібернетика широко застосовується при дослідженні ефективності законодавчого регулювання суспільних відносин, зокрема шляхом соціально-правового моделювання [*Ланде, 2012а*].

Ще однією наукою, з якою тісно пов'язана правова інформатика, є семіотика, в рамках якої досліджуються властивості знакових систем, у тому числі природних і штучних мов. Семіотика має велике прикладне значення при дослідженні і створенні знакових систем, що застосовуються в процесах обігу інформації.

Особливу увагу у цій монографії присвячено зв'язку правової інформатики з наукою про мову – лінгвістикою. Як і уся інформатика, так і її складова – правова інформатика – базується на таких поняттях як алфавіт, слово, речення, текст.

Все більше значення у вирішенні завдань, що стоять перед правовою інформатикою як перед науковим напрямком, так і перед її практичними застосуваннями, приділяється лінгвістичному забезпеченню, а саме такій його частині, як

комп'ютерна лінгвістика, що охоплює проблематику використання природної мови у системах автоматичної обробки інформації, зокрема у галузі правової інформатики.

Розвиток комп'ютерної техніки дозволив на цей час автоматизувати такі трудомісткі процеси як статистичну обробку текстів, ведення словникових і лексикографічних баз даних. Методи лінгвістики застосовуються в правовій інформатиці, зокрема, для вирішення задач індексування і автоматичного реферування текстової інформації правової спрямованості, при упорядкуванні термінології: створенні відповідних словників, тезаурусів, онтологій. За останній час у зв'язку з бурхливим розвитком інформаційних технологій, глобальних комп'ютерних мереж та їх змістовної складової, зокрема, правової спрямованості, у галузі комп'ютерної лінгвістики було отримано значні наукові і практичні результати, створено: інформаційно-пошукові системи і сховища даних великих обсягів (Big Data), системи глибинного аналізу текстової інформації (Text Mining), системи машинного перекладу, системи синтезу усного мовлення, здійснено спроби аналізу усного мовлення тощо.

Комп'ютерна лінгвістика як складова прикладної лінгвістики доповнюється напрямками лінгвістичної експертизи (наприклад, у судовій практиці) і політичною лінгвістикою (аналіз суспільно-політичного дискурсу).

До основних проблем комп'ютерної лінгвістики відноситься проблема моделювання процесу розуміння текстів (переходу від тексту до формалізованого подання його змісту) і проблема синтезу мови (перехід від формалізованого подання змісту до текстів природною мовою). Ці проблеми виникають при вирішенні ряду практичних задач, таких як автоматичне виявлення і виправлення помилок при введенні текстів у бази даних, автоматичний аналіз і синтез мови, автоматичний переклад усної мови, пошук документів у повнотекстових базах даних.

Основні зусилля фахівців у галузі сучасної комп'ютерної лінгвістики націлені на створення потужних словників одиниць мови, вивчення їхньої семантико-синтаксичної структури,

розробку базових процедур морфологічного, семантико-синтаксичного і концептуального аналізу і синтезу текстів.

Таким чином, перед комп'ютерною лінгвістикою стоять, насамперед, задачі лінгвістичного забезпечення процесів збору, накопичення, обробки та пошуку інформації, серед яких найважливіші для правової інформатики наступні:

- автоматизація створення і обробки лексикографічних ресурсів (словників, тезаурусів, онтологій);
- корпусна лінгвістика (Corpus Linguistics) – створення і використання електронних корпусів текстів.
- автоматизація процесів виявлення і виправлення помилок при введенні текстів у бази даних;
- автоматичне індексування документів в базах даних;
- автоматична класифікація текстових документів;
- автоматичне реферування (Automatic Text Summarization) текстових документів і документальних масивів;
- виявлення подібних документів, плагіату;
- лінгвістичне забезпечення процесів пошуку інформації;
- автоматичний машинний переклад текстів;
- побудова природномовних інтерфейсів між користувачами і автоматизованими системами, розробока систем типу «питання-відповідь», зокрема, у галузі права;
- витяг (екстрагування) інформації (Information Extraction) із неструктурованих і слабко структурованих текстів (як елемент технології Text Mining);
- автоматичне розпізнавання мови.

Деякі з цих задач розглянуто в рамках цієї монографії, перший розділ якої присвячено питанням автоматичного аналізу текстових документів, зокрема, документів правової спрямованості; у цьому розділі також наводяться різні критерії значимості окремих лексичних одиниць. Розглянуто підходи до оцінки дискримінантної сили слів у текстах з правової тематики. Підходи перевірені на колекції законодавчих актів України та масиві новинних повідомлень. Запропоновано метод візуалізації рівня нерівномірності входження слів у тексти. Вводиться поняття мереж мови, розглядаються різні параметри цих мереж, що можуть застосовуватися в подальших процедурах лінгвістичної обробки, описується оригінальний метод побудови графів горизонтальної видимості для мереж мови, надаються приклади застосування цього методу до нормативно-правових документів України.

У другому розділі надається огляд методів, які активно застосовуються в задачах порівняльного аналізу текстових документів, що мають відношення до ряду таких технологічних напрямків, як інформаційний пошук, узагальнення та групування інформації. Розглянуто деякі теоретичні засади порівняльного аналізу, надаються та обговорюються основні теоретичні положення, що застосовуються у сучасних методах, переваги і недоліки систем, побудованих з урахуванням їх особливостей. Описано практичні реалізації підходів до порівняльного аналізу електронних текстів, пошуку подібних документів в системах контент-моніторингу, порівняльного аналізу різномовних текстів, побудованого на застосуванні систем статистичного перекладу, а також елементи побудови системи антиплагіату.

Третій розділ присвячено технології створення і ведення електронних словників і енциклопедій MediaWiki. Розглянуто деякі застосування цієї технології у задачах колективного документування, в корпоративних рішеннях, при створенні мережевої енциклопедії «Вікіпедія». Наведено необхідні дані щодо вікі-розмітки. Наведено опис технологій, що застосовуються у суспільно-значущому проекті WikiLeaks.

Четвертий розділ присвячено опису реалізованої у Науково-дослідному інституту інформатики і права

Національної Академії правових наук України діючої моделі електронної енциклопедії законодавства України, яка охоплює понятійно-категоріальний апарат, визначений законами, підзаконними актами та міжнародно-правовими документами. Використання технології MediaWiki в цій моделі відкриває широкі перспективи розширення набору інструментальних засобів для розробки уніфікованих довідкових і лексикографічних систем, що забезпечують можливість накопичення та обміну даними без розробки складних програмних комплексів.

У післямові наведено деякі висновки щодо особливостей застосування технологій комп'ютерної лінгвістики у галузі правової інформатики з урахуванням стрімкого зростання попиту до результатів застосування цих технологій.

Автор виражає щиру подяку В.В. Жигалу за внесок у розробку системи статистичного машинного перекладу, описаного у цій монографії, С.М. Брайчевському, за ідейну підтримку і адаптацію технології MediaWiki при створенні електронної енциклопедії законодавства України, А.О. Снарському, співавтору багатьох робіт щодо дослідження складних мереж, зокрема, мовних мереж, В.М. Фурашеву і Н.А. Савіновой за доречні зауваження щодо застосування методів комп'ютерної лінгвістики саме до галузі права, рецензентам В.А. Широкову, К.І. Бєлякову, В.В. Мохору за конструктивні зауваження, які було враховано у цій роботі.

# 1. Комп'ютерний аналіз нормативно-правових документів

Аналіз текстів нормативно-правових документів є, безумовно, актуальною задачею як правознавства, так і лінгвістики. Застосування засобів комп'ютерної техніки, як апаратних, так і програмних, безумовно, має сприяти підвищенню ефективності цих процесів. Якщо слідувати такій ієрархії: слово, словосполучення, документ, корпус документів, то розгляд процедур комп'ютерного аналізу логічно розпочати саме з комп'ютерної лексикографії.

## 1.1. Комп'ютерна лексикографія в правовій інформатиці

***Комп'ютерна лексикографія – частина комп'ютерної лінгвістики***

Традиційно лексикографія розглядалася як наука про упорядкування лексики – укладання словників, сам процес створення словників або як сукупність словникових творів. Але це розуміння лексикографії на цей час вже вважається занадто вузьким. Ще у 1936 році Л.В. Щерба писав, що робота лексикографа «повинна мати науковий характер і аж ніяк не зводитися до механічного зіставлення якихось готових елементів» [*Щерба, 1936*]. Сьогодні загальновизнано, що лексикографія – це самостійна наукова дисципліна, яка має свій предмет дослідження, свої наукові принципи, власну теоретичну проблематику [*Широков, 1998*], [*Широков, 2005*].

Зокрема, у роботі [*Широков, 2011*] сформульовано принципи, що ґрунтуються на базі теорій семантичних станів і лексикографічних систем, які змушують мовну субстанцію набувати словникової форми.

Комп'ютерна лексикографія може розглядатися, з одного боку, як один із напрямків комп'ютерної лінгвістики, а з іншого – як інструментально-орієнтована гілка загальної лексикографії, задачею якої є представлення словникової інформації у

комп'ютерних системах та забезпечення її функціонування у сучасному інформаційному просторі.

Безпосереднє відношення до становлення комп'ютерної лексикографії як окремої наукової гілки має так званий «лексикографічний ефект», який пояснює наявність системоутворюючих інваріантів лексикографічних систем. Цей ефект можна охарактеризувати як феноменологічний, оскільки він базується на загальних інформаційних властивостях систем і не прив'язаний до їх конкретної будови. Поряд з цим, цей ефект було виявлено і сформульовано саме для лексикографічних систем. У результаті спостережень та узагальнення поведінки різних систем було відзначено спільну для всіх відомих процесів ознаку фундаментального характеру [*Широков, 2011*], а саме те, що в процесі еволюції системи будь-якої природи в її структурі індукується деяка підсистема відносно сталих дискретних сутностей, які відіграють роль елементарних інформаційних одиниць всієї системи. При цьому всі інші феномени системи являють собою певним способом організовані комбінації цих інформаційних одиниць.

Зазначена вище підсистема має властивості, споріднені з властивостями лексичної системи природної мови: вона генерує в своїй структурі щось на зразок тезаурусу і граматики. При цьому сукупності елементарних інформаційних одиниць, як правило, мають відносну стабільність своїх характеристик. Саме ці явища й становлять зміст лексикографічного ефекту, який має значний потенціал операціональності, дозволяючи виявляти в системах відповідні комплекси елементарних інформаційних одиниць.

В контексті цієї роботи лексикографічний ефект виступає насамперед як феноменологічна і методологічна основа виявлення у якості елементарних інформаційних одиниць лексем, ключових слів, що мають складати основу побудови таких підсистем природної мови (або її сегментів), таких як термінологічні системи – словники, тезауруси, онтології.

Лексикографії відводиться особливе місце в правовій інформатиці як інструменту нормування юридичної лексики, зокрема в задачах індексування документів, реалізації

автоматичного пошуку і глибинного аналізу даних, накопичених у масивах нормативно-правової інформації, тощо. Комп'ютерна лексикографія в галузі права знаходить свою практичну реалізацію насамперед у термінологічних системах.

Необхідність побудови термінологічних системи в галузі правової інформатики обумовлюється необхідністю:

– коректного тлумачення термінів при застосуванні законів та інших нормативних актів з метою запобігання помилок, різних правових колізій;

– отримання законодавчого тлумачення термінів;

– отримання довідкової інформації з тих питань, на які існують відповіді в нормативно-правових і законодавчих актах;

– експертного аналізу термінології з нормативно-правових і законодавчих актів щодо дублювання, суперечностей, прогалин для подальшого їх усунення.

***Тезауруси***

Як інструмент, який часто використовується у складі лінгвістичного забезпечення інформаційних систем можна розглядати тезауруси [*Добров, 2009*] – структуровані списки ключових слів, призначених для однозначного подання концептуального змісту документів і запитів. Тезаурус упорядковується так, щоб встановити прозорі еквівалентні, гомографічні, ієрархічні та асоціативні зв'язки між термінами.

Тезаурус містить:

– дескриптори – слова та словосполучення, які однозначно позначають поняття з теми тезаурусу;

– недескриптори – слова та словосполучення, які у природній мові позначають ті самі поняття, що і дескриптори, або еквівалентні поняття;

– семантичні зв'язки (зв'язки на основі значень) між дескрипторами і не-дескрипторами, а також між самими дескрипторами.

Проблема омонімічності у тезаурусі вирішується тим, що кожне ключове слово надається у контексті, який робить це слово однозначним. Для вирішення проблеми синонімічності один із синонімів обирається як дескриптор, а іншим синонімам надається статус недескрипторів. Тільки дескриптори можуть використовуватись при індексуванні та формулюванні запитів, при цьому недескритори допомагають користувачам вибрати дескриптор. Якщо встановлено відповідність між ідентичними поняттями в різних мовах, користувач багатомовного тезаурусу може формулювати запити рідною мовою і шукати документи незалежно від мови, якою вони були індексовані.

Автором було розроблено основу Функціонального Класифікатора з питань державної служби [*Ланде, 1999*], який застосовується як класифікатор відповідного документального масиву нормативно-правових актів. При цьому Функціональний Класифікатор містить окремі терміни (слова і словосполучення) і їх визначення, які пов'язані з документами, в яких вони визначаються (містяться). Терміни пов'язані парадигматичними зв'язками з іншими термінами.

Як основа цього класифікатора застосовувався тезаурус [*ГОСТ 7.24-90*], [*ДСТУ 4032-2001*], в який були включені такі типи лексичних одиниць:

– окремі слова (іменники, прикметники, дієслова, прислівники);

– іменні словосполучення;

– лексично вагомі компоненти складних слів;

– абревіатури;

– скорочення слів та словосполучень.

Словосполучення включалися до словника, якщо опорним словом в них був іменник та виконувалася одна з умов:

– значення словосполучення не витікає із значень його складових;

– хоча б одна із складових словосполучення не використовується в складі інших словосполучень;

– словосполучення є стійким;

– окремі слова словосполучення мають надто широке значення;

– поділ словосполучень на окремі складові веде до втрачання важливих для пошуку парадигматичних зв’язків.

При побудові функціонального класифікатора з вибраної колекції документів були відокремлені необхідні слова та словосполучення, що відповідають предметній галузі класифікатора. Побудова понятійних та словникових статей передбачала, що лексичним одиницям приписуються покажчики, які відповідають стандарту ISO 2788, або відповідні українські позначки. У табл.1. наведено перелік таких покажчиків.

Табл. 1. Основні покажчики з ISO 2788

| Тип покажчика | Значення | Аналог по ISO 2788 |
|---|---|---|
| Посилання від аскриптора до дескриптора | дивись | USE |
| Посилання від дескриптора до синонімічного дескриптора або до аскриптора | синонім | UF (used for) |
| Посилання від аскриптора до комбінації дескрипторів | використай комбінацію | USE ... + ... |
| Посилання від дескриптора до вищого дескриптора | вище | BT (broader term) |
| Посилання від дескриптора до вищого родового дескриптора | вище-рід | BTG (broader term generic) |
| Посилання від дескриптора до вищого дескриптора, який означає ціле | вище-ціле | BTP (broader term partitive) |
| Посилання від дескриптора до нижчого дескриптора | нижче | NT (narrower term) |

| Тип покажчика | Значення | Аналог по ISO 2788 |
|---|---|---|
| Посилання від дескриптора до нижчого видового дескриптора | нижче-вид | NTG (narrower term generic) |
| Посилання від дескриптора до нижчого видового дескриптора, який означає частину | нижче-частина | NTP (narrower term partitive) |
| Посилання від дескриптора до дескриптора, який зв'язаний асоціативно | асоціація | RT (related term) |

Покажчик визначає зв'язки між лексичними одиницями або поняттями та є результатом виконання таких операцій, як:

- усування неоднозначностей лексичних одиниць;
- встановлення відносин синонімії;
- вибір дескриптора, який відповідає за весь клас синонімії при індексуванні;
- встановлення ієрархічних та асоціативних відносин дескрипторів.

Множина комп'ютерних інструментальних засобів лексикографії поділяється на: 1 – програми підтримки лексикографічних робіт і 2 – автоматичні словники, тезауруси, онтології, що базуються на застосуванні спеціальних алгоритмів і лексикографічних баз даних [*Amsler, 1982*]

Існують програмні комплекси, що поєднують властивості першої і другої груп, наприклад, «Вікісловник» – лексикографічне середовище Wiktionary, до опису якого звернемося нижче.

Ще один такий приклад – WordNet — електронний тезаурус/семантична мережа, що разом з відповідним програмним забезпеченням з вільною ліцензією була розроблена для англійської мови у Прінстонському університеті, США, а

зараз знайшла розвиток для різних мов [*Fellbaum, 2005*]. Слова в ній організовані в синонімічні групи (синсети – синонімічні ряди слів, що виражають спільне значення); групи зв'язані одна з одною відносинами антонімії, гіперонімії, гіпонімії та інше, тобто, це інформаційний ресурс, що відображає складні відносини між лексичними одиницями мови. Зупинимося на ньому детальніше.

В WordNet словниковими статтями є синсети – множини слів-синонімів, що позначають деякий концепт у заданому контексті. Кожна словникова стаття має тлумачення, що не допускає неоднозначного розуміння. Для синсета явно вказуються частина мови й тлумачення. Кожне слово, що входить до складу синсета, може додатково мати ряд атрибутів, наприклад, ознаки домінантності, посилання типу «ідіома», «близьке значення» і т.ін. Для кожного слова може бути наведено приклад його вживання в заданому контексті – визначається набір висловів і фразеологізмів, також визначаються тлумачення.

При формуванні «синсетів» частотність вживання використається для впорядкування елементів синсета: виділяється «домінанта» – найбільше часто використовуване нейтральне слово для вираження лексикалізованого поняття – і другорядні елементи синсета, які істотно поступаються домінанті в частоті використання.

Статистико-комбінаторні характеристики контекстів застосовуються для виявлення типових для цього варіанта слова схем сполучуваності, вони заносяться в WordNet у вигляді переліків валентностей, які задаються у формально-граматичному, значеннєвому та синтаксичному планах.

Європейські проекти EuroWordNet і BalkanNet забезпечують роботу з WordNet практично всіма основними європейськими мовами. При цьому зв'язок різних мовних версій WordNet здійснюється через міжмовний індекс (Inter-Lingual-Index – ILI), загальний для всіх версій. На цей час вже існує підхід для побудови багатомовного, у тому числі й україномовного WordNet загального призначення (Розпорядження КМ України від 17 липня 2003 року № 415-р

«Про затвердження плану заходів щодо створення української лінгвістичної системи в мережі Інтернет: український варіант системи WordNet (UkrWordNet)») [*Анисимов, 2005*] [*Якименко, 2005*].

WordNet також може бути представлений як лексична онтологія – одна з компонентів Семантичного вебу – технології W3C. Таке представлення дозволяє різним програмним агентам інтерпретувати дані з різних систем.

Семантичні словники WordNet можуть бути описані засобами мови OWL (Ontology Web Language) і представляти один з ресурсів Семантичного вебу.

Jur-WordNet (Jur-WN) – розширення італійського WordNet – ItalWordNet (IWN) – лексикографічної бази даних, що забезпечує багатомовний інтерфейс доступу до джерел правової інформації [*Sagri, 2004*]. Застосовуючи засоби штучного інтелекту в галузі права в IWN реалізована спроба використання мови фреймів для концептуального представлення правових знань.

Проект EuroWordNet зберігає основну базову конструкцію WordNet, при цьому забезпечується розширення набору лексичних відношень. IWN на цей час складається з 70 тис. слів, організованих у 50 тис. синсетів. Онтологія в правовій сфері – jur-WN. Jur-WN – це багатошаровий лексикографічний ресурс, який містить великий набір семантичних відношень (успадкованих від лінгвістичної конструкції загальної бази даних IWN). Jur-WN являє собою лінгвістичну онтологію для правової галузі, що забезпечує можливості якісного пошуку правової інформації (законодавства, судових справ, політики) в багатомовних джерелах. Jur-WN охоплює юридичну лексику (11 тис. ключових слів, 12 тис. біграм), що дозволяє коректно обробляти інформацію з урахуванням таких мовних явищ як полісемія і синонімія (містить 2000 синсетів), забезпечує взаємодію з правовими ресурсами користувачів, які не є юристами.

При аналізі текстів, зокрема, нормативно-правових актів, необхідно в автоматизованому режимі особливим чином ідентифікувати відмінні особливості одиниць тексту. Для цього

найважливішим засобом є конкорданс – лексикографічний твір, що є переліком усіх випадків вживання кожного слова у визначеному тексті. Кожен випадок слововживання доповнюється інформацією про контекст, про позицію лексичної одиниці, про її словесне оточення. Конкордансі можуть використовуватися для дослідження сполучень лексичних одиниць, нюансів значень, як джерело для лексикографічних прикладів застосувань лексичних одиниць. З іншого боку, конкорданс є спеціалізованою лінгвістичною прикладною програмою, за допомогою якої здійснюється автоматична вибірка заданих мовних одиниць з електронних текстів. Функцію конкордансу є аналіз відразу багатьох текстів або корпусів електронних текстів, при цьому конкорданс надає користувачеві інформацію щодо контексту використання заданих мовних одиниць. Конкорданс може надавати інформацію про частотність вживання і сполучуваності тієї або іншої мовної одиниці, а також звертатися до тексту, в якому був знайдений приклад.

Характеристики конкордансів залежать від таких параметрів, як повнота опису, організація контексту, мовні або понятійні підходи тощо.

Якісно розрізняються конкорданси типу KWIC (Keyword In Context) – ключове слово у контексті та типу KWOC (Keyword Out of Context) – ключове слово поза контекстом [*Дубичинский, 2008*].

На погляд автора, правова наука, є «передовим рубежем» застосування комп'ютерної лексикографії. Невизначеність, багатозначність, навіть суперечливість значень окремих слів, термінів, понять у нормативно правових документах суттєво ускладнюють умови існування, взаємодії людей, суспільства, держави практично в усіх галузях життя.

Особливість юридичної лексики, зокрема, полягає у використанні значного числа антонімів, оскільки право регулює інтереси, що відрізняються своєю протилежною спрямованістю. Крім того, в правовій лексиці присутня велика кількість синонімів, які у ряді випадків мають різне смислове навантаження. Поширеною також є омонімія (однакові за

формою слова, які мають різні граматичні та/або лексичні значення) і полісемія (коли одні й ті ж юридичні терміни мають декілька різних значень).

До того ж додається потреба гармонізації нормативно-правових документів нашої держави із відповідними міжнародними документами, що визначається сучасними світовими інтеграційними процесами. Саме для цього створено і адаптовано в Україні міжнародний тезаурус EUROVOC – багатомовний політематичний інформаційно-пошуковий тезаурус, визнаний як міжнародний термінологічний стандарт. Він реалізований відповідно до стандартів ISO 2788-1986 «Guidelines for the establishment and development of monolingual thesauri» («Посібник з введення і розробки одномовних тезаурусів») та ISO 5964-1985 «Guidelines for the establishment and development of multilingual thesauri» («Посібник з введення і розробки багатомовних тезаурусів»).

EUROVOC охоплює всі основні теми, важливі для діяльності європейських інституцій: політика, міжнародні відносини, європейські співтовариства, законодавство, економіка, торгівля, фінанси, соціальні питання, освіта і комунікації тощо. Цей тезаурус реалізований всіма офіційними мовами Європейського Союзу.

EUROVOC має дворівневу ієрархію, верхній рівень якої складають теми – їм відповідають двохсимвольні коди. Нижній рівень організовано як сукупність мікротезаурусів (позначених чотирма цифрами). Нумерація тем і мікротезаурусів єдина для всіх мов.

У програмній реалізації у середовищі Windows на екрані системи EUROVOC одночасно представлені дві панелі, які ілюструють вибраний рівень ієрархії: список тем і мікротезаурусів, або список мікротезаурусів і зміст вибраного мікротезаурусу, або мікротезаурус і його окремий дескриптор.

На рівні окремих дескрипторів і недескрипторів структура EUROVOC залежить від семантичних відношень, встановлених між ними. Передбачено такі їх типи:

«SN» (Scope Note, примітка щодо можливих значень) – визначення, що уточнює значення дескриптора, або вказівка, як використовувати дескриптор при індексуванні документа та формулюванні запитів;

«MT» (Microthesaurus, мікротезаурус) – посилання на мікротезаурус, до якого належить дескриптор (недескриптор);

«UF» (Used For, використаний для) та «USE» (використовує) – зв'язок еквівалентності між дескриптором і недескриптором (-ами), (UF), або між недескриптором і дескриптором, який замінює цей недескриптор (USE). Фактично зв'язок еквівалентності охоплює кілька типів зв'язків:

- повної синонімічності або ідентичного значення;
- близької синонімічності або схожого значення;
- антонімії або протилежного значення;
- включення, коли дескриптор охоплює одне або більше понять, яким надано статус недескрипторів;
- ієрархічні зв'язки між дескрипторами.

Існують такі зв'язки між дескрипторами:

«BT» (Broader Term, ширший термін) – між певним і родовим (більш узагальненим) дескриптором – зазначається з числом, яке показує кількість кроків за ієрархією між ними;

«NT» (Narrower Term, більш вузький термін) — між родовим і видовим (більш вузьким) дескриптором — зазначається з числом, яке показує кількість кроків за ієрархією між ними;

«RT» (Related Term, взаємозв'язані терміни) — асоціативні зв'язки між дескрипторами. Асоціативний зв'язок показує, що існує інший релевантний дескриптор. Передбачено асоціативні зв'язки таких типів: причини та наслідку; органу або інструменту; ієрархії (оскільки, як сказано вище, поліієрархія не припускається, втрачені ієрархічні зв'язки можна замінити асоціативними); супроводження; послідовності у часі або

просторі; входження до складу; характерної риси; об'єкта дії або процесу; розташування; подібності; антонімії.

Асоціативні зв'язки мають такі істотні характеристики:

- вони симетричні;
- вони несумісні з ієрархічними зв'язками – якщо два дескриптори пов'язані ієрархією, між ними не можна встановити асоціативний зв'язок і навпаки;
- між дескрипторами, які мають спільний термін верхнього рівня, не може бути встановлено асоціативні зв'язки.

Навігація за тезаурусом здійснюється за допомогою посилань. Дескриптор можна вибрати, набравши на клавіатурі першу літеру його назви. Також у програмному забезпеченні EUROVOC реалізовано повнотекстовий пошук і пошук за ключовими словами.

Серед успішних спроб створення електронних лексикографічних ресурсів правової спрямованості ще можна назвати, російський ресурс «Толковый словарь современной информационно-правовой лексики» (www.morepc.ru/informatisation/ dic.html), Словопедія – «Словник термінів, уживаних у чинному Законодавстві України» (http://slovopedia.org.ua/), ЛІГА:ЗАКОН: Термінологічний словник юриста (http://liga.biz-plan.com.ua/resursyi/spravochniki-i-instrumentyi/ terminologicheskij-slovar-yurista.html), НАУ: Словник законодавчих термінів (http://zakon.nau.ua) тощо.

***Онтології***

Термін «онтологія» використовується в декількох областях знань і має два різних значення:

- філософська дисципліна, яка вивчає найбільш загальні характеристики буття і сутностей;
- в інженерії знань: артефакт, структура, модель знань, що описує значення елементів деякої системи.

У філософії онтологією називають теорію про природу буття і види сутностей. В інженерії знань онтологічний рівень формалізує накопичені знання, визначаючи і поєднуючи термінологію різних предметних сфер. Таким чином, чіткої взаємної обумовленості між значеннями терміна «онтологія» у філософії і в інженерії знань не простежується. Зв'язок між ними носить скоріше асоціативний характер.

Незважаючи на існування великої кількості напрацювань у галузі представлення знань, не існує єдиного чіткого визначення онтологій. Під онтологією в рамках цієї роботи розумітимемо систему понять предметної галузі, яка представляється як набір сутностей, що об'єднані різними відношеннями.

Онтології отримують досить широке поширення в задачах представлення знань, семантичної інтеграції інформаційних ресурсів, інформаційного пошуку і т.ін. У науці про «штучний інтелект» онтологія – це «специфікація концептуалізації предметної області», або спрощено, документ або файл, що формально задає зв'язки між поняттями. Це свого роду словник понять предметної сфери і сукупність явно визначених припущень щодо змісту цих понять. Найчастіше онтологія представляється як ієрархія понять, пов'язаних відношеннями певних видів. Розвинуті онтології формалізуються засобами мов логіки і допускають можливості формування логічних тверджень.

Онтології використовуються для формальної специфікації понять і відношень, які характеризують певну область знань. Перевагою онтологій як способу представлення знань є їх формальна структура, яка спрощує їх комп'ютерну обробку.

Терміну «онтологія» задовольняє широкий спектр структур, що представляють знання про ту чи іншої предметної області. Так до онтології можна віднести ряд структур, що відрізняються різним мірою формалізації:

- глосарій;
- проста таксономія;

- тезаурус (таксономія з термінами);
- понятійна структура з довільним набором відношень;
- повністю аксіоматизована теорія.

У загальному вигляді структура онтології являє собою набір елементів чотирьох категорій:

- поняття;
- відношення;
- аксіоми;
- окремі екземпляри.

Дослідники виділяють прикладні онтології, онтології області знання, загальні (родові) онтології і репрезентаційні онтології (йдеться щодо онтологій метарівня, що включають в себе репрезентаційні першоелементи).

Онтології можуть бути також розділені на одномовні і багатомовні. Вже існує ряд онтологій, орієнтованих на представлення знань на декількох мовах, наприклад, EuroWordNet, MikroKosmos і деякі інші.

Також виділяється особливий тип онтологій – лексичні онтології (або лінгвістичні). Відмітною властивістю таких онтологій є «фіксація в одному ресурсі понять (слів) разом з їх мовними властивостями». Такі онтології тісно взаємопов'язані з семантикою граматичних елементів (слів, іменних груп та ін.) Основним джерелом понять в онтологіях цього типу є значення мовних одиниць. Їх також відрізняє своєрідний набір відношень, зазвичай властивий для мовних елементів: синонімія, гіпонімія, меронімія, а також ряд інших. До лінгвістичних онтології автори [*Соловьев, 2006*] відносять WordNet, MikroKosmos, Sensus, РуТез та інші. Коло завдань, що вирішуються такими онтологіями, тісно взаємопов'язане з обробкою природної мови.

Онтології, зокрема, можна ефективно використовувати для підвищення точності інформаційного пошуку – пошукова система буде видавати тільки такі документи, де потрібне

поняття згадується точно, а не ті, у текстах яких зустрілося задане ключове слово.

У Стенфордському університеті розроблена програмна платформа – редактор онтологій Protégé (http://protege.stanford.edu), а також організовано співтовариство ентузіастів, що налічує декілька тисяч учасників, які поповнюють базу онтологій для самих різних предметних галузей. Заслуговує на увагу також проект Estrella (www.estrellaproject.org, рис. 1), в рамках якого розроблено онтологію LKIF (Legal Knowledge Interchange Format) – мову для подання юридичних знань та обміну між правовими інформаційними базами.

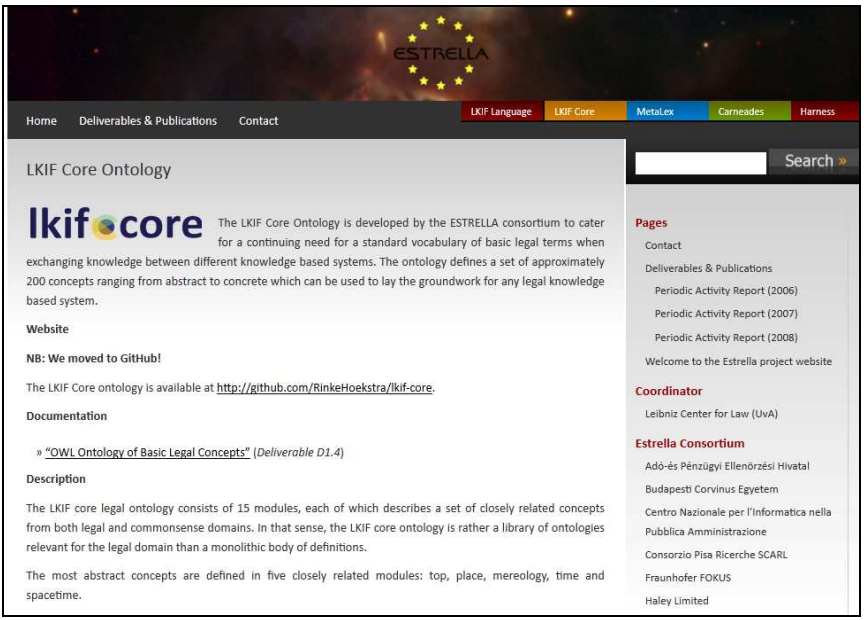



Рис. 1 – Сторінка LKIF на веб-сайті проекту Estrella

На даний момент для розробки систем, заснованих на знаннях, є актуальною задача об'єднання різних репрезентативних підходів з метою забезпечення найбільш повного подання знань у правовій сфері. У рамках розробки базової правової онтології LKIF-Rus на базі міжнародної

онтології LKIF-Core (LKIF – Legal Knowledge Interchange Format) було створено онтологію цивільного права LKIF-CivilRus.

Сучасні засоби онтологічного моделювання дозволяють частково впровадити продуктивний підхід в процес розробки онтології. Для цього, наприклад, можна використовувати SWRL-правила (SWRL – Semantic Web Rule Language), підтримка яких включена в середовище розробки Protégé.

У рамках онтології LKIF-CivilRus розроблена група SWRL-правил, які регулюють інститут дійсності угод у цивільному праві – залежно від дієздатності та волі суб'єктів, а також дотримання встановленої законом форми угоди.

Онтологічне дослідження основ кримінального права і розробка узагальненої онтологію цієї предметної області LKIF-CrimRus відбувалося шляхом розширення російськомовної онтологію верхнього рівня для системи російського права LKIF-Rus, в свою чергу заснованої на базовій юридичної онтології LKIF-Core. Розроблена онтологія заснована на правових нормах, що містяться в частині першій Кримінального кодексу Російської федерації. Онтологія розроблена за допомогою онтологічного редактора Protégé 3.4.4. Мовою опису онтології в Protégé є OWL. Робота з SWRL-правилами реалізована в плагіні SWRLTab для Protégé з використанням апарату логічного виводу Jess Rule Engine. Для візуалізації онтології було використано плагін Ontoviz на основі генератора діаграм Graphviz. Для виконання запитів до онтології використовується мова запитів SPARQL.

Формалізація правових норм є важливим засобом досягнення збігу або кореляції словника конкретного суб'єкта кваліфікації злочинів з уніфікованим словником термінів кримінального законодавства. Зокрема, онтологія, що охоплює ключові поняття і категорії усіх норм кримінального законодавства [*Вороніна, 2010*], забезпечує понятійну сумісність і єдність інформаційно-пошукової мови різних онтологій в області кваліфікації злочинів, індексування в процесі кваліфікації даних і пошук необхідної інформації для оцінки дій осіб, що вчинили суспільно небезпечні діяння.

На сьогоднішній день створено варіант базової правової онтології для системи російського права, що має 8 рівнів ієрархії і включає 127 класів і 108 відношень. При адаптації онтології LKIF-Core для російської правової системи були запозичені основні абстрактні концепції: частина-ціле, просторово-часові відношення, класифікація матеріальних об'єктів.

З правових інструментів використана система правової кваліфікації за допомогою таких сутностей, як «Судження», «Відношення_до_Суждения», «Кваліфікація», «Кваліфікований». Найбільш явні відмінності в правових системах були виявлені при деталізації таких концептів, як «Джерело» і «Суб'єкт». Спадкоємці цих класів визначені виключно російської теорією держави і права.

Існують деякі особливості використання онтологій для подання юридичних знань:

1. По мірі розвитку будь-якої правової системи в нормативні акти вводяться нові або видаляються попередні причинно-наслідкові зв'язки, що може призвести до необхідності перевизначення термінів, зміни їх положення в таксономії. Таким чином онтологію необхідно постійно змінювати. У зв'язку з цим в онтологічному моделюванні є цілий напрямок – управління версіями (versioning).

2. Розробник онтології не може гарантувати, що визначення повністю відобразить сенс юридичного поняття (принаймні, якщо це визначення не з нормативного акту).

3. Не завжди можливим є вираз причинно-наслідкових сутностей правових явищ.

4. Протиріччя між вимогою однозначного визначення термінів у рамках правової інформації та практикою призводить до неузгодженості онтології, що неприпустимо.

Серед проблем, які виникають в процесі розробки міжнародних правових онтологій, найбільш суттєвими можна назвати наступні:

1) відмінності в правових системах і, як наслідок, у правовому понятійному апараті;

2) багатозначність деяких термінів, синонімія;

3) проблеми при позначенні відношень і особливо зворотних відношень.

## 1.2. Комп'ютерний аналіз значимості термінів

На даний час актуальним є завдання визначення того, які з важливих структурних елементів тексту виявляються інформаційно-значущими, такими, що визначають інформаційну структуру тексту. Використання таких елементів як опорних слів дозволяє формувати онтології, тезауруси, пошукові образи, зокрема, при обробці нормативно-правової інформації. Такі елементи можуть використовуватися також для багатьох процедур, що охоплюються концепцією Text Mining, наприклад, пошук подібних документів, виявлення дублікатів, побудова сніпетів, інформаційних портретів, ідентифікації таких компонент тексту, як коллокації, надфразова єдність [*Ягунова, 2012*] тощо.

Ключові слова для пошуку в тексті, опорні слова для автоматичного екстрагування значущих фрагментів текстів або формування автоматичних рефератів, вибираються з урахуванням такої властивості слів, як «розпізнавальна» або дискримінантна сила [*Ланде, 2012*]. При аналізі текстів з правової тематики, зокрема, при вирішенні завдання формування електронної енциклопедії законодавства на основі аналізу всього масиву законодавчих актів України, оцінка дискримінантної сили окремих слів має найважливіше значення.

Опорні слова можуть виділятися шляхом застосування деяких шаблонів-маркерів (сигнальних слів), що знаходяться в тексті, граматичного розбору текстів і побудови синтаксичних мереж слів, або на основі деяких статистичних ознак.

Як приклади сигнальних слів для нормативно правових актів можна навести такі: «під … розуміється …»; «до ... належать ...»; «до ... відноситься ...»; «термін … означає ...»; «термін … складає: ...» тощо.

Статистичні підходи базуються на тому, що якщо слово відносно рівномірно розподілено по тексту документа, то воно

навряд чи може використовуватися для ефективного змістовного пошуку або служити основою вибору якогось значущого фрагмента, який може розглядатися як деяка надфразова єдність. При аналізі текстів з правової тематики, зокрема, при вирішенні завдання формування електронної енциклопедії на основі аналізу всього масиву законодавчих актів України, оцінка дискримінантної сили окремих слів має найважливіше значення.

Більшість відомих інформаційно-пошукових систем і систем класифікації інформації тією чи іншою мірою ґрунтуються на використанні векторно-просторової моделі пошуку (Vector-Space Model), і, відповідно, опису даних [*Salton, 1975*], [*Salton, 1983*], [*Salton, 1988*].

У векторно-просторовій моделі пошуку для урахування дискримінантної сили слів було введено поняття інверсної частоти появи слова в окремих документах масиву. Запропонований метод зважування слів має сьогодні стандартне позначення – TF IDF, де TF вказує на частоту появи слів у документі, а IDF – на величину, зворотну до кількості документів у масиві, що містять дане слово (точніше, логарифм, монотонну функцію від цієї величини):

Дана модель є класичною алгебраїчною моделлю. У рамках цієї моделі документ описується вектором у деякому евклідовому просторі, у якому кожному терму, який використовується в документі, ставиться у відповідність його вага (значимість), що визначається на основі статистичної інформації про його появу у окремому документі або в документальному масиві. Опис запиту, що відповідає необхідній користувачу тематиці, також являє собою вектор у тому ж евклідовому просторі термів. У результаті для оцінки близькості запиту та документа використовується скалярний добуток відповідних векторів опису тематики (запиту) та документа.

У рамках цієї моделі кожному терму $t_i$ у документі $d_j$ (і запиті $q$) зіставляється деяка невід'ємна вага $w_{ij}$. Таким чином, кожен документ і запит можуть бути представлені у вигляді $k$-вимірного вектора $||w_{ij}||_i = 1, ..., k$, де $k$ – загальна кількість різних термів у всіх документах (у словнику). Один з можливих

найпростіших підходів – використати як вагу терму $w_{ij}$ у документі $d_i$ нормалізовану частоту його використання $\text{freq}_{ij}$ у цьому документі, тобто:

$$w_{ij} = tf_{ij} = \text{freq}_{ij} / \max(\text{freq}_i).$$

Цей підхід не враховує частоту окремого терму, який використовується у всьому інформаційному масиві, так звану, дискримінаційну силу терму. Тому у випадку, коли доступна статистика використань термів у всьому інформаційному масиві, більш ефективне наступне правило обчислення ваги:

$$w_{ij} = tf_{ij} \log N / n_i,$$

де $n_i$ – число документів, у яких використовується терм $t_j$, а $N$ – загальне число документів у масиві.

Звичайно значення ваги $w_{ij}$ нормуються, що дозволяє розглядати документ як ортонормований вектор. Такий метод зважування термів має стандартне позначення – TF IDF, де TF вказує на частоту появи терміна в документі (Term Frequency), а IDF – на величину, обернену числу документів масиву, що містять даний терм (Inverse Document Frequency).

Коли виникає завдання визначення тематичної близькості двох документів або документа і запиту, в цій моделі використовується простий скалярний добуток $sim(d_1,d_2)$ двох векторів $\|w_{i1}\|_{i=1,..,k}$ і $\|w_{i2}\|_{i=1,..,k}$ який, очевидно, відповідає косинусу кута між векторами – образами документів $d_1$ і $d_2$. Очевидно, $sim(d_1,d_2)$ належить діапазону [0,1]. Чим більша величина $sim(d_1,d_2)$ – тим більш близькі документи $d_1$ й $d_2$. Для будь-якого документа $d$ маємо $sim(d,d)=1$. Аналогічно, мірою близькості запиту $q$ документа $d$ уважається величина $sim(q,d)$.

Оцінка нерівномірності входження слів можлива і на основі чисто статистичних, дисперсійних оцінок. В роботі [*Ortuño, 2003*] запропонована така оцінка дискримінантної сили слова:

$$\sigma_i = \frac{\sqrt{\langle d^2 \rangle - \langle d \rangle^2}}{\langle d \rangle},$$

де: $\langle d \rangle$ – середнє значення послідовності $d_1$, $d_2$,..., $d_n$, $n$ – кількість появ слова $t_i$ в інформаційному масиві.

Якщо позначити координати (номери) входження слова $t_i$ в інформаційний масив як $e_1$, $e_2$,..., $e_n$, то $d_k = e_{k+1} - e_k \quad (e_0 = 0)$.

Для візуалізації нерівномірності входження слів в тексти в [*Ortuño, 2003*] було запропоновано технологію спектограм, які зовні нагадують штрих-коди товарів [*Carpena, 2009*], разом з тим не дозволяють розглядати входження слів у різних масштабах вимірювань, як це робиться, наприклад, у вейвлет-аналізі [*Чуи, 2001*].

Автором запропоновані та реалізовані інструментальні засоби, що дозволяють візуалізувати щільність появи слова в тексті в залежності від ширини вікна спостереження. Через веб-інтерфейс відповідної програми вводиться текст і слово для аналізу. У результуючій спектрограмі по горизонталі відкладаються номери входження слів у тексті, а по вертикалі – ширина вікна спостереження. Одному входженню слова відповідає світло-сірий колір. Якщо у відповідне вікно спостереження потрапляє кілька цільових слів, то воно зафарбовується більш темним відтінком. Експерт – прикладний лінгвіст за зовнішнім виглядом відразу може визначити міру рівномірності входження в текст слова, що аналізується [*Ландэ, 2009*].

Коефіцієнти нерівномірності входження окремих слів у добірці законодавчих актів України, що відносяться до формування та розвитку інформаційного простору держави (закони України «Про доступ до публічної інформації», «Про Основні засади розвитку інформаційного суспільства в Україні на 2007-2015 роки», «Про телекомунікації», «Про захист персональних даних», «Про основи національної безпеки України»), які було розраховано автором, наведено у Табл. 2, а відповідні спектрограми – на рис. 2 – 6.

При розрахунку коефіцієнта $w_i$ використовувався штучний прийом, вихідний текст розбивався на фрагменти фіксованої довжини по 500 слів, які при розрахунках TF IDF

розглядаються як окремі фрагменти документів. Як видно, нерівномірність входження окремих слів, що точно виражається в коефіцієнтах $w_i$ і $\sigma_i$, може бути визначена візуально у спектрограмах. Однак монотонність зростання значень $w_i$ порушується в одному випадку (слова «Безпека» і «Електронний»), що пояснюється різними підходами, що застосовуються для розрахунку $w_i$ та $\sigma_i$ і частою появою першого слова.

Табл. 2. Значення коефіцієнтів нерівномірності для окремих слів у добірці законодавчих актів України

| **Слово** | **Входження** | $w_i$ **(TF IDF)** | $\sigma_i$ |
|---|---|---|---|
| Технології | 46 | 53,62 | 1,99 |
| Оприлюднення | 33 | 53,98 | 2,21 |
| Безпека | 102 | 96,15 | 2,41 |
| Електронний | 50 | 85.24 | 3.22 |
| Регулювання | 220 | 129,92 | 3,62 |

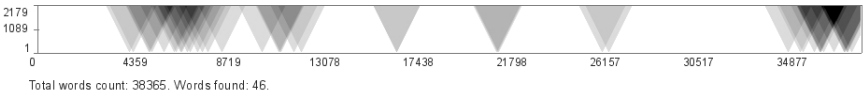



Рис. 2 – Спектрограма входження слова «Технології»

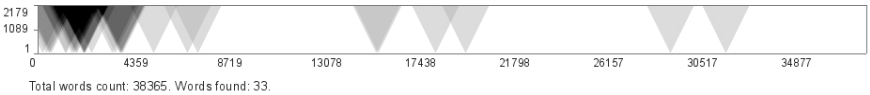



Рис. 3 – Спектрограма входження слова «Оприлюднення»

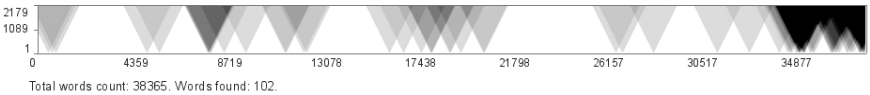



Рис. 4 – Спектрограма входження слова «Безпека»

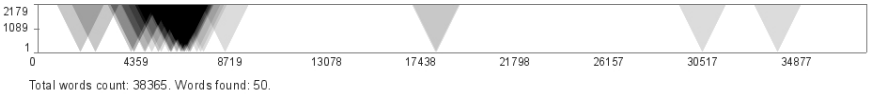



Рис. 5 – Спектрограма входження слова «Електронний»

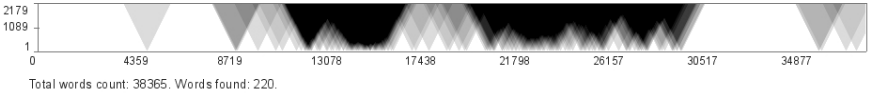



Рис. 6 – Спектрограма входження слова «Регулювання»

Аналогічні розрахунки були проведені для масиву з 50 новинних веб-публікацій 2012 р. з тематикою, яка визначається запитом до системи контент-моніторингу InfoStream [*Ланде, 2007*] (табл. 3, рис. 7 – 11):

**(захист~персональн~даних) | кібербезпек | (інформац~безпек).**

У цьому випадку монотонність зростання значень по відношенню до слова «Безпека» порушується.

Слід звернути увагу, що дискримінантна сила окремих слів на двох розглянутих добірках істотно розрізняється, що пов'язано з стилем і змістом відповідних текстів.

Табл. 3. Значення коефіцієнтів нерівномірності для окремих публікацій у масиві новинних повідомлень

| **Слово** | **Входження** | $w_i$ **(TF IDF)** | $\sigma_i$ |
|---|---|---|---|
| Регулювання | 16 | 33,07 | 1,26 |
| Оприлюднення | 18 | 35,49 | 1,50 |
| Безпека | 56 | 71,59 | 1,75 |
| Електронний | 28 | 50.53 | 2.02 |
| Технології | 39 | 61,45 | 2,06 |

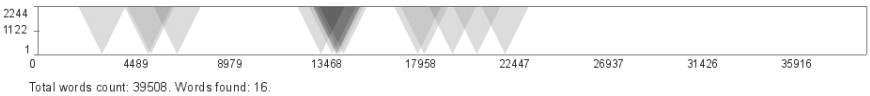



Рис. 7 – Спектрограма входження слова «Регулювання»

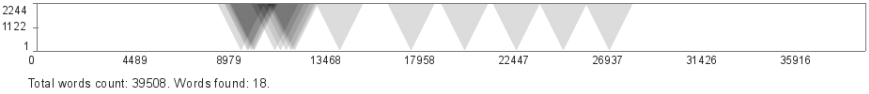



Рис. 8 – Спектрограма входження слова «Оприлюднення»

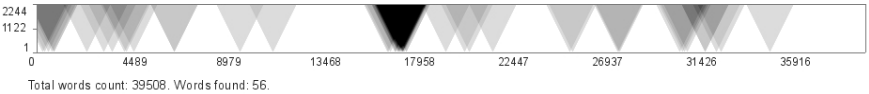



Рис. 9 – Спектрограма входження слова «Безпека»

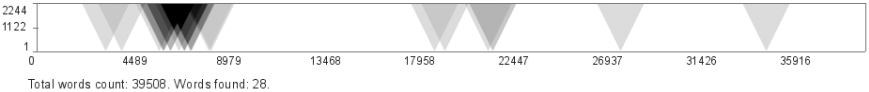



Рис. 10 – Спектрограма входження слова «Електронний»

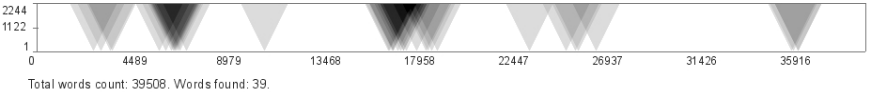



Рис. 11 – Спектрограма входження слова «Технології»

Крім традиційного підходу до оцінки дискримінантної сили слів у текстах, запропонованого Солтоном, дисперсійний аналіз дає близькі за якістю результати. Незважаючи на те, що підхід TF IDF за останній час пройшов ряд трансформацій, доповнюється допоміжними параметрами, зокрема, отримав популярність метод BM25, що враховує довжину документів, дисперсійний аналіз виявляється досить перспективним.

Розглянуті приклади показали, що штучний прийом, що полягає в тому, що вихідний текст великого розміру розбивався на фрагменти фіксованої довжини, цілком виправдався, результати багато в чому збіглися з результатами, отриманими іншим методом.

Наведені приклади показують, нерівномірність слів в масивах новинних повідомлень і в офіційних документах має близьку, багато в чому аналогічну природу, проте дискримінантна сила окремих слів на двох розглянутих добірках істотно розрізняється, що пов’язано з стилем і змістом відповідних текстів.

І, нарешті, запропонований метод візуалізації нерівномірності входження слів, у порівнянні з існуючими, додав ще один вимір – величину вікна спостереження, що виявилося зручним при розгляді текстових (у тому числі документальних) масивів великих обсягів. Техніка спектрограм дозволяє експертам без додаткових зусиль якісно оцінювати значення окремих слів при формуванні так званих надфразових єдностей, екстрагуванні фрагментів текстів для формування довідкових документів.

## 1.3. Складні мережі і задачі комп'ютерної лінгвістики

Основною причиною уваги до теорії складних мереж (Complex Networks) є результати сучасних робіт з опису реальних комп'ютерних, біологічних і соціальних мереж. Такі мережі мають характеристики, не властиві мережам з рівномірно ймовірною зв'язністю вузлів, а будуються на основі зв'язних структур і вузлів-концентраторів.

Разом з цим, практично усі сучасні мережі можна вважати складними. Так, наприклад, задача синтезу мереж із слів тексту (мереж мови) допускає комбінаторний підхід, що спирається на представлення мережі у вигляді кінцевого графа, вершини якого відповідають вузлам мережі, а ребра – лініям зв'язку.

В той же час, використання методів перерахування графів навіть для відносно невеликих мереж вважається неперспективним, оскільки необхідно дослідити величезну кількість можливих варіантів з'єднання вузлів лініями зв'язку.

Наприклад, у мережі з 10 вузлів існує $2^{45}$ варіантів з'єднання. Для 10 вузлів теоретично можливо $C_{10}^2 = \frac{10 \cdot 9}{2} = 45$ ліній сполучення. Кожна з цих можливих ліній зв'язку може реально існувати – стан «1», або не існувати – стан «0», тобто існує $2^{45}$ можливостей.

Для меншої кількості вузлів (наприклад, $n = 3$) варіанти можуть бути реально перебрані ($2^{\frac{32}{2}} = 8$) (рис. 12).

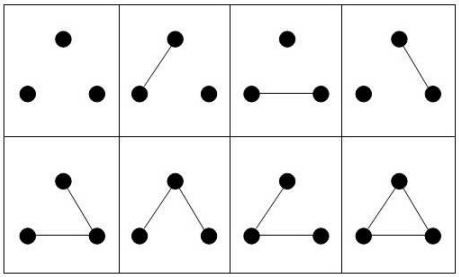


Рис. 12 – Варіанти мереж при $n = 3$

***Параметри складних мереж***

Теорія складних мереж як область дискретної математики вивчає характеристики мереж, враховуючи не лише їх топологію, але і статистичні феномени, розподіл ваги окремих вузлів і ребер, ефекти протікання в таких мережах струму, рідини або інформації. Виявилось, що властивості багатьох реальних мереж істотно відрізняються від властивостей класичних випадкових графів.

Не дивлячись на те, що в розгляд теорії складних мереж потрапляють різні мережі, найбільший внесок у розвиток цієї теорії внесли дослідження соціальних мереж. Термін «соціальна мережа» означає зосередження соціальних об'єктів, які можна розглядати як мережу (або граф), вузли якої – об'єкти, а зв'язки – соціальні стосунки. Цей термін було введено в 1954 році соціологом з «Манчестерської школи» Дж. Барнсом (J. Barnes) в роботі «Класи і збори в норвезькому острівному приході». У другій половині XX століття поняття «Соціальна мережа» стало популярним серед західних дослідників, при цьому як вузли соціальних мереж стали розглядати не лише представників соціуму, але і інші об'єкти, яким притаманні соціальні зв'язки. У теорії соціальних мереж отримав розвиток такий напрям, як аналіз соціальних мереж (Social Network Analysis, SNA). Сьогодні термін «соціальна мережа» означає поняття, що виявилося ширше за свій соціальний аспект, воно включає, наприклад, багато інформаційних мереж, а також так званих «мереж мови» у комп'ютерній лінгвістиці.

У рамках теорії складних мереж розглядають як статистичні, так і динамічні мережі, для розуміння структури яких необхідно враховувати принципи їх еволюції.

У теорії складних мереж виділяють три основні напрями: дослідження статистичних властивостей, які характеризують поведінку мереж; створення моделі мереж; прогнозування поведінки мереж при зміні структурних властивостей. У прикладних дослідженнях зазвичай застосовують такі типові для мережевого аналізу характеристики, як розмір мережі, мережева щільність, міра центральності і т.ін.

Про «структуру співтовариства» в складній мережі можна говорити тоді, коли існує фрагмент мережі – група вузлів, які мають високу щільність ребер між собою, при тому, що щільність ребер між окремими фрагментами – низька. Традиційний метод для виявлення структури співтовариств – кластерний аналіз. Існують десятки прийнятних для цього методів, які базуються на різних мірах відстаней між вузлами, зважених шляхових індексах між вузлами тощо. Для великих соціальних мереж наявність структури співтовариств виявилася невід'ємною властивістю.

При аналізі складних мереж як і в теорії графів досліджуються параметри окремих вузлів; параметри мережі в цілому; мережеві підструктури.

***Параметри вузлів мережі***

Для окремих вузлів виділяють наступні параметри:

- вхідний степінь зв'язності вузла – кількість ребер мережі, які входять у вузол;
- вихідний степінь зв'язності вузла – кількість ребер мережі, які виходять з вузла;
- відстань від вибраного вузла до кожного з інших;
- середня відстань від вибраного вузла до інших;
- ексцентричність (eccentricity) – найбільша з геодезичних відстаней (мінімальних відстань між вузлами) від вибраного вузла до інших;

- посередництво (betweenness) – кількість найкоротших шляхів, що проходять через вибраний вузол;
- центральність – кількість зв'язків вибраного вузла по відношенню до інших;
- уразливість – рівень спаду ефективності (визначено нижче) мережі у разі вилучення вузла і усіх суміжних з ним ребер.

***Загальні параметри мережі***

Для розрахунку характеристик мережі в цілому використовують такі параметри, як: кількість вузлів, кількість ребер, геодезична відстань між вузлами, середня відстань від одного вузла до інших, щільність – відношення кількості ребер в мережі до можливої максимальної кількості ребер при заданій кількості вузлів, кількість симетричних, транзитивних і циклічних тріад, діаметр мережі – найбільша геодезична відстань в мережі, уразливість, що розраховується як максимальна уразливість усіх вершин мережі, асортативність як міра кореляції між мірами вузлів і т.ін.

Існує декілька актуальних завдань дослідження складних мереж, зокрема, мереж мови, серед яких можна виділити наступні:

- визначення фрагментів мережі (клік, кластерів), в яких вузли зв'язані між собою сильніше, ніж з вузлами з інших подібних фрагментів;
- виділення фрагментів мережі (компонент зв'язності), які пов'язані усередині і не зв'язані або слабко зв'язані між собою;
- знаходження перемичок, тобто вузлів, при вилученні яких мережа розпадається на незв'язані частини.

***Розподіл степенів зв'язності вузлів***

Для неорієнтованих мереж степінь зв'язності вибраного вузла – це кількість ребер, сполучених з цим вузлом. Відповідно, середній степінь усієї мережі розраховується як середній степінь зв'язності для усіх вузлів мережі.

Важливою характеристикою мережі є функція розподілу степенів вузлів $P(k)$, яка визначається як ймовірність того, що вузол $i$ має степінь $k_i = k$, тобто значення $P(k)$ відповідає долі вершин із степенем $k$.

Для орієнтованих мереж існує розподіл вхідних $P^{in}(k^{in})$ і вихідних $P^{out}(k^{out})$ напівстепенів, а також розподіл загального степеню $P^{io}(k^{in}, k^{out})$, що задає ймовірність знаходження вузла з вхідною $k^{in}$ і вихідною $k^{out}$ напівстепенями.

Мережі, що мають різні $P(k)$, демонструють дуже різну поведінку. $P(k)$ у деяких випадках може бути розподілами Пуассона ($P(k) = e^{-m} m^k / k!$, де $m$ – математичне очікування), експоненціальним ($P(k) = e^{-k/m}$) або степеневим ($P(k) \sim 1/k^{\gamma}$, $k \neq 0$, $\gamma > 0$).

Важливою особливістю багатьох реальних мереж є розподіл степенів вузлів $P(k)$ за степеневим законом.

Мережі із степеневим розподілом степенів зв'язності вузлів звуться безмасштабними (scale-free). При степеневому розподілі можливе існування вузлів з дуже високим степенем, що практично не спостерігається в мережах з розподілом Пуассона.

***Шлях між вузлами***

Якщо два вузли $i$ та $j$ можна з'єднати за допомогою послідовності з $m$ ребер, то таку послідовність називають маршрутом між вузлами $i$ та $j$, а $m$ називають довжиною маршруту.

Говорять, що вузли $i$ та $j$ зв'язні, якщо існує маршрут між ними. Відношення зв'язності транзитивне, тобто якщо вузол $i$ зв'язаний з вузлом $j$, а $j$ зв'язний з $k$, то $i$ зв'язаний з $k$.

При цьому маршрут, у якого початок і кінець знаходяться в одному і тому ж вузлі, а всі інші вершини використовуються лише один раз, називається циклом.

Відстань між вузлами визначається як довжина маршруту від одного вузла до іншого. Природно, вузли можуть бути сполучені прямо або опосередковано.

Шляхом між вузлами $d_{ij}$ зветься найкоротша відстань між ними. Для всієї мережі можна ввести поняття середнього шляху, як середнє за усіма парами вузлів найкоротшої відстані між ними:

$$l = \frac{2}{n(n+1)} \sum_{i \geq j} d_{ij} ,$$

де $n$ – кількість вузлів, $d_{ij}$ – найкоротший шлях між вузлами $i$ та $j$.

Угорськими математиками П. Ердешем (P. Erdős) і А. Реньї (A. Rényi) було показано, що середня відстань між двома вершинами у випадковому графові росте як логарифм від числа вершин [*Erdős, 1959*], [*Erdős, 1960*].

Деякі мережі можуть виявитися незв'язними, тобто в них знайдуться вузли, відстань між якими є нескінченною. Відповідно, середній шлях може виявитися також нескінченним. Для урахування таких випадків вводиться поняття глобальної ефективності мережі як середнього інверсного шляху між вузлами, що розраховується за формулою:

$$E = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}} ,$$

де сума враховує усі пари вузлів. Ця характеристика відповідає ефективності мережі при пересилці інформації між вузлами (передбачається, що ефективність при пересилці інформації між двома вузлами $i$ та $j$ зворотно пропорційна відстані між ними).

Зворотна величина глобальної ефективності – середнє гармонійне геодезичних відстаней:

$$h = \frac{1}{E} .$$

Оскільки ця формула знімає проблему розбіжності при визначенні середнього шляху, то ця характеристика краще підходить для графів з декількома компонентами зв'язності.

Ефективна відстань між двома вузлами в загальному випадку більша, ніж найкоротша відстань.

Мережі також характеризуються таким параметром як діаметр або максимальний по довжині шлях, тобто шлях, рівний максимальному значенню з усіх $d_{ij}$.

### *Коефіцієнт кластеризації*

Д. Уаттс (D. Watts) і С. Строгатц (S. Strogatz) визначили такий параметр мереж, як коефіцієнт кластеризації [*Watts, 1998*], який характеризує рівень зв'язності вузлів в мережі, тенденцію до утворення груп взаємозв'язаних вузлів, так званих клік (clique).

Для конкретного вузла коефіцієнт кластеризації показує, скільки найближчих сусідів цього вузла є також найближчими сусідами один одному. Коефіцієнт кластеризації для окремого вузла мережі визначається за наведеним нижче правилом. Нехай з вузла виходить $k$ ребер, які сполучають його з $k$ іншими вузлами, найближчими сусідами. Якщо припустити, що усі найближчі сусіди сполучені безпосередньо один з одним, то кількість ребер між ними складала б $\frac{1}{2}k(k-1)$. Тобто це число, що відповідає максимально можливій кількості ребер, якими могли б з'єднуватися найближчі сусіди вибраного вузла. Відношення реальної кількості ребер, які сполучають найближчих сусідів цього вузла до максимально можливого (такого, при якому усі найближчі сусіди цього вузла були б сполучені безпосередньо один з одним) називається коефіцієнтом кластеризації вузла $i$ – $C(i)$. Природно, ця величина не перевищує одиниці.

Існує ще один спосіб обчислення коефіцієнта кластеризації мережі, що базується на такій формулі:

$$C = \frac{3N_{\Delta}}{N_3},$$

де $N_{\Delta}$ – кількість 3-циклів у мережі, а $N_3$ – кількість зв'язних 3-компонент.

3-цикл визначається при цьому як множина трьох вузлів з ребрами між кожною парою вузлів. Зв'язкова 3-компонента – множина, що складається з трьох вузлів, в якому кожен вузол досяжний з іншого вузла, безпосередньо або опосередковано. Таким чином, у 3-компоненту центральний вузол має бути інцидентний двом іншим. Множник 3 введений з урахуванням варіантів різних 3-компонент для кожного 3-цикла. Справедливо:

$$N_{\Delta} = \sum_{k>i>j} a_{ij} a_{ik} a_{jk};$$

$$N_3 = \sum_{k>i>j} \left( a_{ij} a_{ik} + a_{ji} a_{jk} + a_{ki} a_{kj} \right),$$

де $a_{ij}$ – елементи матриці суміжності $A$, відповідної мережі, сума береться за усіма компонентами різних вузлів $i$, $j$ та $k$ тільки один раз.

Коефіцієнт кластеризації може визначатися як для кожного вузла, так і для усієї мережі. Відповідно, рівень кластеризації усієї мережі визначається як нормована по кількості вузлів сума відповідних коефіцієнтів окремих вузлів.

Різниця між двома підходами до визначення кластеризації полягає у тому, що, усереднивши за вершинами, у другому випадку враховується однаковий вплив для кожного трикутника в мережі, а в першому випадку враховується рівний внесок для кожного вузла.

Це призводить до різних значень коефіцієнта кластеризації, тому що вузли з великими степенями з більшою вірогідністю входять до складу більшої кількості трикутників, ніж вершини з меншими степенями.

***Посередництво***

До втрати зв'язності в інформаційній мережі може привести розрив зв'язків між її компонентами, наприклад, при усуненні з інформаційного простору найбільш вагомих компонент (структурних компонент у разі мереж мови), тобто таких, які мають, найбільший коефіцієнт посередництва (betweenness). Цей коефіцієнт для конкретного вузла мережі визначається як сума по усіх парах вузлів мережі співвідношень кількості найкоротших

шляхів між ними, що проходять через заданий вузол, до загальної кількості найкоротших шляхів між ними.

Значення вузла для мережі тим більше, чим в більшій кількості шляхів він задіяний. Тому, вважаючи, що обмін даними найчастіше відбувається по найкоротших шляхах, можна виміряти кількісно значення вузла з точки зору посередництва (betweenness), що визначається кількістю найкоротших шляхів що проходять через цей вузол. Ця характеристика визначає роль цього вузла у встановленні зв'язків в мережі – вузли з найбільшим посередництвом грають головну роль у встановленні зв'язків між іншими вузлами в мережі. Посередництво $b_m$ вузла $m$ визначається за формулою:

$$b_m = \sum_{i \neq j} \frac{B(i,m,j)}{B(i,j)},$$

де $B(i,j)$ – загальна кількість найкоротших шляхів між вузлами $i$ та $j$, $B(i,m,j)$ – кількість найкоротших шляхів між вузлами $i$ та $j$, що проходять скрізь вузол $m$.

Якщо враховувати, що найкоротші шляхи можуть бути невідомі, і замість цього для навігації в мережі використовуються пошукові алгоритми, то посередництво (проміжна центральність) вузла відповідає ймовірності його знаходження пошуковим алгоритмом.

Рівень домінування найбільшого посередника в цьому випадку визначається відповідно до формули:

$$CPD = \frac{1}{n-1} \sum_i (B_{\max} - B_i),$$

де $B_{\max}$ – найбільше в мережі значення рівня посередництва.

Домінування центрального вузла дорівнюватиме 0 для кліки і 1 для зірки, в якій центральний вузол входить в усі шляхи.

***Мережі мови в правовій інформатиці***

Першим кроком у застосуванні теорії складних мереж [*Strogatz, 2001*], [*Albert, 2002*] до текстових документів є

формування мережевої моделі цих документів у вигляді сукупності вузлів і зв'язків, тобто побудова мереж мови (Language Network) [*Головач, 2006*], в яких виявляються найвагоміші вузли, іноді це так звані опірні слова або відповідні словосполучення.

Поряд з послідовним аналізом текстових документів, побудова мереж, вузлами яких є такі елементи, як слова або словосполучення, тобто фрагменти природної мови, дозволяє виявляти структурні елементи текстів, без яких тексти втрачають свою зв'язність. Відомо декілька підходів до побудови мереж з текстів і різні способи інтерпретації вузлів і зв'язків, що приводить, відповідно, до різних видів представлення таких мереж. Вузли можуть бути сполучені між собою, якщо відповідні їм слова стоять поряд у тексті [*Ferrer-i-Cancho, 2001*], [*Dorogovtsev, 2001*], належать до одного речення або абзацу [*Caldeira, 2005*], сполучені синтаксично [*Ferrer-i-Cancho, 2004*], [*Ferrer-i-Cancho, 2005*] або семантично [*Motter, 2002*], [*Sigman, 2002*].

Збереження синтаксичних зв'язків між словами призводить до представлення тексту у вигляді спрямованої мережі (Directed Network), де напрям зв'язку відповідає підпорядкуванню слова.

Якщо поставити у відповідність кожному слову вузол мережі і з'єднати кожні два вузли зв'язком тоді, коли відповідні ним слова стоять у реченні поруч, то таке представлення називають *L*-простором. У *L*-просторі, так само як і в інших наведених нижче мережевих моделях, при виникненні кратних зв'язків прийнято зберігати лише один з них.

Традиційно розрізняють чотири різновиди мереж мови (просторів):

1. *L*-простір. Зв'язуються сусідні слова, які належать до одного речення. Кількість сусідів для кожного слова (вікно слова) визначається радіусом взаємодії R, найчастіше розглядається випадок R = 1.

2. *B*-простір. Розглядаються вузли двох видів, відповідні реченням і словам, що належать до них.

3. *P*-простір. Усі слова, які належать до одного речення, зв'язуються між собою.

4. *C*-простір. Речення зв'язуються між собою, якщо у них застосовуються однакові слова.

У випадку *L*-простору зв'язки можуть враховувати не лише «найближчих сусідів», але і групи з декількох слів, які знаходяться на певній відстані один від одного. Для цього вводиться поняття «радіусу дії» $R$: при $R = 1$ зв'язок існує лише між найближчими сусідами, при $R = 2$ – між найближчими і наступними близькими сусідами і т. д. Змінна $R$ може приймати значення від $R = 1$ до $R_{\max}$, де $R_{\max} + 1$ – загальна кількість слів у реченні. Зростання «радіусу взаємодії» $R$ у цьому випадку призводить до зростання кількості зв'язків, досягаючи насичення при $R = R_{\max}$.

Ще один спосіб представити текст у вигляді мережі полягає у використанні дводольних (bipartite) графів. У такому представленні (*B*-простір) розглядаються вузли двох видів. Один вид відповідає реченням, другий – словам. Зв'язок між різними вузлами означає, що слово належить реченню.

У *P*-просторі усі слова, що належать одному реченню, вважаються зв'язаними між собою.

У *C*-просторі вузли відповідають реченню, а зв'язок між вузлами-реченнями встановлюється у тому випадку, якщо у відповідних реченнях є загальні слова.

Для мережі, побудованої на підставі Британського національного корпусу (*L*-простір мови, $R = 1$) виявилось, що ця мережа англійської мови безмасштабна, а поведінка степеню $P(k)$ характеризується двома режимами степеневого розподілу із значеннями відповідних степеневих показників $\gamma = 1,5$ для $k < 2000$ і $\gamma = 2,7$ для $k > 2000$.

Згідно з визначенням, якщо середня довжина найкоротшого шляху росте з розміром (кількістю вузлів) мережі повільніше за будь-яку степеневу функцію, то мережа є «малим

світом». Мережі малого світу надзвичайно компактні. Для згаданої вище мережі англійської мови довжина найкоротшого шляху складає лише $< l >= 2,63$. Оскільки зростання $R$ призводить лише до додавання нових зв'язків, то $< l >$ зменшуються із зростанням $R$.

Специфічною формою кореляції в мережах є утворення кластерів. Коефіцієнт кластеризації $C$ характеризує схильність мережі до утворення сполучених трійок вузлів. Відомо, що для повного графа $C = 1$, а для мережі у формі дерева $C = 0$.

Відношення середнього коефіцієнта кластеризації досліджуваних мереж до коефіцієнта кластеризації класичного випадкового графа свідчить про те, що мережі мови є добре корельованими структурами. Такі кореляції ростуть із зростанням «радіусу взаємодії» $R$.

Для Британського національного корпусу на підставі аналізу текстів, які містили $\approx 10^7$ слів, набуте значення коефіцієнта кластеризації $< C >= 0,687$.

У випадку розгляду $P$-простору кожне слово-вузол пов'язано з усіма іншими словами, які належать спільному реченню. Таким чином, кожне речення тексту входить в мережу як повний граф – кліка взаємозв'язаних вузлів. Різні речення-кліки об'єднуються в мережу завдяки загальним словам. У $L$-просторі слова зв'язуються в межах вікна, розмір якого характеризуються величиною $R$. Коли розмір вікна $R$ стає рівним розміру речення, то представлення цього речення в $L$- і в $P$-просторах співпадають. Відповідно, коли розмір вікна стає рівним розміру найбільшого речення тексту ($R = R_{\max}$), то представлення усього тексту в $L$- і в $P$-просторах співпадають.

На практиці доведено, що мережа мови є сильно корельованим безмасштабним малим світом (Scale-Free Small World). Існує ряд праць, в яких зроблена спроба пояснити властивості мереж мови за допомогою сценарію переважного приєднання (Preferential Attachment [*Albert, 1999*]), розглядаючи їх як результат процесу зростання, коли нові вузли-слова з більшою ймовірністю приєднуються до вузлів-хабів, що мають багато зв'язків.

### *Мережі горизонтальної видимості*

У рамках концепції складних мереж запропоновано декілька методів побудови мереж на основі часових рядів, серед яких можна назвати декілька методів побудови графів видимості [*Nunez, 2012*], зокрема, так званий граф горизонтальної видимості (Horizontal Visibility Graph – HVG) [*Luque, 2009*], [*Gutin, 2011*]. Ці підходи також дозволяють будувати мережеві структури на підставі текстів, в яких окремим словам або словосполученням деяким спеціальним чином поставлені у відповідність числові вагові значення. Як функція, що ставить у відповідність слову число, можна розглядати, наприклад, порядковий номер унікального слова у тексті, довжину слова, загальноприйняту оцінку TF IDF (у канонічному виді, рівну добутку частоти слова у фрагменті тексту – Term Frequency – на двійковий логарифм від величини, зворотної кількості фрагментів тексту, в яких це слово зустрілось – Inverse Document Frequency) або її варіанти, а також інші вагові оцінки.

При побудові мереж слів в цій роботі також використовується дисперсійна оцінка ваги слів [*Ortuño, 2003*], формулу для розрахунку якої було наведено вище.

Ряди з цифрових значень, відповідних словів, перетворяться в графи горизонтальної видимості, в яких вузлам відповідають не лише цифрові значення, але самі слова, що мають певне змістовне значення. Мережа мови з використанням алгоритму горизонтальної видимості будується в три етапи [*Lande*, 2013]. На першому на горизонтальній осі відзначається ряд вузлів, кожен з яких відповідає словам в порядку появи в тексті, а по вертикальній осі відкладаються вагові чисельні оцінки (візуально – набір вертикальних ліній, рис. 13).

На другому етапі будується традиційний граф горизонтальної видимості [*Luque, 2009*]. Для цього між вузлами встановлюється зв’язок, якщо вони знаходяться в «прямій видимості», тобто якщо їх можна з’єднати горизонтальною лінією, що не перетинає ніяку вертикальну лінію, що розміщена між цими вузлами.

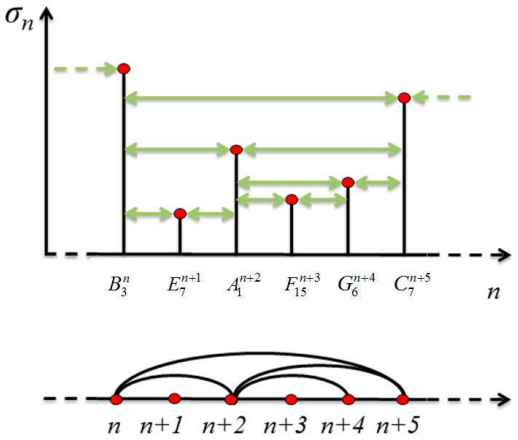



Рис. 13 – Приклад побудови графа горизонтальної видимості

Алгоритм побудови графа горизонтальної видимості можна представити зручним для обчислення способом. Так наприклад, на рис.1 для вузла-слова $A_1^{n+2}$ (верхній індекс – номер слова у тексті, нижній – номер появи конкретного слова, у цьому випадку – слова $A$) суміжними в мережі вважаються слова $B_3^n$ та $C_7^{n+5}$ і встановлюються ребра-зв'язки, такі що $B_3^n$ – найближче зліва від $A_1^{n+2}$ слово, з ваговою оцінкою $\sigma_n = \sigma_B$, що перевищує вагову оцінку слова $A$ ($\sigma_{n+2} = \sigma_A$), а $C_7^m \left(m = n+5\right)$ – найближче справа від $A_1^{n+2}$ слово, для якого $\sigma_{105} > \sigma_{102}$.

На третьому, завершальному етапі, отримана на попередньому етапі мережа компактифікується. Усі вузли з цим словом, наприклад словом $A$, об'єднуються в один вузол. Усі зв'язки таких вузлів також об'єднуються. Важливо відзначити, що між будь-якими двома вузлами при цьому залишається не більше ніж один зв'язок – кратні зв'язки вилучаються. Зокрема це означає, що міра (число зв'язків) вузла не перевищує суми степенів $\sum_k A_k^n$.

У результаті виходить нова мережа слів – компактифікований граф горизонтальної видимості (КГГВ) – рис. 14.

Як тексти при побудові мереж мови автором розглядалася добірка законодавчих актів України, що відносяться до формування та розвитку інформаційного простору держави (Закони України «Про доступ до публічної інформації», «Про Основні засади розвитку інформаційного суспільства в Україні на 2007–2015 роки», «Про телекомунікації», «Про захист персональних даних», «Про основи національної безпеки України»).

Як тексти при побудові мереж мови автором розглядалася добірка законодавчих актів України, що відносяться до формування та розвитку інформаційного простору держави (Закони України «Про доступ до публічної інформації», «Про Основні засади розвитку інформаційного суспільства в Україні на 2007–2015 роки», «Про телекомунікації», «Про захист персональних даних», «Про основи національної безпеки України»).

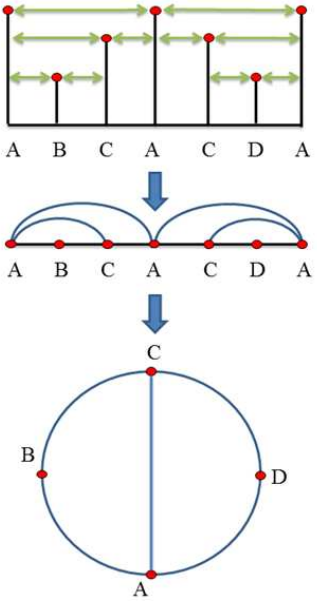



Рис. 14 – Етапи побудови компактифікованого графу горизонтальної видимості

Для усіх побудованих КГГВ-мереж слів було визначено розподіл степенів вузлів, який виявився близьким до

степеневого ($P(k) = Ck^{\alpha}$), тобто ці мережі є безмасштабними. Були проведені розрахунки параметрів мереж для усіх розглянутих літературних творів. В результаті виявилось, що для усіх з них коефіцієнт $\alpha$ змінювався в діапазоні від – 0,95 до – 1,05.

До складу вузлів з найбільшими степенями в КГГВ-мережі, разом з особистими займенниками і іншими службовими словами (частки, прийменники, союзи і таке інше), потрапили слова, що визначають інформаційну структуру тексту [*Giora, 1983*], [*Ягунова, 2010*].

Для порівняння була додатково досліджена поведінка простих мереж мови, коли не першому етапі побудови мережі зв'язуються сусідні слова, що входять в текст ($L$-простір, $R$ = 1), а на другому відбувається компактифікація мережі. Очевидно, вага вузлів в цій мережі відповідає частоті появи слів, а їх розподіл – закону Ципфа [*Zipf, 1949*]. При цьому найбільші степені мають вузли, що відповідають словам з найбільшою частотою, – союзам, прийменникам, займенникам тощо, що мають велике значення для зв'язності тексту, але є малоцікавими з точки зору дослідження інформаційної структури.

Якщо позначити $\Psi$ – множину із $N$ різних слів (наприклад, $N$ = 100), що відповідають найбільш вагомим вузлам наведеної простої мережі мови, а $\Lambda$ – множину слів, що відповідають найбільш вагомим вузлам КГГВ, то множина $\Omega = \Lambda \setminus \Psi$ відповідає інформативним словам, що мають, крім того, важливе значення і для зв'язності тексту. Автором при дослідженні зіставлялися 100 найбільш вагомі вузли для трьох даних типів мереж слів за текстами Законів України «Про телекомунікації» і «Про захист персональних даних».

У КГГВ-мережі по тексту Закону України «Про телекомунікації» з урахуванням значень TF IDF до складу множини $\Omega$ потрапили такі слова, як «Державне», «Регулювання», «Ринку», «Інтернет», «Провайдер», «Трафік». У КГГВ-мережі для цього ж тексту за ваговими значеннями слів, які відповідають дисперсійним оцінкам, додатково до складу

множини Ω потрапили такі слова, як «Суб'єкт», «Ресурс», «Переоформлення», «Рішення», «Споживачів» та інші.

При аналізі тексту Закону України «Про захист персональних даних» до множини Ω (для КГГВ-мережі з урахуванням вагових значень слів за алгоритмом TF IDF) потрапили такі слова, як «Інформація», «Відстрочення», «Орган», «Баз», «Виключено».

У КГГВ-мережі для тексту цього законодавчого акту за ваговими значеннями слів, що відповідають дисперсійним оцінкам, до множини Ω потрапили додатково такі слова, як «Використання», «Прав», «Уповноважений», «Особа».

Таким чином:

1. На основі послідовності дисперсійних оцінок слів тексту і КГГВ, побудовано мережі слів різних текстів.

2. Для текстів нормативно-правових актів серед вузлів відповідних КГГВ з найбільшими степенями були присутні слова, що визначають не лише структури тексту, забезпечують зв'язність, але й ті, що визначають його інформаційну структуру, відбивають семантику законодавчих актів.

3. Алгоритм визначення ваги слів, що базується на дисперсійній оцінці виявився у цьому випадку ефективнішим для визначення інформаційно-значущих слів, що грають важливе значення для структурної зв'язності в текстах законодавчих актів, ніж алгоритм TF IDF [*Ланде, 2013*].

## 1.4. Корпусна лінгвістика в правовій інформатиці

Корпусна лінгвістика – це той розділ мовознавства, що вивчає створення, обробку та використання текстових корпусів. В рамках комп'ютерної лексикографії словники часто формуються шляхом сканування і подальшої обробки паперових словників, проте на цей час все частіше початковим матеріалом для отримання необхідної лексики є текстові корпуси.

Текстовим корпусом у лінгвістиці називають сукупність текстів, зібраних відповідно до певних принципів, розмічених за

відповідними правилами (морфологічною, акцентною, синтаксичною і іншою розміткою) і, як правило, доповнених пошуковою системою. Іноді корпусом ("корпус першого порядку") називають просто будь-які колекції текстів, об'єднаних якоюсь загальною ознакою (мовою, жанром, автором тощо). Так у роботі [*Шаров, 2003*] визначається найпростіший лінгвістичний корпус (корпус першої позиції) як будь-яка колекція текстів з певної тематики, які є доступними в електронній формі. Виходячи з цього визначення, база даних «Законодавство України» (http://zakon.rada.gov.ua), що містить нормативно-правові документи України вже є лінгвістичним текстовим корпусом, нехай і найпростішим за структурою.

Разом з цим, уявленням автора відповідає визначення В.П. Захарова [*Захаров, 2002*]: «Під лінгвістичним корпусом текстів розуміється великий, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, представлений в електронному вигляді й призначений для вирішення різних лінгвістичних завдань».

Нині існують сотні різних текстових корпусів для різних мов, за різною тематикою, з різним рівнем розмітки. Розмічені текстові корпуси створюються і використовуються як для лінгвістичних досліджень, так і для налаштування (навчання) моделей і процесорів за допомогою відомих математичних методів машинного навчання. Так, машинне навчання застосовується для налаштування методів вирішення лексичної неоднозначності, анафоричних посилань, розпізнавання частини мови [*Широков, 2005а*].

Останнім часом все частіше як найповніший лінгвістичний ресурс розглядаються тексти з мережі Інтернет, який стає найрепрезентативнішим джерелом зразків сучасної мови, проте його використання як корпусу вимагає деяких, обмежень, розробки спеціальних технологічних засобів.

Доцільність створення текстових корпусів пояснюється:

– представленням лінгвістичних даних в реальному контексті;

– великою повнотою даних (при великому обсязі корпусу);

– можливістю багатократного використання корпусу, створеного один раз, для вирішення різних лінгвістичних завдань.

Робота з текстовими корпусами на цей час є одним з основних методів лінгвістичних досліджень. Ще у 1960-і роки було створено Брауновський корпус (США), який включає 1 млн. слів; у 1970-і було створено корпус LOB (Великобританія, Норвегія), який також включає 1 млн. слів. У 1980-і роки почали створюватися такі корпуси, як: Машинний Фонд російської мови, Уппсальський корпус російської мови (Швеція), корпус The Bank of English, Birmingham. У 1990-і створено British National Corpus, який включає 100 млн. слів, а також інші національні корпуси (угорський, італійський, хорватський, чеський, японський). На цей час корпус The Bank of English, Birmingham включає 600 млн. слів. На початку XXI ст. створювалися такі корпуси, як American National Corpus, 100 млн. слів і Gigaword corpora (англійський, арабський, китайський), що включає 1 млрд. слів.

Лінгвістичний текстовий корпус характеризується такими основними параметрами: він повинен бути достатньо великого обсягу; бути структурованим або розміченим; бути наведеним в електронному вигляді; доповнюватись спеціальним програмним забезпеченням для роботи з ним.

Цінність текстового лінгвістичного корпусу вбачається в наступному:

– одного разу зроблений корпус може використовуватися при рішенні різних задач;

– корпус містить мовні дані в їх реальному оточенні, що дозволяє досліджувати лексичну і граматичну структуру мови, а також безперервні процеси мовних змін, що відбуваються впродовж певного часу;

– корпус характеризується показністю, або збалансованим складом текстів, що дозволяє

використовувати його для тестування інформаційно-пошукових систем, відповідних морфологічних модулів, систем перекладу тощо;

- корпус може використовуватися при вивченні мови, оскільки за допомогою корпусу можна перевіряти особливості вживання мовних одиниць;

- корпусом можуть користуватися перекладачі. За наявності сумнівів про те, яким чином слід перекласти ту або іншу фразу, можна задати запит пошуковій системі корпусу і порівняти частоту вживань наявних варіантів.

Повнотекстовий пошук у типовому лінгвістичному текстовому корпусі на цей час доповнено такими параметрами, як:

- урахування порядку слів;
- пошук у вибраній підмножині об'єктів;
- використання нормалізації слів;
- підключення синонімічних ланцюжків до слів;
- урахування морфологічних і граматичних параметрів слів.

Розмітка (приписування тексту певної інформації для зручнішого аналізу) – суттєва характеристика сучасного текстового корпусу, яка відрізняє його від електронних колекцій і енциклопедій.

Існують різні типи розмітки:

- метатекстова (автор, назва, обсяг, тематика тощо), що характеризує текст у цілому;
- структурна розмітка є інформацією щодо структури тексту, яка дозволяє відокремити одне слово від іншого, виділити межі словосполучення, речення, тексту;

– лінгвістична розмітка, яка полягає в приписуванні одиницям тексту певної лінгвістичної інформації.

Вважається, чим багатша і різноманітніша розмітка, тим вищою є наукова і навчальна цінність корпусу.

Простір електронних текстових корпусів дав можливість результативного створення і використання електронних конкордансів. Конкордансі відкривають перспективи моделювання мовної картини світу на основі статистичної властивості мови, що проявляються лише на великих обсягах інформації та знаходять своє відображення в лінгвістичних текстових корпусах.

Таким чином, створення подібних корпусів є важливим науковим і практичним завданням. Створення репрезентативного корпусу є важливим національним завданням, що заохочує розвиток досліджень в області цієї мови, його використання. Якнайповнішим розміченим корпусом в Росії, наприклад, є Національний корпус російської мови (НКРМ, www.ruscorpora.ru). Він містить в собі тексти різної спрямованості, у яких кожному слову приписані його лексичні характеристики. Для частини корпусу знята омонімія, виконана велика робота з вибору лексичних параметрів. В рамках НКРМ реалізовано унікальний сервіс – побудова динаміки входження слів в тексти корпусу з 1800 року і по цей час. На рис.15 представлена динаміка входження слів «уголовный» і «административный» за період 1840-2012 рр.

В Українському мовно-інформаційному фонді НАН України створено Український національний лінгвістичний корпус (УНЛК) [*Широков, 2005а*] обсягом понад 36 млн. слів. УНЛК призначений для:

– надання текстової інформації за заданими критеріями;

– створення потоків лінгвістичної інформації для дослідницьких систем;

– інтеграції різних лінгвістично-програмних засобів у єдиному середовищі.

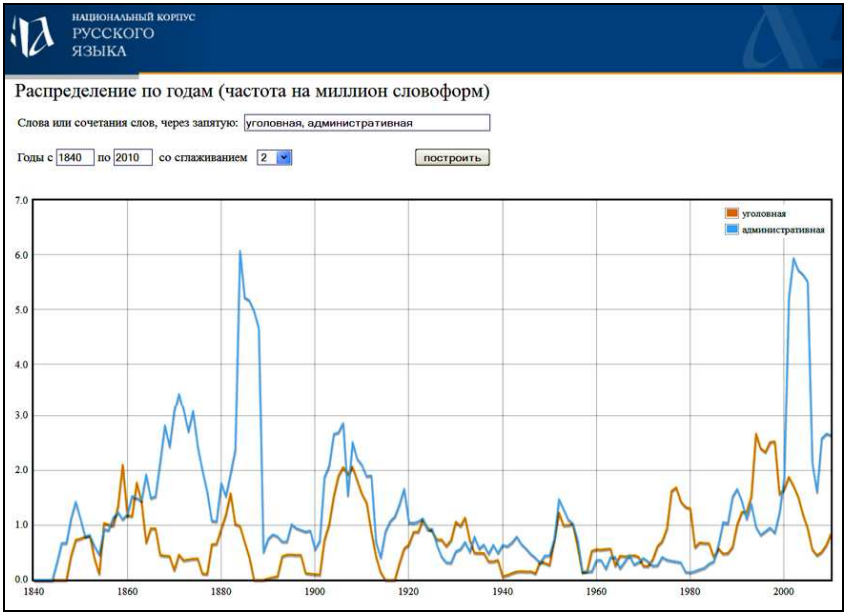



Рис. 15 – Динаміка входження заданих слів у тексти НКРМ

УНЛК як сервіс акумулює такі функціональні можливості, як сервіс електронної бібліотеки, повнотекстового пошуку, граматичного словника, лексикографічної системи «Словники України». Все це забезпечує проведення різнопланових фундаментальних лінгвістичних досліджень.

Для створення багатомовних лексикографічних систем, зокрема статистичних систем машинного перекладу застосовуються так званні паралельні корпуси текстів (Parallel Text Corpora), документів, речень, *N*-грам. Під паралельним текстовим корпусом розуміють великі зібрання паралельних текстів, тобто текстів однією мовою разом з їх перекладом іншою мовою.

На сьогодні існують алгоритми створення паралельних корпусів документів, які можна умовно розділити на дві групи: традиційні і статистичні.

До першої групи можна віднести алгоритми, за допомогою яких створювалися такі паралельні корпуси, як Корпус CRATER (www.comp.lancs.ac.uk/linguistics/crater/corpus.htm); Паралельний корпус перекладів «Слова о полку Игореве» (http://nevmenandr.net/slovo/); паралельний російсько-англійський корпус входить до складу Національного корпусу російської мови (http://www.ruscorpora.ru/corpora-biblio.html); паралельний російсько-словацький корпус [*Гарабик, 2006*] і таке інше. Створення цих корпусів пов'язане з тим, що початкові дані завідомо паралельні.

До другої групи можна віднести паралельні корпуси, створені за допомогою статистичних алгоритмів, засновані на аналізі сторінок багатомовних веб-сайтів, заздалегідь підготовлених фрагментарних масивів і т.ін. [*Resnik, 2002*], [*Ma, 1999*]. Автором було запропоновано новий підхід до створення паралельних корпусів документів, заснований на алгоритмі пошуку дублікатів документів, наведених різними мовами [*Ланде, 2009б*].

До завдань корпусної лінгвістики належить і створення із вже існуючих документів спеціальних текстових корпусів, які надають інструментарій для проведення досліджень у відносно нових для мовознавства галузях, серед яких можна назвати юридичну лінгвістику, що перебуває на межі лінгвістики і правознавства. Саме в цій галузі спостерігається значне підвищення активності.

До відомих міжнародних документальних корпусів з питань права можна назвати, наприклад, такі:

- EU Parliament in XML (http://politicalmashup.nl/eu-parliament-in-xml/, включає усі документи Європейського Парламенту, що містяться в офіційній базі даних, з інструкціями з їх обробки);
- DutchParl: корпус парламентських документів датською мовою (http://politicalmashup.nl/dutchparl/, містить усі доступні у електронному вигляді парламентські документи датською мовою з Нідерланд, Фландрії і Бельгії);

– Парламентські документи з Іспанії (http://politicalmashup.nl/ spanishparliament/, містить всі парламентські документи з Іспанії з 1977 по 2009 р.)

Усталення демократії та поширення формально судових методів розв'язання конфліктів, супроводжується розширенням сфери застосування лінгвістичного аналізу та експертизи текстів документів. Засоби автоматизації лінгвістичної експертизи, що базуються на застосуванні наявних текстових корпусів підвищують ефективність семантичного аналізу текстів (процедура передбачена законодавством України) як самих законодавчих актів, так і текстів, носіями яких виступають суб'єкти та об'єкти правовідносин [*Широков, 2005а*].

## 2. Засади порівняльного аналізу документів

Проблема порівняльного аналізу електронних текстів постала практично одночасно з появою можливостей обробки текстів комп'ютерною технікою. Друга половина XX-го століття характеризувалася становленням цього напрямку, що обумовлюється бурхливим розвитком формальної і комп'ютерної лінгвістики.

Разом з тим, слід відзначити, що фундаментальні роботи у галузі порівняння текстів проводилися ще значно раніше, наприклад, видатна робота з цього приводу була надрукована ще у 1915 р. [*Морозов,1915*].

На даний час з розвитком мережі Інтернет задачі порівняльного аналізу електронних текстів мають відношення до ряду таких технологічних напрямків, як інформаційний пошук, узагальнення та групування інформації. Розвиток Інтернету також визвав значне зростання дубльованої та запозиченої інформації у різних сферах: освіти, засобах масової інформації, іноді в науці [*Шарапов, 2011*], в практиці законотворчості, що пов'язано з проблемою гармонізації вітчизняного і міжнародного законодавства.

На цей час визначено п'ять таких напрямків порівняльного аналізу електронних текстів [*Osipovs, 2009*]:

- аналіз унікальності документів інформаційно-пошуковими системами в Інтернеті для запобігання зайвої індексації однакових документів;
- виявлення передруку, випадків плагіату, перевірка коректності запозичень у нормативно-правових документах;
- архівування документів (зменшення обсягів даних на комп'ютерних носіях);
- кластеризація документів за їх подібністю, виявлення кластерів близьких за змістом

документів, виявлення основних тематик в інформаційних потоках.

– пошук та фільтрація спаму.

Динаміка інформаційних потоків у веб-просторі, соціальних та піринговых мережах, обумовлює і ряд проблем, що виникають при порівнянні електронних текстів. Серед цих проблем можна назвати: великі обсяги інформації (звідки випливає задача оптимізації обчислень, знаходження раціональних алгоритмів порівнянь), засилля інформаційного шуму, недостовірної інформації, наявність інформаційних дублікатів наведених різними мовами тощо.

Задача виявлення нечітких дублікатів є однією з актуальніших і складніших, особливе практичне значення вона приймає, зокрема, при інтеграції інформаційних ресурсів, боротьбі з плагіатом, визначенні спам-розсилань тощо.

У багатьох випадках в основу пошуку нечітких дублікатів, зокрема, плагіату покладено порівняння текстів [*Ашура, 2006*], [*Stone, 2003*]. Спочатку проводиться порівняння текстів в цілому, а далі відбувається розбивка на абзаци і потім пошук конкретних фрагментів тексту в інших документах. В інших же випадках використовується пошук за ключовими словами або ж словосполученнями. Враховуючи кількість термінів з малою частотою, знайдених при перевірці, як результат отримують підтвердження того, що документ або фрагмент є плагіатом, або навпаки. Багато систем виконують пошук не тільки подібних фрагментів, але й проводять достатньо складний аналіз тексту, що включає, наприклад, використання методу Флешу, який дозволяє обчислити індекс «легкості» тексту [*Нейл, 2005*]. Аналізуючи показники індексів абзаців у роботі, що перевіряється, можна ідентифікувати аномалії, найбільш ймовірні абзаци плагіату, пошук яких потім проводиться у внутрішній базі даних, або ж у веб-просторі.

Серйозне спрощення названої задачі може бути отримано за рахунок застосування формальних методів, наприклад, математичної статистики, сигнатурних алгоритмів, репутаційних підходів (наприклад, шляхом ранжирування першоджерел, тематик, опорних слів тощо).

На формальному рівні, шляхом зіставлення фрагментів текстів, шинглів, лінгвістичних сигнатур тощо, сьогодні з успіхом виявляються нечіткі дублікати, що формуються шляхом копіювання, прямого перекладу з іноземних мов, компіляції тощо.

Однак нечіткі дублікати включають також результати переробки оригіналів на змістовному рівні, наприклад, перекази тих самих подій, текстів, ідей, опис різних аспектів різними авторами. Крім того, нечіткі дублікати можуть бути представлені в різних медіа-середовищах. В даний момент подібність узагальнених таким чином документів не завжди може бути визначена за допомогою нижченаведених методів та алгоритмів. Саме тому порівняльний аналіз електронних текстів являє собою нині відкриту науково-практичну проблему.

## 2.1. Проблематика порівняльного аналізу документів

Якщо деякий документ повністю збігається за текстом з іншим документом, то кажуть що має місце чіткий дублікат. Якщо документ за текстом співпадає не повною мірою, але є збіг за змістом, то кажуть про «майже дублікат», «нечіткий дублікат», або «текстовий синонім» [*Никконен, 2007*], [*Bourdaillet, 2007*], [*Широков, 2005*]. Текстовий синонім є розширенням поняття синонімії звичайних слів або словосполучень до повних текстів, які мають властивості семантичної подібності. Виходячи з цього якісного розуміння у наведеній роботі було надано визначення: «назвемо текст $T_1$ слабко синонімічним текстові $T_2$, якщо ці тексти відрізняються за формою, тобто за формальним зовнішнім виглядом, але близьким за змістом».

Запобігання використанню інформації, що явно дублюється, не становить проблем, проте подібні за змістом документи знаходяться не так легко. На практиці явні дублікати виявляються за допомогою так званих сигнатурних механізмів: контрольних сум, хешів, але цей підхід не вирішує всіх проблем користувачів, для яких частіше за все не має значення, з чим вони мають справу, з прямим передруком або з перефразуванням.

Загальну традиційну схему виявлення нечітких дублікатів документів наведено на рис. 16.

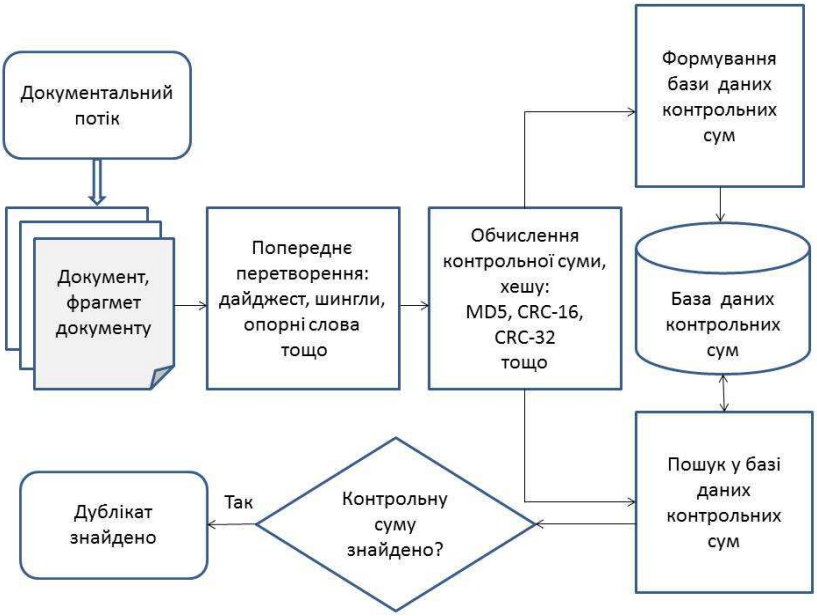



Рис. 16 – Загальна схема технології виявлення нечітких дублікатів

Важливою вимогою при реалізації алгоритмів виявлення нечітких дублікатів документів є їхня стійкість по відношенню до невеликих змін змісту документів, що аналізуються. При цьому слід відзначити, що в технологічному ланцюжку загальної схеми виявлення дублікатів документів застосовується обчислення контрольних сум, хешів (MD5, SHA-1, CRC-32 тощо), які були розроблені для задач практичної криптографії.

Необхідно відзначити, що відповідні алгоритми орієнтовані на прямо протилежну задачу, а саме, розсіювання інформації, тобто невеликі зміни у змісті вихідних документів мають призводити до кардинальних змін у контрольних сумах, хешах. Саме це протиріччя, є однією з передумов виникнення

помилок, які зустрічаються у сучасних системах виявлення нечітких дублікатів.

В рамках цієї роботи не будуть окремо обговорюватись питання точного пошуку, що фактично основані на операції порівняння двох символів. Достатньо детально вони обговорюються роботах, які вже стали класичними, наприклад [*Кормен,2006*], [*Гасфилд,2003*].

У якості найбільш вживаних сучасних алгоритмів порівняння строкових даних розглянемо метод Карпа-Рабіна [*Karp,1987*] і сигнатурні методи, у тому числі метод шинглів, запропонований А. Брьодером [*Broder,2000*] та його розвиток – метод супершинглів. Велике практичне значення мають методи лексичних сигнатур, які враховують мовну природу документів, що порівнюються. Також розглядатимуться деякі практичні реалізації розглянутих методів, певний обсяг яких наведено, наприклад, у [*Сегалович, 2007*].

Методи, засновані на урахуванні повторень ланцюжків слів, наприклад, метод «шинглів», докладно описано в роботах [*Manber, 1994*], [*Broder, 1997*] та [*Broder, 2000*]. Однак ці ефективні в багатьох випадках методи пошуку «майже дублів» виявилися не дуже чутливими для невеликих правових текстів з великим обсягом перефразувань.

Сьогодні стало звичайним звернення до статистичних підходів. У 2002 році представники компанії Яндекс оприлюднили свою методику виявлення дублікатів, засновану на аналізі *N* найбільш «якісних» слів. При цьому якість слів визначалася експертами, а відповідний математичний апарат одержав назву «нечіткої цифрової сигнатури». При цьому використовується так званий «наївний підхід» (множення ймовірностей залежних подій – слів у повідомленнях), а також елементи «ручного» відбору значимих слів (очевидно, важливість окремих слів може змінюватися у часі).

## 2.2. Формалізація відношення подібності

Першим кроком формалізації відношення подібності або слабкої синонімії (умовно будемо вважати – семантичної

близькості, але, зрозуміло, що друге поняття – більш загальне) є введення відповідної функції $F(X,Y)$, яка ставить у відповідність деякій парі документів $(X,Y)$ деяке дійсне число. Функцію $F(X,Y)$ визначено в околі [0, 1], тобто $0 \le F(X,Y) \le 1$. Необхідною і достатньою умовою співпадіння $X$ та $Y$ є $F(X,Y)=1$.

Важлива якість відношення подібності – несиметричність, тобто, у загальному випадку:

$$F(X,Y) = F(Y,X).$$

У подальшому викладенні перейдемо до скорочених позначень, а саме $F(X,Y)>0$ позначимо як $X \prec Y$ ("$\prec$" – відношення подібності), а $F(X,Y)=1$ позначимо як $X \equiv Y$ ("$\equiv$" – відношення точного дублювання).

Очевидно, що для відношення подібності і точного дублювання справедливі правила рефлексивності:

$$A \prec A, \quad A \equiv A,$$

де $A$ – довільний документ.

Відношення подібності не має властивості симетричності. Із подібності документа $A$ документу $B$ не випливає зворотне, тобто:

$$A \prec B \not\Rightarrow B \prec A.$$

Також не виконується умова транзитивності:

$$A \prec B, \quad B \prec C \not\Rightarrow A \prec C.$$

Дійсно, наприклад, окремий документ може бути подібний до тексту з добірки, яка його включає, але сама добірка може не бути подібною до цього документа. Або документ може бути подібний до двох документів, з яких він скомпільований, але самі оригінали можуть суттєво відрізнятися від нього.

Для відношення дублювання, навпаки, симетричність і транзитивність виконуються:

$A \equiv B \Rightarrow B \equiv A,$

$A \equiv B, \quad B \equiv C \Rightarrow A \equiv C.$

Зауважимо, що відношення, яке має властивості рефлексивності, симетричності і транзитивності є відношенням еквівалентності [*Шнейдер, 1971*], у нашому випадку, відношенням змістовного збігу або дублювання.

Як було відзначено, властивість дублювання документів є більш жорстким критерієм подібності, наприклад, збіг 3, 4 або 5 термів свідчить про деяку змістовну близькості, тобто можна записати:

$"\prec" \Rightarrow "\equiv".$

## 2.3. Алгоритми виявлення подібних документів

***Алгоритм Карпа-Рабіна***

Перш, ніж безпосередньо перейти до розгляду алгоритму Карпа-Рабіна введемо деякі позначення.

Очевидно, будь-який рядок у пам'яті комп'ютера представляється послідовністю байтів, кожний з яких є послідовністю бітів, тобто двійкових значень. Позначимо $T$ – двійковий рядок довжиною $|T| = m$; $P$ – двійковий шаблон для пошуку довжиною $|P| = n$. Введемо функцію:

$$H(P) = \sum_{i=0}^{n} 2^{n-i} P(i),$$

де $P(i)$ – $i$-й біт рядку $P$.

Також визначається функція від підрядку $T$ довжиною $n$ ($T_r^n$), яка починається з $r$–го біту:

$$H(T_r^n) = \sum_{i=0}^{n} 2^{r+n-i} T_r^n (r+i-1).$$

З того, що будь-яке ціле число можна єдиним чином представити у вигляді суми позитивних степенів двійки,

випливає, що підрядок $T_r^n$ входить у $T$ починаючи з позиції $r$ у тому й тільки у тому разі, коли $H(P) = H(T_r^n)$.

При пошуку підрядку у рядку, таким чином, при порівнянні $H(P)$ і $H(T_r^n)$, з того, що представлення $2^n$ числа вимагає $n$ бітів, випливає, що необхідні для порівняння числа експоненційно великі, тобто задача порівняння виявляється експоненційно складною, тобто практично не здійсненою.

У 1987 році Р. Карп і М. Рабін оприлюднили алгоритм, який отримав назву методу рандомізованих дактилограм, в якому суттєво знижена складність обчислень. Але цей алгоритм дозволяє стверджувати про входження підрядку у рядок не абсолютно точно, а з деякою високою ймовірністю. Надамо стислий зміст цього алгоритму.

Нехай $H_p(P) = H(P) \mod p$ – залишок від ділення $H(P)$ на $p$. У випадку, коли $p$ – просте число ( $p << n$ ), залишок від ділення $H_p(P)$ і $H_p(T_r^n)$ на $p$ називають дактилограмами $P$ і $T_r^n$ по модулю $p$. Звісно, $0 \le H_p(P),\ H_p(T_r^n) \le p-1$.

Якщо $P$ входить до $T$, починаючи з позиції $r$, то $H_p(P) = H_p(T_r^n)$, але зворотне, не вірно. Кажуть, що коли $H_p(P) = H_p(T_r^n)$, але $P$ не входить до $T$, то має місце помилковий збіг $P$ і $T_r^n$ з позиції $r$. Для оцінки ймовірності відсутності помилкових збігів введемо позначення $\pi(q)$ – кількість простих чисел, що не перевершують $q$. Доведено, що має місце наступна нерівність:

$$\frac{q}{\ln q} \le \pi(q) \le 1.26\frac{q}{\ln q}.$$

Крім того, має місце теорема, що справедлива для $P$ і $T$, при умові $nm \ge 29$ ($n = |P|$, $m = |T|$). Якщо $I$ – будь яке позитивне число, а $p$ – випадковим чином вибране просте число, що не перевищує $I$, то ймовірність помилкового збігу $P$ і $T$ не перевищує $\pi(nm) / \pi(I)$.

Звідси випливає такий алгоритм випадкової дактилограми для пошуку входжень $P$ до $T$:

1. Вибрати позитивне ціле число $I$.
2. Випадковим чином вибрати просте число $p$, що не перевершує $I$, та обчислити $H_p(P)$.
3. Для кожної позиції $r$ у $p$ обчислити $H_p(T_r^n)$ і зіставити з $H_p(P)$. Якщо вони рівні, то або оголосити про ймовірний збіг, або у явному вигляді перевірити збіг $P$ з $T$, починаючи з позиції $r$.

Оскільки кожне $H_p(T_r^n)$ можна обчислити за визначений постійний час з $H_p(T_{r-1}^n)$, то алгоритм дактилограми реалізується за час $O(m)$, виключаючи час явної перевірки оголошеного збігу.

Іноді виявляється зайвою перевірка збігу через занадто низьку ймовірність помилки. Справді, відомо, що коли $I = nm^2$, то ймовірність помилкового збігу не перевищує 2,53/$m$. Дійсно:

$$\frac{\pi(nm)}{\pi(nm^2)} \le 1,26\frac{nm}{nm^2}\frac{\ln(nm^2)}{\ln(nm)} = 1,26\frac{1}{m}\left(\frac{\ln n + 2\ln m}{\ln n + \ln m}\right) \le \frac{2,53}{m}.$$

Наприклад, якщо $n = 250$, $m = 4000$ і, відповідно $I = 4\times10^9$, то ймовірність того, що хоча б один з виявлених збігів буде помилковим менша за 2,53/4000, тобто не перевищить 0,001.

***Алгоритм «шинглів»***

Через десять років після оприлюднення алгоритму Кара-Рабіна А. Брьодер його співавтори [*Broder, 1997*] представили свій алгоритм, що базується на оцінці подібності документів за збігом послідовностей з визначеної кількості $n$ сусідніх слів. Такі послідовності автори назвали шинглами (від англ. *Shingles* – «*лусочки*»). Необхідно зазначити, що різні шингли формуються з послідовностей слів в нахліст, а не впритул, тобто наступний шингл починається з наступного слова, а не із слова з номером, більшим на довжину шинглу. Два документи

вважаються дублікатами, якщо множини їх шинглів суттєво перетинаються.

Таким чином, розбиваючи текст на послідовності слів (їх ще називають *N*-грамами), ми отримуємо набор шинглів у кількості $N-n+1$, де $N$ – кількість слів у документі.

При цьому на шингли розбивається кожний екземпляр документів, що порівнюються. Оскільки кількість шинглів, відповідних кожному документу є, таким чином, досить великою, було запропоновано декілька методів їх зменшення для отримання репрезентативних підмножин для порівняння.

Перший запропонований метод полягав у тому, що розглядалися лише ті шингли, чиї дактилограми, що обчислювалися за алгоритмом Карпа-Рабіна, ділилися без залишку на деяке число $m$. Основний недолік цього підходу – залежність вибірки від довжини документа, і тому невеликим за розміром документам відповідають дуже короткі вибірки, що призводить до зменшення якості виявлення дублікатів.

У відповідності з іншим запропонованим методом відбиралася лише фіксована кількість $S$ шинглів з найменшими значеннями дактилограм або залишалися всі шингли, якщо їх загальна кількість не перевищувала $S$.

Подальшим розвитком методу шинглів є методи «супершинглів» та «мегашинглів» [*Broder, 2000*].

Метод супершинглів полягає у тому, що для кожного документа вибираються випадкові набори шинглів у кількості 84. Для кожного вибраного шинглу підраховується дактилограма Карпа-Рабіна. Після цього 84 шингли розбиваються на 6 груп (супершинглів) по 14 шинглів у кожній. У результаті кожний документ представляється 6 супершинглами. Виявляється, що при умові подібності документів на рівні 95%, ймовірність збігу 2-х супершинглів становить приблизно 90%, а якщо подібність між документами лише 80%, то ймовірність збігу щонайменш двох шинглів становить лише 2,6%. Таким чином, для ефективного порівняння документів виявляється достатнім дослідити збіг лише однієї пари шинглів.

Ідея об'єднання шинглів отримала подальший розвиток в методі мегашинглів, який полягає у тому, що для кожного документа розглядаються усі пари його супершинглів. Кількість таких пар дорівнює $C_6^2 = \frac{6 \cdot 5}{2} = 15$. Стверджується, що два документи є щонайменше нечіткими дублікатами, якщо у них співпадає хоча б один мегашингл.

Необхідно зазначити, що методи, пов'язані з обчисленням шинглів, не є ефективним при порівнянні невеликих за розміром документів.

***Методи лексичних сигнатур***

При порівнянні документів має сенс враховувати мовну природу останніх, чого недостає у двох наведених вище формальних алгоритмах. Урахування цієї особливості, у деяких випадках, підвищує ефективність виявлення нечітких дублікатів або ж, за іншим визначеннями, слабко синонімічних текстів.

У цих випадках, на відмінність від алгоритму шинглів, у якості основних одиниць вимірювання використовуються слова з документів. Цей клас алгоритмів передбачає фокусування на семантичній подібності документів, не приділяючи уваги аналізу їхньої структури. Найбільш часто використовується побудова словників опорних слів і подальше порівняння словників окремих документів між собою.

У роботі [*Широков, 2005*] пропонується такий найпростіший критерій слабкої синонімії: якщо більше заданого відсотку ($n$) тексту $T_1$ присутні в тексті $T_2$ в тих самих формах, якщо більше деякого відсотку ($m$) пар слів, що стоять поруч в тексті $T_1$ присутні в тексті $T_2$ в тих самих формах, і заданий відсоток ($k$) трійок слів текстів $T_1$ і $T_2$ збігаються, то тексти є слабко синонімічними. Числа $m$, $n$, $k$ залежать від довжини досліджуваних текстів, тобто, є функціями від довжини, розміру словника текстів, жанру тощо.

Також існує декілька підходів для отримання числових значень міри близькості між документами. Якщо позначити множину опорних слів документа $T$ як $S(T)$, а $|S(T)|$ як міру

появи слів в документі, що визначається як кількість опорних слів, то міри близькості обчислюються наступним чином:

Косинус: $simCos(A,B) = \frac{|S(A) \cap S(B)|}{\sqrt{|S(A)|^2 + |S(B)|^2}}$;

Коефіцієнт Жаккарда: $simJaccard(A,B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$;

Коефіцієнт Дайса: $simDais(A,B) = \frac{|S(A) \cap S(B)|}{|S(A)| + |S(B)|}$.

В роботі [*Антонова, 2011*] показано, що у деяких випадках ефективною є несиметрична спрощена міра близькості:

$$simNSL(A,B) = \frac{|S(A) \cap S(B)|}{|S(A)|},$$

а також симетрична, що обчислюється як сума двох несиметричних:

$$simSSL(A,B) = \frac{|S(A) \cap S(B)|}{|S(A)|} + \frac{|S(A) \cap S(B)|}{|S(B)|}.$$

У цьому випадку $|S(A)|$ – кількість слів у тексті $A$, $S(A) \cap S(B)$ – перетин множин слів текстів $A$ і $B$, тобто слова із $A$ і $B$, що збігаються.

Наведений вище найпростіший критерій слабкої синонімії текстів використовує деякі числові характеристики пар текстів. Як критерій лексичної відстані двох текстів $T_1$ та $T_2$ у цьому випадку застосовують величину $simNSL(T_1,T_2)$. Із визначення випливає, що $simNSL(T_1,T_2) \neq simNSL(T_2,T_1)$, $simNSL(T_1,T_2) \leq 1$, якщо $T_1$ та $T_2$ не збігаються та $simNSL(T_1,T_2) = 1$, якщо $T_1 = T_2$. Позначимо $T_1^2$ – множину пар словоформ тексту $T_1$, що стоять поруч, а $T_1^3$ – множину трійок словоформ, і, узагальнюючи, $T_1^n$ – множину кортежів з $n$ словоформ. Лексичну відстань $n$-го порядку визначимо таким чином:

$$simNSL_n(T_1, T_2) = \frac{\left|T_1^n \bigcap T_2^n\right|}{\left(\left|T_1\right| - (n-1)\right)}.$$

Тоді наведений вище критерій можна переформулювати таким чином: якщо $simNSL(T_1, T_2) < k$, $simNSL_2(T_1, T_2) < m$, $simNSL_3(T_1, T_2) < n$, то тексти $T_1$ та $T_2$ є слабкосинонімічними.

Наведений критерій має працювати лише тоді, коли тексти мають спільні фрагменти, тобто близькі як за змістом, так і за формою. Якщо ж тексти, що порівнюються, різні за формою, пропонується застосовувати лексико-граматичну відстань між текстами $T_1$ та $T_2$, що визначається формулою:

$$simLG_n(T_1, T_2) = \frac{\sum_{i=1,|T_1|,\, j=1,|T_2|} v_i^n v_j^n \lambda^n}{\left(\left|T_1\right|\right)},$$

де $\lambda$ визначається в залежності від подібності слів ($\lambda = 1$, якщо слова співпадають $v_i$, $v_j$, ½, якщо слова $v_i$, $v_j$ – точні синоніми тощо).

Ще один з підходів, що базується на розрахунку лексичної дактилограми, має назву I-Match і був запропонований в роботі [*Chowdhury, 2002*]. Для цього для вихідної множини документів будується словник, який включає слова з середніми значеннями інверсної частоти документа IDF.

Значення IDF обчислюються за формулою:

$$IDF = \log \frac{\left|D\right|}{\left|t_j \in d_j\right|},$$

де $\left|D\right|$ – кількість документів у колекції; $\left|t_j \in d_j\right|$ – кількість документів, де зустрічається $t_j$.

Слова з середніми значеннями IDF, як правило, забезпечують більш точні результати при виявленні нечітких дублікатів. Після цього для кожного документа вибираються слова, які також входять до сформованого словника. Множина

цих слів впорядковується і розраховується відповідна контрольна сума (застосовується хеш-функція SHA-1), яка й має назву I-Match. Два документи вважаються нечіткими дублікатами, якщо в них відповідні значення I-Match співпадають. Наведений алгоритм з погляду обчислювальної складності більш ефективний, ніж алгоритм шинглів, крім того, він може застосовуватися для порівняння невеликих за розміром документів. Однак застосування жорсткої хеш-функції SHA-1 робить його нестійким до невеликих змін змісту документів.

Ще один сигнатурний метод, що застосовується у службі «Яндекс.Новини» для виявлення сюжетів новин, базується на визначенні опорних слів [*Ilyinsky*, 2002]. У відповідності з цим алгоритмом вибирається множина з $N$ опорних слів. Після цього кожному документу ставиться у відповідність $N$-вимірний двійковий вектор, координати якого відповідають присутності слів (приймають відповідні значення 0 або 1). Цей двійковий вектор називають сингантурою документа. Два документи вважаються подібними, якщо їхні сигнатури збігаються.

Зауважимо, що принципи вибору опорних слів можуть розрізнятися у різних реалізаціях.

## 2.4. Практика виявлення подібних документів

### *Дублювання документів у веб-просторі*

Одним з ключових аспектів розвитку сучасних інформаційних технологій є специфіка взаємин між інформаційними агентствами (ІА), які традиційно грають роль постачальників інформації, і ЗМІ, які є основним її споживачем. Мабуть, ці взаємини значною мірою застаріли і потребують серйозних коректив як у технологічному плані, так і в плані організаційному, включаючи законодавче регулювання.

Головна причина такого стану справ полягає у швидкому розширенні впливу на інформаційні процеси мережевих технологій і, зрозуміло, у першу чергу Інтернет. Розвиток цих технологій призвело до якісних змін у структурі всього процесу інформування громадськості на всіх його ланках, в результаті чого ситуація потребує кардинального перегляду основних

механізмів, що лежать в основі функціонування медійних засобів.

Інформаційні агентства постачають своїм передплатникам інформацію на умовах, які на сьогоднішній день виглядають щонайменше дивно. Зокрема, типовою умовою щодо використання матеріалів ІА є заборона на розмноження та розповсюдження їх у будь-який спосіб. Таким чином агентства намагаються захистити свою продукцію від копіювання, часто посилаючись на законодавство про авторські права. У статті 10 Закону України «Про авторське право і суміжні права» та аналогічному Законі Російської Федерації «Об авторском праве и смежных правах» у статті 8 передбачено, що повідомлення про новини або поточні події не охороняються авторським правом. Таким чином, умови, декларовані більшістю ІА з посиланням на законодавство про авторські права, є неправомірними, принаймні, по відношенню до їх основної продукції – інформаційних повідомлень.

Не краща ситуація і зі змістовним аспектом проблеми. Ніхто, безумовно, не ставить під сумнів авторські права на ті матеріали, які дійсно мають автора в звичайному сенсі (інтерв'ю, аналітичні розробки, ексклюзивні репортажі тощо). Але говорити про авторські права на повідомлення про офіційний візит глави держави або набуття чинності нового закону явно позбавлене конкретного сенсу. Ми вже не говоримо про тексти законів, указів тощо, для яких законодавчо передбачено порядок оприлюднення.

Як завжди в подібних ситуаціях, нові тенденції починають прокладати собі дорогу, не чекаючи офіційних рішень, що неминуче призводить до перерозподілу не тільки ресурсів, але і функціональних ролей учасників комунікації. Тому для вироблення обґрунтованих рекомендацій бажано було б розібратися у тому, що і як відбувається насправді.

У зв'язку з цим як науковий, так і практичний інтерес викликає питання, якою мірою матеріали, доступні платним передплатникам основних ІА, стають доступними у відкритому доступі на інформаційних веб-сайтах. Цінність інформаційних повідомлень багато в чому визначається оперативністю, тому

окремим завданням є оцінка запізнювання публікацій в Інтернет у порівнянні з часом розсилки відповідних повідомлень. Забігаючи наперед, скажемо, що майже в третині розглянутих випадків час затримки виявися негативним, тобто ІА копіювали повідомлення з веб-сайтів, та ще й зі значним запізненням.

При проведенні досліджень автор мав унікальну можливість доступу до передплатних матеріалів провідних ІА, представлених в українському інформаційному просторі. Крім того, у розпорядженні автора перебувала система контент-моніторингу InfoStream, за допомогою якої в реальному масштабі часу скануються тисячі інформаційних веб-сайтів, представлених в українському та російському сегментах веб-простору. Таким чином, в ході дослідження розглядалися два текстових корпуси (точніше, набори «словесних сигнатур» текстів [*Ландэ*, 2006], представлених у цих корпусах) – повідомлень ІА і текстів з веб-простору.

Як репрезентативна множина текстів розглядалися повідомлення ІА із загальнополітичної тематики, що надходили впродовж 5-25 листопада 2007 року. Їх обсяг дорівнював 8955 документів. Ці повідомлення порівнювалися з текстами, з веб-простору, що сканувалися впродовж листопаду 2007 року, кількість яких склала понад 1 млн. документів.

Технічно завдання знаходження дублікатів (у цьому випадку мова йшла саме щодо дублікатів, а не повідомлень з тієї ж тематики, враховувалися лише передруки з незначними спотвореннями) вирішувалось методом, який описано в роботі [*Ландэ, 2006*]. Цей метод належить до групи методів знаходження «нечітких» дублікатів, заснованих на виділенні деякої множини опорних слів.

Аналіз новинних повідомлень показав, що при передруці матеріалів найчастіше залишаються без змін декілька перших речень тексту або перший абзац. Багато з недобросовісних видань передруковують зміст повідомлень, просто змінюючи назви (робота так званих «хедлайнерів») – цей чинник був також враховано. Такий вид дублювання успішно виявляється шляхом використання сигнатур: шинглів, хешів, тощо. (але вже без урахування заголовків).

Як деякі «інваріанти» для окремих повідомлень використовувалися лінгвістичні сигнатури: ланцюжки з 12 опорних слів, які пройшли процедуру морфологічної обробки (стемінгу). Така невелика кількість опорних слів в ланцюжку, який є своєрідною словесною сигнатурою, пояснюється невеликою середньою довжиною новинних повідомлень (2000–3000 символів).

В результаті проведених досліджень вдалося отримати такі дані:

- з 8955 повідомлень ІА на веб-сайтах було опубліковано 5567 повідомлень (62%);
- загальна кількість передруків на різних веб-сайтах склала 39901 (456%). Відповідний розподіл, який виявився гіперболічним, наведено на рис. 6;
- кількість передруків з позитивним часом запізнювання (з матеріалів ІА – на веб-сайти) склало 28 933 (73%);
- кількість передруків з негативним часом запізнювання (передруків з Інтернет у стрічки ІА) склало 10 968 (27%).

Ранжируваний графік розподілу повідомлень ІА за часом затримки публікацій наведено на рис. 6, на якому чітко видно екстремальні відхилення у початковій і кінцевій області. Відхилення у початковій області характеризує великий час затримки включення до стрічки ІА матеріалів, розміщених, як правило, на сайтах органів державної влади (інертність ІА, відсутність у них коштів для здійснення моніторингу веб-простору).

Відхилення в кінцевій області пояснюються затримками передруків на веб-сайтах повідомлень (рис. 17), які отримали з часом деяке нове продовження. Разом з тим центральна область графіка (від 1000-го по 5000-е повідомлення) має стабільний характер з середнім значенням запізнення близько півгодини (рис. 18).

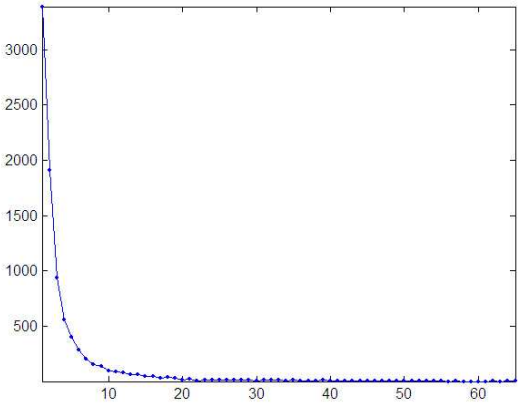



Рис. 17 – Кількість повідомлень ІА (вісь ординат), ранжируваних за кількістю передруків на веб-сайтах (вісь абсцис)

Масовий характер передруків дозволяє робити висновки, що практично всі повідомлення, цікаві адміністраторам відповідних веб-сайтів, передруковуються. Мабуть, приблизно 38% повідомлень ІА виявилися недостатньо цікавими.

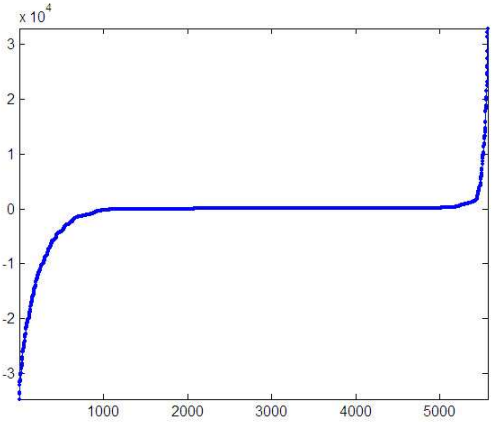



Рис. 18 – Розподіл повідомлень ІА (вісь абсцис) за часом запізнювання в хвилинах (вісь ординат)

У результаті дослідження виявилося, що методи визначення нечітких дублікатів повідомлень, які були розвинені в останні роки, виявилися корисними у цьому застосуванні.

Результати змушують замислитися, за що ж платять передплатники інформаційним агентствам сьогодні, коли більша частина інформації з мінімальною затримкою доступна в Інтернеті, а повноту можуть забезпечити системи контент-моніторингу? Мабуть, за аналітичний підбір цієї інформації, репрезентативність і достовірність. Тобто, інформаційне агентство, якщо воно бажає вижити у сучасних умовах, має приділяти підвищену увагу саме аналітичній обробці інформації, перетворюючись на агентство інформаційно-аналітичне.

### *Виявлення подібних документів у системі InfoStream*

Для виявлення подібних документів у системі контент-моніторингу InfoStream [*Григорьев*, 2007] застосовується досить потужний інформаційний ресурс – ретроспективна база даних цієї системи обсягом понад 100 млн. документів з понад 7000 джерел. Слід зазначити, що відсоток повідомлень, що дублюються, у системі InfoStream значно менше, ніж у всьому веб-просторі. Це пояснюється підбором джерел для сканування, до числа яких входять лише ті, що переважно публікують оригінальні матеріали. Виявлення опорних слів у системі InfoStream здійснюється після морфологічної обробки (стемінгу). Морфологічна обробка базується на використанні морфологічних частотних словників, до яких увійшли іменники, загально відомі прізвища та назви фірм і організацій. Обчислення вагових коефіцієнтів слів проводиться на підставі вагового підходу, а саме, модифікації стандартного підходу TF IDF – Okapi BM25.

Виявлення дублікатів і подібних документів (текстових синонімів) повідомлень у системі InfoStream виконується на основі методу, що полягає у знаходженні в різних документах загальних опорних слів, з ланцюжків яких утворюються лінгвістичні сигнатури.

Визначення змістовно подібних документів і дублікатів, що застосовується в системі InfoStream на цей час, полягає в тому, щоб вважати дублікатами документи, в яких $n$ ($n$ = 6) опорних слів одного документа збігаються опорними словами іншого документа з $N$ ( $N$ = 12) можливих. Слід зазначити, що

застосування більш «м'якого» критерію (меншої кількості опорних слів *n*, що збігаються) до множини відібраних термів дозволяє реалізувати режим «пошуку подібних документів».

Нижче буде описано підхід до виявлення дублікатів документів, наведених різними мовами, при цьому використовуються перекладні еквіваленти опорних слів різними мовами, отриманих за допомогою нового підходу до виявлення дублікатів.

***Визначення інформаційних сюжетів***

Визначимо інформаційний сюжет як множину документів (інформаційних повідомлень), присвячених одній тематиці або одній події. «Середовищем існування» інформаційних сюжетів є інформаційний простір, що сьогодні цілком репрезентативно представляється мережею Інтернет, що проте не обмежує авторів розглядом тільки цієї мережі. Інформаційні сюжети можна трактувати як документальні або контентні системи (від англ. *Content – «зміст»*).

Інформаційні сюжети можуть бути представлені як мережеві структури, так звані динамічні мережі. Поточний стан інформаційного сюжету може бути представлений у вигляді графа $< M, L >$, де $M$ – це множина документів, що входять до сюжету, а $L$ – множина ребер – зв'язків подібності, цитування, посилань тощо.

На практиці при пошуку новинної інформації завжди виникає завдання виявлення інформаційних сюжетів, які складаються з окремих документів, і їх ранжирування за деякими ознаками, що повинне забезпечити, не лише виявлення найважливішої теми, але і багатоаспектне висвітлення усіх найбільш значущих аспектів. Це завдання, що вирішується в багатьох системах з використанням різних підходів і алгоритмів. При цьому незмінним залишається технологічний ланцюжок: побудова семантичної мережі з інформаційних повідомлень, кластеризація – виявлення найбільш взаємозв'язаних груп, тобто інформаційних сюжетів, «зважування» (оцінка важливості, актуальності) і наочна візуалізація найвагоміших з них [*Ландэ, 2005*].

При виділенні сюжетних ланцюжків для визначення попарної близькості окремих документів, як правило, використовуються алгоритми виявлення подібних документів, що застосовуються в пошукових системах.

Для пред'явлення користувачам інформаційні сюжети мають бути ранжирувані. Основні чинники, що впливають на ранжирування за важливістю, – оперативність інформації і розмір сюжетного ланцюжка. Під оперативністю розуміється деяка функція від часу публікації усіх документів в інформаційному сюжеті, а розмір сюжету відбиває загальний інтерес до конкретної теми. В усіх цих підходах центральне завдання полягає в ототожненні документів, що відносяться до одного сюжету і відокремлення документів, що слабко «перетинаються».

На рис. 19 представлено типовий алгоритм виявлення інформаційних сюжетів. Останнє за часом генерації інформаційне повідомлення порівнюється з попередніми, оцінюється рівень їх подібності. Якщо рівень подібності з деяким раніше сканованим документом перевищує деякий поріг, документ, що аналізується вважається належним до інформаційного сюжету, до якого відноситься раніше прийнятий документ. Якщо подібних документів не знаходиться, фіксується новий сюжет, який складається спочатку з одного документу.

Для результуючого відображення кожного інформаційного сюжету використовуються відібрані за змістовною близькістю документи з різних джерел, відсортовані в хронологічному порядку. При цьому сюжети можуть розглядатися як дайджести, що інтегрують унікальну інформацію, яка міститься в окремих документах.

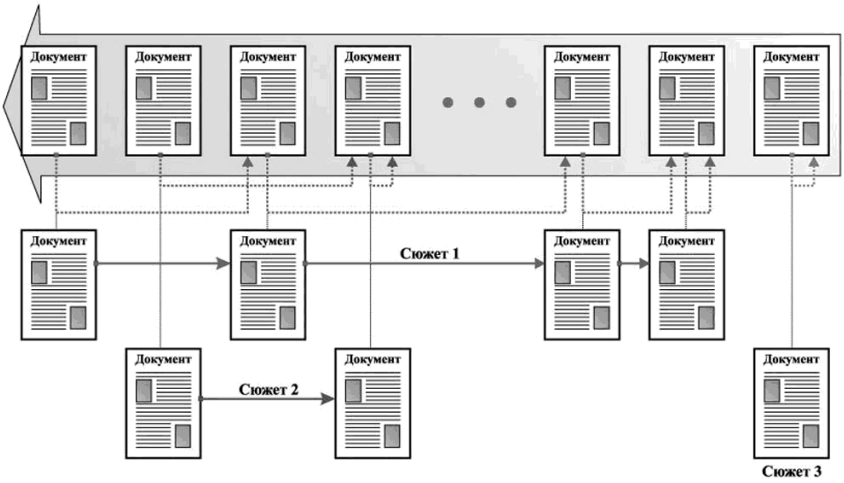



Рис. 19 – Типовий алгоритм виявлення інформаційних сюжетів

У системі «Яндекс.Новини» (http://news.yandex.ru) для виділення інформаційного сюжету будується матриця попарної близькості документів, яка обробляється алгоритмом кластеризації з емпірично підібраними параметрами. Для того, щоб збільшити зв'язність великих сюжетів, в «Яндекс.Новини» додатково використовується кластеризація, що забезпечує групування атомарних кластерів до більших. Усі повідомлення в результатах пошуку у системі «Яндекс.Новини» згруповані, при цьому ранжирування інформаційних сюжетів побудовано на стандартних для Яндексу принципах ранжирування видачі. Воно засноване на часі публікації та кількості новин усередині новинних сюжетів, при цьому ранг однієї новини визначається як її «свіжість» з урахуванням пріоритетів задоволення критеріїв пошуку.

У системі InfoStream (http://infostream.ua) тематична близькість документів визначається на основі нормованих послідовностей найбільш вагомих термінів, що входять до кожного документа. Послідовності подібних (з певним коефіцієнтом взаємної близькості, що перевищує деякий встановлений емпірично рівень) документів утворюють ланцюжки. При цьому кожен документ потрапляє до деякого ланцюжку, що у крайньому випадку складається тільки з нього

самого. Потім ланцюжки зважуються за довжиною та оперативністю, після чого користувачеві пред'являється певна кількість самих важливих тематичних інформаційних сюжетів. Для репрезентації сюжетного ланцюжка, заголовки документів також зважуються відносно ключових слів, відповідних сюжету, а потім з усіх заголовків вибираються найбільш «вагомі» для відображення (рис. 20).

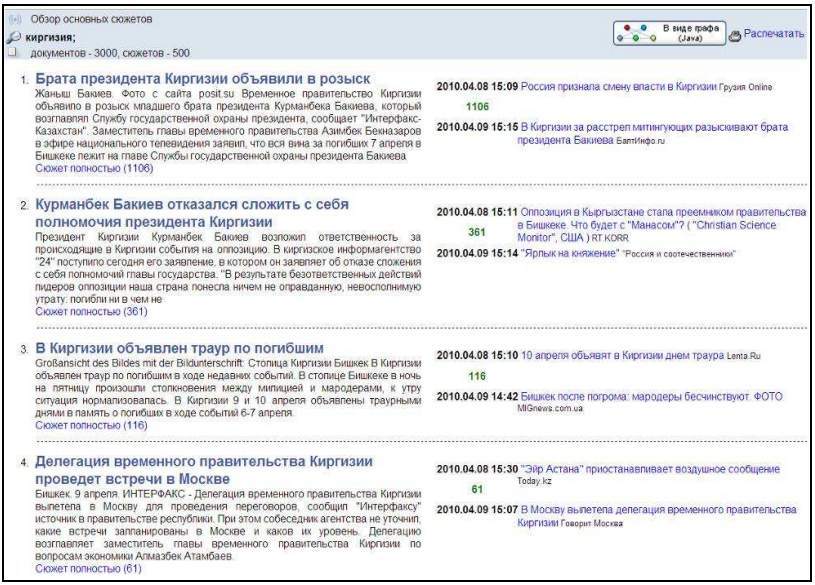



Рис. 20 – Приклад відображення інформаційних сюжетів у системі InfoStream

### *Підходи до визначення нових подій і виявлення спаму*

Для часткового вирішення завдань визначення нових подій і виявлення спам-повідомлень в інформаційних потоках реалізовано окремий сервіс, близький за ідеологією до режиму «пошуку подібних документів» в системі контент-моніторингу InfoStream.

В рамках системи InfoStream повідомлення вважається подібним до вихідного, якщо містить певну кількість найбільш опорних слів з нього.

Під спам-популярністю повідомлення будемо розуміти кількість так само подібних йому повідомлень в текстовому корпусі спаму. Під ЗМІ-популярністю розуміється кількість подібних повідомлень в ретроспективній базі електронних засобів масової інформації. Масив повідомлень, завідомо точно визначених авторами як спам, був ранжируваний за спам-популярністю. Отримана залежність «спам-популярність – кількість повідомлень» виявилася близькою до гіперболічної. Для кожного з повідомлень, ранжируваних за значенням ЗМІ-популярності, також побудована залежність «ЗМІ-популярність – кількість повідомлень».

Було виявлено деяку кількість повідомлень, що характеризуються великим співвідношенням спам-популярності до ЗМІ-популярності. Цей факт дозволяє судити про сукупність термінів, що визначають спам-популярність, як про ще один фільтр, який можна реалізувати в антиспамівському програмному забезпеченні. Повідомлення, у яких ЗМІ-популярність перевищує спам-популярність, але все ж були спамом, виявляються несанкціонованими розсилками інформаційно-аналітичних матеріалів, які представляють певний інтерес для інформаційних агентств.

Таким чином, представлений підхід до виявлення спам-повідомлень, додаткової селекції спаму. При цьому здається істотним виявлення близькості досліджуваного повідомлення не тільки до корпусу спаму, але й до корпусу електронних ЗМІ.

Під локальною популярністю повідомлення в інформаційному потоці розуміється кількість йому подібних повідомлень за той період (годину, день, тиждень), коли з’явилося вихідне повідомлення. Під глобальною – кількість подібних повідомлень за значний ретроспективний період.

Особливий інтерес представляють повідомлення, що характеризуються великим співвідношенням локальної популярності до глобальної. Цей факт дозволяє судити щодо подій, які описуються в даних повідомленнях, як про нові. Таким чином маємо алгоритм виявлення документів, що отримали популярність тільки останнім часом, який є окремим рішенням

актуальної проблеми виявлення нових подій із документальних потоків [*Снарский, 2007*].

Загальноприйнята технологічна схема вирішення задачі виявлення нових подій з потоку новин, як правило, передбачає, що нові події описуються в документах, для яких в часовій ретроспективі за допомогою окремих програмних модулів формуються ланцюжки подібних документів (сюжетні ланцюжки). Документи, що відображають різні нові події можуть бути основою нових груп взаємопов'язаних документів – кластерів (групування подій). У свою чергу, кожен з цих кластерів з часом може стати основою формування повноцінного сюжетного ланцюжка.

Наведемо деякі припущення, що стосуються документів, які містять інформацію про нові події:

а) мінімальний час, що минув з моменту публікації документа;

б) близькість лексичного складу документа до лексичного складу масиву документів за невеликий проміжок часу (оперативність новин);

в) істотна відмінність лексичного складу документа від лексичного складу масиву документів за значний період часу – вікна спостереження;

г) наявність у документі слів, що входять до плюс-словника, який включає важливі для змісту новин слова типу «теракт», «конфлікт», «сенсація» тощо);

д) високий ранг репутації джерела, а також допустимість лексики заголовків новин (що визначається експертами);

е) відсутність дублювання інформації.

Введемо позначення: $N$ – величина вікна спостереження потоку новин; $n$ – величина масиву оперативних новин ($n < N$); $D_i$ – $i$-й документ; $PlusDic$ – плюс-словник; $sim(D_i, D_j)$ – міра близькості документа $i$ документу $j$; $sim(D_i, PlusDic)$ – міра близькості документа $i$ плюс-словнику; $Rang_i$ – ранг джерела, яке відповідає $i$-му документу.

У цих позначеннях міра близькості лексичного складу документа $D_i$ від лексичного складу масиву останніх новинних повідомлень розраховується наступним чином:

$$\sum_{j=1}^{n} sim(D_i, D_j),$$

Відповідно міра близькості лексичного складу документа від лексичного складу усього масиву документів із вікна спостереження обчислюється таким чином:

$$\sum_{j=n}^{N} sim(D_i, D_j), \quad N >> n.$$

При цьому міра близькості окремих документів, обислюється як введена вище α-подібність.

Формула для обчислення рангу документа як «носія» інформації щодо нових подій з урахуванням умов а) – е) може бути записана наступним чином:

$$N(D_i) = \frac{Rang_i \cdot sim(D_i, PlusDic) \cdot \sum_{j=1}^{n} sim(D_i, D_j)}{\log(i+1) \cdot \sum_{j=n}^{N} sim(D_i, D_j)},$$

з урахуванням наведених вище позначень, а також того, що якщо нумерація документів з потоку проводиться у зворотному порядку, значення $\log(i+1)$ в знаменнику відображає внесок часу, що пройшов з моменту публікації події.

На основі наведеної формули може відбуватися ранжирування документів, що надходять до системи інтеграції новин.

Даний алгоритм реалізує прогнозно-аналітичну модель, основна методологія оцінки достовірності якої в даний час полягає в експертному порівнянні виявлених нових подій з основними сюжетами, отриманими через певний інтервал часу. Для настройки алгоритму експертами використовувалися такі «важелі», як параметри *N*, *n*, плюс-словник, масив рангів джерел інформації, масив винятків для заголовків та адрес.

В даний час в багатьох популярних системах інтеграції новин завдання виявлення нових подій замінюється виявленням

основних новинних сюжетних ланцюжків. Такий підхід, звичайно, частково вирішує назване завдання, однак, надаючи користувачам відповідь на питання «про що найбільше пишуть останнім часом», фактично відрізняється цільовою функцією.

Було проведено ретроспективне дослідження з метою оцінки, наскільки сьогоднішні події, що визначаються відповідно до запропонованого підходу, стануть основою сюжетів наступного дня. Виявилося, що таких подій не більше 20%. Найчастіше більша частина сюжетів наступного дня повторює сюжети дня попереднього. Доводиться визнавати, що не всі нові події однакові за важливістю і породжують в подальшому значні кластери подібних документів.

Запропонований підхід, звичайно ж, не можна вважати остаточним рішенням поставленого завдання. Наприклад, не завжди зміна розмірів вікон спостереження та обсягів оперативних масивів може привести до адекватного виявлення нових обставин, які мають свою передісторію. Розглянутий плюс-словник вимагає постійного супроводу, а в деяких випадках «персоналізації».

Однак, отримані практичні результати показали свою ефективність як істотне доповнення до пошукових режимів. При цьому найважливіше, мабуть те, що користувач прив'язується не до нових повідомлень, а до нових подій реального світу.

***Якість виявлення подібних документів***

Нижче описуються критерії якості виявлення подібних документів, що базуються на аналізі таких властивостей матриць подібності, як симетричність і транзитивність [*Ландэ, 2006*], [*Ландэ, 2009а*] розглядаються аналітичні вирази для розрахунку цих критеріїв, а також наводяться результати експериментів на багатомовних текстових корпусах, які формуються за допомогою системи контент-моніторингу.

На практиці кожному документу $D_i$ з контрольного документального корпусу за алгоритмом збігу термів в сигнатурах (у різних експериментах змінювалась необхідна кількість термів, що збігаються) ставився у відповідність вектор з елементами:

$$a_{ij} = \begin{cases} 1, & D_i \equiv D_j, \\ 0, & D_i \not\equiv D_j. \end{cases}$$

Умова симетричності у цих позначеннях записується таким чином:

$$\forall \; i, \; j: \; a_{ij} = a_{ji},$$

а умова транзитивності такі:

$$\forall \; i, \; j, \; k: \; a_{ij} = 1, \; a_{jk} = 1 \Rightarrow a_{ik} = 1.$$

Згідно з наведеними міркуваннями були запропоновані критерії, що базуються на обчисленні коефіцієнтів симетричності ( $S$ ) і транзитивності ( $T$ ) для матриці подібності. На контрольному документальному корпусі, змінюючи кількість порівнюваних в сигнатурах термів, були отримані різні значення відповідних коефіцієнтів. Коефіцієнт симетричності обчислюється наступним чином:

$$S = 2\frac{\sum_{i}^{N}\sum_{j\neq i}^{N} a_{ij}a_{ji}}{\sum_{i}^{N}\sum_{j\neq i}^{N} a_{ij}},$$

а коефіцієнт транзитивності визначається за формулою:

$$T = \frac{\sum_{i}^{N}\sum_{j\neq i}^{N}\sum_{\substack{k\neq i \\ k\neq j}}^{N} a_{ij}a_{jk}a_{ik}}{\sum_{i}^{N}\sum_{j\neq i}^{N}\sum_{\substack{k\neq i \\ k\neq j}}^{N} a_{ij}a_{jk}}.$$

де $N$ – кількість документів у контрольному корпусі.

Очевидно, що коефіцієнт симетричності, який обчислюється таким чином, асоціюється з точністю при

визначенні дублікатів документів, а рівень транзитивності – з повнотою.

Разом з тим слід відзначити, що перевірка коефіцієнтів асиметричності і транзитивності може використовуватися лише для формальної перевірки наближення відношень до властивостей еквівалентності. Саме визначення того, що ця еквівалентність – змістовне дублювання, підтверджується аналітиками-експертами. Наведений вище алгоритм крім свого емпіричного підтвердження має ту перевагу, що дозволяє варіювати деяким числом (кількістю порівнюваних термів в сигнатурах), значення якого можна підібрати з урахуванням оптимізації двох названих коефіцієнтів.

## Порівняльний аналіз різномовних документів

### *Вибір моделі представлення текстів*

Нижче описано запропонований у [*Ланде, 2009а*] підхід, за допомогою якого можливо виявити нечіткі дублікати документів, наведених різними, а саме, українською та російською мовами. Процедуру виявлення дублікатів, що детально розглядається, побудована на використанні методів витягу опорних слів, що базується на статистичних властивостях тексту, використання частотного морфологічного словника, а також двомовних словників перекладів.

Для виділення ключових слів з тексту в рамках підходу, що розглядається, використовується векторна модель документа, у відповідності до якої, кожному слову документа приписується його ваговий коефіцієнт. Чим більше вага слова, тим більше це слово характеризує документ. В рамках цієї роботи було перевірено два підходи обчислення вагових коефіцієнтів слів.

Відомо, що стандартний метод TF IDF не зовсім коректно працює на великих текстових масивах. Як альтернативу запропоновано його модифікацію – Okapi BM25 [*Salton, 1988*]:

$$TF \cdot IDF = \sum_{i=1}^{n} IDF(w_i) \frac{f(w_i, D) \cdot (k_1 + 1)}{f(w_i, D) + k_1 \cdot (1 - b + b \cdot |D| / avgdl)}$$

де $f(q_i,D)$ – частота слова $w_i$ у документі $D$; $|D|$ – довжина документа $D$ (число слів); *avgdl* – середня довжина документа в колекції; $k_1$ і $b$ – вільні параметри, зазвичай обрані як $k_1 = 2.0$ і $b = 0.75$. $IDF(q_i)$ – IDF інверсна частота документа, що обчислюється за формулою:

$$IDF(q_i) = \log \frac{N - n(w_i) + 0.5}{n(w_i) + 0.5},$$

де $N$ – загальна кількість документів у масиві, $n(w_i)$ – кількість документів, що містять термін $w_i$.

Суть цього підходу полягає у тому, що на відміну від первинного підходу TF IDF в Okapi BM25 береться до уваги довжина документа.

***Загальний опис алгоритму***

Створення системи виявлення різномовних дублікатів можна представити у вигляді декількох етапів (рис. 21), а саме:

- створення морфологічних словників;
- створення частотних словників – навчання системи;
- створення словників перекладів;
- побудова програмами пошуку ключових слів;
- створення процедури пошуку дублікатів, наведених різними мовами.

Спочатку створюються морфологічні словники, які для кожної словоформи містять її імовірну нормальну форму для подальшої нормалізації знайдених словоформ. Для побудови електронних морфологічних словників за основу бралися електронна версія словника Залізняка, який налічує близько 93 тис. слів у нормальній формі (для російської мови) та безкоштовний словник ispell, що налічує близько 1 мільйона словоформ, відповідно, для української мови. Морфологічні словники були доповнені відомими власними іменами, назвами установ і організацій, яких у них не було.

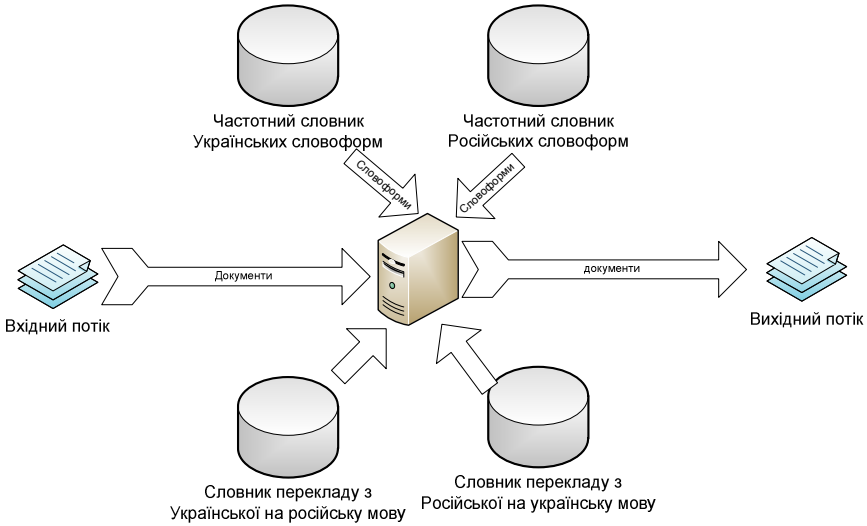



Рис. 21 – Інформаційні потоки процедури пошуку та перекладу опорних слів

Після цього на базі морфологічного словника, в якому записується частота кожної словоформи, знайденої в процесі «навчання» частотного словника на тестовому масиві документів, створюється частотний словник. У частотному словнику для кожного слова записана кількість його поява у деякому великому масиві документів, а також кількість документів, у яких знайшлося це слово. Для створення частотного словника використовувався корпус новинних документів, які скануються з Інтернет системою контент-моніторингу InfoStream. Корпус складався з текстів веб-публікацій українською (1 344 086 документів) і російською (2 399 367 документів) мовами. При машинному навчанні частотного словника з кожного документа в корпусі витягалися словоформи, які (з певною ймовірнісною похибки) були приведені до нормальної форми. При цьому підраховувалася кількість, як словоформ, так і нормальних форм у документах, а також підраховувалася кількість документів, у яких зустрілася словоформа або нормальна форма.

Для пошуку опорних слів у результуючі словники заносилися тільки ті слова, що зустрічалися у документальному корпусі понад двох разів.

Навчання словника, проходило у три етапи. Перший етап полягав в поділі документів на словоформи і запису отриманих словоформ у тимчасовий файл. На другому етапі, створений файл сортується за словоформою та відповідним номером документа. Далі підраховується кількість входжень однієї словоформи та кількість документів, у яких вона зустрічалася. Знайдені значення частот записуються у частотний словник, після чого відбувається пошук нормальної форми. Остаточна нормальна форма і номери документів записуються в окремий файл-словник.

Для вирішення завдання побудови паралельних текстових корпусів у результуючі словники відбираються тільки словоформи іменників.

Однією з основних проблем при автоматичному аналізі текстів є омонімія. Існуючі підходи зняття омонімії можна розділити на два основні типи: детерміновані та імовірнісні. До детермінованих можна віднести методи, вживані, наприклад, в системі «ЕТАП» [*Цинман, 2000*], де використовується «фільтровий метод» синтаксичного аналізу, система «Диалинг» [*Сокирко, 2005*], або морфологічний аналізатор англійської мови ENGTWOL [*Jurafsky, 2000*], що засновані на правилах зняття неоднозначності на основі контекстних правил. Імовірнісний підхід до подолання омонімії широко обговорювався в роботах російських дослідників [*Зеленков, 2005*], [*Баглей, 2007*], [*Зинькина, 2005*], застосовувався для зняття неоднозначності у іменників шляхом використання розмічених вручну текстових корпусів і вибору лексичних і граматичних ключів ще у 80-х роках XX століття в системі М. Харста [*Hearst, 1991*].

Описаний нижче підхід передбачає використання алгоритму зняття контекстної неоднозначності, так як омонімія є суттєвою проблемою при визначенні опорних слів документа, наприклад, слово «села», яке у практиці російської мови може бути множиною від слова «село», а також похідною від дієслова «садиться», може некоректно перекладатися і використовуватися

українською мовою, тому що слово «село» перекладається на українську як «село», а слово «садиться» – «сідати».

Неправильний вибір нормальної форми може призвести до того, що в однакових за інформаційним змістом документах, наведених різними мовами, будуть використані різні опорні слова. Для вирішення цієї проблеми може застосовуватися, як виявилося пізніше, ефективний і досить швидкий алгоритм, що особливо важливо, оскільки етап навчання частотних словників і етап їх використання пов'язані з обробкою великих обсягів текстової інформації.

У табл. 4 показано приклад навчання частотного словника для російських слів «садиться» і «село». Запропоновано правило, відповідно до якого, якщо в систему надійшла словоформа, яка на практиці може призводити до декількох нормальних форм (наприклад, для словоформи «села» допустимі нормальні форми «село» і «садиться»), то так звані «індекси нормальних форм» для цієї словоформи збільшуються на одиницю.

У табл. 4 показано приклад, коли у російськомовному текстовому корпусі словоформа «села» зустрілося 20 разів, словоформа «село» – 50 разів, словоформа «сели» – 10 разів, а словоформа «селом» – 30 разів. У результаті навчання, у словники потрапляють слова «село» з індексом нормальної форми 100 і «садиться» з індексом 80, відповідно, у подальшому при відборі опорних слів перевага буде надаватися слову «село».

У рамках даних досліджень використовувалися словники перекладів з російської на українську, і з української на російську мови. Вихідні дані для побудови словників перекладів були отримані шляхом перекладу іменників в нормальній формі існуючими програмами перекладу текстів. Якщо одному слову відповідало декілька перекладів, то вибиралось найбільш уживане значення у відповідності з частотним словником.

Для кожного документа, який зчитується із вхідного потоку, відбувається його розподіл за словоформами. Після цього відбувається пошук нормальної форми для кожної словоформи. У випадку омонімії, вибирається та нормальна форма, що є найбільш частотною за словником. Далі відбувається підрахунок кількості словоформ.

Табл. 4. Приклад навчання системи

| Словоформа | Кількість | Індекс нормальних форм |
|---|---|---|
| села | 20 | садиться → +20<br>село → +20 |
| село | 50 | садиться → +50<br>село → +50 |
| сели | 10 | садиться → +10 |
| селом | 30 | село → +30 |
| | | село = 100<br>садиться = 80 |

***Формування опорних слів документів та їх перекладів***

Опорні слова витягаються за допомогою формули Okapi BM25. Після обчислення вагових коефіцієнтів відбувається ранжирування слів і вибираються перші $N$ ($N$ = 12). Отримані $N$ опорних слів перекладаються з однієї мови на іншу за допомогою словників перекладів. Всі опорні слова і слова-переклади приписуються до документа.

Одним з методів оцінки якості витягу опорних слів є виявлення кількості небажаних для перекладу слів (омонімів), що потрапляють до їх складу при перегляді документів різної довжини, створених за допомогою різних алгоритмів. Для проведення оцінки бралися 1000 довільно обраних документів, які мають різну довжину. Після цього відбувалося обчислення опорних слів відразу за двома алгоритмами та за допомогою експертних оцінок вираховувалась загальна кількість омонімів. У результаті було виявлено що класичний підхід TF IDF поводиться стабільно на відносно малих документах, що містять 100–150 слів. При перевищенні цієї межі в опорні слова потрапляли небажані омоніми. На відміну від TF IDF, застосування Okapi BM25 не привело до витягу небажаних слів, практично, на всіх документах.

***Пошук різномовних дублікатів документів***

Пошук різномовних дублікатів в рамках підходу, що розглядається, здійснюється у два етапи. На першому етапі

проводиться пошук дублікатів документів, наведених різними мовами, за допомогою системи InfoStream. Системі подавалося п'ять опорних слів з документів, наведених українською мовою, що являли собою, перекладені опорні слова з української на російську мову (рис. 22).

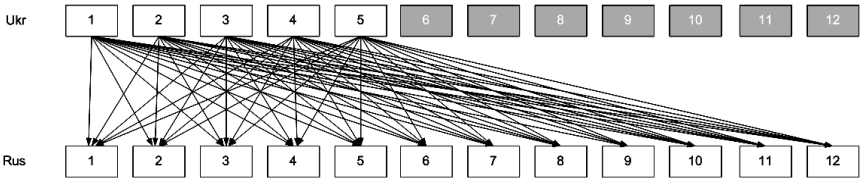



Рис. 22 – Порівняння опорних слів

Далі проводиться порівняння поданих опорних слів із дванадцятьма опорними словами документів, наведених російською мовою. Після цього проводилася фільтрація небажаних, «неповних» дублікатів документів. Для цього були використані такі додаткові критерії відсіювання не повних дублікатів:

- загальна кількість слів у перекладеному варіанті не повинна відрізнятися більше ніж на 10%;
- кількість слів, які починаються з великої букви (не на початку рядку), не повинно відрізнятися більше ніж на 3 слова;
- кількість чисел у документах не повинна вирізнятися більше ніж на два;
- знайдені числа в документах не повинні відрізнятися більш ніж на 15%.

У результаті пошуку дублів новинних документів було створено паралельний двомовний корпус документів [*Lande, 2008*] об'ємом приблизно 30 тис. документів (рис. 23).

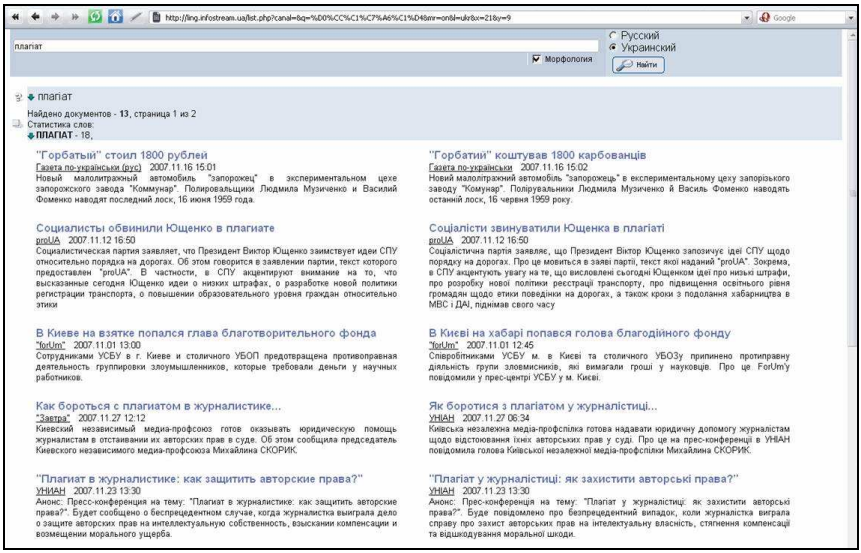



Рис. 23 – Інтерфейс системи пошуку у двомовному корпусі

З отриманого корпусу документів було вибрано 1000 випадкових документів, які було надано вивченню експертами. Аналіз показав, що у середньому 98% змісту кожного документа мають різні доповнення та зміни, наприклад, посилання на інше видавництво, або ж інший заголовок. Також аналіз показав, що з 1000 обраних документів знайшовся лише один документ, який відповідав документу, наведеному іншою мовою. Відмінність складалася лише у тому, що в документі перекладу були більш докладно описані подробиці первинної статті, а довжина первинної статті була дуже малою – близько 40 слів.

Таким чином, наведений алгоритм дозволяє проводити пошук дублікатів, представлених не тільки мовою, якою було написано первинний документ, але й іншою мовою. Тобто використовуючи механізм підключення інших мов до системи, можливий пошук дублікатів, представлених відразу декількома мовами.

Як приклад пошуку дублікатів, авторами було створено двомовний паралельний корпус, який на цей час налічує понад 2,6 млн. пар речень, до якого надається вільний доступ (http://ling.infostream.ua/).

На практиці не завжди можливо виявити дублікат документу, якщо він був створений на основі об'єднання декількох текстів, ця ситуація найчастіше виникає при пошуку плагіату. У таких випадках розглянутий алгоритм, якщо його застосовувати без модифікацій, буде малоефективним, але ситуацію можливо поліпшити шляхом ведення пошуку опорних слів не на рівні всього тексту, а на рівні декількох абзаців. Після цього цей алгоритм можна застосовувати без обмежень.

Для оцінки якості виявлення дублікатів використовувалися також і формальні методи. Так на рис. 24 зображені графіки, які показують коефіцієнти близькості і відмінності опорних слів паралельних документів.

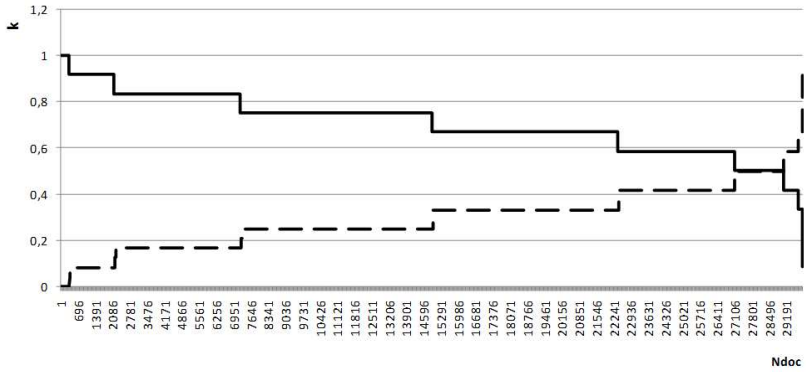



Рис. 24 – Ранжируваний список коефіцієнтів близькості (суцільна лінія) і відмінності (пунктирна лінія) документів, поданих російською мовою та їх дублікатів українською мовою

Коефіцієнт близькості обчислюється за такою формулою:

$$k_1 = \frac{N_z}{b},$$

де $N_z$ – кількість загальних ключових слів у паралельних документах, наведених українською і російською мовами; $b$ – максимальна кількість однакових опорних слів, дорівнює 12.

Коефіцієнт відмінності:

$$k_2 = \frac{N_v}{c},$$

де $N_v$ – кількість відмінних опорних слів у документах; $c$ – максимальна кількість різних опорних слів у обох документах, яка приймається рівною 24.

З рис. 24 видно, що перетин графіків відбувається при $k_1 = 0.5$ і $k_2 = 0.5$.

При цьому середнє значення загальних опорних слів при пошуку російськомовних документів, близьких до україномовних складає 8,45. Середнє значення загальних опорних слів при пошуку україномовних документів, близьких до російськомовних складає 8,97.

Приблизний коефіцієнт помилки обчислення опорних слів складає 0,52. Цей коефіцієнт обчислюється за формулою:

$$E = \frac{1}{N}\left|R_1 - R_2\right|,$$

де $N$ – загальна кількість документів; $R_1$ – загальна кількість опорних слів при пошуку дублів російськомовних документів, подібних україномовним; $R_2$ – кількість загальних опорних слів при пошуку дублів україномовних документів, подібних російськомовним.

У процесі пошуку опорних слів для документів у результати попадали такі опорні слова, для яких не було пари у паралельному документі, або ж це слово було перекладено синонімом.

Аналізуючи паралельний корпус, було визначено, що найбільш перекладеними документами з російської на українську і навпаки були документи, видані тими ж самими джерелами (табл. 5). Зокрема, можна навести агентства Укрінформ і УНІАН, які видають новини одразу декількома мовами.

Як показала статистика, не всі паралельні документи видаються одним джерелом. Трапляються і такі документи, які іншим видавництвом перекладені іншою мову, на відміну від

першоджерела. На рис. 25 показано розподіл джерел в корпусі по співвідношенню виданих ними документів. Більш жирна лінія – це видавництва, які друкували документи російською мовою. Тонка лінія – видавництва які надрукували документи українською мовою.

Табл. 5. 10 найбільш частотних джерел з паралельного корпусу

| № з/п | Українські документи | | Російські документи | |
|---|---|---|---|---|
| | **Джерело** | **Кількість статей** | **Джерело** | **Кількість статей** |
| 1 | УкрІнформ | 2919 | УкрИнформ | 2973 |
| 2 | Газета по-українськи | 2821 | Газета по-українськи | 2449 |
| 3 | УТРО-Україна | 1933 | УТРО-Украина | 1917 |
| 4 | УНІАН | 1621 | УНИАН | 1340 |
| 5 | NEWSru.ua | 1548 | Газета «Хрещатик" | 1323 |
| 6 | forUm | 1358 | NEWSru.ua | 1270 |
| 7 | Газета «Хрещатик" | 1073 | Корреспондент.net | 1087 |
| 8 | proUA | 1027 | "Украинская правда" | 1071 |
| 9 | РБК-Україна | 1016 | forUm | 1064 |
| 10 | INTV | 1014 | РБК-Украина | 1057 |

Як показала статистика, не всі паралельні документи видаються одним джерелом. Трапляються і такі документи, які іншим видавництвом перекладені іншою мову, на відміну від першоджерела. На рис. 25 показано розподіл джерел в корпусі по співвідношенню виданих ними документів. Більш жирна лінія – це видавництва, які друкували документи російською мовою. Тонка лінія – видавництва які надрукували документи українською мовою.

Як показала статистика, не всі паралельні документи видаються одним джерелом. Трапляються і такі документи, які іншим видавництвом перекладені іншою мову, на відміну від першоджерела. На рис. 25 показано розподіл джерел в корпусі

по співвідношенню виданих ними документів. Більш жирна лінія – це видавництва, які друкували документи російською мовою. Тонка лінія – видавництва які надрукували документи українською мовою.

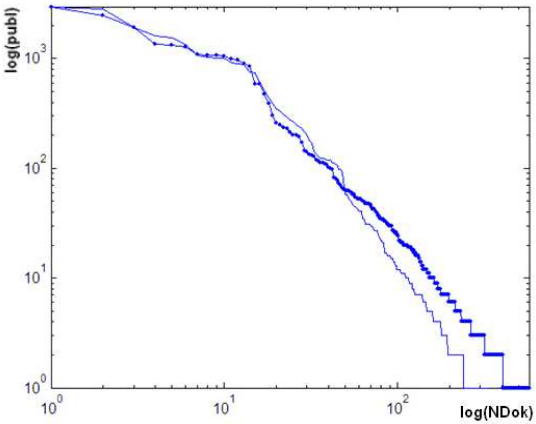



Рис. 25 – Розподіл джерел по кількості обсягів документів, що видаються, в паралельному корпусі у подвійній логарифмічній шкалі

Як видно з табл. 5 майже всі видавництва, які писали українською мовою мають своє місце в рейтингу видавництв, які писали російською.

Також було визначено, що до корпусу входять документи з 574 видавництв, які пишуть російською мовою, і з 328 видавництв, які пишуть українською мовою.

Дослідження статистики використання опірних слів в паралельних масивах текстів дозволило отримати наступні результати:

– кількість слів в українськомовному масиві склала 5595591, з них унікальних 181453 слова;

– в російськомовному масиві кількість слів склала 5641695, загалом з них 174 640 унікальних слів.

На рис. 26 показано відношення частот слів до їхнього рангу в текстових масивах досліджуваного паралельного корпусу.

Як відомо, цей показник відповідає степеневому закону розподілу (закон Ципфа-Мандельброта), при цьому для кожної з мов параметри такого розподілу різні. Однак дослідження показало, що для розглянутих паралельних текстових масивів параметри розподілу практично збігаються.

На рис. 26 показано відношення частот слів до їхнього рангу в текстових масивах досліджуваного паралельного корпусу.

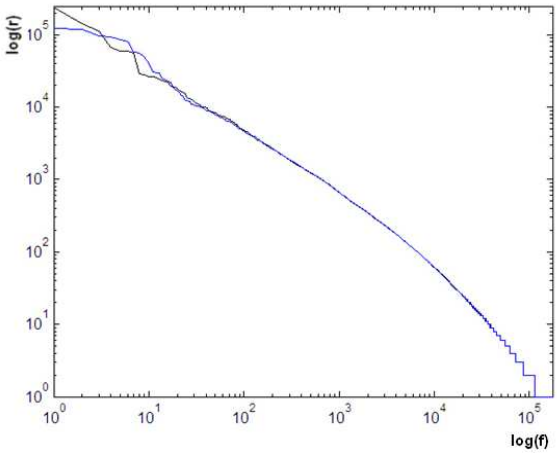



Рис. 26 – Співвідношення частот слів до їхнього рангу для російсько- та україномовних документів, в логарифмічній шкалі

Як відомо, цей показник відповідає степеневому закону розподілу (закон Ципфа-Мандельброта), при цьому для кожної з мов параметри такого розподілу різні. Однак дослідження показало, що для розглянутих паралельних текстових масивів параметри розподілу практично збігаються.

Представлені алгоритми і підходи в даний час використовуються в системі контент-моніторингу InfoStream, зокрема, на етапі індексування документа в цій системі до нього приписується декілька опорних слів, які перекладаються іншою мовою за допомогою словників перекладів. Для пошуку дублікатів береться певна кількість із знайдених опорних слів з

вихідного документу і порівнюються з усіма перекладеними опорними словами інших документів.

Використовуючи механізм підключення до системи контент-моніторингу різних мов, можна знаходити подібні документи або дублікати в багатомовних базах даних, вирішувати проблеми тематичного пошуку, а також пошуку передруків.

## 2.5. Системи статистичного машинного перекладу

Сучасні системи машинного перекладу призначені для вирішення низки завдань, що мають важливе значення для таких застосувань, як автоматичне виявлення опорних (ключових) слів у документах, створення словників, виявлення дублікатів (плагіату), представлених різними мовами, створення масивів різних мовних версій одних й тих самих документів, і, нарешті, створення автоматичних онлайн-перекладачів, налаштованих на визначені предметні галузі, зокрема, на правознавство.

Сьогодні практично всім відомі онлайн-служби, що забезпечують швидкий і безкоштовний переклад фрагментів текстових документів. Зокрема, Google (http://translate.google.com) надає можливість перекладу з 57 мов. Найбільшим конкурентом Google є мережева служба Bing Translator (http://www.microsofttranslator.com/), за якою стоїть корпорація Microsoft, що забезпечує в даний час переклад з двадцяти мов. При цьому обидва мережевих гіганта використовують таку гілку технологій машинного перекладу як статистичний машинний переклад, що базується на гігантських обсягах інформаційних ресурсів і простих та ефективних алгоритмах [*Hutchins, 2005*], [*Hutchins, 2007*].

Для систем статистичного перекладу характерне використання масивів текстів, представлених одночасно двома мовними версіями (паралельних корпусів). Чим більше об’єм паралельного корпусу, а так само, чим якісніше переклад текстів, що містяться в ньому, тим краще перекладає статистичний перекладач.

В якості теоретичної основи технології статистичного машинного перекладу використовується модель, що базується на теоремі Байєса. Дана модель надає можливості поліпшити переклад, використовуючи найбільш частотні слововживання різними мовами, враховуючи відповідні частоти при перекладі документа. Теорема Байєса у цьому випадку виражається простою формулою:

$$p(e \mid f) = \frac{p(e)(f \mid e)}{p(f)},$$

де *f* – визначений фрагмент оригіналу (слово, *n* слів, що йдуть одне за одним, речення тощо), *e* – фрагмент перекладу, $p(e \mid f)$ – умовна ймовірність того, що перекладом вихідного фрагменту *f* буде фрагмент *e*, $p(f \mid e)$ – умовна ймовірність того, що перекладу *e* відповідає вихідний фрагмент *f*.

Використовуючи формулу Байєса, формально записується правило знаходження найбільш ймовірного перекладу:

$$\arg\max_e p(e \mid f) = \arg\max_e p(e)(f \mid e).$$

У наведеному виразі була проігнорована ймовірність *p*(*f*), тому, що вона однакова для будь-якого вихідного тексту *e*. У граничному випадку ймовірність *p*(*e*) пропорційна тому, наскільки велика частота появи конкретного фрагменту тексту в масиві, який представлено мовою перекладу. Ймовірність $p(f \mid e)$ відповідає моделі перекладу. Загалом, чим більше вихідних текстів перекладається в конкретний фрагмент *e*, тим гірша якість перекладу. Ймовірність *p*(*e*) визначається досить легко – за масивом можливих фрагментів перекладу. По суті, такий підхід дозволяє розділити завдання на дві частини – спочатку застосувати модель пошуку речень мовою перекладу, в яких згадуються ті ж поняття, що й мовою оригіналу, а потім скористатися частотною оцінкою імовірності *p*(*e*), щоб вибрати найкращий варіант перекладу.

Загальноприйнята методологія створення систем статистичного машинного перекладу охоплює такі основні етапи (рис. 27):

– створення корпусу паралельних документів;
– створення корпусу паралельних речень;
– створення масивів паралельних *N*-грам;
– створення індексних файлів системи перекладу, що базується на *N*-грамах;
– безпосереднє створення модулів статистичного перекладача.

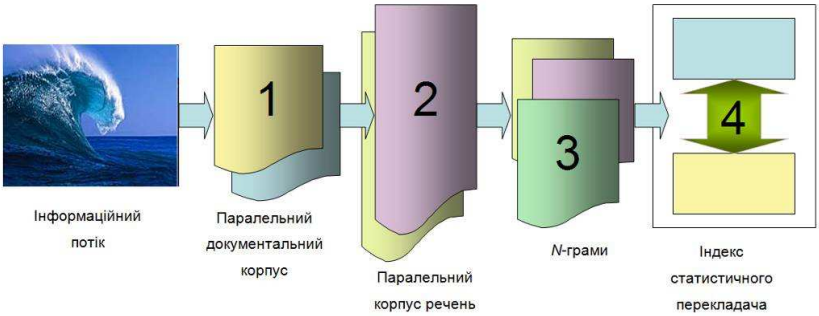



Рис. 27 – Дані, що відповідають чотирьом етапам формування статистичної системи машинного перекладу

Як джерела даних для створення статистичних перекладачів використовують паралельні текстові корпуси, що містять різні мовні версії одних і тих самих документів. Джерелами таких документів є збірки перекладеної художньої літератури відомих письменників, тексти парламентських засідань країн і організацій, наприклад, парламентські звіти Канади видаються двома мовами, офіційні документи Європейського економічного співтовариства видаються 11 мовах, тощо.

Для створення паралельних і мультипаралельних корпусів також використовуються повідомлення інформаційних агентств, сторінки веб-сайтів, що мають декілька мовних версій.

При побудові паралельних документальних корпусів для забезпечення більшої точності використовуються додаткові

критерії, наприклад, підраховується кількість речень, цифр, назв, довжини фрагментів текстів, тощо.

Вирівнювання документальних корпусів на рівні речень, тобто побудова паралельних корпусів речень, виконується на основі головного постулату систем статистичного перекладу – принципу монотонності [*Потьомкін, 2008*]. Цей принцип полягає у тому, що різні мовні версії одного й того ж документа містять речення, розміщені в одному і тому ж порядку, тобто друге речення йде після першого, третє – після другого тощо.

Наступним етапом формування бази даних статистичного перекладача є формування масиву *N*-грам. *N*-грамою називається послідовність з *N* слів одного тексту, що йдуть одне за одним. Наприклад, у фразі «*найпростішою моделлю перекладу є дослівний переклад*» містяться такі триграми:

– «найпростішою моделлю перекладу»;
– «моделлю перекладу є»;
– «перекладу є дослівний»;
– «є дослівний переклад».

*N*-грами – це традиційний об’єкт дослідження комп’ютерних лінгвістів. Перше практичне застосування *N*-грами отримали в програмах визначення мов, якими написані тексти, перевірки правопису, а потім вже в технологіях статистичного машинного перекладу. *N*-грами, що відповідають одномовним текстовим корпусам, сьогодні є комерційними продуктами, які створюються і пропонуються на ринку. Наприклад, масив англійських 5-грам (пентаграм) пропонується компанією LDC (рис. 28). Цей масив містить 5 триліонів записів і розміщується у архівованому вигляді на 6 DVD-дисках. Технології *N*-грам, крім того, широко використовуються і пропагуються компаніями Google і Bing (рис. 29–30).

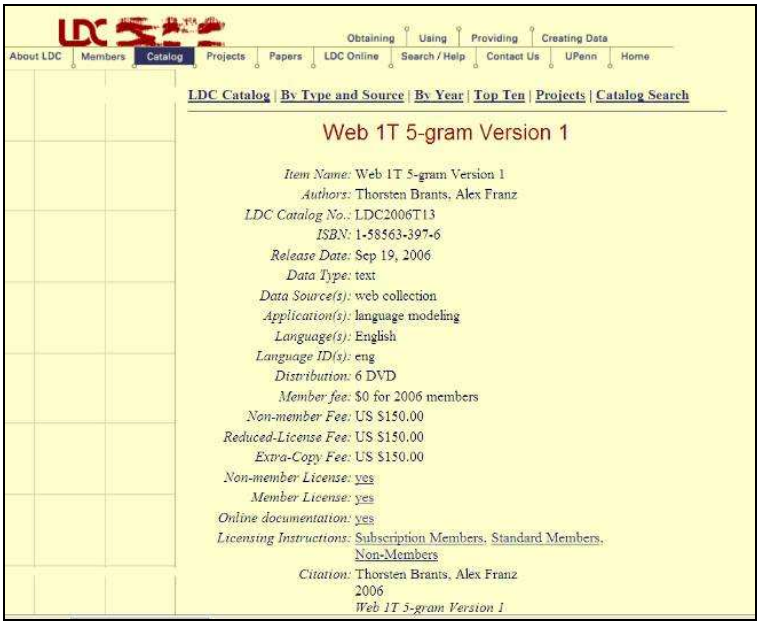


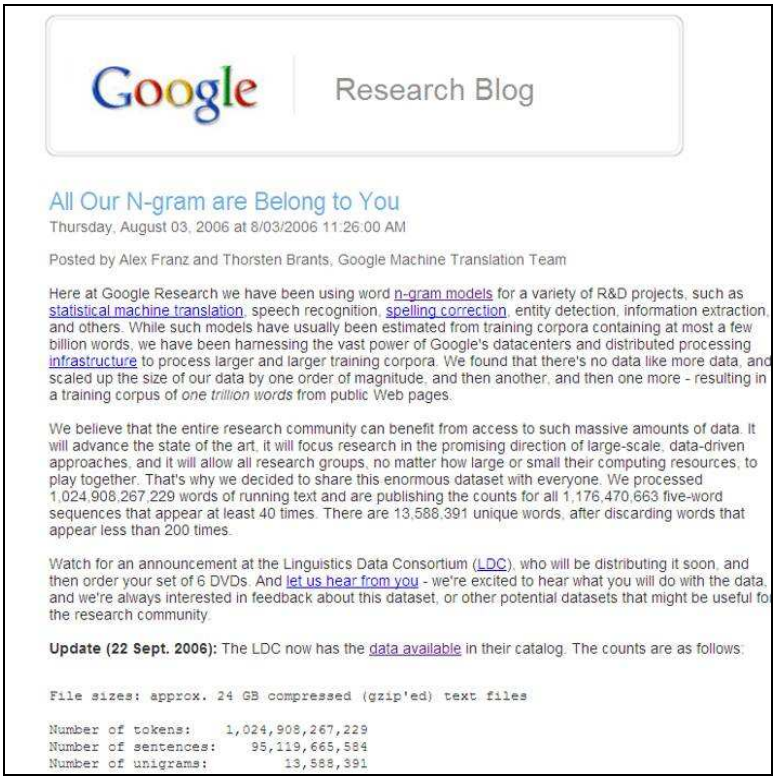

Рис. 28 – Фрагмент опису корпусу пентаграм компанії LDC

При побудові баз даних сучасних систем перекладу створюються масиви *N*-грам (найчастіше – пентаграм).

Для цих масивів у рамках технологій статистичного машинного перекладу використовують паралельні двомовні корпусу речень. Для кожної пари речень будують *N*-грами однією мовою, яким відповідають (за місцем у відповідному реченні) *N*-грами іншої мови. Наприклад, якщо представити речення, в якому функцію слів виконують латинські літери: «a b c d e f g h», а перекладом (відповідним реченням з паралельного корпусу) буде: «а б в г д е є ж», то для них будуть побудовані пари пентаграм:

$$\mathbf{П_{1eng}(abcde)\sim П_{1rus}(абвгд),\ П_{2eng}(bcdef)\sim П_{2rus}(бвгде),}$$

$$\mathbf{П_{3eng}(cdefg)\sim П_{3rus}(вгдеє).}$$



Google Research Blog

All Our N-gram are Belong to You

Thursday, August 03, 2006 at 8/03/2006 11:26:00 AM

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects, such as statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, and others. While such models have usually been estimated from training corpora containing at most a few billion words, we have been harnessing the vast power of Google's datacenters and distributed processing infrastructure to process larger and larger training corpora. We found that there's no data like more data, and scaled up the size of our data by one order of magnitude, and then another, and then one more - resulting in a training corpus of *one trillion words* from public Web pages.

We believe that the entire research community can benefit from access to such massive amounts of data. It will advance the state of the art, it will focus research in the promising direction of large-scale, data-driven approaches, and it will allow all research groups, no matter how large or small their computing resources, to play together. That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

Watch for an announcement at the Linguistics Data Consortium (LDC), who will be distributing it soon, and then order your set of 6 DVDs. And let us hear from you - we're excited to hear what you will do with the data, and we're always interested in feedback about this dataset, or other potential datasets that might be useful fo the research community.

**Update (22 Sept. 2006):** The LDC now has the data available in their catalog. The counts are as follows:

```
File sizes: approx. 24 GB compressed (gzip'ed) text files

Number of tokens:    1,024,908,267,229
Number of sentences:    95,119,665,584
Number of unigrams:         13,588,391
```

Рис. 29 – Корпус *N*-грам в технологіях Google

Далі проводився підрахунок кількості *N*-грам, які зустрічаються в паралельному корпусі речень. У разі якщо *N*-грамі однієї мови відповідає кілька *N*-грам іншої мови, то вибирається найбільш частотна *N*-грама.

Типовий алгоритм роботи статистичного перекладача наступний. На вхід модуля перекладу подається документ мовою оригіналу, який відразу ж розбивається на речення. Для кожного речення будуються *N*-грами та виконується пошук їхніх перекладів іншою мовою. У разі якщо переклад якоїсь *N*-грами не вдається виявити, шукається відповідна (*N*–1)-грама і виконується пошук її перекладу тощо. Якщо програма не виявляла перекладу для біграм, виконувався пошук слова у словнику перекладів окремих слів. Якщо виникає ситуація, що

деяке слово не перекладається і таким чином, то воно залишається на мові оригіналу або (в деяких системах) проводився «псевдо переклад» (транслітерація, тощо). Після перекладу всіх пропозицій документ форматується за шаблоном, і виводиться користувачеві як результат.

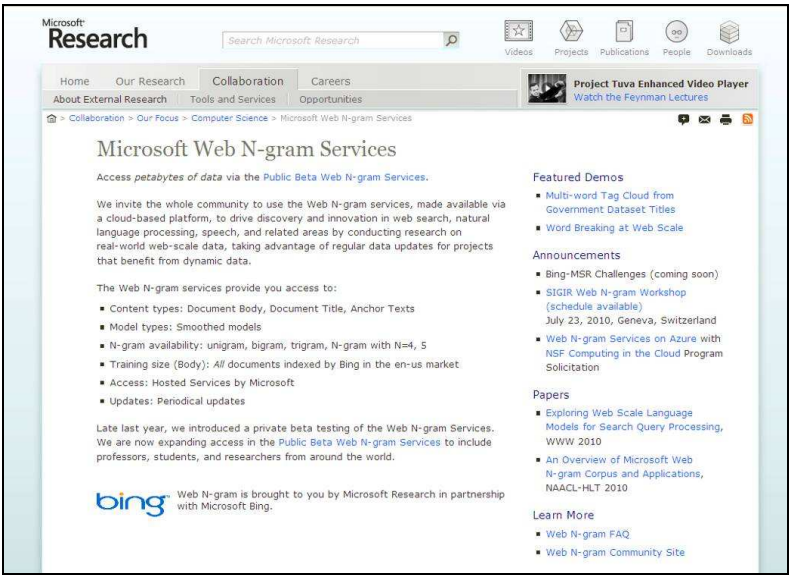



Рис. 30 – *N*-грами в технології Bing Translate

Далі проводився підрахунок кількості *N*-грам, які зустрічаються в паралельному корпусі речень. У разі якщо *N*-грамі однієї мови відповідає кілька *N*-грам іншої мови, то вибирається найбільш частотна *N*-грама.

Типовий алгоритм роботи статистичного перекладача наступний. На вхід модуля перекладу подається документ мовою оригіналу, який відразу ж розбивається на речення. Для кожного речення будуються *N*-грами та виконується пошук їхніх перекладів іншою мовою. У разі якщо переклад якоїсь *N*-грами не вдається виявити, шукається відповідна (*N*–1)-грама і виконується пошук її перекладу тощо. Якщо програма не виявляла перекладу для біграм, виконувався пошук слова у словнику перекладів окремих слів. Якщо виникає ситуація, що

деяке слово не перекладається і таким чином, то воно залишається на мові оригіналу або (в деяких системах) проводився «псевдо переклад» (транслітерація, тощо). Після перекладу всіх пропозицій документ форматується за шаблоном, і виводиться користувачеві як результат.

На сьогоднішній день реалізовано і доступно користувачам декілька статистичних перекладачів, в основі яких закладена технологія перекладу за допомогою *N*-грам. Це вже названі раніше Google Translate, Bing Translator і ще не функціонуючий, але вже розрекламований статистичний перекладач компанії IBM, частина функціоналу якого запроваджено на сайті компанії в рамках технології n.Fluent з метою спрощення спілкування розробників різними мовами.

У табл. 6 надано порівняльний приклад перекладу на російську мову оригінального тексту різними онлайн-перекладачами (як статистичними, так і тих, що базуються на правилах).

Табл. 6. Приклад перекладу сучасними програмами-перекладачами

| **Оригинал** | **Статистичні перекладачі** | **Перекладачі, що базуються на правилах** |
|---|---|---|
| | **Bing Translator** | **BabelFish** |
| Businesses that hire students can save 8 percent by using a foreign worker over an American because they don't have to pay Medicare, Social Security and unemployment taxes. | Предприятия, которые нанимают студентов можно сохранить 8 процентов с помощью иностранный работник над американцем, поскольку они не придется платить медицинского обслуживания, социального обеспечения и безработицы налоги. | Дела которые нанимают студентов могут сохранить 8 процентов путем использование иностранного рабочего над американцом потому что они don' t должен оплатить тягла Medicare, социального обеспечения и незанятости. |

| Оригинал | Статистичні перекладачі | Перекладачі, що базуються на правилах |
|---|---|---|
| | **Google Translate** | **PROMT** |
| | Бизнесы, которые нанимают студентов может спасти 8 процентов с помощью иностранного работника по американской, потому что они не должны платить медицинской помощи, социального обеспечения и безработицы налогов. | Фирмы, которые нанимают студентов, могут спасти 8 процентов при использовании иностранного рабочего по американцу, потому что они не должны заплатить Бесплатную медицинскую помощь, социальное обеспечение и налоги безработицы. |

Як видно з прикладу, переклад тексту системами статистичного машинного перекладу (Statistical Machine Translation, SMT) близький за якістю з системами перекладу, що базуються на правилах (Rule-Based MT). Як приклади систем Rule-Based MT було взято PROMT (http://translate.ru/) і BabelFish (http://babelfish.yahoo.com/). Разом з тим, якість перекладу всіх систем, у цьому випадку, (перш за все, це сиосується довгих речень) залишає бажати кращого.

У статистичному російсько-українському перекладачі InfoStream є ряд відмінностей від традиційної моделі статистичного машинного перекладу, що дозволяє, на думку її авторів, здійснювати більш якісні та швидкі переклади [*Жигало, 2010*].

Для побудови статистичного перекладача потрібен великий паралельний корпус. Як джерело для побудови паралельного корпусу документів були взяті новинні повідомлення із системи контент-моніторингу InfoStream. При побудові первинного паралельного документального корпусу в підсистемі машинного перекладу системи InfoStream, на відміну від загальновідомих систем, використовувалися лінгвостатистічні алгоритми, які застосовуються до результатів контент-моніторингу мережевих ЗМІ – масивів новин. В основу

алгоритму визначення паралельних документів було узято метод порівняння опорних слів, описаний вище.

При побудові корпусу паралельних документів з первинного набору паралельних документів проводилася фільтрація і відсіювання зайвих дублів. На підставі наведеного вище алгоритму було створено паралельний україно-російський корпус документів.

Наступний крок до створення паралельного перекладача – вирівнювання паралельного російсько-українського корпусу документів на речення. Поділ паралельних документів на речення здійснювався з урахуванням таких критеріїв як:

- ознаки кінця речення були взяті символи (.,!,?,;,:);
- було введено додаткове обмеження, якщо в тексті зустрічалося скорочення або ініціали з точкою то вони не вважається кінцем речення.

Далі проводився підрахунок кількості речень в паралельних документах. Якщо дані документи за кількістю речень, були однаковими, то ці речення використовувалися в подальшій обробці. Кожне речення було розділено на слова. Також накладалися додаткові обмеження на визначення слова в кожній із мов. Наприклад, ті слова перекладу українською мовою, перед якими згадувалися слова: «який», «яка», «що», «котрий», визначалися як складні і розглядалися як одне слово. Далі проводився підрахунок кількості слів у паралельних реченнях. В результаті, до паралельного корпусу речень, вибиралися лише ті речення, які за кількістю слів не відрізнялися більш ніж на одне слово.

Наступний етап побудови статистичного перекладача полягав у створенні статистичних словників триграм (рис. 31), біграм і статистичного словника перекладів слів з паралельного корпусу речень, який було побудовано раніше. Для кожної пари речень були побудовані триграми однієї мовои, які відповідали триграмам іншої мови.

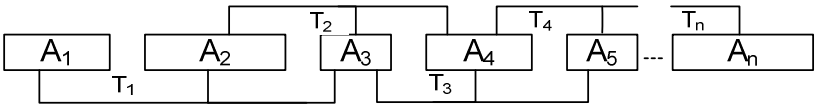



Рис. 31 – Поділ речення на триграми. $A_1$–$A_n$ – слова. $T_1$–$T_n$ – триграми

Для визначення слова, що входить до складу *N*-грам, використовувалися ті ж правила поділу речення на слова, які застосовувалися для вирівнювання паралельного корпусу. В основу словників увійшли пари – *N*-грама однієї мови і переклад її іншою мовою. Проводився підрахунок кількості *N*-грам, які зустрічалися в паралельному корпусі речень. У разі, якщо *N*-грамі відповідало декілька *N*-грам іншою мовою, то вибиралася найбільш частотна *N*-грама.

Технологічний ланцюжок статистичного перекладача InfoStream починався з того, що поданий документ розділявся на речення. Для кожного речення будувалися триграми і виконувався пошук перекладу 3-грами паралельною мовою. У тому випадку, коли не вдавалося виявити переклад, використовувалася біграма, і далі виконувався пошук перекладу цієї біграми. Якщо не виявлялось перекладу для триграми і біграми, виконувався пошук слова у словнику перекладів слів. Якщо якесь слово не було відомо перекладачеві проводився псевдопереклад слова за визначеними морфологічними правилами. Після перекладу всіх речень документ форматувався за шаблоном як вхідний документ і виводився користувачеві як результат.

Основні проблеми, пов’язані з побудовою словників – це великі обсяги масивів даних і забезпечення швидкого доступу до них. На даний момент загальний обсяг словників займає понад 2 Гб, прямий пошук у такому словнику не є ефективним – для пошуку в словниках триграм і біграм було застосовано метод бінарного пошуку.

Для здійснення бінарного пошуку було побудовано індексний файл для кожного словника. Як ключ був узятий хеш

від кожного слова, а як функція хешування – хеш-функція CRC-32, яка повертає 4-х байтове додатне число. При цьому підході очевидні колізії, що являють собою однакові хеш-значення різних за написанням слів. Для подолання колізій, використовувався спеціальний словник, який також завантажувався в оперативну пам'ять.

На відміну від 5-грам, що використовуються в Google і Bing Translator, для побудови російсько-українського перекладача виявилося достатнім використовувати триграми та деякі синтаксичні і морфологічні правила, що значно підвищило продуктивність системи. У табл. 7 надано порівняльний переклад одного й того ж фрагмента тексту взятого із системи контент-моніторингу InfoStream.

Табл. 7. Порівняльний переклад програмами-перекладачами

| Оригінал | Статистичні перекладачі | Перекладачі, що базуються на правилах |
|---|---|---|
| | **InfoStream перекладач** | **RutaPlay** |
| Жертвы акул получат по $50 тыс<br>Что будет с остальными туристами?<br>В Москву прибыла делегация из Египта с целью сообщить о текущей ситуации в связи с нападением акул на туристов в Шарм-Эль-Шейхе. Представители египетских туристических властей, в частности, советник министра по туризму страны Хелми Бадир, 8 декабря сообщили журналистам, что все россиянине, пострадавшие от морских хищников, получат от правительства Египта по 50 тыс. долларов в качестве компенсации. | Жертви акул отримають по $50 тис<br>Що буде з іншими туристами?<br>До Москви прибула делегація з Єгипту з метою повідомити щодо поточної ситуації в зв'язку з нападом акул на туристів у Шарм-Ель-Шейху. Представники єгипетських туристичних органів влади, зокрема, радник міністра з туризму країни Хелми Бадір, 8 грудня повідомили журналістам, що всі росіянина, постраждалі від морських хижаків, отримають від уряду Єгипту по 50 тис. доларів на якості компенсації. | Жертви акул одержать по $50 тис<br>Що буде з іншими туристами?<br>У Москву прибула делегація з Єгипту з метою повідомити про поточну ситуацію у зв'язку з нападом акул на туристів у Шарм-Ель-Шейху. Представники єгипетських туристичних влади, зокрема , радник міністра по туризму країни Хелми Бадир, 8 грудня повідомили журналістів, що все росіянині, що постраждали від морських хижаків, одержать від уряду Єгипту по 50 тис. доларів як компенсації. |
| | **Google Translate** | **Pragma** |

| | | |
|---|---|---|
| | Жертви акул отримають по $ 50 тис<br>Що буде з рештою туристами?<br>До Москви прибула делегація з Єгипту з метою повідомити про поточну ситуацію в зв'язку з нападом акул на туристів в Шарм-ель-Шейху.<br>Представники єгипетських туристичної влади, зокрема, радник міністра з туризму країни Хелм Бадір, 8 грудня повідомили журналістам, що всі росіянина, постраждалі від морських хижаків, отримають від уряду Єгипту по 50 тис. доларів в якості компенсації. | Жертви акул отримають по $50 тыс<br>Що буде з іншими туристами?<br>У Москву прибула делегація з Єгипту з метою повідомити про поточну ситуацію у зв'язку з нападом акул на туристів в Шарм-эль-шейху.<br>Представники єгипетської туристичної влади, зокрема, радник міністра по туризму країни Хелми Бадир, 8 грудня повідомили журналістів, що усе росіянинові, постраждалі від морських хижаків, отримають від уряду Єгипту по 50 тис. доларів в якості компенсації. |

З наведеної таблиці можна зробити висновок що на сьогоднішній день статистичні перекладачі трохи програють у якості перекладу перекладачам, що базуються на правилах. У цьому випадку якість перекладу перекладача InfoStream не поступається якості перекладача Google Translate.

Запропонована методологія дозволила створити систему з елементами самонавчання (після робіт, зроблених на етапі ініціалізації) статистичного перекладу, орієнтовану на масовий переклад текстової інформації з інформаційних потоків, представлених російською та українською мовами (рис. 32).

В межах цієї системи документи, з інформаційного потоку надходять до модулю перекладу, здійснюється виявлення дублікатів, представлених різними мовами, доповнюється паралельний документальний корпус, потім відбувається вирівнювання на рівні речень, побудова масивів 3-, 2-грам і слів, формується індекс оновленого перекладача, який використовується модулем перекладу, на вхід якого надходять як запити окремих користувачів (у режимі онлайн), так і весь сканований системою InfoStream інформаційний потік.

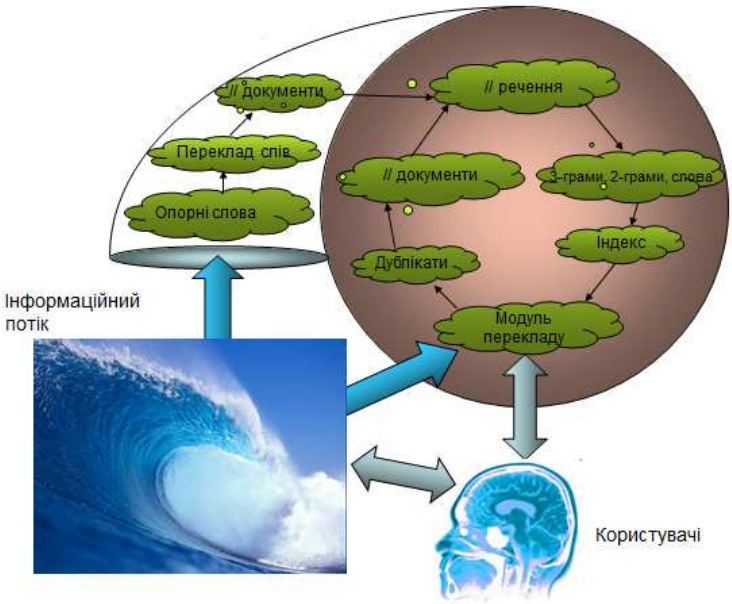



Рис. 32 – Статистичний перекладач – система із зворотним зв’язком

Розроблене програмне забезпечення в даний час знаходиться в стадії комплексного налагодження і тестування та доступно за адресою http://docsbundle.info/t.php (рис. 33). Подібний підхід, очевидно, буде ефективним, практично для будь-яких близьких за статистичними параметрами мов.

Сьогодні можна констатувати, що проблема створення прийнятних за якістю (забезпечують розуміння текстів користувачами) онлайн-перекладачів вирішена. Крім того, стає очевидним, що саме гілка систем статистичного машинного перекладу є найбільш ефективною і якісною.

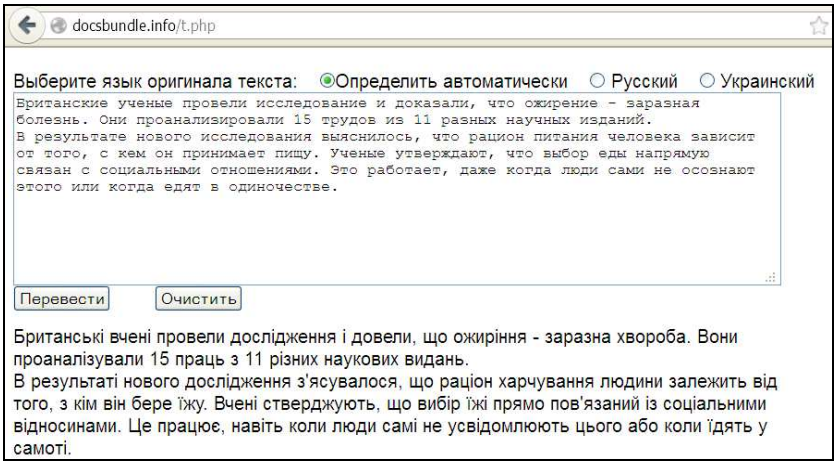



Рис. 33 – Інтерфейс бета-версії онлайн-перекладача системи InfoStream

Разом з тим, всім системам статистичного перекладу притаманний ряд проблем, серед яких можна назвати:

1. Брак необхідного обсягу вихідних перекладів (паралельних текстових корпусів).
2. Практично всі рафіновані системи статистичного машинного перекладу неякісно обробляють довгі фрази, що перевищують 10-12 слів.
3. Існують проблеми швидкості, які не дозволяють створювати повноцінні потокові перекладачі.
4. Можна визнати, що для забезпечення якості перекладу необхідно застосовувати хоча б елементарні правила. До систем машинного перекладу для поліпшення якості вводяться деякі загальні правила, тим самим перетворюючи чисто статистичні системи в гібридні.

Безумовно, додавання деяких правил, тобто створення гібридних систем, трохи поліпшує якість перекладів, особливо при недостатньому обсязі вхідних даних, що використовуються при побудові індексу машинного перекладача.

Перед системами машинного перекладу постає ряд актуальних задач, що диктуються потребами користувачів:

1. Підвищення точності та адекватності перекладу. Як показано численними прикладами, вирішення задачі досягається шляхом збільшення обсягів відповідних баз даних паралельних текстів, а також урахування деякого набору правил для кожної з пар мов.

2. Переклад повних документів, а не лише фрагментів. Теоретично сучасні системи статистичного машинного перекладу спроможні забезпечити цей сервіс, проте для надання його в масовому масштабі необхідно досягти високої продуктивності програм-перекладачів, що на цей час не завжди можливо. Друга причина – комерційна – повні версії перекладачів надаються зареєстрованим користувачам за плату.

3. Переклад документальних потоків, а не окремих документів. Подібний сервіс може бути затребуваний, наприклад, при обробці потоків судових рішень, нормативно-правових документів, міліцейських протоколів тощо, при створенні мовних версій сайтів.

4. Знаходження інформаційних дублікатів, представлених різними мовами. Вирішення цього завдання забезпечить, наприклад, можливість аналізу межі проведення інформаційних операцій, виявлення передруків, плагіату, рівня гармонізації законодавства.

5. Створення інформаційно-пошукових систем, з інтерфейсом на мові користувача, що надають результати тією ж самою мовою, але забезпечують пошук в іншомовних сегментах веб-простору. Наприклад, створення україномовної пошукової системи за іншомовним правовим веб-ресурсом (сьогодні, наприклад, через мовні проблеми величезний сегмент веб-простору є по суті

«прихованим веб» (Deep Веб) для нашого користувача).

6. Надання можливості автоматичної підготовки аналітичних документів на основі інформації, представленої різними мовами. Це завдання стикується із завданням автоматичного реферування, узагальнення документів, проте належить до обробки вихідної інформації, представленої різними мовами.

7. Генерація систем, що забезпечують правильне звукове відтворення перекладеної інформації. Такі системи можуть бути корисними людям із обмеженими можливостями, а також потребують інформації в екстремальних умовах (наприклад, водіям автотранспорту, подорожуючим за кордоном, співробітникам правоохоронних органів тощо).

# 3. Технологія MediaWiki

## 3.1. Електронні словники і енциклопедії

Одним з важливих напрямків розбудови інформаційного суспільства є створення умов для гарантованого надання його членам доступу до інформації, зокрема правової. Ця задача не може бути ефективно вирішена без тієї чи іншої системи розподілу інформації по широкій мережі. Відповідно, розподілені системи, як правило, відзначаються високою стабільністю по відношенню до зовнішніх впливів [*Горбулін, 2009*].

Сьогодні використання традиційного веб-простору для отримання необхідної інформації та знань можна вважати проблематичним. Дійсно, веб-простір переповнений неструктурованими ресурсами, пошукові системи не можуть або не устигають охопити усе, що від них вимагають, «прихований» веб, що не охоплюється пошуковими системами живе своїм життям, а перспективна ідея Семантичного вебу досі чекає на своє втілення.

Як же шукають знання в Інтернеті сьогодні? Користувач, звертається до найбільш популярних пошукових серверів, таких як Google, Yahoo, Яндекс, після чого довго і наполегливо перегортає десятки сторінок з формально релевантними, але не завжди дійсно потрібними йому результатами. Тому завдання доступності необхідної користувачеві інформації в режимі он-лайн і її оперативної зміни сьогодні актуальна як ніколи раніше.

В той же час, в Інтернеті існують, нехай як рідкісні виключення, проекти, які реально утілюють в собі мрії творців веб-технологій. Один з таких проектів – це «народна енциклопедія» – Вікіпедія (www.wikipedia.org). Сьогодні – це дійсно інтерактивний продукт спільної творчості. Вікіпедія досить точно відповідає тій концепції WWW, яка замислювалася, зокрема, Тімом Бернерсом-Лі (Tim Berners-Lee). Вікіпедія це найбільша із безкоштовних он-лайн бібліотек в Інтернеті, найдинамічніша, самооновлювана, найдоступніша для

оновлень, проте, не найнадійніша, що пояснюється її ідеологією [*Ландэ, 2005а*], [*Ландэ, 2005б*].

Слово «енциклопедія» увійшло до ужитку в XVIII столітті завдяки діяльності французьких просвітників Дідро і Д'Аламбера, що написали 28 томів праці «Енциклопедія, або тлумачний словник наук, мистецтв і ремесел». «Енциклопедія» стала першою колективною працею багатьох європейських учених, що створювали статті шляхом активного листування з редакцією.

У мережі Інтернет постійно відбуваються спроби збору і систематизації рафінованих знань, накопичених людством упродовж віків. Звичайно ж, нині існують сотні і тисячі енциклопедій, у тому числі і електронних. Велика частина авторитетних енциклопедій є платними, при цьому паперові видання в десятки разів дорожче електронних. У вільному ж доступі знаходяться як правило застарілі або обмежені версії, наприклад, «Британніка» 1911 року (http:// 1911encyclopedia.org) або Microsoft Encarta (4,5 тисячі статей з 60 тисяч). У Росії Яндекс дає доступ до Великої радянської енциклопедії 1978 року і Словника Брокгауза і Ефрона 1907 року. Мегаенциклопедія Кирила і Мефодія (mega.km.ru) практично є електронною версією Сучасного Енциклопедичного словника 1997 року. Досить актуальна енциклопедія «Кругосвет» (krugosvet.ru), що містить лише близько 10 тисяч статей.

Електронні мережеві енциклопедії мають одну безперечну перевагу перед традиційними паперовими, в яких усі допущені помилки залишаються до наступного «виправленого і доповненого» видання. Один з недоліків традиційних енциклопедій полягає також в низькому рівні оперативності – вони створюються дуже повільно, але застарівають досить швидко. Наприклад, для забезпечення актуальності «Британніка», у ній кожні два роки переписується до 35 % статей.

В той же час, намітилися в явному виді тенденції до створення відкритих «народних» енциклопедій. Сьогодні певної критичної позначки досягли такі онлайнові енциклопедії, до яких доброзичливо ставляться навіть у науковому світі. До таких проектів належить, передусім, Вікіпедія, але не лише. Як приклад

можна назвати енциклопедію Everything2 (http://everything2.com), ідеологія якої багато в чому успадкувала риси живих журналів і онлайнових форумів. У цій енциклопедії користувач може змінювати тільки власний текст, при цьому він не може редагувати тексти, написані іншими людьми. Система повноважень в цій енциклопедії базується на досвіді (eXPerience, коефіцієнт XP), який автори енциклопедії набувають в процесі роботи. Досвідчені користувачі можуть оцінювати статті інших учасників.

Найбільшим словниковим ресурсом в Інтернеті на цей час є Вікісловник (англ. Wiktionary) – багатофункціональний, багатомовний словник і тезаурус, що вільно поповнюється і заснований на програмному забезпеченні MediaWiki. Вікісловник – один з проектів фонду «Вікімедія», що з'явився 12 грудня 2002 року.

У Вікісловнику знаходяться граматичні описи, тлумачення і переклади слів, а також опціональнро може бути представлена інформація про етимологію, фонетичні властивості і семантичні зв'язки слів. Таким чином, Вікісловник є агрегованим електронним словником, одночасно граматичним, тлумачним, етимологічним і багатомовним, а також тезаурусом.

Концепція Вікісловника допускає повний, усебічний опис усіх лексичних одиниць усіх природних (і основних штучних) мов, що мають писемність, проте на практиці повнота може варіюватися в різних мовних розділах проекту. У кожному конкретному мовному розділі усі статті пишуться виключно цією мовою, при цьому передбачаються переклад слів і інших одиниць цієї мови максимально можливим числом інших мов. При описі морфології передбачається максимально повне урахування словозмін.

Вікісловник як тезаурус містить наступні семантичні відношення: синоніми, антоніми, гіпероніми, гіпоніми, согіпоніми, холоніми, мероніми, пароніми.

Вікісловник надає інформацію, відсутню у Вікіпедії: словосполучення, приказки, абревіатури, акроніми, опис помилок правопису, спірні випадки вживання тощо, доповнюючи основний проект Вікімедія.

Україномовний розділ проекту Вікісловника – Український Вікісловник (http://uk.wiktionary.org/). В ньому зібрано й повсякчас поповнюються тлумачення й переклади українських слів, а також переклади слів і висловів з інших мов (рис. 34).

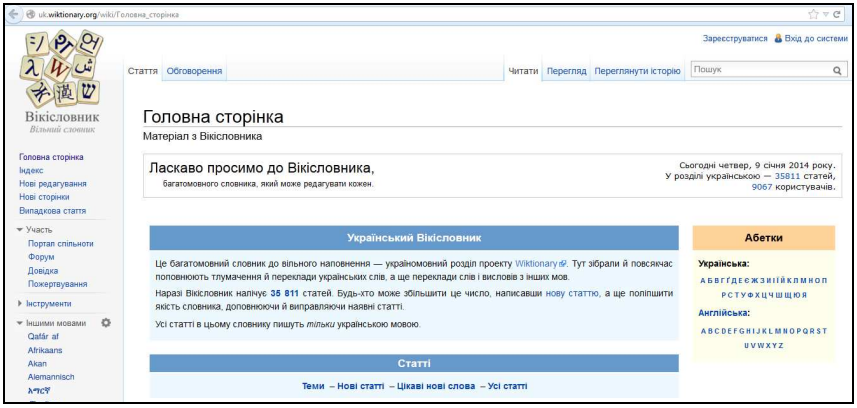



Рис. 34 – Фрагмент титульної сторінки веб-ресурсу Українського Вікісловника

Приклад словникової статті з Українського Вікісловника наведено на рис. 35.

На кінець 2013 року Український Вікісловник налічував близько 36 тис. статей, посідаючи 40-е місце серед 170 мовних версій Вікісловника. Найбільші Вікісловники – англійський (2,8 млн. статей) та французький (2,1 млн. статей).

Вікісловники активно використовуються при вирішенні різних завдань, пов’язаних з машинною обробкою тексту і мови:

- машинний переклад;
- створення електронних словників, що інтегрують відкриті лінгвістичні ресурси, як приклади: Англійський Вікісловник, WordNet і VerbNet [*McFate, 2011*];

– розпізнавання і синтез мови (Вікісловник виступає як джерело даних для автоматичної побудови словника вимов) [*Schlippe, 2011*];

– побудова відображення онтологій і баз знань;

– розмітка тексту за частинами мови [*Li, 2012*];

– аналіз тональності тексту [*Chesley, 2006*].

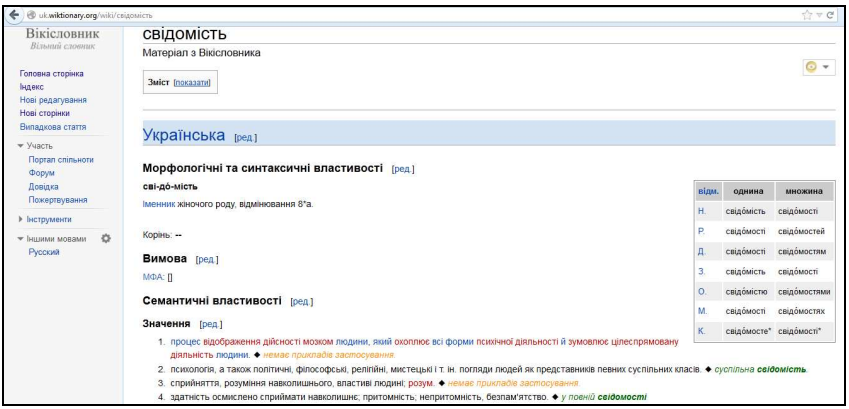



Рис. 35 – Словникова стаття з Українського Вікісловника

На базі технології Вікі, на якій зупинемось нижче, на цей час побудовано багато словників і довідників нормативно-правової спрямованості.

Як приклад можна назвати російську систему «Клерк» (wiki.klerk.ru, рис. 36). На сайті системи «Клерк» розміщується довідкова інформація для бухгалтерів і підприємців. Будь-який охочий може реєструватися і розмістити свій матеріал або доповнити наявні матеріали. Зараз в енциклопедії понад 15 тис. статей.

Щоб знайти матеріал, можна скористатися режимом пошуку – ввести запит. Інший спосіб дістатися до матеріалу – скористатися рубрикатором – списком, що охоплює декілька сотень категорій.

Ще один приклад – система «ВикиПроцесо» (www.civilprocess.ru/wiki, рис. 37) – це багатофункціональна платформа, створена для усіх, хто вивчає цивільний процес або практикує у російських судах з цивільних справ – це і учбовий курс, і енциклопедія, і задачник, і форум.

Технічно проект виконаний на MediaWiki, тобто за технологією Вікі. Проект знаходиться в постійному розвитку.

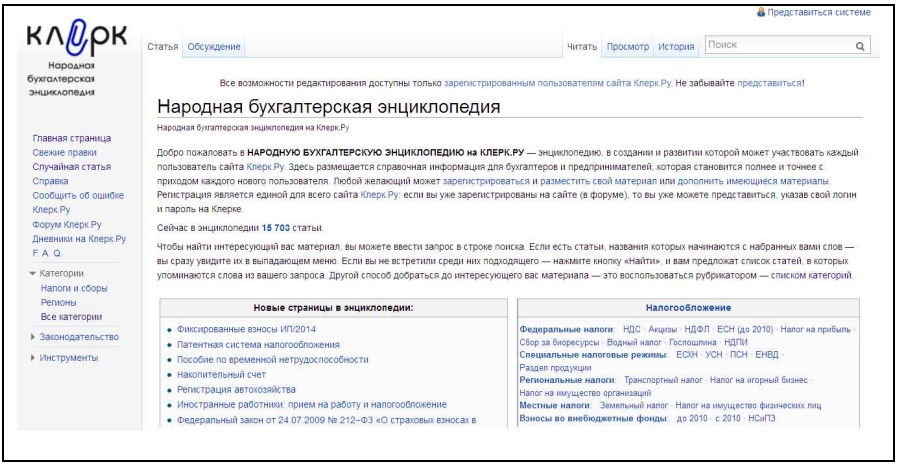



Рис. 36 – Фрагмент головної сторінки веб-сайту «Клерк»

«ВикиПроцесо» складається з п’яти блоків, покликаних з одного боку, відповідати потребам різних груп користувачів, а з іншого побудувати єдину логічну систему знань щодо цивільного процесу:

- гамми – базові твердження цивільного процесу;
- бакалавр – статті з основ цивільного процесу;
- основні – основний масив статей;
- дискусія – статті наукові дискусії, що описують;
- проблема – статті з практичних проблем і шляхів їх рішення.

Безумовно, цікавим для фахівців і початківців в галузі права є англомовний веб-сайт Wiki Law School («Правова школа», США, www.wikilawschool.net), також побудований на базі технології MediaWiki (рис. 38). Головна мета цього веб-ресурсу – змістовне заповнення кожного розділу правової науки. До категорій (Cases) на веб-сайті Wiki Law School належать: антитрастове законодавство, цивільні процедури, конституційне право, кримінальне право, авторське право тощо.

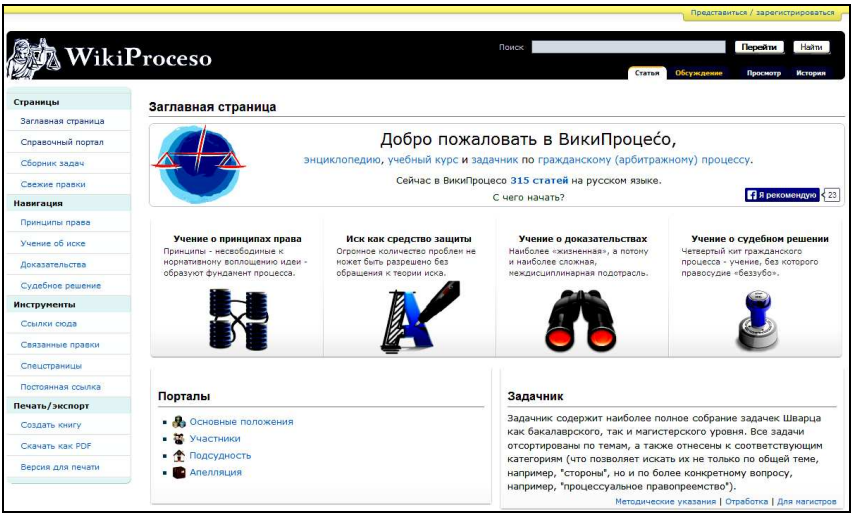



Рис. 37 – Головна сторінка веб-сайту «ВикиПроцесо»

## 3.2. Технологічні рішення

Технологія Вікі, на якій, зокрема, базується Вікіпедія, народилася вже 10 років тому як підхід до колективного ведення проектів з розробки програмного забезпечення, підтримки в належному порядку технічних завдань і специфікацій. Основна ідея вікі-технології полягає в забезпеченні можливості колективної роботи з документами – будь-який документ з електронної бібліотеки підлягає редакції будь-яким користувачем – у відкритих веб-системах будь-яким користувачем з Інтернету, в

корпоративних застосуваннях – будь-яким користувачем корпоративної мережі.

Багато в чому Вікі нагадує Open Source (відкритий код), навіть ліцензійний режим у них схожий, проте Вікі служить для створення не програмного забезпечення, а контенту – в першу чергу, текстів. В той же час, за допомогою вікі-систем, як з'ясувалося, можна успішно розробляти програмне забезпечення колективного авторства.

Вікі-системи – це веб-сайти, що працюють за принципом Вікі, тобто які можна не лише читати, але і змінювати в режимі он-лайн. Такі сайти мають вільну, вручну створювану структуру: окремі сторінки і статті зв'язуються гіперпосиланнями.

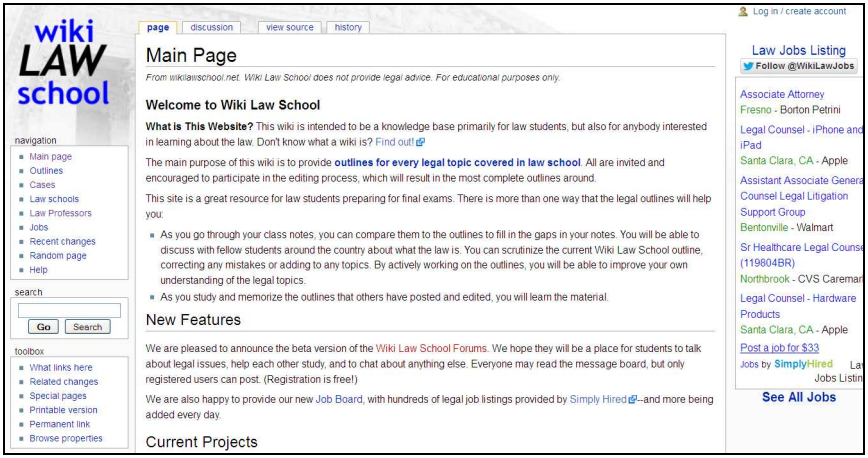



Рис. 38 – Фрагмент веб-сайту «Wiki Law School»

Існує досить багато програм різних розробників, що реалізовують ідеологію Вікі. Це досить не пов'язані програми, що надають навіть різний синтаксис розмітки документів, починаючи від HTML (як найскладніший варіант), і закінчуючи обумовленими правилами, які іноді важко навіть назвати мовою, проте які зручні для вузького завдання – створення документу для розміщення у вікі-середовищі, оформлювальний дизайн при цьому відходить на другий план. Найчастіше синтаксис Вікі задає легкий, практично не вимагаючий особливої розмітки варіант редагування, в якому

учасники можуть навіть не використовувати HTML для створення власних веб-сторінок.

Сьогодні більшість вікі-систем розміщуються на публічних серверах і є об'єктом інформаційного вкладу будь-якого відвідувача.

Вікіпедія застосовує спеціальний синтаксис, що відрізняється особливою простотою і прозорістю.

Технологію Вікі було створено задовго до ідеї Вікіпедії ще у 1995 році Уордом Каннінгемом (Ward Cunningham), і призначалася для колективної розробки документації (текстової і гіпертекстової). Ще Вікі розшифровують як «технологію швидкої правки», проте частіше говорять про неї як про «технологію живих документів».

Перше вікі-середовище було винайдено Каннінгемом для веб-вузла Pattern Languages Community з метою спрощення спільного створення і ведення програмних документів (вільне доступне програмне забезпечення Вікі на PHP доступне за адресою http://wikipedia.sourceforge.net, а перша його реалізація доступна за адресою http://c2.com/cgi/wiki). Вікі-технологія – це основа побудови веб-систем, призначених для колективної розробки, зберігання, структуризації тексту, гіпертексту, файлів, у тому числі мультимедіа. Усі сторінки вікі-сайту є статтями, зміст яких – це текст, в якому можна використовувати просту вікі-розмітку або теги HTML.

Проект Вікі виявився настільки успішним, що вийшов за рамки програмного продукту і перетворився на концепцію.

Можливість редагувати вміст вікі-сайту будь-яким відвідувачем, з одного боку, дозволяє без зусиль накопичувати і систематизувати інформацію, але, з іншого боку, створює велике поле для внесення помилок і вандалізму. Щоб запобігти останньому усі вікі-сайти використовують технологію, що забезпечує збереження кожної версії документа. Якщо документ піддається вандалізму, або до нього внесена помилка при редагуванні, користувач Вікі може легко відновити стару версію, тобто при необхідності можна швидко повернутися до правильного варіанту

тексту. Програмне забезпечення Вікі також дозволяє обмежити доступ і права редагування сторінок до певного кола користувачів.

Назвемо лише деякі ідеї, вікі-технології:

- можливість редагування вікі-статей певним кругом користувачів;
- зберігання усіх версій вікі-статей з моменту їх створення;
- швидка і проста генерація гіперпосилань між документами, а також підтримка цілісності гіперпосилань;
- спрощення процесу публікації тексту.

Одна з найвідоміших вікі-систем – це The Portland Pattern Repository, яка також містить велику колекцію посилань й на інші пакети програмного забезпечення Вікі.

Можна також відзначити, що програмне забезпечення Вікі багато в чому схоже з «живими журналами», при цьому воно має аналогічні плюси і мінуси. До безумовних плюсів належить простота роботи користувача через звичайний браузер. При цьому Вікі у багатьох відношеннях простіше, хоча чимало з більш багатофункціональних і складних вікі-систем дозволяють робити резервне копіювання, авторизувати користувачів, забезпечувати безпеку контенту.

Зазвичай вікі-системи використовують доступ до документів через браузер і видають їх в HTML, але, щоб писати вікі-документи, мову HTML знати необов'язково. Вікі-розмітка як правило спрощена: виділений текст укладають в зірочки, гіперлінки – в квадратні дужки тощо. Іноді в сторінки для редагування вбудовують і спеціальні скрипти, щоб документ можна було правити у звичний спосіб – натисненням відповідних іконок. В той же час у Вікі є особлива власна розмітка – передусім для організації посилання на інші вікі-тексти.

### *Корпоративні рішення на базі MediaWiki*

Технологія Вікі використовується не лише в онлайнових енциклопедіях. Так в статті творця Вікі Уорда Каннінгема «The Wiki Way» повідомляється, що такі великі компанії як Motorola і The New York Times використовують Вікі в усіх робочих процесах

колективного ведення проектної документації. Спочатку система Каннінгема використовувалася для створення бази даних для потреб архітектури і міського планування, яку ентузіасти намагалися розширити на інші області діяльності.

Те, що розроблялося компанією Microsoft для порталу-сервера Sharepoint також можна кваліфікувати як вікі-технологію. Sharepoint пропонує масу шляхів забезпечення спільної роботи з документами. При цьому загальне використання файлів далеко йде від колишнього «розподілу документів» (sharing). Відбувається перехід до спільних бібліотек, звідки документи можна легко дістати, попрацювати з ними і покласти на місце зручнішим, ніж раніше, чином.

***Вікіпедія***

Назва Вікіпедія пішла від слів «вікі» і «енциклопедія», а слово вікі, у свою чергу, – від гавайського слова «wikiwiki» – «якнайшвидше». Проект Вікіпедія орієнтований на створення онлайнової енциклопедії, написаної самими користувачами. Кожен відвідувач сайту енциклопедії може внести свій посильний внесок: підправити статтю, додати або видалити інформацію. Основний принцип Вікіпедії – щонайповніша демократія, тому автор-початківець може брати участь в проекті разом з досвідченими експертами.

Проект Вікіпедія існує з 15 січня 2001 року, як публічна універсальна інтернет-енциклопедія, в якій автором статті може бути будь-який охочий. Проект є некомерційним і розвивається на кошти спонсорів, які здебільшого є простими учасниками Вікіпедії. Координатором проекту є американський фонд Wikimedia Foundation Inc. Творці Вікіпедії Ларрі Сенгер (Larry Sanger) і Джиммі Уейлз (Jimmy Wales) почали роботи із створення безкоштовної бібліотеки у 2000 році, яку назвали «Нупедія». За задумом це була безкоштовна колективно створена енциклопедія, зміст якої міг би вільно поширюватися за ліцензією GNU FDL (GNU Free Documentation License). Навесні 2000 року було створено сайт NuPedia.com, програмне забезпечення якого також було вільним і поширювалося у вигляді відкритих кодів (Open Source) також за ліцензією GNU, згідно якої дозволяється

повномасштабне або часткове використання або тиражування без будь-яких фінансових відрахувань.

Проте схема підготовки статей в Нупедію, як показав час, виявилася невдалою, передусім, зважаючи на жорстку модерованість. Передбачалися редактори, куратори тематик (більшість з яких були відомими ученими), які досить жорстко селекціонували авторів за кваліфікацією. Відібрані автори могли починати писати статті, які потім редагували, узгоджувалися, а лише потім публікувалися на сайті. Ця технологія забезпечувала найвищу якість, проте привела до того, що в 2003 році, проіснувавши близько двох років, в Нупедії накопичилося усього лише 23 готових статей і 68 незавершених матеріалів. Очевидно, що цей етап створення загальнодоступної енциклопедії виявився проваленим і був закритий. Проте ідея залишалася живою, і автори від неї не відмовилися.

В той же час, творці Нумедії побачили, що головна перевага вікі-систем полягає не в тому, що вони не допускають помилок, а у тому, що вони дозволяють виправляти помилки дуже швидко. Тому наступним проектом після Нумедія, став проект онлайнової бібліотеки з декілька скоригованої ідеологією – Вікіпедія. Цей проект з самого початку узяв на озброєння вікі-технологію, що на практиці виявилася ідеальною для нього. Дійсно, якщо запустити вікі-систему, доступну абсолютно усім, і почати створювати на її базі енциклопедію, яка по мірі заповнення буде абсолютно прозора для усіх, можна притягнути до її написання багатьох користувачів Інтернету. За задумом авторів весь світ писатиме туди статті, доповнюватиме їх за необхідністю, а також виправлятиме помилки. Вікіпедія дозволяє публікувати статті усім бажаючим без всяких редакторів. До того ж правити статтю може теж будь-хто, а не тільки автор. Мало того, історія правок зберігається вічно, і якщо будь-яка стаття буде зіпсована, завжди можна повернутися до старого варіанту. Новий сайт проекту отримав адресу – Wikipedia.com.

Сьогодні Вікіпедія – це повноцінна онлайнова енциклопедія з 30 млн. статей 285-ма мовами (передусім англійською – 4 млн. статей – для порівняння: в «Британніці» – близько 120 тисяч, а в останньому виданні Великої Радянської Енциклопедії було 100 тисяч статей). Російська Вікіпедія (тобто

розділ Вікіпедії російською мовою) займає 7-е місце за кількістю статей серед усіх мовних розділів Вікіпедії, маючи на кінець 2013 року у своєму складі мільйон статей.

Українська Вікіпедія, на цей час, охоплює близько півмільйона статей. У Китаї Вікіпедія офіційно заборонена, але входить до першої сотні веб-ресурсів за популярністю.

***Принципи Вікіпедії***

Кожен користувач Інтернету може стати автором статті Вікіпедії. Більше того, кожен може виправити чужу статтю. У Вікіпедії для більшості статей немає модераторів, тому усі зміни відразу ж стають помітні іншим відвідувачам – варто тільки натиснути на кнопку «Записати сторінку» під вікном правки. Для цього не потрібно навіть авторизуватися (хоча рекомендується).

Здавалося б, Вікіпедія є хаотичною системою, в якій зловмисники і вандали можуть грати вирішальну роль. Така онлайнова енциклопедія могла б дуже швидко перетворитися в середовище для конфліктів. Проте нічого подібного не сталося – у тому й полягає основний феномен Вікіпедії: це система, що самоорганізується. Статті у Вікіпедії написані досить грамотно і часто навіть перевершують за точністю і повнотою викладу статті з традиційних енциклопедій, перевершуючи останні за оперативністю.

Існують три взаємодоповнюючі пояснення цьому феномену:

- у системі продумані прості і чіткі правила складання і редакції статей;
- у системі передбачено «громадське» адміністрування зі своєю ієрархією адміністраторів, які призначаються виходячи з їхнього внеску у розвиток системи;
- вандалів, зловмисників (чи, як їх називають у Вікіпедії, «тролів») насправді в десятки разів менше, ніж учасників проекту, що вносять позитивний вклад.

Правила написання статей у Вікіпедії дійсно прості: стаття має бути нейтральною і не повинна бути оригінальним

дослідженням. Тобто Вікіпедія – це джерело вторинної інформації, а не науковий журнал. Припускається, що статті для Вікіпедії слід писати літературною мовою в науковому стилі. Писати можна тільки від третьої особи без просторічних і сленгових виразів, а також без особистих підписів і зауважень.

Крім того, рекомендується писати статті неупереджено і об'єктивно, тобто без емоцій, які в об'єктивній і точній енциклопедії є зайвими. Рекомендується також писати з нейтральної точки зору: якщо твердження, що висловлене в статті спірне або не є загальновизнаним, це треба вказати, причому дуже бажано посилатися на джерело спірного твердження.

Для учасників Вікіпедії все ж найважливішим принципом є «нейтральна точка зору». У відповідних рекомендаціях вказано, що статті онлайнової енциклопедії повинні «представляти ідеї і факти так, щоб прибічники і супротивники цієї точки зору могли прийти до згоди».

«Ми не пропагуємо якусь точку зору, неважливо, чи вважаємо ми її вірною або ні, – ми прагнемо розповісти про усі існуючі – я вважаю, що це одна з найважливіших особливостей нашого проекту, яка допомагає людям бути вільними, не піддаватися тоталітаризму і пропаганді», – сказав в інтерв'ю один з адміністраторів Вікіпедії Максим Вотяков (МахіМахіМах).

Звичайно, передбачені деякі заходи по недопущенню профанації проекту. Концепцією проекту передбачено формування певної ієрархії громадських адміністраторів, що стежать за дотриманням певних правил, що полягають в нейтралітеті оцінок, дотриманні «енциклопедичності», умінні знаходити консенсус. Ряди адміністраторів проекту розширюються, практично автоматично, за рахунок додавання редакторів-спостерігачів з числа найактивніших зареєстрованих авторів. Найвищий орган Вікіпедії – Wikimedia Foundation, неприбуткова організація, створена для експлуатації проекту. Нагадаємо, що увесь зміст Вікіпедії знаходиться під дією ліцензії GNU Free Documentation License. Таким чином, будь-які тексти з Вікіпедії можуть вільно копіюватися і поширюватися за умови виконання принципу «спадкоємства прав», що отримав назву «копілефт» (калька з англійського «copyleft», протилежність «copyright»).

Тексти для Вікіпедії повинні створюватися або запозичуватися з ліцензійно чистих джерел, наприклад, з документів, що поширюються по тій же GNU FDL, або документів, що становлять «громадське надбання» (public domain). До таких документів належаться, зокрема, видання, у яких закінчився охоронний термін авторських прав. Вітаються і переклади статей, які відсутні у національних версіях Вікіпедій.

При цьому абсолютно справедливо декларується те, що Вікіпедія не може ніяким способом гарантувати правильність наведених в ній даних. Під час прочитання вони могли бути змінені, зіпсовані, або написані тим, чия думка відрізняється від загальноприйнятої у відповідній галузі знань. Тому в ліцензії зафіксовано: «Вікіпедія не може нести відповідальність за будь-який завданий збиток, оскільки є добровільним співтовариством, вільно організованим для створення відкритих освітніх, культурних і інформаційних ресурсів. Інформація надається як акт доброї волі і не існує угоди або акту про наміри між користувачами і Вікіпедією стосовно використання або зміни інформації, не передбаченого GNU Free Documentation License. Також ніхто у Вікіпедії не несе відповідальності за зміну, редагування або видалення будь-якої інформації, доданої користувачами у Вікіпедію або інші пов’язані з нею проекти».

І нарешті, у Вікіпедії виправляти усе набагато простіше, ніж псувати, оскільки вікі-середовище зберігає історію усіх змін усіх сторінок сайту. Будь-який користувач, побачивши вандалізм, може кликнути на посилання «Історія» і відновити будь-яку попередню версію сторінки. Усі зміни в статтях відстежуються добровольцями, і нікому ще не вдавалося надовго зіпсувати статті або зловмисно внести туди недостовірну інформацію. Відповідний експеримент було проведено журналістами New York Times, які спеціально зіпсували декілька статей. Менш ніж за п’ять хвилин усі зміни були виправлені. І ще одне міркування: навряд чи хтось стане писати енциклопедичні статті, не знаючи предмета.

Тих же користувачів, які занадто часто порушують правила, може заблокувати адміністратор. Разом з цим, використання Вікіпедії, особливо на початковому етапі розвитку проекту викликало певні побоювання у фахівців: все ж таки Вікіпедія не має наукової редакції, різні статті Вікіпедії нерівноцінні, більше

увага в ній приділяється актуальним подіям, ніж сталій інформації. Принцип створення цієї енциклопедії породжує і, мабуть, головну проблему: тексти Вікіпедії по суті, відбивають думки більшості користувачів енциклопедії. При цьому можуть бути пропущені альтернативні, іноді правильніші погляди. Тобто контент Вікіпедії відповідає уявленням «середньостатистичного» читача. Але це і є відмінність енциклопедії від наукового дослідження.

Проте, незважаючи на усі побоювання і недоліки, Вікіпедія, особливо її англомовна версія, є сьогодні, схоже, найбільш точним і повним мережевим ресурсом інформаційно-довідкового характеру.

***Wiki-розмітка***

Сторінки Вікі-сайту є статтями, вміст яких – це звичайний текст, в якому можна використовувати теги HTML або особливу вікі-розмітку, зручнішу для текстових документів, ніж HTML. Скориставшись посиланням або кнопкою, будь-який відвідувач вікі-сайту може відредагувати і зберегти змінений варіант тексту будь-якої існуючої сторінки або створити нову. Процедура публікації тексту у Вікіпедії зведена до двох кнопок – редагувати і зберегти.

Певна частина статей у Вікіпедії є створеними автоматично «заготівками». Коли будь-який автор позначає в тексті термін або вираз як посилання на неіснуючу статтю, у Вікіпедії автоматично генерується нова стаття-шаблон, така, що містить текст щодо того, що стаття ще не написана і її можна написати.

Для публікації, наприклад, математичних формул у Вікіпедії використовується TeX – мова розмітки спеціального призначення, яка широко застосовується при підготовці текстів науково-технічного характеру, творцем якої був Д. Кнут.

Для того, щоб цією мовою представити формулу:

$$\sum_{n=0}^{\infty} \frac{x^n}{n!},$$

у статті досить записати код:

<math>\sum_{n=0}^\infty \frac{x^n}{n!}</math>.

Залежно від бажання автора і складності формул, що записуються у текстах статей, генеруються або зображення формул у форматі PNG, або простий код HTML. Передбачається, що із зростанням функціональності браузерів генеруватиметься розширений код HTML або MathML.

Історія правок усіх вікі-статей зберігається в базі даних (кнопка «Історія» на інтерфейсі редагування статті), будь-яка редакція статті може бути викликана на екран і збережена, як остання.

Майже будь-яку сторінку цього сайту можна відредагувати. Для того, щоб внести правки до статті:

- клацніть по кнопці «Правити» або по посиланню «Правити» (для розділу статті), після чого відкриється форма з текстом статті (розділу);
- внесіть бажані правки до тексту статті;
- з метою контролю проглянути перелік внесених правок (клавіша «Внесені правки»);
- переконатися в коректності правок проглянувши заздалегідь текст статті (клавіша «Попередній перегляд»);
- Заповніте поле «Коротко опишіть суть вашої правки»;
- натисніть кнопку «Записати сторінку».

У тексті статті для поліпшення читабельності і створення гіперпосилань (вікі-посилання) можна використовувати вікі-розмітку, що відрізняється простотою і прозорістю.

У правих колонках наведених нижче таблиць представлені тексти у вікі-розмітці, а в лівих – результати інтерпретації цієї розмітки вікі-системою.

| **При перегляді** | **При редагуванні** |
|---|---|
| Для того, щоб створити посилання у Wiki, укладіть потрібне вам слово в подвійні квадратні дужки. | Для того, щоб створити посилання у [[Wiki]], укладіть потрібне вам слово в подвійні квадратні дужки. |
| Якщо слово за змістом тексту стоїть не в називному відмінку і однині (не співпадає з назвою статті), то після вертикальної риски напишіть його з потрібним вам закінченням. | Якщо слово за змістом тексту стоїть не в [[Називний відмінок\|називному відмінку]] і однині (не співпадає з назвою статті), те після вертикальної риски напишіть його з потрібним вам закінченням. |
| **Підрозділ**<br><br>**Під-підрозділ** | === Підрозділ ===<br><br>==== Під-підрозділ ==== |
| • щоб почати ненумерованый список – поставте на початку рядка зірочку<br>o підсписок | * * щоб почати ненумерований список – поставте на початку рядка зірочку<br>** ** підсписок |
| Поодинокий переклад рядка ніяк не відбивається на результуючому тексті і може допомогти розділити речення у рамках одного параграфа, що допомагає при редагуванні і відстеженні змін.<br><br>Порожній рядок починає новий параграф. | Поодинокий переклад рядка ніяк не відбивається на результуючому тексті і може допомогти розділити речення у рамках одного параграфа, що допомагає при редагуванні і відстеженні змін.<br><br>Порожній рядок починає новий параграф. |
| *акцентування*,<br>**сильне акцентування**,<br>***дуже сильне акцентування*** | ‘‘акценування’’,<br>‘‘‘сильне акцентування’’’,<br>‘‘‘‘‘дуже сильне акцентування’’’’’ |

Наведемо приклад HTML-розмітки початкового фрагменту статті «Леонардо да Вінчі»:

```
<table class="infobox" style="width: 22em; font-size: 90%; text-align: left;
align: right; padding:0.2em;">
<tr>
<td style="font-size: larger; text-align: center;" colspan="2"
<b>Леонардо да Вінчі</b></td>
```

```
</tr>
<tr>
<td style="text-align: center;" colspan="2"><a
href="/wiki/%D0%86%D1%82%D0%B0%D0%BB%D1%96%D0%B9%D
1%81%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%
B0" title="Італійська мова">італ.</a> <i><span lang="it"
xml:lang="it">Leonardo da Vinci</span></i></td>
</tr>
<tr>
<td colspan="2" style="text-align: center;"><a
href="/wiki/%D0%A4%D0%B0%D0%B9%D0%BB:Possible_Self-
Portrait_of_Leonardo_da_Vinci.jpg" class="image"><img alt="Possible
Self-Portrait of Leonardo da Vinci.jpg"
src="//upload.wikimedia.org/wikipedia/commons/thumb/9/9e/Possible_Self
-Portrait_of_Leonardo_da_Vinci.jpg/225px-Possible_Self-
Portrait_of_Leonardo_da_Vinci.jpg" width="225" height="340"
srcset="//upload.wikimedia.org/wikipedia/commons/9/9e/Possible_Self-
Portrait_of_Leonardo_da_Vinci.jpg 1.5x,
//upload.wikimedia.org/wikipedia/commons/9/9e/Possible_Self-
Portrait_of_Leonardo_da_Vinci.jpg 2x" /></a><br />
Ймовірно автопортрет Леонардо да Вінчі</td>
</tr>
<tr>
<th>Дата народження</th>
<td><span style="white-space:nowrap;"><a
href="/wiki/15_%D0%BA%D0%B2%D1%96%D1%82%D0%BD%D1%8
F" title="15 квітня">15 квітня</a> <a href="/wiki/1452"
title="1452">1452</a><span style="display:none">(<span
class="bday">1452-04-15</span>)</span></span></td>
</tr>
<tr>
<th>Місце народження</th>
<td>поблизу Вінчі, <a
href="/wiki/%D0%A4%D0%BB%D0%BE%D1%80%D0%B5%D0%BD%
D1%86%D1%96%D1%8F" title="Флоренція">Флоренція</a></td>
</tr>
<tr>
<th>Дата смерті</th>
<td><span style="white-space:nowrap;"><a
href="/wiki/2_%D1%82%D1%80%D0%B0%D0%B2%D0%BD%D1%8F"
title="2 травня">2 травня</a> <a href="/wiki/1519"
title="1519">1519</a><span style="display:none">(<span
class="dday">1519-05-02</span>)</span></span> (67&amp;#160;років)</td>
```

```
</tr>
<tr>
<th>Місце смерті</th>
<td>замок <a
href="/w/index.php?title=Clos_Luce&amp;action=edit&amp;redlink=1"
class="new" title="Clos Luce (ще не написана)">Clos Luce</a>, поблизу
<a
href="/wiki/%D0%90%D0%BC%D0%B1%D1%83%D0%B0%D0%B7"
title="Амбуаз">Амбуаза</a>, <a
href="/wiki/%D0%A2%D1%83%D1%80%D0%B5%D0%BD%D1%8C"
title="Турень">Турень</a>, <a
href="/wiki/%D0%A4%D1%80%D0%B0%D0%BD%D1%86%D1%96%D
1%8F" title="Франція">Франція</a></td>
</tr>
<tr>
<th>Національність</th>
<td><a
href="/wiki/%D0%86%D1%82%D0%B0%D0%BB%D1%96%D0%B9%D
1%86%D1%96" title="Італійці">італієць</a></td>
</tr>
<tr>
<th>Навчання</th>
<td><a
href="/wiki/%D0%90%D0%BD%D0%B4%D1%80%D0%B5%D0%B0_%
D0%92%D0%B5%D1%80%D1%80%D0%BE%D0%BA%D1%96%D0%
BE" title="Андреа Верроккіо" class="mw-redirect">Андреа
Верроккіо</a></td>
</tr>
<tr>
<th>Напрямок</th>
<td>Багато галузей, зокрема <a
href="/wiki/%D0%9C%D0%B8%D1%81%D1%82%D0%B5%D1%86%D
1%82%D0%B2%D0%BE" title="Мистецтво">мистецтво</a> і <a
href="/wiki/%D0%A2%D0%B5%D1%85%D0%BD%D1%96%D0%BA%
D0%B0" title="Техніка">техніка</a></td>
</tr>
<tr>
<th>Вплив на</th> <td><a
href="/wiki/%D0%A0%D0%B0%D1%84%D0%B0%D0%B5%D0%BB%
D1%8C_%D0%A1%D0%B0%D0%BD%D1%82%D1%96"
title="Рафаель Санті">Рафаель Санті</a></td>
</tr>
</table>
```

Той самий фрагмент у форматі вікі-розмітки має вигляд:

```
{{Художник
| Ім'я              = Леонардо да Вінчі
| Оригінал імені     = {{lang-it|Leonardo da Vinci}}
| Фото               = Possible Self-Portrait of Leonardo da Vinci.jpg
| Ширина             =
| Підпис             = Ймовірно автопортрет Леонардо да Вінчі
| Ім'я при народженні =
| Дата народження    = 15.4.1452
| Місце народження   = поблизу Вінчі, [[Флоренція]]
| Дата смерті        = 2.5.1519
| Місце смерті       = замок [[Clos Luce]], поблизу [[Амбуаз]]а,
[[Турень]], [[Франція]]
| Національність     = [[італійці|італієць]]
| Релігія            = [[Католицизм]]
| Громадянство       =
| Жанр               =
| Навчання           = [[Андреа Верроккіо]]
| Напрямок           = Багато галузей, зокрема [[мистецтво]] і [[техніка]]
| Роки творчості     =
| Член КПРС з        =
| Покровитель        =
| Вплив              =
| Вплив на           = [[Рафаель Санті]]
| Премії             =
| Твори              =
| Сайт               =
}}
```

Фрагмент зображення у веб-браузері, що відповідає цим кодам, наведено на рис. 39.


Рис. 39 – Фрагмент зображення у веб-браузері

У Вікіпедії є швидка і зрозуміла навіть новачкам автоматизована генерація і підтримка цілісності гіперпосилань між документами на усьому сайті. Проблеми з недозволеними посиланнями у Вікіпедії просто не існує. Для того, щоб створити гіперпосилання на статтю, не треба запам'ятовувати складний синтаксис, як в HTML (<a href="http://адреса посилання">Назва посилання</a> і т. д.), досить просто при внесенні вікі-розмітки укласти назву статті, на яку треба послатися, в подвійні квадратні дужки – [[Назва статті]].

Якщо статті, на яку вказує посилання, не існує, то посилання у будь якому разі створиться, але текст буде червоного кольору, а не звичайного синього. Активізувавши це посилання можна перейти до шаблону статті і написати її, фактично відкоригувавши шаблон. Таким чином, у Вікіпедії існує три види посилань – на існуючі статті, на статті ще ненаписані і на зовнішні веб-ресурси, наприклад, після внесення розмітки [http://rada.gov.ua/ «Верховна Рада України»], можна побачити у вікні браузеру блакитне посилання Верховна Рада України.

У вікі-середовищі, природно, можна використовувати і таблиці, їх форматування також спрощене в порівнянні з HTML.

Більшість статей Вікіпедії, як вже говорилося, можуть виправляти і доповнювати анонімні користувачі. Для обговорення окремих текстів можна скористатися кнопкою «Обговорення».

Фонд Вікімедія, некомерційна організація, яка підтримує роботу Вікіпедії, також підтримує й інші проекти зі створення вільних матеріалів у Інтернеті. До них, зокрема, належать:

- вільна бібліотека – ВікіТека;
- вільні підручники – ВікіПідручник;
- словник – ВікіСловник;
- збірники цитат – ВікіЦитати;
- вільні новини – ВікіНовини;
- вільний каталог біологічних видів – ВікіВиди;
- збірка медіа-файлів – ВікіСховище.

Вікі-проекти Фонду Вікімедія [Barrett, 2009], найбільшим з яких є Вікіпедія, за станом на березень 2013 року підтримувалися роботою декількох кластерів що складаються з 974 серверів.

Як було вже відзначено, у Вікіпедії багато спільного, як в ідеології, так і в технологічній частині з «живими журналами» (LiveJournal). Одна із загальних рис технології Вікі та живих журналів – це те, що документи в них зберігаються в базі даних, і жодна з версій не знищується. Навіть при мінімальній зміні в документі створюється нова версія. Поточною версією вважається завжди остання, проте завжди можливе повернення на будь-яку кількість кроків назад.

Вікіпедія, будучи феноменом, сама по собі, породила ряд похідних феноменів. Наприклад, парадоксальні результати були отримані в результаті порівняння Вікіпедії, проекту Everything і традиційної енциклопедії. Незважаючи на нічим не обмежену свободу авторів, статті Вікіпедії у своїй більшості є стилістично однорідними, представленими у єдиному форматі і розкривають, як правило, одне, основне значення слова. Автори ж Everything2, навпаки, зловживають сленгом. Статті, крім того, неоднорідні за своєю структурою. На здивування учених, незважаючи на найліберальніші правила, з точки зору синтаксису і семантики мови Вікіпедія практично не відрізняється від традиційної енциклопедії. Лінгвісти вважають, що визначальну роль в цьому зіграли два основні чинники – по-перше, правила, визначені для авторів (формальний опис, нейтральність і послідовність), і, по-друге, діяльність відданих учасників співтовариства. Доброзичливих користувачів в десятки і сотні разів більше, ніж недобросовісних і вандалів. Тому велика частина неякісного контенту вилучається ще до того, як бачить загал користувачів Інтернету.

### *Вікипедія і Семантичний веб*

Основна ідея Семантичного вебу – побудова мережі знань, щоб користувач, звертаючись до Інтернету із запитом, отримував не посилання на об’єкт пошуку, а його опис, параметри, основні відомості, тобто знання. Досягається це шляхом відмови від можливостей оформлення, переходу від HTML до XML,

використання семантичної розмітки, побудови, так званих, онтологій. При цьому представлення інформації на сайтах стають більш орієнтованими на комп'ютери, ніж на людей. Відповідно до концепції Семантичного вебу, сайти, що входять в нього повинні спеціально адаптуватися, – вони повинні переформатуватися, має бути додана спеціальна семантична розмітка.

У цьому плані Вікіпедія є альтернативним підходом для досягнення тієї ж мети – отримання знань. Пошук у Вікіпедії навіть за звичайними ключовими словами призводить до отримання змістовних матеріалів, що містять сконцентровані знання.

Багато із сучасних пошукових систем в Інтернеті врахували переваги Вікіпедії. Сьогодні результати з Вікіпедії видаються першими в пошукових системах Yahoo!, Google, Answers.com. Зокрема, на сайті Answers.com одним клацанням миші можна отримати інформацію з будь-якої теми, не посилання на інші веб-сторінки, а саме інформацію. Це викликало здивування-захват у користувачів і преси, адже практично відбувається пошук у гігантській електронній енциклопедії, обсягом понад 1 млн. словарних статей. Усю інформацію сайт Answers.com отримує в режимі реального часу з різних джерел, у тому числі з різних спеціалізованих енциклопедій (до цього переліку входить і Вікіпедія), словників, атласів тощо. Всього до переліку входить понад 100 ретельно відібраних редакторами авторитетних інформаційних джерел (замість сотень мільйонів сайтів, індексованих, наприклад, пошуковою системою Google).

За час свого розвитку проект Вікіпедії, незважаючи на побоювання, пов'язані з непрофесійністю авторів, можливим вандалізмом, спонтанністю створення окремих статей, дозволив створити досить якісний продукт – повну і об'єктивну, вільно доступну усім, багатомовну енциклопедію.

Успіх Вікіпедії продемонстрував, що користувачі Інтернет потребують достовірної енциклопедичної інформації. Тому проект мережевої енциклопедії вже сьогодні є серйозним джерелом довідкової інформації, знань, що має, на відміну від традиційних джерел, ще одну особливість – оперативність.

Завдяки зростанню популярності проекту, на нього стали звертати увагу провідні пошукові системи і інтегратори контенту.

### *WikiLeaks*

Центральною проблемою створення розподілених систем незалежних інформаційних ресурсів є їх наповнення та організація обміну наборами даних великого обсягу.

Тут слід зазначити два суттєвих моменти. По-перше, необхідно створювати відповідні програмні комплекси, що є достатньо складним процесом, який вимагає не лише високої кваліфікації розробників, але й значного часу на створення придатного для експлуатації продукту. По-друге, такі програмні продукти мають бути належним чином уніфіковані в плані пакетної обробки інформаційних масивів, що необхідно для реалізації швидкого перенесення даних з одного ресурсу до іншого. Саме завдяки застосуванню технології MediaWiki було вирішено дві окреслені проблеми при реалізації соціально-значущого проекту WikiLeaks [*Ландэ, 2011а*].

WikiLeaks – назва, утворена з двох слів, – Wiki і Leak. Перше причетне до технології Вікі, друге перекладається як «витік». Найбільший за усю історію витік секретних матеріалів через цю службу буквально перевертає свідомість усього людства, яке вимушено йде до «ери відкритості». Технологія WikiLeaks не засекречена, проте рівень впливу на усю решту інформаційного простору – питання, яке залишається доки відкритим.

Портал WikiLeaks.org спочатку входив в систему wikipedia.org. Тепер, очевидно, це не так. Незважаючи на свою назву, WikiLeaks не є вікі-сайтом: читачі, що не мають відповідного дозволу, не можуть міняти його зміст. Після перезапуску проекту у травні 2010 року користувачі втратили можливості брати участь в обговоренні за допомогою функціонала Віка на сайті проекту, а опис проекту на сайті було виправлено: якщо у 2007 році на ньому говорилося, що WikiLeaks є «Вікіпедією» без цензури, то в 2010 році вже повідомлялося, що WikiLeaks – не «Вікіпедія», і функціональності «Вікіпедії» у проекту немає.

З технічної точки зору WikiLeaks, за ідеєю творців, повинна була забезпечити:

1) можливість редагування матеріалів і розміщення їх на веб-ресурсах;

2) децентралізоване зберігання даних при збереженні анонімності їх власників;

3) багатоступінчаста анонімна передача інформації так, щоб проміжні ланки в ланцюзі передачі інформації не знали адрес відправників і адресатів;

4) надійне шифрування інформації.

Велику увагу при створенні WikiLeaks приділено захисту ресурсу і забезпеченню анонімності джерел даних. Інформація на сайт надходить як від відомих джерел, так і за допомогою спеціального анонімного електронного ящика. WikiLeaks забезпечує захист анонімності джерел і, у разі потреби, надає послуги адвоката. Редактори між собою листуються зашифрованими каналами зв'язку для захисту даних, які передаються, від перехоплення. Новини не підлягають цензурі, але є можливість видалити або затримати публікацію деяких деталей з оригінальних документів, через обсяг інформації, що надходить, і мотивів безпеки.

Автори проекту також потурбувалися щодо запобігання можливого видалення даних, розміщених на ресурсі. Це досягається за рахунок існування близько трьохсот копій сайтів-дзеркал. WikiLeaks побудовано таким чином, що після публікації контент відразу ж відображається на декількох сотнях дзеркал, внаслідок чого видалити матеріали з сайту стало практично неможливо.

Відповідно, WikiLeaks базується на використанні таких технологій (рис. 40):

1) MediaWiki (основна система управління контентом усіх вікі-проектів);

2) Freenet (децентралізоване анонімне сховище даних, де ніхто не знає, що зберігає);

3) Tor (мережа «цибульної» маршрутизації);

4) PGP (від англ. Pretty Good Privacy) – спосіб шифрування інформації.

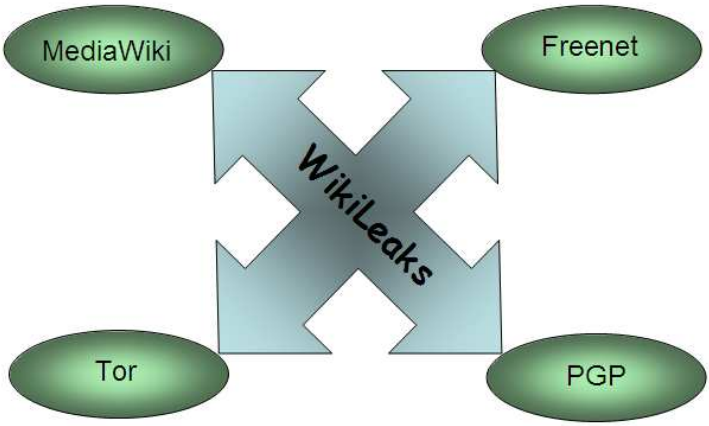



Рис. 40 – Основні компоненти технології WikiLeaks.org

Для навігації є можливим захищений доступ з використанням захищеного HTTPS – протоколу (протокол передачі гіпертексту з шифруванням).

MediaWiki (МедіаВікі) – це програмний механізм управління контентом веб-сайтів, що працюють за технологією Вікі, яку написано спеціально для WikiPedia і яка використовується в багатьох інших проектах фонду «WikiMedia», приватних і державних організаціях. MediaWiki – це вільне програмне забезпечення, що поширюється на умовах Громадської ліцензії GNU.

MediaWiki написано мовою програмування PHP, для зберігання даних вона використовує реляційну базу даних (можна використовувати MySQL, PostgreSQL, SQLite); підтримує використання програм кешування.

MediaWiki надає інтерфейс роботи з множиною веб-сторінок, розмежування прав доступу до адміністрування системи, можливість обробки тексту, завантаження зображень і інших файлів. При цьому користувачі мають можливість додавати власні нові можливості і програмні інтерфейси.

У MediaWiki передбачено спеціальний інтерфейс прикладного програмування, що забезпечує прямий доступ до інформації з баз даних. Клієнтські програми можуть

використовувати API для авторизації, отримання даних і відправки змін.

Freenet (www.freenetproject.org) – це однорангова мережа, призначена для децентралізованого розподіленого зберігання даних без можливості їх цензури, створена з метою надати користувачам електронну свободу слова шляхом забезпечення їх строгої анонімності. Freenet може розглядатися як величезний пристрій зберігання інформації. Коли ви зберігаєте файл в цей пристрій, ви отримуєте ключ, за допомогою якого ви можете отримати інформацію назад. Коли ви пред'являєте Freenet ключ, вона повертає вам збережений файл (якщо він існує). Це засіб зберігання даних розподілений по усіх вузлах, підключених до Freenet.

Tor (The Onion Router) – вільне програмне забезпечення для реалізації так званої «цибульної маршрутизації» – гібридної анонімної мережі. Це система, що дозволяє встановлювати анонімне мережеве з'єднання, захищене від прослуховування. Tor – це анонімна мережа, що надає передачу даних в зашифрованому вигляді.

PGP (англ. Pretty Good Privacy) – комп'ютерна програма, так само бібліотека функцій, що дозволяє виконувати операції шифрування (кодування) і цифрового підпису повідомлень, файлів та іншої інформації, представленої в електронному вигляді. Шифрування PGP здійснюється послідовно хешуванням, стискуванням даних, шифруванням з симетричним ключем, і, нарешті, шифруванням з відкритим ключем, причому кожен етап може здійснюватися одним з декількох підтримуваних алгоритмів.

***Динаміка витоків***

На рис. 41 наведено гістограму динаміки появи в інформаційному просторі слова Wikileaks, що отримано за допомогою системи контент-мониторинга новин InfoStream. Динаміка розраховувалася починаючи з 1 січня 2007 р., практично з того моменту, коли з'явилися перші публікації про діяльність Wikileaks, до кінця січня 2011 р.

Динаміка свідчить про неухильне, постійне зростання популярності публікацій щодо WikiLeaks, проте при найближчому розгляді виявляється, що насправді події листопаду, пов'язані з WikiLeaks, – типова інформаційна операція, просто з досить тривалим періодом підготовки. На рис. 42 наведено деталізовану гістограму, отриману за тією ж технологією за коротший проміжок часу (листопад 2010 року – січень 2011 року).

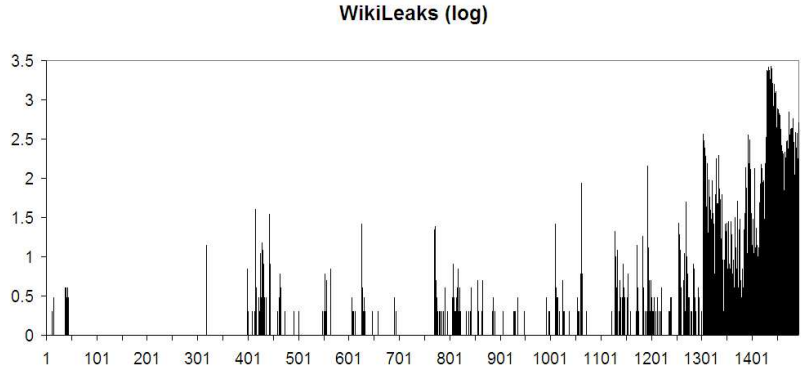



Рис. 41 – Динаміка публікацій в RUNet і UANet інформації щодо Wikileaks. Горизонтальна вісь – порядковий номер дня, вертикальна вісь – значення $\log(N+1)$, де $N$ – кількість публікацій за добу

Аналіз показує, що WikiLeaks, що виникла як маловідома служба, досягла свого максимуму в 3% новинного інформаційного простору в листопаді-грудні 2010 року, потім стабілізувалася рівні 0,5%, і увійшла до нашого життя, як один з передвісників відкритого інформаційного суспільства. Служба WikiLeaks стала потужним майданчиком для здійснення багатьох інформаційних операцій в глобальному масштабі.

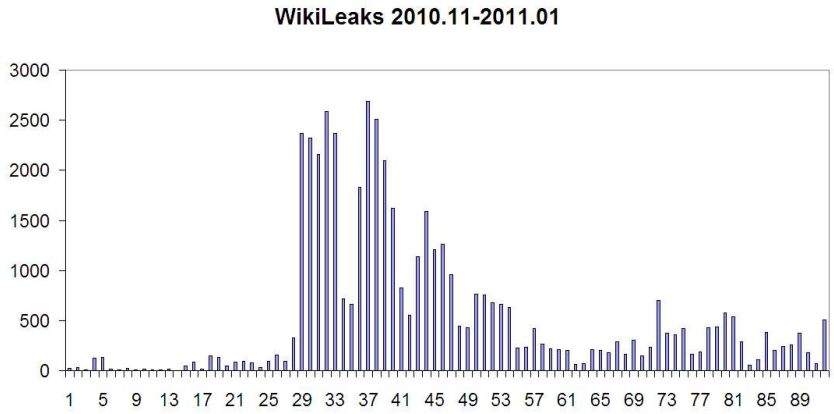



Рис. 42 – Динаміка публікацій у RUNet і UANet інформації щодо Wikileaks за три місяці. Горизонтальна вісь – порядковий номер дня, вертикальна вісь – кількість публікацій за добу

З появою порталу Wikileaks значно збільшився вплив інтернет-простору на, практично, усі сторони життя як окремо взятої країни, так усієї світової спільноти.

Слід чекати прискорення якісно нового розвитку інформаційних технологій як в області забезпечення загальнодоступності інформації (незважаючи на всякого роду «грифи», спроб захисту закритої і конфіденційної інформації від витоків і, тим більше, масового поширення), так і в області її захисту. Ця боротьба здійснюватиметься на усіх можливих рівнях: законодавчому, адміністративному, технологічному.

# 4. Модель електронної енциклопедії законодавства України

На цей час в Національній академії правових наук України створюється Електронна енциклопедія законодавства (ЕлЕнЗ), яка має охоплювати понятійно-категоріальний апарат, визначений законами, підзаконними актами та міжнародно-правовими документами. ЕлЕнЗ містить термінологію, зв'язки з нормативно-правовими документами, аналітичними та оглядовими матеріалами, практикою застосування. При цьому передбачається можливість розширення проекту за рахунок нових інформаційних компонент. З цієї точки зору перспективним є використання технології MediaWiki. Ця технологія поширюється вільно і є доступною широкому колу організацій та окремих осіб, що працюють в сфері інформаційних технологій.

З самого початку технологія MediaWiki призначалася для створення енциклопедій, статті яких формувались безпосередньо в інтерактивному режимі за допомогою спеціальних вбудованих засобів. Власне, ці засоби інтерактивного наповнення системи практично необмеженою кількістю користувачів і розглядалися як ядро вікі-технології. Але згодом виявилось, що отримана технологія може застосовуватись в значно ширшому спектрі проблем, що стимулює швидкий розвиток технологічних ідей.

Технологія MediaWiki – це основа побудови веб-систем, призначених для колективної розробки, зберігання, структуризації тексту, гіпертексту, файлів, у тому числі мультимедіа. Усі сторінки MediaWiki-сайту є статтями, вміст яких являє собою текст, в якому можна використовувати просту вікі-розмітку або теги HTML.

На цей час загальновідомий проект Вікі (на базі технології MediaWiki) виявився настільки успішним, що вийшов за рамки програмного продукту і перетворився на концепцію.

Саме технологію MediaWiki було вибрано як базову для створення ЕлЕнЗ виходячи з таких 7 головних причин:

1. Застосування мережевої технології, успішність якої з самого початку була підтверджена практикою.
2. Вільна ліцензія, не потрібна власна розробка ПЗ.
3. Можливість включення великої кількості інформаційних адміністраторів.
4. Можливість організації гіперпосилань як на зовнішні ресурси, так і на окремі свої статті.
5. Можливість створення різних категорій, збільшення видів інформаційних ресурсів.
6. Можливість «підключення» мультимедійних матеріалів.
7. Можливість пакетного завантаження вже накопичених матеріалів.

Головною перевагою технології MediaWiki є, безперечно, комплекс вбудованих інструментальних засобів, які дозволяють виконувати штатні операції з даними в інтерактивному режимі через стандартний веб-інтерфейс.

Насправді в сучасному стані технологія MediaWiki дозволяє виконувати набагато ширший спектр робіт з даними, що зберігаються в системі [*Barrett, 2009*]. Перш за все, це розвинені операції експорту та імпорту даних без суттєвих обмежень обсягу, що дозволяє тиражувати комплекси та утворювати «клони» (якщо є така потреба). Обмін даними здійснюється в форматі XML, який допускає створення потрібних наборів даних різноманітними технічними засобами від пакетних конверторів до звичайних XML-редакторів. Формуючи набір тегів, маємо можливість безпосередньо керувати склад та структуру даних, що завантажуються в wiki-систему.

Слід зазначити, що використання формату XML створює умови для зручної обробки досить складної структурованої інформації великих обсягів. Головна перевага такої технології полягає в тому, що формат XML дозволяє точно описати та відобразити будь-яку, навіть дуже складну структуру даних. Опис конкретної вікі-системи виконується прозоро і містить в

собі всю наявну інформацію – від характеристик форматування головної сторінки до параметрів ревізій кожної статті. Такий підхід дозволяє створювати та модифікувати контент вікі-системи шляхом незалежної обробки кожного елемента загального набору даних. При цьому текст кожної статті являє собою окремий тег, який безпосередньо містить всі елементи внутрішньої розмітки Вікі, що спрощує створення нових документів, оскільки не вимагає використання додаткових засобів.

Таким чином, маємо можливість безпосередньо керувати складом та структурою даних, що завантажуються в вікі-систему, просто формуючи набори відповідних тегів.

Технологія MediaWiki дозволяє здійснювати різноманітні операції з модифікації даних, в тому числі з ревізіями статей. Вони можуть виконуватися в двох основних режимах: інтерактивними засобами безпосередньо в системі, та автоматично шляхом підготовки даних у вигляді файлу для пакетного завантаження.

Операції над даними можуть здійснюватись на кількох рівнях: штатними засобами Веб-інтерфейсу системи; штатними засобами з використанням спеціальних утиліт з командного рядка та зовнішніми утилітами, які містяться в бібліотеках мовами Perl, PHP, Python. Зовнішні утиліти поділяються на дві категорії. Перша категорія передбачає безпосереднє використання API MediaWiki, а друга працює на рівні ботів.

Стандартні операції обміну даними можуть бути виконані штатними утилітами, що входять до комплекту MediaWiki – це вивантаження даних в файл у форматі XML (можна вивести у вихідний файл всі ревізії статей, або лише останні) та завантаження даних з формату XML, що має таку ж структуру, як і у випадку вивантаження. Система розрахована на отримання інформаційних наборів даних, що надходять з різних джерел у різних форматах. Отже, система забезпечує можливість уніфікованого конвертування поширених форматів у обумовлений формат, який використовується в ній для пакетного завантаження. З цієї точки зору технологія MediaWiki

може застосовуватись також для створення різноманітних інтеграторів довідкової інформації.

Слід також зазначити, що технологія MediaWiki, дозволяє використовувати різні інформаційно-пошукові системи, що забезпечує додаткову гнучкість створюваних продуктів.

Особливий інтерес становить можливість формувати ієрархічний гіпертекст, що значно спрощую доступ до потрібної користувачеві інформації. Забезпечення інформаційних потреб користувача у цьому випадку складається з двох фаз. Перша фаза полягає в звичайному інформаційному пошуку за ключовими словами (статистика свідчить, що принаймні 80% запитів складаються з одного слова). Друга фаза – це розширення інформаційного поля шляхом отримання додаткових статей за гіперпосиланнями, які можуть міститися в кожній статті.

На цей час створено діючу модель електронної енциклопедії законодавства України, що містить 6000 статей, доповнених як зовнішніми посиланнями на законодавчі акти, кодекси України, так і внутрішніми – посиланнями з одних статей із ЕлЕнЗ на інші (рис. 43). Передбачається надання системи, що має бути створена на базі цієї моделі, у доступ користувачам як безпосередньо за гіперпосиланнями, так і у режимі повнотекстового пошуку.

Можливості програмного забезпечення дозволяють організувати колективну роботу з ведення/супроводження ЕлЕнЗ у модерованому режимі – головний адміністратор ЕлЕнЗ має повноваження призначення інформаційних адміністраторів, управління ботами – програмами, що здійснюють пакетне завантаження даних у базу даних системи, та отримання статистики роботи користувачів і наповнення бази даних. Інформаційні адміністратори мають здійснювати введення та коригування даних, зареєстровані корпоративні користувачі також мають можливість введення та коригування введених ними даних, відвідувачі з мережі Інтернет мають право доступу до баз даних ЕлЕнЗ у режимі «читання» та пошуку. Користувачі

мережі Інтернет звертаються до веб-сайту ЕлЕнЗ через стандартний браузер.

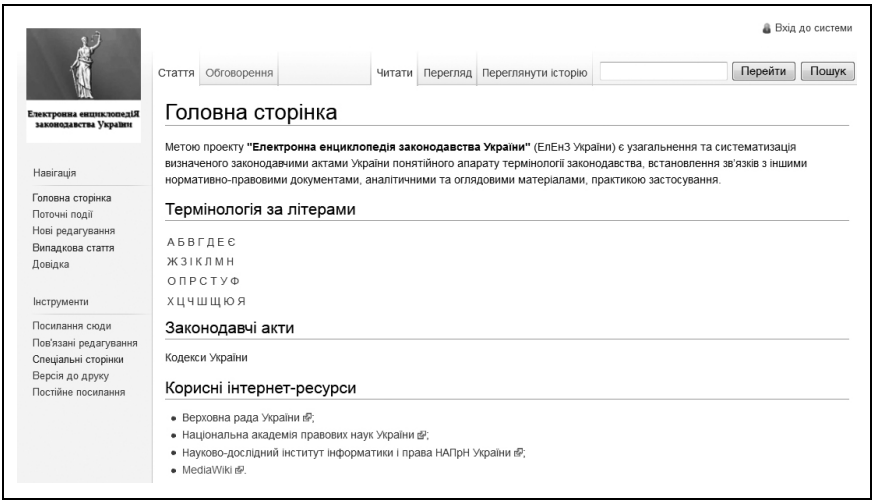



Рис. 43 – Інтерфейс моделі ЕлЕнЗ

При розробці електронної енциклопедії законодавства України враховувалось:

1. Необхідність синхронізації із поточними змінами у законодавстві. Це потребує технологічного зв'язку із відповідними базами даних.
2. Необхідність «ручної» роботи із коригування відібраних документів.
3. База даних ЕлЕнЗ має доповнюватися у автоматизованому режимі повними документами, гіперпосиланнями, мультимедійними матеріалами.

Інформаційне сховище (база даних) комплексу ЕлЕнЗ складається із окремих документів, в яких містяться опис понять, визначений законодавчими актами. В окремих документах інформаційного сховища передбачається встановлення зв'язків (гіперпосилань) як між окремими документами інформаційного сховища, так і з іншими

(зовнішніми по відношенню до ЕлЕнЗ) нормативно-правовими документами.

Інформаційне сховище (база даних) комплексу ЕлЕнЗ складається із окремих документів, в яких містяться опис понять, визначений законодавчими актами. В окремих документах інформаційного сховища передбачається встановлення зв'язків (гіперпосилань) як між окремими документами інформаційного сховища, так і з іншими (зовнішніми по відношенню до ЕлЕнЗ) нормативно-правовими документами.

Як інформаційне заповнення бази даних комплексу ЕлЕнЗ використовуються текстові документи – фрагменти декларативних частин законодавчих документів, в яких визначається понятійний апарат. Передбачається два шляхи завантаження документів в інформаційне сховище – діалоговий (стандартними засобами MediaWiki) і пакетний через зовнішній файл, який формується у визначеному в комплексі ЕлЕнЗ зовнішньому системному форматі.

Можливості технології MediaWiki дозволяють організувати колективну роботу з ведення/супроводження ЕлЕнЗ у модерованому режимі (рис. 44).

У відповідності з наведеною схемою головний адміністратор ЕлЕнЗ має повноваження призначення інформаційних адміністраторів, управління ботами та отримання статистики роботи користувачів і наповнення бази даних. Інформаційні адміністратори мають здійснювати введення та коригування даних, зареєстровані корпоративні користувачі також мають можливість введення та коригування введених ними даних, відвідувачі з мережі Інтернет мають право доступу до баз даних ЕлЕнЗ у режимі «читання» та пошуку.


Рис. 44 – Схема організації колективної роботи

Електронна енциклопедія законодавства України як комплекс комп'ютерних програм реалізує інформаційне сховище (базу даних), з самого початку орієнтоване на збереження, повнотекстовий пошук і надання у доступ користувачам визначеного законодавчими актами України понятійного апарату термінології законодавства, яку реалізовано у вигляді окремих документів – складових інформаційного сховища. При цьому в окремих документах передбачається встановлення зв'язків (гіперпосилань) як між окремими документами інформаційного сховища, так і з іншими (зовнішніми по відношенню до ЕлЕнЗ) нормативно-правовими документами. ЕлЕнЗ є інтелектуальною надбудовою над програмою MediaWiki, що розповсюджується під Загальнодоступною громадською ліцензією GNU, створено на апаратній основі (окремому сервері), яка обладнана загальносистемним та прикладним програмно-технологічним забезпеченням. Операційне середовище, необхідне для функціонування ЕлЕнЗ, включає операційну систему типу UNIX (FreeBSD), інтерпретатор мови програмування Perl (версії 5), СУБД

MySQL, веб-сервер Apache, який забезпечує можливість підключення програм функціонування веб-сайтів за допомогою інтерфейсу PHP, а також програмне забезпечення MediaWiki.

Цей комплекс складається з таких основних компонент:

1. Загальносистемна компонента, реалізована на базі програми MediaWiki, яка включає, серед іншого, реляційну базу даних (MySQL) і інформаційно-пошукову систему.
2. Модуль екстрагування вихідних даних із законодавчих документів, орієнтований на вхідний формат представлення даних у Верховній Раді України.
3. Модуль приведення документів до внутрішнього системного формату.
4. Модуль реалізації зовнішніх гіперпосилань на базові законодавчі акти із документів, що входять до інформаційного сховища.
5. Модуль реалізації внутрішніх гіперпосилань між документами, що входять до інформаційного сховища комплексу ЕлЕнЗ.
6. Модуль завантаження підготовлених даних у пакетному режимі, який побудовано з використанням прикладного програмного інтерфейсу (API) програми MediaWiki.
7. Модуль виведення змісту інформаційного сховища у внутрішньому системному форматі для зовнішнього редагування, а також проставлення зовнішніх і внутрішніх гіперпосилань.
8. Інтерфейс користувача, реалізований з використанням засобів розмітки MediaWiki.

При цьому реалізовано повномасштабне інформаційне сховище, що враховує особливості визначеної проблематики, надає релевантну інформацію за запитами, накопичує та надійно зберігає інформацію для використання в аналітичній роботі.

Комплекс комп'ютерних програм «Електронна енциклопедія законодавства України» на цей час захищений авторським свідоцтвом [*Пилипчук, 2013*], встановлено на сервері Науково-дослідного інституту інформатики і права НАПрН України. Модель електронної енциклопедії законодавства України доступна у режимі вільного доступу для користувачів мережі Інтернет за адресою http://dict.ippi.org.ua.

При розробленні Електронної енциклопедії законодавства України на базі технології MediaWiki враховується:

1. Необхідність синхронізації із поточними змінами у законодавстві. Це потребує технологічного зв'язку із відповідними БД.

2. Необхідність «ручної» роботи, яка має виконуватися за рахунок підключення зацікавлених осіб: студентів, правників.

3. Застосування де-факто стандарту MediaWiki здешевлює проект, робить його доступним у методичному плані.

4. Для підвищення якості ЕлЕнЗ необхідно підключення до проекту фахівців-лексикографів.

5. Термінологічна база ЕлЕнЗ має доповнюватися повними документами, гіперпосиланнями, мультимедійними матеріалами.

Використання технології MediaWiki відкриває широкі перспективи розширення набору інструментальних засобів для розробки уніфікованих систем, що забезпечують можливість накопичення та обміну даних без розробки складних програмних комплексів.

# Післямова

Правова інформатика – це прикладна галузь загальної інформатики, що з одного боку застосовується в умовах існуючої правової системи для потреб цієї системи, а з іншого – вивчає правові проблеми обігу правової інформації.

У свою чергу, комп'ютерна лінгвістика надає прикладні методи опису і обробки мови для комп'ютерних систем, серед яких і системи, що належаться до правової інформатики.

Найважливішими напрямками комп'ютерної лінгвістики, пов'язаними з системами правової інформації, є комп'ютерна лексикографія (створення електронних словників, тезаурусів, онтології у галузі правової інформації, оцінка дискримінантної сили слів у текстах з правової тематики для застосування в інформаційному пошуку, навігації в правових системах), корпусна лінгвістика (створення та використання електронних корпусів текстів правової спрямованості, як одномовних, так і багатомовних, інколи паралельних), автоматичний переклад текстів, витяг, екстрагування фактів із текстів (Fact Extraction), усе, що стосується концепції глибинного аналізу текстів (Text Mining), методи порівняльного аналізу текстових документів – виявлення інформаційних дублікатів, подібних документів в системах контент-моніторингу, порівняльного аналізу різномовних текстів, побудованих на застосуванні систем статистичного перекладу. Наведені підходи ілюструються їх застосуванням на колекції законодавчих актів та масивах новинних повідомлень.

Ці питання знаходять широке практичне застосування, стали вже традиційними для цієї галузі, і відповідно, знайшли своє відображення у монографії.

Особливу увагу у цій монографії було приділено опису технологіям створення і ведення електронних словників, зокрема, засобами системи MediaWiki. Ця технологія на цей час знаходить широке застосування в практичній лексикографії, а також у створенні систем управління знаннями, мережевих енциклопедій (найвідоміший проект – Вікіпедія).

На базі технології MediaWiki за участю автора реалізовано діючу модель електронної енциклопедії законодавства України, яка охоплює понятійно-категоріальний апарат, визначений законами, підзаконними актами та міжнародно-правовими документами. Використання MediaWiki в цій моделі відкриває широкі перспективи розширення набору інструментальних засобів для розробки уніфікованих систем, що забезпечують можливість накопичення та обміну даних без розробки складних програмних комплексів.

Нажаль, обсяги цього видання не дозволили зупинитися на багатьох напрямках, таких як створення систем типу «запитання-відповідь», розпізнавання мови і текстів. Усі ці напрямки сьогодні також втілюються у юридичну практику, і також визначаються як лінгвістичні засади правової інформатики.

# Література


[*Анисимов, 2005*] *Анисимов А.В., Тарануха В.Ю.* Возможности применения WordNet и других лингвистических онтологий в современных информационных системах // Автоматика-2005: Материалы 12-й международной конференции по автоматическому управлению. – Харьков: Изд. НТУ «ХПИ», 2005. – **3**. – С. 56-57.

[*Антонова, 2011*] *Антонова А.Ю., Клышинский Э.С.* Об использовании мер сходства при анализе документации // Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL'2011», – 2011. – С. 134-138.

[*Ашура, 2006*] *Аушра А.* Научная электронная библиотека как средство борьбы с плагиатом // Educational Technology & Society 9(3), 2006. – С. 270-276.

[*Баглей, 2007*] *Баглей С.Г., Антонов А.В., Мешков В.С., Титов А.В.* Вероятностный подход к задаче разрешения омонимии слов и словарных пар // Труды межд. конф. Диалог'2007. – 2007. – С. 23-28.

[*Воронина, 2010*] *Воронина И.Е., Пигалкова Е.А.* Создание базовой онтологии для российской системы права на основе онтологии LKIF-CORE // Вестник ВГУ, серия: Системный анализ и информационные технологии, 2010. – № 1. – С. 154-159.

[*Гарабик, 2006*] *Гарабик Р., Захаров В.* Параллельный русско-словацкий корпус // Труды международной конференции Корпусная лингвистика – 2006. Sankt-Petersburg: St. Petersburg University Press 2006, s. 81–87.

[*Гасфилд,2003*] *Гасфилд Д.* Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология / пер. с англ. – СПб.: Невский Диалект; БХВ-Петербург, 2003. – 654 с.

[*Головач, 2006*] *Головач Ю., Пальчиков В.* Лис Микита і мережі мови, Журн. Фіз. Досл., 2006. – № 10. – С. 247-291.

[*Горбулін, 2009*] *Горбулін В.П., Додонов О.Г., Ланде Д.В.* Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. – К.: Інтертехнологія, 2009. – 164 с.

[*ГОСТ 7.24-90*] Государственный стандарт Союза ССР. Система стандартов по информации, библиотечному и издательскому делу.

Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению ГОСТ 7.24-90, Москва, 1990.

[*Григорьев*, 2007] *Григорьев А.Н., Ландэ Д.В., Бороденков С.А. и др*. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис. – К.: Старт-98, 2007. – 40 с.

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

[*Додонов, 2010*] *Додонов А.Г., Ландэ Д.В.* Живучесть информационных сюжетов как динамических документальных систем // Реєстрація, зберігання і обробка даних, 2010. – № 2, – 12. – С. 88-102.

[*ДСТУ 4032-2001*] ДСТУ 4032-2001 «Одномовний тезаурус. Методика розроблення»

[*Дубичинский, 2008*] *Дубичинский В.В*. Лексикография русского языка: учеб. пособие. – М.: Наука, Флинта, 2008. – 432 с.

[*Жигало, 2010*] *Жигало В.В., Ландэ Д.В*. Статистический онлайн-переводчик InfoStream // Прикладна лінгвістика та лінгвістичні технології: MegaLing'2010: Зб. наук. пр. – К.: Довіра, 2010. – С. 65-78.

[*Захаров, 2002*] *Захаров В.П.* Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005. – 48 с.

[*Зеленков, 2005*] *Зеленков Ю.Г., Сегалович И.В., Титов В.А*. Вероятностная модель снятия морфологической омонимии на основе нормализубющих подстановок // Труды межд. конф. Диалог'2005. – М.: Наука, 2005.

[*Зеленков, 2007*] *Зеленков Ю.Г., Сегалович И.В*. Сравнительный анализ методов определения нечетких дубликатов для ВЕБ-документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: труды 9-й Всероссийской научной конференции RCDL´2007: сб. работ участников конкурса, 2007. – Т. 1. – С. 166-174.

[*Зинькина, 2005*] *Зинькина Ю.В., Пяткин Н.В., Невзорова О.А*. Разрешение функциональной омонимии в русском языке на основе контекстных правил // Труды межд. конф. Диалог'2005.– М.: Наука, 2005. С. 198-202.

[*Кормен,2006*] *Кормен Т.Х., Лейзерсон Ч., Ривест Р., Штайн К.* Алгоритмы: построение и анализ. – 2-е изд. – М.: «Вильямс», 2006. – 1296с.

[*Ланде, 1999*] *Ланде Д.В., Сороко В.М.* Створення основ функціонального класифікатора з питань державної служби в Україні // Вісник державної служби України, 1999. – № 4. – С. 83-88.

[*Ландэ, 2005*] *Ландэ Д.В.* Поиск знаний в Інтернет. Профессиональная работа – М.: «Диалектика», 2005. – 272 с.

[*Ландэ, 2005а*] *Ландэ Д.В.* За знаниями – к Википедии. Часть 1 // Телеком, 2005. – № 9. – С. 60-64.

[*Ландэ, 2005б*] *Ландэ Д.В.* За знаниями – к Википедии. Часть 2 // Телеком, 2005. – № 11. – С. 60-64.

[*Ландэ*, 2006] *Ландэ Д.В., Дармохвал А.Т., Морозов А.Ю.* Подход к выявлению дублирования сообщений в новостных информационных потоках // Труды 8-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL'2006», 2006. – С. 115-119.

[*Ланде, 2007а*] *Ланде Д.В.* Програмно-апаратний комплекс інформаційної підтримки прийняття рішень: науково-методичний посібник / Д.В. Ланде, В.М. Фурашев, О.М. Григор'єв. – К.: Інжиніринг, 2006. – 48 с.

[*Ланде, 2013*] *Ланде Д.В., Снарський А.О.* Графи горизонтальної видимості як засіб витягу інформаційно-занчущих слів із законодавчих актів // Правова інформатика, N 2 (38), 2013. – С. 13–18.

[*Ландэ*, 2007б] *Ландэ Д.В., Брайчевский С.М. и др.* Выявление новых событий из потока новостей // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». – М.: Изд-во РГГУ, 2007. – С. 349-352.

[*Ландэ*, 2008а] *Ландэ Д.В., Дармохвал А.Т., Жигало В.В.* Статистико-лексикографический подход к индексированию двуязычных текстовых массивов // MegaLing'2008. Горизонти прикладної лінгвістики та лінгвістичних технологій // Доповіді міжнародної конференції / Мовно-інформаційний фонд України. – Сімферополь: «ДиАйПи», 2008. – С. 120-121.

[*Ланде, 2008,*] *Ланде Д.В., Жигало В.В.* Підхід до рішення проблеми пошуку різномовного плагіату // Сб. наукових праць

«Проблеми автоматизації та управління». – К.: НАУ, 2008. – Вип. 2(24). – С. 125-129.

[*Ландэ, 2009а*] *Ландэ Д.В., Дармохвал А.Т., Жигало В.В.* Матричные критерии качества выявления подобных документов в информационных потоках // Збірник наукових праць Інститута проблем моделювання в енергетиці ім. Г.Є.Пухова НАНУ. – Вип. 53, 2009. – С. 156-163.

[*Ландэ, 2009б*] *Ландэ Д.В., Жигало В.В.* Подход к созданию многоязычных параллельных корпусов веб-публикаций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15). – М.: РГГУ, 2009. – С. 278-283.

[*Ландэ, 2011*] *Ландэ Д.В., Дармохвал А.Т., Жигало В.В.* Выравнивание параллельных текстов с использованием словарей *N*-грамм // MegaLing'2011. Горизонти прикладної лінгвістики та лінгвістичних технологій. Доповіді міжнародної наукової конференції. Весняна сесія. 12-16 травня 2011. – Сімферополь: ТНУ ім. В.І. Вернадського. – С. 77.

[*Ландэ, 2011а*] *Ландэ Д.В., Фурашев В.Н.* WikiLeaks – начало переформатирования информационного общества? // Открытые информационные и компьютерные интегрированные технологии: сб. науч. тр. – Х.: Нац. авиокосм. ун-т «ХАИ», 2011. – Вып. 49. – С. 238-247.

[*Ланде, 2012*] *Ланде Д.В.* Методи оцінки рівня дискримінантної сили слів у текстах з правової тематики // Правова інформатика, 2012. – № 3 (35). – С. 5-9.

[*Ланде, 2012+а*] *Ланде Д.В., Фурашев В.М.* Основи інформаційного і соціально-правового моделювання: монографія. – К.: ТОВ «ПанТот», 2012. – 144 с.

[*Ландэ, 2009*] *Ландэ Д.В.* Визуализация статистики вхождения слов // MegaLing'2009. Горизонты прикладной лингвистики и лингвистических технологий: материалы международной конференции 21-26 сентября 2009 г., Украина, Киев. – К.: Довіра, 2009. – С. 63-64.

[*Морозов, 1915*] *Морозов Н.А.* Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд. // Известия отд. русского языка и словестности Имп. Акад. наук, Т. XX, кн. 4, 1915.

[*Нейл, 2005*] *Нейл К., Шанмагантан Г.* Веб-инструмент для выявления плагиата // Открытые системы, 2005. – № 1. – С. 40-44.

[*Никконен, 2007*] *Никконен А.Ю.* Устранение избыточности и дублирования сюжетов новостных сообщений // Интернет-Математика. Сборник работ участников конкурса. – Екатеринбург: Изд-во Урал. Ун-та, 2007. – С. 157-167.

[*Пилипчук, 2013*] Комплекс комп'ютерних програм «Електронна енциклопедія законодавства України» («КПП "ЕлЕнЗ») / Пилипчук В.Г., Фурашев В.М., Ланде Д.В., Брайчевський С.М. // Міністерство освіти і науки України. Державний департамент інтелектуальної власності України. Свідоцтво про реєстрацію авторського права на твір № 51598 від 07.10.2013.

[*Потьомкін, 2008*] *Потемкин С.Б., Кедрова Г.Е.* Выравнивание неразмеченного корпуса параллельных текстов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2008». – Вып. 7 (14). – М.: РГГУ, 2008. – С. 431-436.

[*Снарский, 2007*] *Снарский А.А., Ландэ Д.В и др.* Особенности соотношения локальной и глобальной популярности сообщений электронных СМИ // MegaLing'2007. Горизонты прикладной лингвистики и лингвистических технологий. Доклады международной конференции 24-28 сентября 2007 г. – Симферополь: «ДиАйПи», 2007. – С. 223-224.

[*Сокирко, 2005*] *Сокирко А.В., Ножов И.М.* Описание МаПоста // АОТ «Технологии» Описание МаПоста: http://www.aot.ru/docs/mapost.html (17 октября 2005 г.)

[*Соловьев, 2006*] *Соловьев В.Д.* Онтологии и тезаурусы / В.Д. Соловьев, Б.В. Добров, В.В. Иванов, Н.В. Лукашевич. – Казань: Казанский государственный университет, 2006. – 157 с.

[*Темников, 1963*] *Темников Ф.Е.* Информатика // Изв. высш. учеб. заведений. Электромеханика, 1963. – № 11. – С. 1277.

[*Цинман, 2000*] *Цинман Л.Л.*, Сизов В.Г. Лингвистический процессор ЭТАП: дескрипторное соответствие и обработка метафор // Труды межд. семинара Диалог'2000. – М.: Изд-во РГГУ, 2000. – С. 366-369.

[*Чуи, 2001*] *Чуи К.* Введение в вэйвлеты / К. Чуи. – М.: Мир, 2001. – 416 с.

[*Шарапов, 2011*] *Шарапов Р.В.* Анализ подходов к обнаружению заимствованных текстов // Фундаментальные исследования. – 2011. – № 3 – С. 47-49.

[*Шаров, 2003*] *Шаров С.А.* Представительный корпус русского языка в контексте мирового опыта // НТИ. Сер. 2 Информационные процессы и системы. – 2003. – № 6. – С. 9–17.

[*Широков, 1998*] *Широков В.А.* Інформаційна теорія лексикографічних систем. – К.: Довіра, 1998. – 331 с.

[*Широков, 2005*] *Широков В.А.* Елементи лексикографії.– К.: Довіра, 2005. – 304 с.

[*Широков, 2005а*] *ШироковВ.А., Бугаков О.В., Грязнухіна Т.О.* Корпусна Лінгвістика. – К.: Довіра, 2005. – 471 с.

[*Широков, 2011*] *Широков В.А.* Комп'ютерна лексикографія. – К.: Наукова думка, 2011. – 352 с.

[*Шрейдер, 1971*] *Шрейдер Ю.А.* Равенство, сходство, порядок. – М.: Наука, 1971. – 256 с.

[*Щерба, 1936*] *Щерба Л.В.* Передмова до російсько-французського словника / Русско-французский словарь / Сост. Л. В. Щерба, М. И. Матусевич, М. Ф. Дусс. Под общ. рук. и ред. Л. В. Щербы. – М., 1936. 11 с. без пагинации – 491 с.

[*Ягунова, 2010*] *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.

[*Ягунова, 2012*] *Ягунова Е.В., Ландэ Д.В.* Динамические частотные характеристики как основа для структурного описания разнородных лингвистических объектов // труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции RCDL-2012». – С. 196-205.

[*Якименко, 2005*] *Якименко К.Н.* Основні принципи організації та побудови української системи WORDNET // УСиМ. – 2005. – №1. – С. 62-67.

[*Albert, 1999*] *Albert R., Jeong H., Barabasi A.-L.* Diameter of the world wide веб // Nature (London), 1999. – **401**, 130.

[*Albert, 2002*] *Albert R., Barabasi A.-L.* Statistical mechanics of complex networks // Reviews of Modern Physics, 2002. – **74**. – P. 47.

[*Amsler, 1982*] *Amsler R.A.* Computational lexicology: A research program. In American Federated Information Processing Societies Conference Proceedings. – National Computer Conference, 1982. – P 657-663.

[*Barrett, 2009*] *Barrett D. J.* MediaWiki. – O'Reilly Media Inc., 2009. – 376 p.

[*Bourdaillet, 2007*] *Bourdaillet J.* Alignment of Noisy Unstructured Text Data // IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data. Hyderabad, India – January 8, 2007. – P. 139-146.

[*Broder,1997*] *Broder A., Glassman S.C., Manasse M.S.* Syntactic Clustering of the Веб // WWW6, 1997.

[*Broder,2000*] *Broder A.* Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. – 2000. – P. 1-10.

[*Caldeira, 2005*] *Caldeira S.M.G., Petit Lobao T.C., Andrade R.F.S., Neme A., Miranda J.G.V.* The network of concepts in written texts // Preprint physics/0508066 (2005).

[*Carpena, 2009*] *Carpena P., Bernaola-Galván P., Hackenberg M., Coronado A.V., Oliver J.L.* Level statistics of words: Finding keywords in literary texts and symbolic sequences // Phys Rev E Stat Nonlin Soft Matter Phys. 2009, E 79. – P. 035102-1 – 035102-4.

[*Chesley, 2006*] *Chesley P., Vincent B., Li Xu, Srihari R. K.* Using verbs and adjectives to automatically classify blog sentiment // Training. – 2006. – T. 580. – C. 233-235.

[*Chowdhury, 2002*] *Chowdhury A., Frieder O. etc.* Collection statistics for fast duplicate document detection // ACM Transactions on Information Systems (TOIS), April 2002. – **20**, Issue 2. – P. 171-191.

[*Dorogovtsev, 2001*] *Dorogovtsev S.N., Mendes J. F. F.* Language as an evolving word веб // Proc. R. Soc. Lond. B 268, 2603 (2001).

[*Erdős, 1959*] *Erdős, P., Renyi A.* On Random Graphs. I // Publ. Math., 1959. – **6**. – P. 290–297.

[*Erdős, 1960*] *Erdős P., Renyi A.* On the evolution of random graphs // Publ. Math. Inst. Hungar. Acad, 1960. – Sci. 5. – P. 17–61.

[*Fellbaum, 2005*] *Fellbaum C.* WordNet: An Electronic Lexical Database. – MIT Press, 2005. – 425 p.

[*Ferrer-i-Cancho, 2001*] *Ferrer-i-Cancho, R., Sole R.V.* The small world of human language // Proc. R. Soc. Lond. B 268, 2261 (2001).

[*Ferrer-i-Cancho, 2004*] *Ferrer-i-Cancho R., Sole R.V., Kohler R.* Patterns in syntactic dependency networks // hys. Rev. E 69, 051915 (2004).

[*Ferrer-i-Cancho, 2005*] *Ferrer-i-Cancho, R.* The variation of Zipf's law in human language. // Phys. Rev. E 70, 056135 (2005).

[*Giora, 1983*] *Giora R.* Segmentation and Segment Cohesion: On the Thematic Organization of the Text // Text. An Interdisciplinary Journal for the Study of Discourse Amsterdam. – **3**. – № 2. – P. 155-181 (1983).

[*Gutin, 2011*] *Gutin G., Mansour T., Severini S.* A characterization of horizontal visibility graphs and combinatoris on words // Physica A, – 390 – P. 2421-2428 (2011).

[*Hearst, 1991*] *Hearst M.A.* Noun homograph disambiguation using local context in large text corpora// Processing of the 7th conference on Research and Development in Information Retrieval ACM/SIGIR, pp. 36-47. – UW Centre for the New OED & Text Research Using Corpora, Pittsburgh, PA., 1991.

[*Hutchins, 2005*] *Hutchins W.J.* Current commercial machine translation systems and computer-based translation tools: system types and their uses // International Journal of Translation, 2005. – **17**. – № 1-2. – P. 5-38.

[*Hutchins, 2007*] *Hutchins W.J.* Machine translation: a concise history // To be published in Computer aided translation: Theory and practice, ed. Chan Sin Wai. Chinese University of Hong Kong, 2007.

[*Ilyinsky*, 2002] *Ilyinsky S., Kuzmin M., Melkov A., Segalovich I.* An efficient method to detect duplicates of Веб documents with the use of inverted index // WWW-2002 – Eleventh Intern. World Wide Веб Conference. URL: http://www2002.org/CDROM/poster/187/.

[*Jurafsky, 2000*] *Jurafsky D, Martin J.H.* Speech and Language Processing: An ntroduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall PTR, Upper Saddle River, NJ, 2000.

[*Karp,1987*] *Karp R.M., Rabin M.O.* Efficient randomized pattern-matching algorithms // IBM Journal of Research and Development. – **31** (2), March 1987. 249-260.

[*Lande*, 2007] *Lande D.V., Zhygalo V.V.* About the creation of a parallel bilingual corpora of веб-publications // ePreprint Arxiv (0807.0311).

[*Lande*, 2013] *Lande D.V., Snarskii A.A.* Compactified HVG for the Language Network // International Conference on Intelligent Information Systems: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science, 20-23 aug. 2013, Chisineu, Moldova: Proceedings IIS / Institute of Mathematics and Computer Science, 2013. – P. 108-113.

[*Li, 2012*] *Li S., Graça J. V., Taskar B.* Wiki-ly supervised part-of-speech tagging: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. – Jeju Island, Korea: Association for Computational Linguistics, 2012. – C. 1389-1398.

[*Luque, 2009*] *Luque B., Lacasa L., Ballesteros F., Luque J.* Horizontal visibility graphs: Exact results for random time series // Physical Review E,– P. 046103-1–046103-11 (2009).

[*Ma, 1999*] *Ma Xiaoyi, Liberman M.Y.* BITS: A Method for Bilingual Text Search over the Веб // http://papers.ldc.upenn.edu/ MTSVII1999/BITS.pdf

[*Manber, 1994*] *Manber U.* Finding similar files in a large file system // Proceedings of the 1994 USENIX Conference, January 1994. – P. 1-10.

[*McFate, 2011*] *McFate C., Forbus K.* NULEX: An Open-License Broad Coverage Lexicon // The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA – Short Papers. – The Association for Computer Linguistics, 2011. – C. 363-367.

[*Motter, 2002*] *Motter A.E., de Moura A.P.S., Lai Y.-C., Dasgupta P.* Topology of the conceptual network of language // Phys. Rev. E 65, 2002. – 065102(R).

[*Nunez, 2012*] *Nunez A.M., Lacasa L., Gomez J.P., Luque B.* Visibility algorithms: A short review // New Frontiers in Graph Theory, Y. G. Zhang, Ed. Intech Press, 2012. – Ch. 6. – P. 119 – 152.

[*Ortuño, 2003*] *Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M.* Keyword detection in natural languages and DNA // Europhys. Lett., 2002. – **57**. – P. 759 – 764.

[*Osipovs, 2009*] *Osipovs P., Borisov A.* Practice of Web Data Mining Methods Application // J. Riga Technical University **40**: 101-107 (2009) . – P. 11-18.

[*Resnik, 1998*] *Resnik P.* Parallel strands: a preliminary investigation

into mining the веб for bilingual text. In D. Farwell, L. Gerber and E. Hovy (eds) Machine Translation and the Information Soup, Springer, Berlin, 1998. – P. 72-82.

[*Resnik, 2003*] *Resnik, P. and Smith, N.A.* The Web as a parallel corpus. Comput. Linguist. 29, 3 (Sep. 2003). – P. 349-380.

[*Salton, 1975*] *Salton G, Wong A, Yang. C.A* Vector Space Model for Automatic Indexing // Communications of the ACM, 18(11): 613-620, 1975.

[*Salton, 1983*] *Salton G., McGill M.J.* Introduction to Modern Information Retrieval. – New York: McGraw-Hill, 1983. – 448 p.

[*Salton, 1988*] *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval. // Information Processing and Management, 1988. –**24**: 513-523,

[*Sagri, 2004*] *Sagri M-T., Tiscornia D., Bertagna F.* Jur-WordNet // in Proceedings of the Second Global WordNet Conference, pp. 305-310, Brno, Czech Republic, January 20-23, 2004.

[*Schlippe, 2011*] *Schlippe T., Ochs S., Schultz T.* Grapheme-to-phoneme model generation for Indo-European languages // In Proceedings of The 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, Japan, 25-30 March. – 2012. – С. 4801-4804.

[*Sigman, 2002*] *Sigman M., Cecchi G.A.* Global Properties of the Wordnet Lexicon // Proc. Natl. Acad. Sci. USA, 99, 1742 (2002).

[*Stone, 2003*] *Stone W.R.* Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! // IEEE Antennas and Propagation, Aug. 2003. – **45**. – № 4. – P. 47-49.

[*Strogatz, 2001*] *Strogatz S.H.* Exploring Complex Networks // Nature, 2001. – **410**. – P. 268-276.

[*Watts, 1998*] *Watts D.J., Strogatz S.H.* Collective dynamics of «smallworld» networks // Nature, 1998. – **393**. – P. 440–442.

[*Zipf, 1949*] *Zipf G.K.* Human Behavior and the Principle of Least Effort. – Cambridge, MA: Addison-Wesley Press, 1949. – 573 p.

Наукове видання

**Ланде Дмитро Володимирович**

# ЕЛЕМЕНТИ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ В ПРАВОВІЙ ІНФОРМАТИЦІ

Монографія

Підписано до друку 28.02.2014
Формат 60 х 84/8. Гарнітура Times. Офсетний друк.
Умов. др. арк. 8,5. Наклад 300 прим.
Віддруковано з оригінал-макета у видавництві ТОВ Інжиніринг
м. Київ, вул. Федорова, 9