Міністерство освіти і науки України

Національний університет «Києво-Могилянська академія»

Факультет інформатики

Кафедра інформатики

**Магістерська робота**

освітньо-кваліфікаційний рівень - магістр

на тему:

**«Розробка методу ітеративної побудови термінологій в колекціях наукових текстів»**

Виконав: студент 2-го року навчання

спеціальності

**8.05010101**«Інформаційні управляючі

системи та технології»

Решетньов Ігор Володимирович

Керівник: Глибовець А. М.,

кандидат фізико-математичних наук

Магістерська робота захищена

з оцінкою «\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_»

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2014 р.

Київ 2014

**Постановка завдання**

В рамках даної роботи буде досліджено алгоритм ітеративної побудови термінології у вигляді RDF графу за колекцією наукових текстів спільної тематики. Схема побудови зв’язків між термінами буде враховувати статистичні властивості термінів а також лексикографічні евристики пов’язані зі структурою україномовних документів. Принципова схема алгоритму складається з двох кроків: початкове виділення з тексту важливих термінів за статистичним методом, і послідуюче видобування з тексту пов’язаних термінів за рахунок застосування лексикографічних шаблонів. В рамках дослідження передбачається розробити структурну схему алгоритму, з подальшою формалізацією алгоритму, а також програмне забезпечення для моделювання роботи алгоритму. Кінцевою метою розробки є програма, що дозволить отримати RDF граф термінів колекції документів, наданої у форматі pdf. Результати даної роботи мають вирішити задачу складання термінології по колекції документів і можуть слугувати основою для створення інтелектуальних систем пошуку документів у семантичному вебі.

**Анотація**

Мета даної роботи – розробити метод і відповідне програмне забезпечення для вирішення задачі побудови термінологіі в колекціі текстів наукової тематики.

Дане дослідження є складовою частиною циклу робіт проведених на кафедрі з тематики, присвяченої побудові пошукової системи і репозіторію наукових праць.

З алгоритмічної точки зору результатом буде такий алгорітм, що на вході отримує колекцію документів у форматах pdf | doc, а на виході віддає файл що містить RDF граф з термінологією що зустрічається в даній колекції.

Як складовий модуль системи пошуку і каталогізації наукових праць, на модульному рівні дистрибутив програми буде надано у вигляді jar-архіву з відповідним API.

Для кінцевого користувача і для тестування ефективності побудови термінології буде розроблено веб-інтерфейс до компонентів програми, де в користувача будуть можливості завантажити в систему колекцію документів, запустити алгоритм побудови термінології, передивитись і скачати файл-результат у вигляді RDF, а також продивитись термінологію, її входження в документи колекції і пов'язані з результатом роботи алгоритму метадані термінів.

Функціональна схема роботи алгоритму побудови термінології буде включати в себе наступні етапи і технології:

* початкова індексація документів колекції за допомогою рішень з відкритим кодом (Apache Lucene);
* застосування алгоритмів розбору NLP (лінгвістичних аналізаторів) на реченнях і абзацах текстів, зберігання отриманих метаданих у спеціалізованих індексах;
* експериментальні дослідження щодо формулювання множини ефективних запитів до використаних пошукових систем і сховищ даних, що будуватимуть зв’язні висловлювання щодо термінів, опис і використання лінгвістичних евристик;
* розробка і застосування ітеративного алгоритму побудови термінології, що зможе покращувати результати при додаванні нових документів до колекції, а також після кожної ітерації;
* використання документарної бази MongoDB в якості сховища даних RDF;
* розробка RESTful API і веб-інтерфейсу користувача до досліджуваної системи;
* підбір тестових колекцій наукових документів і написання тестових пакетів для оцінки ефективності і точності алгоритму;
* експериментальні дослідження алгоритму з варіацією параметрів;
* встановлення залежностей між параметрами, публікація найбільш вдалих налаштувань як окремих методів API;
* розгортання готової системи на сервері для публічного доступу.

Зміст

Вступ 5

Розділ I. Огляд існуючих підходів 6

1.1. Роль тезауруса в інформаційному пошуку 6

1.2. Тезаурус як форма подання зв’язаної термінології 7

1.3. Тенденції опису ресурсів в семантичному вебі 8

1.5. RDF як формат публікації тезаурусів 9

1.6. Огляд стандарту JSON-LD як конкретної специфікації RDF. 9

1.7. Дослідження існуючих автоматизованих методів побудови тезаурусів 10

1.7.1 Статистичні методи 11

1.7.2 Лексикографічні методи 13

Розділ II. Розробка методу 25

2.1 Структурна схема алгоритму 25

2.2 Математична модель і формалізація методу 36

Розділ III. Практична частина 36

Реалізація методу побудови термінології 36

Вибір готових рішень для реалізації кроків алгоритму. 39

Архітектура системи обробки документів для побудови термінології 39

Специфікація прикладного програмного інтерфейсу доступу до системи. 39

Розділ IV. Обговорення результатів 39

Схема тестування та оцінка результатів 39

Висновок 42

Використані джерела 42

Додатки 43

# Вступ

# Розділ I. Огляд існуючих підходів

## 1.1. Роль тезауруса в інформаційному пошуку

Зазвичай, інформаційна потреба користувача пошукової системи не відповідає термінам, що зустрічаються в документах, або користувач неправильно розуміє в конкретиний час в конкретному місці термінологію області знань, до котрої він здійснює пошук. За таких умов, одним із методів покращення пошукової видачі є використання тезаурусів термінів предметних областей. Тезауруси є таблицями термінів, що поєднують пов’язані між собою терміни, зазвичай вказуючи тип зв’язку (NT, BT, USE, RT). Інформаційні системи можуть використовувати тезауруси на етапі індексації документів, для того щоб правильніше класифікувати документи по категоріях, або під час пошуку, розширюючи пошуковий запит користувача пов’язаними термінами.

Тут варто згадати експерименти Фурне і Ландауера[[1]](#footnote-1), що показали, як більшість користувачів тільки у 20% випадків вибирали однакові терміни для опису тих самих речей, що свідчить про необхідність для пошукових систем враховувати синонімію термінів, що в свою чергу задається саме в тезаурусах. Більш того, порівняльні заміри релевантності пошукової видачі навпрост і з доданою інформацією з тезаурусу, за думкою Чена[[2]](#footnote-2), показують можливість її покращення. Таким чином, можна вважати обгрунтованим використання тезаурусів в пошукових системах задля покращення їх якості.

Головна проблема складання тезаурусів полягає у тому, що для більшості комерційних баз даних, що поширють наукову інформацію, вони складаються експертами з областей знань, а також фахівцями зі складання тезаурусів, тобто власноруч. Якщо ж розглядати окремі найновіші області знань, де відповідна термінологія тільки формується і випускається велика кількість нових публікацій, то в данному випадку термінологічні словники дуже швидко застарівають, і їх треба оновлючати частіше, тобто знову залучати експертів. На противагу такому підходу, існують методи автоматизованої побудови тезаурусів, що в якості корпусу беруть усі найновіші публікації з теми, і будують на їх основі зв’язки між термінами. Як наслідок, маючи таку систему оновлювати термінологічні зв’язки стає значно простіше і дешевше. В публікації Моніки Лассі[[3]](#footnote-3) було розглянуто основні методи автоматизованої побудови тезаурусів, такі як метод спільного вжитку, метод концептуального простору і метод баєсівських мереж. Дані методи мають як різну ефективність і часову оцінку складності, так і приниципи - статистичний і лексикографічний. Розробці і застосуванню нового методу, що б використовував комбінацію ідей, що зустрічаються в цих підходах, і присвячено дану магістерську тезу.

## 1.2. Тезаурус як форма подання зв’язаної термінології

Тезаурусом називають керований словник, що містить семантичні зв’язки між термінами, і використовується з метою поліпшення процесу пошуку пов’язаних термінів. Означень тезауруса досить багато, починаючи від доволі короткого і загального, наведеного в роботі Шутца і Педерсона[[4]](#footnote-4): *“Ми визначаємо тезаурус як просте відношення між словами і іншими близько пов’язаними словами”,* уточнюючи далі що тезаурус має бути достатньо специфічним щоб надавати користувачу синоніми шуканих слів. На противагу, в роботі Міллера[[5]](#footnote-5) подано більш точне означення тезауруса, що є *“лексико-семантичною моделлю концептуальної реальності або її представника, що виражена у формі системи термінів і їх зв’язків, пропонує доступ за допомогою багатьох аспектів і використовується в якості системи обробки і пошуку всередені модулю інформаційної пошукової системи.”* Цікаво, що автор наголошує на принциповій нерозривності теоретичної моделі тезауруса і практичного використання програмних модулів із такою функціональністю.

Функціонально, зв’язки між термінами в тезаурусах прийнято позначати так: ширший термін (BT), вужчий термін (NT), переважний термін (USE), пов’язаний термін (RT) – для позначення зв’язків, для котрих неможливо точно встановити напрямок зв’язку.

Конвенційний спосіб складання тезаурусів – за допомогою залучення експертів з певних галузей знань – досить коштовний. Одним із способів зменшення вартості складання тезауруса є використання вже існуючих лексикографічних баз даних, таких як WordNet, що має серйозну проблему при застосуванні такого корпусу загальної лексики до доменно-специфічних областей знань[[6]](#footnote-6). Інший підхід – автоматизовані методи створення тезаурусів, що можуть підготувати досить точну основу для подальшого уточнення експертами.

## 1.3. Тенденції опису ресурсів в семантичному вебі

Останні тенденції в галузі опису ресурсів семантичного вебу, зокрема технології Wikipedia, Google Knowledge Graph, навіть WordNet, певним чином актуалізують питання автоматизованих методів побудови зв’язків між термінами, через те що всі технології пропонують мануальний людський спосіб розмічення концептів, за допомогою різних форматів і утиліт.

## 1.5. RDF як формат публікації тезаурусів

Формат RDF[[7]](#footnote-7) є одним із найпоширеніших способів подання даних і метаданих для технологій семантичного вебу. В основі даного формату, не вдаючись до подробиць абстрактного синтаксису і конкретних специфікацій, що ретельно розроблені у відповідних міжнародних стандартах, лежить ідея подання інформації у вигляді триплетів “суб’єкт”-“предикат”-“об’єкт”. Така загальна і проста на перший погляд модель може вдало задовільнити потреби тезауруса для опису його вмісту. Однак поряд із задоволенням потреб на рівні узгодженості моделей постає інша надзвичайно важлива властивість міжнародного формату – широка всесвітня підтримка на рівні реалізації прикладних систем. Як зазначалося раніше, роль тезауруса визначається не лише точністю і об’ємом представлених термінологічних зв’язків, але й практичною застосованістю програмного модулю, простотою доступу і придатністю до машинної обробки. Саме завдяки можливості публікації оброблених за допомогою програмної частини тезауруса даних безпосередньо в мережі Інтернет у міжнародно-прийнятному форматі, нами було запропоновано формат RDF і підлеглу систему веб-сервісів з програмним інтерфейсом в якості кінцевого формату доступу до тезаурусу.

## 

## 1.6. Огляд стандарту JSON-LD як конкретної специфікації RDF.

Серед конкретних специфікацій RDF, формат JSON-LD, поданий у стандарті ISO25964[[8]](#footnote-8), на нашу думку найкращім чином відповідає поставленій задачі публікації ресурсів тезауруса у вигляді веб-сервісу. До визначених нами переваг формату належить:

* реалізація моделі *пов’язаних даних*, що є необхідною властивістю нами означеного тезауруса як системи доступу до даних про зв’язки термінів, і певним чином наслідує ідеї концепції RESTful HATEOAS[[9]](#footnote-9) веб-сервісів, що на нашу думку є одним з найбільш прогресивних на сьогодні індустріальних підходів до підтримки машинної обробки веб-ресурсів;
* підтримка елемента *контексту* для опису доменних обмежень і типів полів тіла документа, що дозволяє значно скоротити запис без втрати в точності визначення обмежень;
* формат JSON є більш прийнятним для читання людиною з-поміж інших альтернатив;
* серіалізацію даних у формат JSON-LD за допомогою веб-сервісу досить легко реалізувати завдяки широкій підтриці базового формату JSON в прикладних бібліотеках з відкритим кодом.

До базових концепцій формату відносяться[[10]](#footnote-10):

1. IRI – інтернаціональні ідентифікатори ресурсів,
2. контекст, що слугує головним чином для задання скорочень до IRI,
3. Ідентифікатори вузлів і типізовані значення.

Запропонованих базових елементів формату має бути достатньо щоб подати дані тезаурусу у мінімально задовільному до стандарту вигляді.

## 1.7. Дослідження існуючих автоматизованих методів побудови тезаурусів

Методи автоматизованого складання тезаурусів можна поділити на два принципових класи: *статистичні*, що інтенсивно використовують частотні і позиційні характеристики термінів в документах в якості основи для різноманітних моделей виявлення зв’язків між термінами, і *лексикографічні,* що використовують відомості зі сфери обробки людської мови для здійснення синтаксичного, морфологічного і інших видів аналізу тексту з метою знаходження семантичних зв’язків на основі отриманої виключно з тексту інформації. При цьому в лексикографічних методах зазвичай набувають широкого вжитку зібрані експертами корпуси мов, що містять правила спільного вжитку слів, словоформи і синонімічні ряди, а також доступні програмні пакети для проведення початкового аналізу вільного тексту, зокрема утіліти тегування за частинами мови, лематизатори, стемери і поготім. В свою чергу, для статистичних методів такою основою є утіліти індексування і ранжування термінів. Подане розбиття методів на два класи по великому рахунку умоглядне і не є взаємовиключним, тому в більшості досліджень зустрічається спільне використання технік з різних підходів, з відданням переваги одному з класів методів.

### 1.7.1 Статистичні методи

Основою для багатьох статистичних методів пошуку залежностей між термінами є створення індексу термінів, що описують зміст документів найкращім чином. В свою чергу, побудова індексу зазвичай потребує ранжування термінів за ступенем важливості, і найбільш вживаними техніками зважування, що беруть свій початок в алгоритмах пошукових систем, є використання частоти терміну (**TF**), зворотньої документарної частоти (**IDF**), а також їх комбінацій. Розглянемо даний підхід детальніше.

**Ранжування термінів за метрикою TFIDF**

Метрика зважування **TFIDF**, у багатьох своїх модифікаціях грала значну роль основної техніки ранжування в пошокувих системах на певному етапі їх розвитку, і велика кількість досліджень з покращення формул[[11]](#footnote-11) і проведених експериментів зробили дану методику досить ефективною.

Розглянемо специфічний зміст даних метрик. Відтак, використовуючи виключно частоту терміна в якості міри зважування, оптимальними термінами в індексі стають ті, що мають середню частоту в документі. Найчастіші слова в документі вважаються неприйнятними для відрізення одного документу від інших в колекції, а найменш вживані слова – вважаються недостатньо впливовими для якісного опису змісту документа.

Натомість, метрика інвертованої документарної частоти бере до уваги появу слів у всій колекції, а не в окремо взятому документі. Таким чином, слова що зустрічються в усіх або принаймні у багатьох документах колекції вважаються неприйнятними для відрізнення документів, і навпаки, ті, що зустрічаються в декількох документах, вважаються прийнятними для опису цих документів.

Комбінація наведених технік призводить до спільного використання сильних сторін обох методів: великої ваги набувають терміни, що зустрічаються з середньою частотою в документах і водночас з малою частотою в колекції.

Базові формули для обчислення таких метрик подані нижче.

\mathrm{TF} = \frac{n_i}{\sum_k n_k}

*де ni відповідає кількості входжень шуканого слова в документ, а в знаменнику – кількість всіх слів у документі.*

\mathrm{IDF} =  \log \frac{|D|}{|(d_{i}\supset t_{i})|}

*де |D| - кількість документів в колекції, в знаменнику – кількість документів, що містить шуканий термін.*

**Метод спільного вжитку**

Даний метод є одним із підходів в інформаційному пошуку до формування багатослівних термінів, що є статистичною альтернатовою лексикографічного методу тегування за частинами мови. Основними елементами для обчислення в методі є частоти входження терміну у певні різні за розміром контекстні рамки, такі як цілий документ, глави документу, параграфи й інші менш дрібні елементи. При цьому чим ближче слова зустрічаються в контексті обраної рамки, тим більшою призначається міра спільного вжитку. В дослідженні Шутца і Педерсона[[12]](#footnote-12) в основу для реалізації даного підходу покладено створення матриці *С*, елементи якої *Cij* показують кількість разів спільного вжитку між словами *i* та *j* в контекстній рамці розміру *k*. З даною реалізацією пов’язані очевидні проблеми ефективності через велику розмірність і тривалий час побудови такої матриці. Стосовно якості знайдених термінологічних зв’язків за даним методом теж виникають сумінів, як у статті Чена[[13]](#footnote-13), що стверджує про неефективність складеного за даним методом тезаурусу при застосуванні до задач пошуку, і пропонує власний метод, що розглянуто далі.

**Метод концептуального простору**

В цьому підході вводяться поняття *коцептуального простору* як мережі термінів і зважених асоціацій між ними, що здатні відтворити концепти і зв’язки між ними у відповідному *інформаційному просторі,* представленому у вигляді колекції документів в базі даних.

Для даного підходу виділяють чотири кроки:

1. збір документів і списку об’єктів інтересу,
2. фільтрація об’єктів і автоматизована індексація,
3. аналіз методом спільного вжитку,
4. асоціативний пошук.

Етап збору документів у методі вважається досить важливим, адже базова колекція документів має бути достатньо повною для репрезентативного опису усіх доменно-специфічних концептів окремо взятої області. На наступному кроці фільтрування об’єктів інтересу визначалася множина термінів, що відповідала б ключовим поняттям предметної колекції. Через те що не всі ключові поняття зазвичай включені в опис предметної області, застосовується техніка автоматизованого індексування колекції для знаходження таких термінів з тексту. В рамках індексування, в методі були задіяні такі кроки:

* звернення до словників для ідентифікації окремих слів у тексті,
* послідуюче фільтрування списку слів за допомогою вилучення стоп-слів, що не підходять в якості кандидатів на включення до індексу,
* стемінг слів, що залишились після фільтрації,
* формування термінологічних фраз шляхом поєднання суміжних слів у документах і включення їх до індексу.

На третьому етапі, аналіз методом спільного вжитку в описаному методі починався з обчислення частот термінів і інвертованих документарних частот, з наданням переваг у зважуванні термінам в заголовках документів, ключовим словам відомим з пердметної області заздалегіть і побудованим багатослівним термінам. Аналіз здійснювався на базі розробленої Ченом і Лінчем[[14]](#footnote-14) асиметричної функції кластеризації, що за доведенням авторів надає кращі результати ніж загальновживана косинусна міра схожості.

Модель асоціативного пошуку, включена в даний метод, наближена до ментальних способів представлення інформаційних потреб користувачів пошукової системи у вигляді мережі термінів і зв’язків між ними, що зазвичай є нечіткими. Авторами методу було розроблено інтерфейс навігації по побудованому тезаурусу, в основу котрого було покладено нейронну мережу Хопфілда, що в результаті своєї роботи по досягненню збіжності уточнювала ваги між термінами, таким чином наближаючи стан зв’язків між термінами до подібного за силою зв’язків стану їх представлення в людській пам’яті[[15]](#footnote-15).

### 1.7.2 Лексикографічні методи

Лексикографічні методи пошуку зв’язків між термінами базуються на приципі прямого зазначення зв’язку між словами за допомогою мовних засобів, причому характер зв’язку можно визначити, виходячи з синтаксичної та лексичної будови висловлювань. Перш ніж розглянути основні позиції даних методів, а також їх безпосередній зв’язок з утворенням термінології, необхідно подати декілька означень.

Так, **терміном** (лат. terminus – «рубіж, межа»), називають слово або словосполучення, що позначає поняття певної галузі науки, техніки тощо. Основними ознаками терміну є : системність, наявність дефініції, тенденція до однозначності в межах свого термінологічного поля, тобто термінології певної галузі.[[16]](#footnote-16) Зазначені ознаки наявні тільки в конкретних терміносистемах, але за їх межами термін втрачає вказані характеристики і стає загальновживаним словом. Слід також зазначити, що основна функція терміна – номінативна та що терміни у більшості випадках позбавлені емоційного забарвлення.

Термін – одиниця лексичного складу мови, завданням котрої є називати реалії даної спеціальності та бути підставою для наукового висловлювання. Без цих найменувань неможливо здійснювати наукову роботу. У науковому тексті термін виступає як стабільний елемент, який приєднують до одиниць загальновживаного лексичного складу мови. Відзначаються дві основні сфери функціонування термінів: сфера внутрішньо-наукова та сфера зовнішніх зв’язків мови.

«Кожен термін є результатом розумової діяльності людини, зокрема діяльності, спрямованої на узагальнення і абстрагування».3

Слід мати на увазі те, що в кожній із галузей конкретний термін має інше семантичне значення, котре може збігатися із значенням у загальній мові або може відрізнятися від нього. Варто також зазначити, що на зміну значення терміна впливають мовні та і позамовні фактори.

Розглядаючи типологію термінів, для наших задач є ключовою типологія за формою, а саме:

**однослівні**: *кластеризація, означення, алгоритм*

**багатослівні**, або **термінологічні словосполучення** (найчастіше двослівні терміни): *часова складність, соціологічне дослідження.*

Для задач пошуку термінів також важливо надати способи утворення термінів. Відомі такі способи утворення термінів[[17]](#footnote-17):

1. утворюванням похідних слів: додавання префіксів та суфіксів до існуючих слів;
2. складанням слів у термінологічні словосполучення
3. об’єднуванням слів у композити – складені слова
4. використанням скорочених слів або скорочень
5. перенесенням значення
6. запозиченням слів з інших мов

В розрізі лексикографічних методів пошуку зв’язків в термінології, найцікавішим для нас є спосіб складання термінологічних словосполучень. Цей спосіб належить до найпродуктивніших способів словотвору. Два або більше слів зливаються в одне словосполучення. Окремі слова втрачають у словосполученні свою самостійність та їх питоме значення зникає. У новому сполученні його члени набувають нового змісту, який є в даній формі однозначний. термінологічні словосполучення, як правило, поділяють на номінативні та дієслівні. Кожне термінологічне словосполучення виступає у мові як певна форма, тобто неможливо його члени пересувати або заміняти синонімами.

З поняття терміну логічно випливає означення термінології:

**Термінологія**  (лат. terminus – «рубіж, межа» і грец. λόγος – «слово, вчення») :

1. Сукупність термінів, що обслуговують певну сферу знань, пов’язаних із системою понять : мистецтво, техніка, виробництво тощо. На думку вчених, слово ***термін*** уперше з’явилося у Німеччині в 1876 р.
2. Розділ лексикології, що займається загально-теоретичними питаннями терміна.[[18]](#footnote-18)

На Україні, крім поняття ***термінологія***, досі часто користуються терміном ***термінознавство***. Дефініція його така – «наука, яка займається загальновживаними питаннями терміна, термінології, номенклатури».[[19]](#footnote-19)

На даному етапі ми можемо побачити прямий зв’язок між тезаурусом, термінологією і лексикографією.

**Лексикографія** (від [грец.](http://uk.wikipedia.org/wiki/%D0%93%D1%80%D0%B5%D1%86%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0) *Λεξικογραφία*; від [грец.](http://uk.wikipedia.org/wiki/%D0%93%D1%80%D0%B5%D1%86%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0) *Λεξικόν* - «словник» і [грец.](http://uk.wikipedia.org/wiki/%D0%93%D1%80%D0%B5%D1%86%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0) *Γράφω* - «пишу») – словникарство, розділ мовознавства, що займається створенням словників та їх теоретичних засад.[[20]](#footnote-20) У рамках лексикографії слід розрізняти лексикографічну теорію, що включає в себе основні проблеми лексикографії, з якими зустрічається лексикограф у своїй роботі та власне лексикографічну практику, тобто практичні питання щодо збору лексикографічного матеріалу, його розробку та з рештою укладання словникового тексту. Лексикографія пов’язана з іншою мовознавчою дисципліною – лексикологією. Головним завданням лексикографії є пояснення незрозумілих слів, яке здійснювалося початково у вигляді глос, тобто тлумачення написів на полях і в тексті рукописних книг. Лексикографія займається словникарським кодифікуванням лексики будь-якої мови.

**Основним об’єктом** лексикографіі є слово. Будь-який словник містить певну кількість слів конкретної досліджуваної мови або мов, тобто необхідно з точки зору лексикографії розглядати слово і мову.

Одним із завдань лексикографії є **укладання словників,** адже «словник є корисним помічником для спеціальної термінологічної роботи, а також для стилізації наукового тексту, тому що в них наочно зібрані спеціальні назви та звороти».[[21]](#footnote-21) словники відображають культуру й мову народу та є джерелом правил її написання, вимови, наголошування тощо.

Важливо, що серед типів словників, складанню котрих присвячена наука і ремесло лексикографії, почесне місце займають саме термінологічні словники.

Отже, зв’язок проблематики складання тезаурусів і викростання лексикографічних методів є вкрай необхідною умовою пошуку вдалих рішень.

Адже в працях Стралковського[[22]](#footnote-22) і Грефенстета[[23]](#footnote-23) зазаначалася одна з найпоширеніших проблем всіх статистичних методів – проблематика індексування фразових термнів, або, в нашому випадку, термінологічних словосполучень. Зокрема, автор зазначав необхідність побудови якісних рішень на основі лінгвістичних особливостей текстів, зокрема з використанням техніки тегування за частинами мови, як одну із головних задач покращення статистичних методів в інформаційному пошуку.

Для того щоб таку методику спробувати застосувати для пошуку термінів, необхідно розглянути можливі способи утворення термінів.

Розглянемо способи утворення термінів шляхом складання словосполучень. Згідно з роботою Кріслової[[24]](#footnote-24), що стосується лінгвістичної україномовної термінології, можна виділити наступні схеми утворення термінологічних словосполучень за частинами мови окремих слів. Тут необхідно зробити припущення, що нижчеподана схема буде певним чином відповідати і способам творення в термінологічних системах інших галузей знань, що можливо подалі необіхідно буде врахувати під час модифікації схеми.

Двослівні словосполучення:

|  |
| --- |
| П р и к м е т н и к + і м е н н и к |

(*когнітивна лінгвістика, односкладне речення, мотивована основа)*

І м е н н и к + і м е н н и к

(*контекстуальність комунікації, методи лінгвістики, оператор порівняння)*

Тричленні термінологічні словосполучення можуть бути побудовані таким способом:

І м е н н и к + п р и к м е т н и к + і м е н н и к

(*розповідь від першої особи, дієслово доконаного виду, актуалізація мовних засобів)*

П р и к м е т н и к + п р и к м е т н и к + і м е н н и к

(*додатковий орієнтаційний момент, дієслівне двоскладне речення, актуальний теперішній час*)

П р и к м е т н и к + і м е н н и к + і м е н н и к

Крім двочленних та тричленних словосполучень, досить часто в термінології виникають чотиричленні терміни, структура котрих може виглядати:

П р и к м е т н и к + і м е н н и к + п р и к м е т н и к + і м е н н и к

(*словниковий склад літературної мови*, *стилістичне транспортування мовних засобів, позачасове вживання теперішнього часу, односкладне речення з допоміжним дієсловом)*

П р и к м е т н и к + і м е н н и к + і м е н н и к + і м е н н и к

(*аналітичне творення ступенів порівняння*)

І м е н н и к + п р и й м е н н и к + і м е н н и к + і м е н н и к

(*графіка із застосуванням диграфів, прикметник із значенням можливості* )

Важливість саме іменникових словосполучень, що відповідали б термінам, підтверджується і в праці Лендау, де зазначається, що серед науково-технічної літератури і в термінологічних словниках більшість термінів виражені іменниками[[25]](#footnote-25).

Подальші схожі роздуми побачимо в методі Г. Грефенстета.

**Метод автоматичної генерації тезаурусу з необробленого тексту з використанням техніки браку знань**

Даний метод був вперше описаний і застосований Грегорі Грефенстетом в одноіменній статті 1993 року[[26]](#footnote-26). Покладені в основу методу техніки видобування термінів і зв’язків між ними для побудови початкової версії тезаурусів з будь-якої колекції специфічних за областю текстів названі техніками з браком знань через те, що для побудови такого тезаурусу наявність будь-яких знань про область знань є зовсім не необхідною, і така побудова виводиться першочергово з наявних текстів. Дослідники застосували даний метод до більш ніж 20 різних корпусів текстів, розміром від 1 до 6 мегабайт, і отримали позитивні результати. Розглянемо ж лексикографічні закономірності, покладені в основу даних технік більш детально.

Відкритий текст колекції документів в першу чергу розбивався на токени за регулярною граматикою. Потім кожен токен аналізувався, і до нього з лексикону мови добиралися теги частин мови і їх основних параметрів, що мав у собі токен. Далі тегований текст позбавляли неоднозначності стохастичним методом, так щоб кожен токен мав тільки одну форму частини мови і один тег. Після цього готовий тект розбирався, і збиралися зазначені у форматі тезаурусу залежності між словами.

Для іменників, збирались як атрибути всі модифікуючі прикметники і дієслова, для котрих дані іменики були суб’єктами або об’єктами, а також всі інші іменники, що відносились до них як додатки або входили в умовні звороти. Схожість між іменниками обчислювалась за допомогою зваженої метрики Жаккарда[[27]](#footnote-27). В результаті порівняннь для кожного іменника зберігався сортований список найбільш схожих на даний термін слів з тексту. Для скорочення такого списку застосовувався фільтр, що залишав тількі ті слова, що є взаємно найбільш схожими, тобто щоб кожен з термінів містився в рамках 10 найбільш схожих слів в обох списках. Цікаво, що з результуючому списку схожих слів автори використали порівняльну частоту входження термінів у якості основи для побудови ієрархічних зв’язків в тезаурусі, зазначаючи що даний обраний метод є надзвичайно слабким і неточним.

Для виокремлення загальновживаних фраз, або термінологічних словосполучень, автори використали тільки модель з двослівних словосполучень, утворених іменниками, спираючись на проблематику пошуку більш довгих підпослідовностей у тексті. Між іншим, серед усіх таких фраз в результуючий тезаурус потрапили тільки 10 найбільш вживаних, кожна з котрих зустрічалася в тексті немодифікованою принаймні 3 рази. Щоб підібрати список синонімічних фраз, для зважування був обраний інший метод, що брав до уваги ширший за синтаксичний контекст, що виходив за рамки окремих словосполучень. Як показали експерименти, для фразових словосполучень, частота вживання котрих значно менша, адекватну оцінку надавав тільки аналіз ширшого контексту, тобто цілих речень.

Окрім окреслених груп слів, в даному методі також відбувався пошук дієслів і сімейств слів. У підсумку, дослідникам вдалося побудувати 20 різних тезаурусів на різних тематичних колекціях текстів, для чого було застосовано 3 різних техніки: використання локальної лексико-синтаксичної інформації для побудови зв’язків близькості між термінами, використання більш ширшого контексту речення для пошуку двослівних фраз, і насамкінець присутність у всьому документі для виявлення сімейств термінів. В реалізації лінгвістичних підходів не було використано складних лінгвістичних закономірностей, таких як семантичні маркери на іменниках, або детальний морфологічний аналіз, і все ж таки отримані результати говорять на користь ефективності застосованих технік.

**Гіпонімія як основа для пошуку ієрархічних зв’язків**

В наступному лексикографічному методі ключову роль грає поняття гіпонімії, котре ми розглянемо тут детальніше.

**Гіперо-гіпонімія** – це родо-видові відношення в лексико-семантичній системі[[28]](#footnote-28). Наприклад: *береза, дуб, клен, явір, сосна - дерево; троянда, рожа, настурція, тюльпан, нарцис - квітка.* Родові слова називають **гіперонімами**, а видові — **гіпонімами**.

Гіперо-гіпонімія близька до синонімії. ЇЇ навіть називають, але на відміну від синонімії, яка допускає двосторонню заміну в тексті (першого синоніма на другий і навпаки), в гіперо-гіпонімії можлива тільки одностороння заміна — заміна гіпоніма на гіперонім.

Зрозуміло, що явище гіпонімії є безпосереднім вказівником в тексті на зв’язок типу “загальне-конкретне” між термінами, котрий є невід’ємною складовою тезаурусів. Розвиваючи дану ідею, розглянемо статтю М. Хеарста[[29]](#footnote-29), котрий описав автоматизований лексикографічний метод видобування гіпонімів з тексту.

**Метод збирання гіпонімів з великих текстових колекцій**

Дві головні цілі, вказані даним підходом – елімінація потреби у попередньо складених знаннях про предметну область і можливість застосування методу на широкому колі текстових колекцій. В роботі було виокремлено множину лексико-синтаксичних шаблонів, що є широко розповсюдженими серед колекцій текстів різних жанрів, що безпосередньо вказують на шукані лексичні залежності, і котрі легко розпізнати в тексті як програмними засобами, так і власноруч.

В центрі методу полягає ідея про вміст великої кількості корисної інформації про предметну область в самому тексті, що може бути виявлена як людиною, так і алгоритмом, не вдаючись до занадто конкретних деталей окреслених явищ і речей, тобто не потребуючи від системи здійснення глибокого лексикографічного чи семантичного аналізу. Наприклад, з речення: *“Мови програмування, такі як Java, C++ і Smalltalk, відносяться до класу імперативних об’єктно-орієнтованих мов”, -* людина чи алгоритм, що не має ніякого уявлення про окремі мови програмування, з тексту знає про відношення між поняттями, узагальнююче поняття, а також окремих представників. Подане в формальній лекико-синтаксичній нотації авторів статті, попереднє висловлючання набуде вигляду:

**NP0, такі як {NP1,NP2 ... , (або | і)} NPn**

де **NPi** – відповідає певному термінологічному іменниковому словосполученню, що потребує, щоб для всіх **NPi, 1 < i < n**, виконувалось твердження

**hyponym(NPi, NP0)**

З речення-прикладу можна зробити висновок, що виконується

**hyponym(“Java”,”мова програмування”), hyponym(“C++”, “мова програмування”), hyponym(“Smalltalk”, “мова програмування”)**

Дану техніку пошуку таксономічних зв’язків автори пов’язують з попередніми доследженнями, зокрема з роботою Aшаві[[30]](#footnote-30), що використав ієрархію шаблонів для інтерпретації означень, що складалися переважно з індикаторів частин мови і символів-масок. Основним недоліком такого підходу автори вважають проблему підбору такої множини шаблонів, що б з однаковою точністю вказували б на спрямованість зв’язку в текстах різних стилів. З основних прикладних напрямків, де застосування пошуку гіпонімії є виграшним, виділяють сферу приросту і перевірки корпусів мов, складених людиною (наприклад, WordNet), сфери пошуку приблизних значень незнайомих іменників в термінологічних словосполученнях і семантичної схожості між словосполученнями, що стоять в одному ряду підпорядкування узагальнюючому гіпероніму.

Із знайдених авторами методу експериментальним чином надійних шаблонів, в статті було запропоновано шість окремих формальних нотацій з прикладами. Не обмежуючись в широті застосування даних шаблонів, використавши граматичні правила розробленої нотації і зробивши переклад ключових слів на українську мову, ми подамо зазначені шість шаблонів у вигляді трьох формальних правил:

1. **такі NP як {NP,}\* {(або | і)} NP**

... роботи таких дослідників як Хеарст, Грефенстет і Ашаві.

*hyponym(“Хеарст”,”дослідник”), hyponym(“Грефенстет”,” дослідник”), hyponym(“Ашаві”,” дослідник”)*

1. **NP {, NP} \* {,} {(або | і) інші NP**

Набряки, гіматоми, рани і інші травми ...

*hyponym(“набряк”,”травма”), hyponym(“гіматома”,”травма”), hyponym(“рана”,”травма”)*

1. **NP {,} (зокрема | особливо) {NP,}\* {або | і} NP**

Всі демократичні країни, зокрема Канада і Великобританія ...

*hyponym(“Канада”,”демократична країна”), hyponym(“Великобританія”,”демократична країна”)*

Серед спостережених проблем реалізації даного методу, дослідники зазначають проблему вилучення інфінітивної форми термінологічного словосполучення з тексту. Адже навіть при використанні лематизації слів, що входять в склад словосполучення, необхідно знову розв’язати задачу узгодження слів в словосполученні, що може потребувати введення великої кількості додаткових правил і винятків. Якщо ж дані поліпшення вилучених з тексту словосполучень не проводити, якість складених у словник предикатів відношень значно втрачається. Ті ж самі, і навіть більш ширші проблеми ми зустрінемо підчас застосування таких шаблонів для україномовних текстів.

Серед описаних проблем узгодження слів і приведення до нормальної форми словосполучень, зустрічалися наступні:

* іменники всередені словосполучень, що подані в множині;
* числові модифікатори прикметників, такі як “деякі”, “інші”, можуть бути усунуті без втрати смислу;
* порівняльні слова, так як “важливий”, “менший” – найкращі кандидати на усунення з фрази.

Щодо доречності усування модифікаторів іменників у словосполученні, автори приходять висновку про залежність даного фактору від широти вживаності тезаурусу, що будується. Якщо в статті основні експеременти проводились на тезаурусі загального вжитку, де було дійдено висновку про найчастіше використання однослівних іменників в якості елементів таксономічних відношень, то для більш спеціалізованих колекцій, зокрема наукових публікацій, більш характерними будуть багатослівні термінологічні словосполучення.

Експериментальні запуски реалізованого методу на енциклопедичних колекціях різної довжини показували взагалом прийнятні результати щодо кількості знайдених співпадінь з шаблонами, і отриманих зв’язків. Тестування точності отриманих зв’язків проводилось за рахунок порівняння з уже розміченими людиною парами значень в корпусі англійської мови WordNet[[31]](#footnote-31).

Підсумовуючи основні здобутки даного методу, можна вказати на порівняльну дешевизну застосування даного методу для автоматизованого збору семантичних лексиічних зв’язків в вільнотекстових документах. Даний метод позіціонується як альтернатива статистичним методам, і має перед ним перевагу в точності роботи на рідких зв’язках між термінами, що зустрічаються в тексті поодиноко, і не можуть буті оброблені вдало статистичними методами. Представлені в дослідженні шаблони і стратегії відсікання модифікаторів іменників не претендують на повноту, і залишають певну свободу для майбутніх доповнень.

# Розділ II. Розробка методу

Розробка ітеративного методу побудови термінології за допомогою комбінації лексикографічних і статистичних методів

## 2.1 Структурна схема алгоритму

**Індексування**

Процес побудови термінології на основі колекції текстів можна розкласти на два принципових кроки:

а) по-перше, видобування зі всіх слів, що зустрічаються в текстах документів, таких, що відповідають термінам в області знань відповідних документів;

б) по-друге – встановлення на множині даних термінів відношень, що використовуються в тезаурусах, зокрема симетричне відношення пов’язаних термінів (RT – *Related Term*), і асиметричні відношення типу “частина-ціле” (BT – *Broader Term*, і NT – *Narrower Term* відповідно).

Завдання виокремлення термінів з множини усіх слів документа смисловим чином є подібним до звичайної операції індексування текстів пошуковими системами. Під час такого індексування здійснюється побудова таблиці відповідності ідентифікатора індексованого документа з певним списком ключових слів, що найкращим чином розкривають зміст даного документа, є його ключовими словами. Індексація може проводитись як експертами власноруч, так і автоматизовано за допомогою різних алгоритмів. Найбільш вживаний алгоритм індексування базується на зважуванні слів на основі їх частоти присутності в документах.

Частота терміну (TF), інвертована документарна частота (IDF), а також їх комбінація **TFIDF** є найбільш широко вживаними техніками завжування.

*(докладніше про зважування тут:)*

Отримавши зважену послідовність слів за такою методикою, відсортовану по спаданню ваги, на початку послідовності будемо мати слова, що найкращим чином характеризують зміст документів, а отже є кандидатами в терміни.

Слід зауважити, що ефективність такого зважування, особливо під час обрахунку метрики IDF, цілком залежить від розміру і різноманіття документів колекції. Якщо ж колекція документів, на основі котрої відбувається пошук термінів, є досить невеликою, даний метод буде надзвичайно чутливим до стилю тексту і окремих вживаних слів певної предметної області, і, таким чином, не зможе виявити кандидатів в терміни з достатньою точністю.

На даному етапі, слід повернутися до поставленої задачі, де ми на першому кроці намагаємось обмежити кількість слів, по потраплять у список термінології, і застосувати загальновживаний метод зважування на основі **TFIDF** для наших потреб.

Отже, для обмеження даного відсортованого списку слів можна ввести обмежуючий оператор, що надав би можливість визначити граничний елемент списку, після котрого починається перелік загальновживаних слів як наукового стилю текстів, так і текстів загальної тематики.

Дана функція може мати наступні запропоновані варіації для нашого методу:

1. “Проходять усі” – хвіст послідовності не відкидається, всі слова документів інтерпретуються як терміни, що беруть участь у пошуку зв’язків. Очікувані недоліки застосування даного оператора – значне збільшення простору пошуку і, відповідно, часової складності обрахунків, а також наявність в результуючому тезаурусі великої кількості зайвих, не маючих практичної цінності зв’язків між словами текстів. Однак, результат застосування може бути використаний на етапі досліджень інших методів обмеження, наведених нижче, для порівняння результатів їх застосування із базовим варіантом.
2. “Стоп-список” – параметризований оператор, що відкидає задану наперед параметром кількість слів у хвості послідовності. Даний метод використовує один з популярних методів вилучення стоп-слів в пошукових системах, проте залишається чутливим до розміру колекції текстів. За допомогою такого відсіювання, наприклад, 100 слів в кінці послідовності, можна позбутися певних загальновживаних слів, проте навряд вдасться здійснити відокремлення загальновживаних слів наукового стилю мовлення.
3. Пропорційний метод – подібний до оператору “стоп-список”, з обмеженням в якості параметру певного відсотку слів в хвісті послідовності. Грунтується на методологічній засаді про відомість статистичного розподілу термінів в колекціях наукових текстів.
4. Метод найшвидшого спуску. Базується на запропонованій нами гіпотезі значного стрибку функції розподілу частот термінів при переході з загальновживаних слів до специфічних термінів. Підхід полягяє в послідовному аналізі підпослідовностей слів розміру *k,* і обранні в якості бар’єру останнього елементу такої підпослідовності, що має найбільший стрибок частоти серед усіх інших.

Під час дослідних експериментів було вирішено зупинитись на пропорційному підході до обмеження вхідного списку термінів, що б надавав принаймні першочергове відкидання загальновживаних слів, і мав би зручний для варіації під час експеринтів параметр відсотку відсіву. Однак для розроблюваного методу даний вибір не є остаточно зафіксованим, і програмна реалізація методу може мати усі наведені вище способі обмеження списку вхідних слів, разом із їх комбінаціями у разі потреби.

**Надійне зважування**

Зрозуміло, що наведені способи обмеження списку слів будуть працювати лише за умови застосування надійної схеми зважування, що в свою чергу, в нашому випадку, буде залежити від способу підрахунку складової документарної частоти термінів, чутливої до складу і розміру колекцій.

В нашій роботі проблему замалих колекцій текстів для надійного зважування запропоновано вирішувати шляхом побудови компоненту довідкової системи документарних частот термінів.

Тут можна виділити два підходи:

* Побудова і індексація великої і різноманітної навчальної колекції текстів наукової тематики, з послідуючим зберіганням отриманих документарних частот як еталонних. В свою чергу, в якості документарної основи для такої колекції, можна запропонувати випадкову вибірку 1000 україномовних публікацій з системи Google Scolar, або повну збірку журналу “Записки НаУКМА” за декілька років.
* Звертання до вже існуючих частот термінів в великих пошукових системах. Зокрема, можна за основу документарної частоти брати параметр кількості знайдених результатів при пошуку ключового слова. До переваг даної методики слід віднести як власне вже готовість до використання (частоти пораховані), так і те, що покладена в основу даної пошукової системи колекція документів надзвичайно велика, таким чином можно зробити припущення про надійне порівняння між собою документарних частот слів, отриманої таким чином. До недоліків належить складність і ненадійність доступу (для роботи алгоритму необхідний постійний широкосмуговий доступ в Інтернет, реалізація пошукового скрипта буде використовувати або забагато зайвих даних під час підвантаження і розбору веб-сторінок, або потраплятиме під квоти використання відповідного програмного інтерфейсу), а також непостійність результату бо система динамічна.

**Пошук зв’язків**

Після отримання першочергового списку термінів, для складання тезаурусу необхідно віднайти характер і направленість зв’язків між термінами.

В даній роботі вводиться поняття характеристичного фрагменту тексту, що є безпосереднім входженням терміну в документ у певному контексті, зокрема як члену відповідного речення. З-поміж багатьох методів розгляду контексту вживання слів, як-от частин оточуючих словосполучень і зворотів, речень, чи взагалі вікон з фіксованим розміром кількості слів, ми обрали саме речення в якості основи для наших досліджень, виходячі з наявних інструментів, що дозволяли б застосувати методику тегування за частинами мови в якості основи для лексикографічних методів.

Отже, наступним кроком є знаходження характеристичних фрагментів тексту для усіх термінів зі списку. Даний пошук може бути здійснений лінійно, проте маючи на увазі можливість масштабування розробленого методу, запропоновано використати одну з пошукових систем з відкритим кодом, що повертала б всі документи з нашої колекції що містять певний термін, таким чином обмежуючи простір лінійного пошуку. Далі, серед знайдених документів здійснюється пошук характеристичних фрагментів – речень, в котрі входить даний термін.

Наступним кроком здійснюється аналіз всіх знайдених характеристичних фрагментів, із застосуванням різних методик для визначення типу зв’язку.

1. Застосування найпростішого методу спільного вживання термінів всередені одного характеристичного фрагменту. Даний метод дозволяє встановити зв’язок пов’язаних термінів (RT), якщо дані терміни входять в характеристичні фрагменти тексту разом з початковим терміном.

2. Застосування множини визначених лексикографічних шаблонів, що дозволяють віднайти зв’язки типів BT, NT і RT.

**Розширення термінології термінологічними словосполученнями**

Застосування лексикографічних шаблонів базується на методі віднайдення іменникових термінологічних словосполучень, що в свою чергу, у разі співпадіння з текстом, дозволяє виокремити не тільки однослівні терміні, але і такі що подаются декількома словами, котрих є набагато більше. Таким чином, побочним продуктом застосування лексикографічних шаблонів є розширення першочергового списку термінів термінологічними словосполученнями, чого не можна було досягти на першому етапі тількі за рахунок індексування в рамках використаних інструментів.

**Методи пошуку співпадіння за лексикографічним шаблоном**

Для застосування визначених нами лексикографічних шаблонів, введемо таку формальну нотацію:

*Лексикографічний шаблон (LP)* – впорядкований список операторів співпадіння.

*Оператор співпадіння –* команда, що вимагає застосування операції пошуку співпадіння типу іменникового словосполучення (NP – Noun Phrase), або конкретного слова чи символу з синонімічного ряду (EW – Exact Word).

*NP* – оператор співпадіння, що виконує пошук іменникового словосполучення за рахунок застосування вказаних для кожного такого оператору списку правил співпадіння по частинах мови. Повертає в якості результату всі знайдені у фразі іменникові словосполучення в порядку даних правил співпадіння, а також позиції знайдених іменникових словосполучень у фразі. До операторів співпадіння даного типу в якості параметру можна задати їх роль (індекси 1 і 0).

*Роль оператора NP -* індекс 1 або 0, що вказує на головну або другорядну роль даного оператору в шаблоні (записується як NP1 або NP0).

*EW –* оператор співпадіння, що здійснює пошук входження конкретного символу або слова в фразу зі списку можливих альтернатив, повертає позиції входжень таких слів.

*W –* оператор вікна, що вказує мінімільні і максимальні рамки розміру вікна, що має роль маски співпадіння з будь-якими підпослідовностями слів в реченні.

*IT* – оператор ітерації, що позначає повторювану послідовність операторів в шаблоні.

*Правило співпадіння (MR) –* задана послідовність тегів частин мови, котрій має відповідати підпослідовність слів у реченні.

*Теги частин мови (N, A, P)* – параметри конфігурації правил співпадіння для виокремлення термінологічних словосполучень, що позначають іменник (N), прикметник (A), і прийменник (P) відповідно.

*Задовільнення шаблону –* знаходження множини задовільняючих оператори співпадіння підпослідовностей слів, де кожна позиція такої підпослідовністі відповідає як порядку входження у фразу, так і порядку оператора, визначеного у шаблоні. Тобто, всі можливі співпадіння по окремим операторам мають бути об’єднані в результуючу множину шляхом обмеження по слідуванню правил.

Наприклад, щоб зафіксувати в нашій формальній нотації лексикографічний шаблон 1, що відповідає за прямі означення з використанням тире, маємо записати наступне.

LP = (NP0(MR<A,N>), EW(“*–*”,”-”), NP1(MR<N,N>))

Даний шаблон має задовільнити наступна фраза:

*Соціологічне дослідження — система процедур для отримання наукових знань про*[*соціальні явища і процеси*](http://uk.wikipedia.org/wiki/%D0%A1%D0%BE%D1%86%D1%96%D0%B0%D0%BB%D1%8C%D0%BD%D0%B5_%D1%8F%D0%B2%D0%B8%D1%89%D0%B5)

При цьому першому оператору співпадіння буде відповідати термінологічне словосполучення “соціологічне дослідження”, оператору співпадіння по слову було надано дві альтернативи – власне символ “тире”, а також дефіс, для обробки випадків заміни даного символа у вхідному тексті, останньому оператору відповідає словосполучення “система процедур”.

Таким чином, оператори співпадіння типу EW у шаблоні грають роль фіксованих точок шаблону, в той час як оператори NP – роль наповнюваних змінних, що видобувають словосполучання з фрази під час задовільнення шаблону.

**Інтерпретація зв’язків у задовільненому шаблоні**

Під час складання шаблону, до параметрів NP додатково вказується параметр головної чи другорядної ролі в шаблоні, що інтерпретують зв’язки між отриманими співпадіннями по NP наступним чином:

* між представниками NP0 і NP1 встановлюється зв’язок BT;
* між представниками NP1 і NP0 встановлюється зв’язок NT;
* між представниками однакових ролей встановлюється зв’язок RT.

Підгрунтям для такої інтерпретації є те, що у більшості шаблонів на відповідних місцях термінологічних словосполучень за частинами речення бувають або однорідні означення чи додатки, або узагальнюючі слова, або, наприклад, у разі співпадіння з шаблоном прямих означень у тексті – відповідно термін і його родова приналежність. Таким чином, в тексті у разі спіпадіння з шаблоном направленість зв’язку є чітку визначеною.

**Пошук фразових іменникових словосполучень**

Для реалізації алгоритму пошуку гіпонімів, спершу необхідно навчити систему розпізнавати фразові словосполучення.

Пропонований підхід - фікскація іменників у реченні, з послідуючим добиранням навколишніх слів за правилами.

За Херстом, відношення між термінами “загальне-чатскове” відповідають гіпонімічним зв’язкам у тексті. І даному досліднику вдалося виокремити підмножну шаблонів, що була б достатньо точною і виконуваною для більшості текстів, враховуючи відмінності в формуванні термінології в окремих сферах інтересів.

Вважаючи на схожість наукового стилю на міжнародному рівні, здається вдалою думка про локалізацію знайдених шаблонів для української мови, з доданням нових.

Для того щоб звузити рамки дослідження і досягти певного результату для специфічних, проте найбільш уживаних способах творення термінології, до розгляду було залучено тільки терміни-іменники і іменникові словосполучення. Таке рішення було прийнято, виходячи з тих припущень, що більшість темінів утворено саме поодинокоми іменниками або термінологічними словосполученнями, тому і зв’язки в тексті слід шукати між іменниками і словосполученнями.

При цьому пропонується використати каскадний підхід, за котрим спочатку буде проводитись пошук найбільш широких і неточних зв’язків між усіма словами в реченні, потім серед іменників в реченні, потім між оточуючими ці іменники словосполученнями.

Під позначенням NP в записі шаблону слід розуміти іменникове словосполучення, що може складатись як з одного іменника, так і набувати набагато складніших форм.

З шаблонів, що відповідають за зв’язки між термінами у реченні, було обрано наступні категорії:

1. Прямі означання і дефініції, з використанням характерних для української мови знаків пунктуації і слів-зв’язок ( тире, слова “є”, “вважається”, “слід розуміти” і т. д.)

NP (*–* | -) (це | є | вважається | означає) NP

1. Шаблони за Херстом, а саме:
   1. такий NP як {NP ,}\* {(і | або)} NP
   2. NP {, NP} \* {,} (або | і) інший NP
   3. NP {,} (включаючи | особливо) {NP,}\* {і | або} NP
2. Шаблони на позначення зв’язків частина-ціле
   1. NP (є частиною | входить в | складається з) NP

Всі подані шаблони розширюються синоімічними і схожими за вживанням словами в формулах шаблону. Під час співставлення речень з шаблоном відбувається приведення всіх слів до нормільної форми, що дозволяє зменшити необхідну кількість варіацій шаблону.

**Етапи розробки лексикографічних підходів**

* + - 1. Відповідно до оглянутої літератури і відомостей з лексикографії, побудова і тестування на простих прикладах-реченнях шаблонів, що дозволять виокремити термінологічні словосполучення
      2. Тестування на простих реченнях шаблонів, що виокремлюють зв’язки між термінами (темрінологічними словосполученнями)
      3. Тестування побудованих шаблонів на визначеннях з термінологіних словників. Оцінка точності запропонованих методів.

1. Тестування на тематичних колекціях

**Деталі реалізації методу співставлення з шаблоном.**

Під час застосування правил, враховується їх черговість, таким чином в першу чергу віднаходяться і потрапляють в якості елементів співпадіння ті іменникові словосполучення, що є ширшими за кількістю слів, а отже рідшими за вживаням.

Таким чином для налаштування підсистеми пошуку термінологічних словосполучень, необхідно правильно підібрати не тільки правила їх формування по частинах мови, а ще й послідовність застосування.

## 2.2 Математична модель і формалізація методу

**Формальні позначення**

– множина текстових документів.

- множина лексикографічних шаблонів.

- множина термінів тезауруса.

– відсортований за метрикою і обмежений функцією список важливих однослівних термінів колекції

– множина багатослівних термінологічних словосполучень.

- множина зв’язків тезауруса,

- множина характеристичних фрагментів тексту для терміна *t.*

- множина речень характерестичного фрагменту *С.*

*–* множина лематизованих слів речення *S*

– послідовність співпадінь термінологічних словосполучень з лексикографічним шаблоном.

– функція обмеження відсортованого списку термінів.

- функція видобування термінів з документу.

- функція, що будує послідовність відсортованих термінів документа за спаданням метрики TFIDF.

- функція обчислення частоту терміну в документі.

– функція, що ставить кожному терміну у відповідність його інвертовану документарну частоту з еталонної колекції.

- функція пошуку характеристичних фрагментів терміна.

- функція розбиття характеристичного фрагменту текста на речення.

- функція видобуття послідовності лем з речення.

- функція задовільнення шаблону, що повертає множину послідовностей співвавших термінологічних словосполучень в порядку слідування позицій шаблону.

- функція встановлення зв’язків на множині послідовностей співпадінь шаблону.

**Псевдокод розробленого методу**

1. :
2. :
   1. :
      1. :
         1. : якщо :
         2. :
            1. ;

**Правила співпадіння з лексикографічним шаблоном**

# Розділ III. Практична частина

## Реалізація методу побудови термінології

**Збір даних**

Етап початкового збору даних є необхідним для побудови початкової довідкової системи документарних частот термінів, а також знадобиться під час проведення тестування алгоритму на точність складання тезаурусів. Тому для сбору документів з відкритого доступу, зокрема випусків журналу “Наукові записки НаУКМА”, було розроблено спеціальниї скрипт, що обходив сторінки сайтів архіву[[32]](#footnote-32) і зберігав всі документи в директоріях по роках і темах. Розбиття по темах є необхідним для створення різних тематичних тезаурусів на етапі тестування. Всього з даної колекції було взято 2926 повнотекстових документів за 1996-2013 роки видання.

**Розробка довідкового компоненту документарної частоти**

**Рішення на основі зовнішньої пошукової системи**

Документарну частоту по невеликій колекції, а там паче по 1 документу, побудувати дуже важко. Набагато простіше взяти вже існуючу документарну частоту терміну з пошукової системи. Як приклад було обрано пошукову систему Google, що разом з результатами звичайного пошуку завжди повертає приблизну кількість результатів пошукової видачі. Даний показник і було взято за документарну частоту терміна. Під час побудови допоміжних методів щодо видобування даного показника з пошукової видачі, було прийнято рішення, що звичайний парсинг веб-сторінок пошукової видачі буде громіздким рішенням, оскільки задля одного числа прийдеться запитувати і розбирати цілу сторінку, що містить багато зайвих даних. Тому в рамках окремного програмного модулю дипломної роботи було написано споживач Google Custom Search RESTful API[[33]](#footnote-33), з налаштуванням всього необхідного в консолі сервісу Google. Як результат було отримано метод, що за даним терміном повертає його документарну частоту, причому кількість інтернет трафіку для таких запитів є мінімальною. Проте виникає інша проблема для корустування Google API: квоти вільного використання, що становлять 100 запитів на день. Звичайно, щоб обробляти великі документи і знайти, які терміни в них є важливими, необхідно дізнатись документарну частоту кожного. Тому було прийнято рішення, по-перше, зберігати в локальному кеші всі знайдені документарні частоти, і в першу чергу звертатись до нього, що з бігом часу призведе до досить повної бази термінів, і по-друге, залишити можливість обробляти звичайні веб-сторінки пошуку, на випадок вичерпання квоти.

Кеш, як і інші дані проміжної роботи алгоритму, було вирішено зберігати в документарній базі MongoDB.

**Рішення на основі індексації еталонної колекції документів**

В якості еталонної колекції було обрано всі випуски журналу “Наукові записки НаУКМА”, отримані на етапі збору даних статей в откритому доступі.

Попередньо структура каталогів еталонної колекції була вирівняна, і всі документи було зібрано в одній директорії. Для здійснення індексації колекції було використано рішення від Apache Lucene, так само як і в фінальному варіанті робочої системи. Доступ до програмного інтерфейсу біліотеки індексації здійснювався з Java-коду, і спеціально для даної задачі створення початкового індексу документарних частот термінів була створена окрема конфігурація запуску програми. Основні особливості і кроки роботи даної програми наступні:

1. Зчитування параметрів директорії що містить документи колекції, а також окремої тимчасової конфігурації документарної бази даних MongoDB для зберігання проміжних результатів пошуку документарних частот.
2. Створення нового індексу Apache Lucene.
3. Подокументний розбір PDF і додавання текстів в індекс.
4. Перегляд індексу і зберігання документарних частот всіх термінів в документарну базу.
5. Дамп документарної бази для подальшого використання в якості початкового наповнення колекції документарних частот при звичайній роботі методу.

Зібраний дамп документарних частот зберігається в вихідних кодах програмної частини роботи у вигляді файлу даних із розширенням **.bson** і метаданих колекції термінів у файлі з розширенням **.json**. Дані файли мають бути імпортовані в нову базу на початку роботу алгоритму на новій системі.

Взагалом після індексації колекції і фільтрації термінів, в дампі нараховується близько 143000 термінів.

### Вибір готових рішень для реалізації кроків алгоритму.

## Архітектура системи обробки документів для побудови термінології

## Специфікація прикладного програмного інтерфейсу доступу до системи.

# Розділ IV. Обговорення результатів

Застосування алгоритму для побудови термінології у вигляді RDF схеми з використанням синтетичних та реальних даних україномовної наукової періодики.

## Схема тестування та оцінка результатів

Для перевірки результатів роботи алгоритму було вирішено побудовати тезаурусів на основі різних тематичних розділів україномовного журналу Записки НаУКМА, зібравши колекції текстів по різних темах за декілька років, а також 2 контрольних тезауруса, не пов’язаних з даним журналом. Контрольні тезаурусу планується наповнювати документами з пошукової видачі Google Scolar на 2 різних теми, по 30 документів на кожну. Контрольні тезауруси необхідні для того щоб мінімізувати вплив початкового фільтру важливих термінів, що в даній роботі побудований на основі всіх тем журналу Записки НаУКМА за всі роки, і тому завідомо краще працюючому, ніж з довільною колекцією.

В звідну таблицю, по кожному тезаурусу було зібрано наступну статистику:

* кількість документів в базовій колекції текстів,
* кількість знайдених термінів при початковому фільтруванні,
* кількість доданих термінологічних словосполучень і загальна кількість термінів,
* кількість знайдених зв’язків,
* середня кількість зв’язків одного терміна з іншими,
* кількість зв’язків по типах RT, BT і NT,
* час, витрачений на початкове індексування колекції
* час, витрачений на пошук зв’язків
* характеристики обладнання, на котрому проводились тестові запуски,
* виміряний коефіціент точності методу.

Коефіцієнт точності методу вимірювався наступним чином. З предметної області тестованої колекції обиралися 10 термінів, до кожного терміна добиралося три пов’язаних терміна, а також 1 ширший і 1 вужчий термін.

Коефіцієнт точності мав наступне правило обрахунку:

Де

* N1 відповідає кількості вдало розпізнаних термінів з N,
* N2 – кількості правильно розпізнаних зв’язків типу RT в тезаурусі, з 3N очікуваних,
* N3 – кількості розпізнаних зв’зяків типу BT і NT відповідно, з 2N очікуваних.

Введені в формулу обрахунку вагові коефіцієнти мають підкреслити першочергову важливість знаходження простих зв’язків.

В додатках подано відформатовані для сприйняття людиною результати у вигляді побудованих тезаурусів у форматі RDF.

Налаштування конфігурації бази для тестування

<http://bits-and-kites.blogspot.com/2014/01/spring-mongodb-tutorial.html>

Для тестування роботи алгоритму було вирішено розробити систему конфігурацій для швидкого прогону різних варіацій алгоритму на різних даних, та в одночас з необхідною оптимізацією швидкості таких тестувань, а також із заміром основних характеристик виконання окремих кроків алгоритму.

Найбільше для даного випадку підходить реалізація окремих інтеграційних тестів, що містимуть незалежні конфігурації оточення, такі як реалізація документарної бази у памяті.

1. Тестування алгоритму на словниках. Беремо термінологічний словник, зв’язвний, і дивимось чи знаходить алгоритм ці терміни і зв’язки

2. Тестування на вікіпедії. Там ми вже маємо готові зв’язки між термінами. Порівняти ті зв’язки з тими, що побудує алгоритм по документу з вихідного тексту.

3. Тестування на реальних даних.

Схема тестування є необхідною як для дослідження якості знайдених темінологічних зв’язків між термінами, так і для розгляду модифікацій алгоритму задля віднайдення найбільш ефективного рішення, що буде включено в заключну реалізацію системи як веб-сервісу.

Для цього є пропозиція віднайти:

* єдиний формат теазурусу для порівняння.
* зафіксувати модифікації власного алгоритму
* налаштувати тестове середовище, що дозволяло б запускати прогони алгоритмів.
* знайти сукупність реальних текстових наукових матеріалів для порівняння роботи на них алгоритмів.
* якщо алгоритм допускає недетермінованість, роботи декілька прогонів, і потім у порівннянні використовувати усереднену метрику
* розробити метрику порівнняння.

**Результати тестування методу на різних тематичних колекціях**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Назва колекції | Документів,  В колекції |  |  |
| 1 | Компьютерні науки | 96 |  |  |
| 2 | Філософія і релігієзнавство | 45 |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# Висновок

(досягнені результати згідно з поставленої мети)

# Використані джерела

**"Designing and Implementing RESTful Web Services with Spring." *Tutorial ·*. Web. 12 May 2014. <https://spring.io/guides/tutorials/rest/2/>.**

**"JSON-LD 1.0." *JSON-LD 1.0*. Web. 12 May 2014. <http://www.w3.org/TR/json-ld/#data-model>.**

**Landau, Sidney I. *Dictionaries: The Art and Craft of Lexicography*. New York: Scribner, 1984. Print.**

**"RDF 1.1 Concepts and Abstract Syntax." *RDF 1.1 Concepts and Abstract Syntax*. Web. 12 May 2014. <**[**http://www.w3.org/TR/2014/PR-rdf11-concepts-20140109/**](http://www.w3.org/TR/2014/PR-rdf11-concepts-20140109/)**>.**

**Towards a Comprehensive Theory of Lexicographic Definitions.**

[**http://wordnet.princeton.edu/**](http://wordnet.princeton.edu/)

[**http://www.w3.org/wiki/RdfThesaurus**](http://www.w3.org/wiki/RdfThesaurus)

[**http://www.imsglobal.org/vocabularies/iso2788\_relations.xml**](http://www.imsglobal.org/vocabularies/iso2788_relations.xml)

[**http://www.niso.org/schemas/iso25964/**](http://www.niso.org/schemas/iso25964/) **- newest thesauri schema standard**

[**http://www.niso.org/schemas/iso25964/iso25964-1\_v1.4.xsd**](http://www.niso.org/schemas/iso25964/iso25964-1_v1.4.xsd)

КОВАЛЬ, А.П. *Науковий стиль сучасної української мови : структура наукового тексту*. Київ: Видавництво Київського університету, 1970.

# Додатки

1. Chen, H., T. D. Ng, et al. (1997). "A Concept Space Approach to addressing the vocabulary problem in scientific Information Retrieval: An experiment on the Worm Community System." Journal of the American Society for Information Science 48(1): p. 17 [↑](#footnote-ref-1)
2. Chen, H., T. D. Ng, et al. (1997). "A Concept Space Approach to addressing the vocabulary problem in scientific Information Retrieval: An experiment on the Worm Community System." Journal of the American Society for Information Science 48(1): p. 178 [↑](#footnote-ref-2)
3. Lassi M. *Automatic thesaurus construction*. GSLT course Linguistic Resources, 2002. [↑](#footnote-ref-3)
4. Schütze, H. and J. O. Pedersen (1997). "A cooccurrence-based thesaurus and two applications to Information Retrieval." Information Processing and Management 33(3): p.307. [↑](#footnote-ref-4)
5. Miller, U. (1997). "Thesaurus construction : problems and their roots." Information Processing and Management 33(4): p.489 [↑](#footnote-ref-5)
6. Schütze, H. and J. O. Pedersen (1997). "A cooccurrence-based thesaurus and two applications to Information Retrieval." Information Processing and Management 33(3): p.308. [↑](#footnote-ref-6)
7. http://www.w3.org/TR/rdf-primer/ [↑](#footnote-ref-7)
8. <http://www.niso.org/schemas/iso25964/iso25964-1_v1.4.xsd> [↑](#footnote-ref-8)
9. HATEOAS – Hypermedia as the Enginge of Application State, http://en.wikipedia.org/wiki/HATEOAS [↑](#footnote-ref-9)
10. http://www.w3.org/TR/json-ld/#basic-concepts [↑](#footnote-ref-10)
11. Anderson, J. D. & Pérez-Carballo, J. (2001). "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part 2: Machine indexing, and the allocation of human versus machine effort." Information Processing and Management 37: p. 260 [↑](#footnote-ref-11)
12. Schütze, H. and J. O. Pedersen (1997). "A cooccurrence-based thesaurus and two applications to Information Retrieval." Information Processing and Management 33(3): 307-318. [↑](#footnote-ref-12)
13. Chen, H., T. Yim, et al. (1995). "Automatic thesaurus generation for an electronic community system." Journal of the American Society for Information Science 46(3): p.178. [↑](#footnote-ref-13)
14. Chen, H & Lynch, K.J. (1992). Automatic construction of networks of concepts characterizing document databases. IEEE Transacations on Systems, Man and Cybernetics. 22(5), 885-902. [↑](#footnote-ref-14)
15. Chen H., Lynch, K.J., Basu, K., & Ng, D.T. (1993). Generating, integrating, and activating thesauri for concept-based document retrieval. IEEE Expert. Special Series on Artificial Intelligence in Text-based Information Systems. 8(2), 885-902. [↑](#footnote-ref-15)
16. *Українська мова:* *Енциклопедія.* Редактори В. М. Русанівський, О. О. Тараненко та інші. Київ: Українська енциклопедія, 2000. C. 629. [↑](#footnote-ref-16)
17. *Становлення української лінгвістичної термінології. Магістерська робота. М. Кріслова, 2009* [↑](#footnote-ref-17)
18. *Українська мова:* *Енциклопедія.* Редактори В. М. Русанівський, О. О. Тараненко та інші. Київ: Українська енциклопедія, 2000. С.231. [↑](#footnote-ref-18)
19. *Українське термінознавство.* Львів: Світ, 1994. С. 143. [↑](#footnote-ref-19)
20. *Лексикографія.* [online] [актуалізація 2014-05-25] [цит.2014-05-25]< http://uk.wikipedia.org/wiki/Лексикографія> [↑](#footnote-ref-20)
21. DOSTÁL, J. Termín a jeho definice ve výkladových, terminologických a naučných slovnících a encyklopediích. *Modernizace výuky v technicky orientovaných předmětech a oborech. Dodatky.* 1.vyd. Olomouc: Votobia, 2003. Str. 57. [↑](#footnote-ref-21)
22. Strzalkowski, T. (1995). "Natural language information retrieval." Information Processing and Management 31(3): p. 401 [↑](#footnote-ref-22)
23. Grefenstett G. Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques, 1993, p. 2. [↑](#footnote-ref-23)
24. *Становлення української лінгвістичної термінології. Магістерська робота. М. Кріслова, 2009, С.49*  [↑](#footnote-ref-24)
25. Словники: мистецтво та ремесло лексикографії. Сідні І. Лендау. К.І.С, 2012, С. 173 [↑](#footnote-ref-25)
26. [↑](#footnote-ref-26)
27. G. Adamson, 1974. The use of an association measure based on character structure to identify semantically-related pairs of words. P. 260 [↑](#footnote-ref-27)
28. * Кочерган М. П. Вступ до мовознавства: Підручник. — Вид.2-ге.-К.: [ВЦ «Академія»](http://uk.wikipedia.org/wiki/%D0%92%D0%B8%D0%B4%D0%B0%D0%B2%D0%BD%D0%B8%D1%87%D0%B8%D0%B9_%D1%86%D0%B5%D0%BD%D1%82%D1%80_%C2%AB%D0%90%D0%BA%D0%B0%D0%B4%D0%B5%D0%BC%D1%96%D1%8F%C2%BB), 2008. — 368 с. (Альма-матер). C. 206

    [↑](#footnote-ref-28)
29. Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the 14th International Conference on Computational Linguistics, 1992. [↑](#footnote-ref-29)
30. Ashavi. H. (1987) Processing dictionary definitions with phrasal pattern hierarchies. *American Journal of Computational Linguistics,* 13(3): 195-202. [↑](#footnote-ref-30)
31. http://wordnet.princeton.edu/ [↑](#footnote-ref-31)
32. <http://nz.ukma.edu.ua/index.php?option=com_content&task=section&id=10&Itemid=47> і <http://www.ekmair.ukma.kiev.ua/> [↑](#footnote-ref-32)
33. https://developers.google.com/custom-search/json-api/v1/overview [↑](#footnote-ref-33)