Національний університет «Києво – Могилянська академія»

Магістеріум факультету інформатики



**Магістерська теза**

на тему:

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

*Підготував:*

студент ІУСТ-2,

Решетньов І.В.

*Науковий керівник:*

А.М.Глібовець

**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 6

Розділ I 6

Роль автоматизованої побудови тезаурусів як предмету дослідження в галузі інформаційного пошуку 6

1.1. Проблеми інформаційного пошуку 6

1.2. Розвиток онтологічних і семантичних систем аналізу текстів 7

1.3. Сучасний стан опису ресурсів в семантичному вебі: WordNet, Google Knowlrdge Graph, Wikipedia. 7

1.4. RDF граф як формат опису залежностей між ресурсами в семантичному вебі 7

1.5. Огляд концептів і абстрактного синтаксису RDF 7

1.6. Огляд стандарту JSON-LD як конкретної специфікації RDF. 7

1.7. Дослідження існуючих методів побудови термінології і складання тезаурусів 7

1.7.1 Статистичні методи 8

1.7.2 Лексикографічні методи 9

Розділ II 17

2.1 Структурна схема і формалізація алгоритму 17

2.1.1 Структурна схема алгоритму 17

2.1.2 Математична модель і формалізація алгоритму 19

2.2 Оцінка складності алгоритму 25

Розділ III 25

Реалізація алгоритму 25

Вибір готових рішень для реалізації кроків алгоритму. 25

Реалізація системи обробки документів для побудови термінології 25

Реалізація прикладного програмного інтерфейсу доступу до системи. 25

Реалізація графічного інтерфейсу користувача у вигляді веб-сервісу. 25

Розділ IV 25

Схема тестування та оцінка результатів 25

Висновок 26

Використані джерела 27

Додатки 27

# 

# Вступ

# Розділ I

# Роль автоматизованої побудови тезаурусів як предмету дослідження в галузі інформаційного пошуку

# 1.1. Проблеми інформаційного 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 сосою 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, що показували, як люди тільки у 20% випадків вибирали однакові терміни для опису тих самих речей, а також заміри релевантності пошукової видачі з доданою інформацією з тезаурусу, що мають більші показники. Таким чином, можна вважати доведеним і очевидним, що використання тезаурусів в пошукових системах значно покращує їх якість.

Проблема полягає у тому, що для більшості комерційних баз даних, що поширють наукову інформацію, тезауруси складаються експертами з області 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 побулови тезаурусів, що в якості корпусу беруть усі найновіші публікації з теми, і будують по них зв’язки між термінами. Як наслідок, маючи таку систему оновлювати термінологічні зв’язки значно простіше і дешевше. В публікації Моніки Лассі було розглянуто основні методи автоматизованої побудови тезаурусів, такі як CO, LSI, SVD та метод баєсівських мереж. Дані методи мають як різну ефективність і часову оцінку складності, так і приниципи - статистичний і лексикографічний. Розробці і тестуванню нового методу, що б використовував комбінацію ідей, що зустрічаються в цих методах, і присвячено дану магістерську тезу.

## 1.2. Розвиток онтологічних і семантичних систем аналізу текстів

?

## 1.3. Сучасний стан опису ресурсів в семантичному вебі: WordNet, Google Knowlrdge Graph, Wikipedia.

Ворднет- корпус англійської мови, розмічений експертами.

## 1.4. RDF граф як формат опису залежностей між ресурсами в семантичному вебі

## 1.5. Огляд концептів і абстрактного синтаксису RDF

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

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

Використати для контексту JSON-LD.

<http://json-ld.org/playground/> - ресурс для перевірки валідності формату і конвертації між видами серіалізації.

## 1.7. Дослідження існуючих методів побудови термінології і складання тезаурусів

**CO, LSI, SVD, Bayesian Networks.**

Chen et al. (1995, p. 178) declare that the research in the area of generating terms by co-occurrence analysis

done in the 1980’s and 1990’s have not given the results hoped for. Some experiments have resulted in poor

retrieval when terms were generated completely automatically. An experiment by Ekmekcioglu, Robertson &

Willetii in 1992 employed a user-directed approach where users were suggested terms for expanding their

queries. Four different approaches were used; original queries; query expansions generated by co-occurrence

data; Soundex codes, where the same phonetic code is assigned to words that sound the same; and strong

similarity measure based on similar character microstructure. The results showed no significant difference in

retrieval effectiveness between the initial queries and the expansions made. There was, however, a very small

degree of overlap between the relevant documents retrieved by the initial queries and the relevant documents

retrieved by the co-occurrence approach. The small degree of overlap of retrieved documents show that by

adding to the query terms from a thesaurus generated by co-occurrence analysis, **an increase in the number of**

**relevant documents retrieved can be accomplished**. (ibid, p. 178f)

### 1.7.1 Статистичні методи

Idf = 1/df

Метрика, що дозволяє сортувати терміни по важливості і специфічності в документах.

*Тут навести приклади досліджень, що сходяться до того, що проста метрика tf-idf настількі ж ефективна, як і складні ймовірностні моделі.*

Документарну частоту по невеликій колекції, а там паче по 1 документу, побудувати джуе важко. Набагато простіше взяти вже існуючу докумнтарну частоту терміну з пошукової системи. Як приклад було обрано пошукову систему Google, що разом з результатами звичйного пошуку завжди повертає приблизну кількість результатів. 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 інших даних. Тому в рамках окремного програмного модулю дипломної роботи було написано споживач Google Custom Search RESTful API, 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.

### 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.[[1]](#footnote-1) 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[[2]](#footnote-2):

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.[[3]](#footnote-3)

На Україні, крім поняття ***термінологія***, досі часто користуються терміном ***термінознавство***. Дефініція його така – «наука, яка займається загальновживаними питаннями терміна, термінології, номенклатури».[[4]](#footnote-4)

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.[[5]](#footnote-5) У рамках лексикографії слід розрізняти лексикографічну теорію, котра 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» (*переклад –– М.К.*).[[6]](#footnote-6) словники відображають культуру й мову народу та є джерелом правил її написання, вимови, наголошування тощо.

///

Також, 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 ще в праці Стрцалковского[[7]](#footnote-7) 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 сатистичних методів в інформаційному пошуку.

/////////

Гіпоніми і гіперніми

<http://en.wikipedia.org/wiki/Hyponymy_and_hypernymy>

Першим для пошуку гіпонів лексикографічні патерни використав Херст в 1992 році.

<http://en.wikipedia.org/wiki/Noun_phrase>

За Gregory Grefenstette, можна вибирати тільки сполучення іменник-іменник, як варіант.

Варто обмежетись іменниковими словосполученнями і фразами як у роботі Херста, під час пошуку і транслювання лексикографічних патернів.

weighted Jaccard measurement

Патерни (з роботи херста)

such NP as {NP ,}\* {(or | and)} NP

NP {, NP} \* {,} or other NP

NP {, NP}\* {,} and other NP

NP {,} (including | especially) {NP,}\* {or | and} NP

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

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

**Етапи впровадження лексикографічних підходів**

в даній роботі наступні:

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.

4. Тестування методів на текстах вікіпедії. Порівняння з вже розміченими зв’язками між термінами.

5. Тестування на вільних 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 підисистеми пошуку термінологічних словосполучень, необхідно правильно підібрати не тільки правила їх формування по частинах мови, а ще й послідовність застосування.

В якості тестового словника використано <http://ipp.lp.edu.ua/Library/004/004.html>

*Кизименко Л.Д., Бєдна Л.М.* Словник-довідник соціального працівника

*Для студентів та соціальних працівників*

# Розділ II

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

## 2.1 Структурна схема і формалізація алгоритму

### 2.1.1 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.

**Методу 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.

*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" \o "Соціальне явище)*

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.

2.1.2 Математична модель і формалізація алгоритму

## 2.2 Оцінка складності алгоритму

# Розділ III

Практична частина

## Реалізація алгоритму

### Вибір готових рішень для реалізації кроків алгоритму.

## Реалізація системи обробки документів для побудови термінології

## Реалізація прикладного програмного інтерфейсу доступу до системи.

## Реалізація графічного інтерфейсу користувача у вигляді веб-сервісу.

# Розділ IV

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

## Схема тестування та оцінка результатів

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:

* реалізації алгоритмів CO, LSI, SVD, Bayesian Networks, для того щоб будувати тезауруси за їх допомогою, і потім порівнюват між собою.
* 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.

# 

# 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/>>.**

**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. *Українська мова:* *Енциклопедія.* Редактори В. М. Русанівський, О. О. Тараненко та інші. Київ: Українська енциклопедія, 2000. C. 629. [↑](#footnote-ref-1)
2. *Становлення української лінгвістичної термінології. Магістерська робота. М. Кріслова, 2009* [↑](#footnote-ref-2)
3. *Українська мова:* *Енциклопедія.* Редактори В. М. Русанівський, О. О. Тараненко та інші. Київ: Українська енциклопедія, 2000. С.231. [↑](#footnote-ref-3)
4. *Українське термінознавство.* Львів: Світ, 1994. С. 143. [↑](#footnote-ref-4)
5. *Лексикографія.* [online] [актуалізація 2014-05-25] [цит.2014-05-25]< http://uk.wikipedia.org/wiki/Лексикографія> [↑](#footnote-ref-5)
6. 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-6)
7. Strzalkowski 1995, p. 401

   Strzalkowski, T. (1995). "Natural language information retrieval." Information Processing and Management 31(3): 397-417. [↑](#footnote-ref-7)