МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Нестеров Євгеній Сергійович

УДК 004.912

ПРОГРАМНИЙ ІНСТРУМЕНТАРІЙ ВИОКРЕМЛЕННЯ СЕМАНТИЧНИХ СУТНОСТЕЙ У НАВЧАЛЬНОМУ МАТЕРІАЛІ НА ПРИНЦИПАХ TEXT MINING

Спеціальність 8.050103

«Інженерія програмного забезпечення»

АВТОРЕФЕРАТ

дисертації на здобуття освітньо-кваліфікаційного рівня
магістр

Київ – 2016

Робота виконана на кафедрі автоматизації проектування енергетичних процесів та систем НТУУ «КПІ» Міністерства освіти і науки України.

Науковий керівник: кандидат технічних наук, доцент
Гагарін Олександр Олександрович
доцент кафедри автоматизації проектування
енергетичних процесів і систем НТУУ «КПІ» (м. Київ)

Рецензенти:

Захист відбудеться __ червня 2016 р., на засіданні ДЕК кафедри АПЕПС НТУУ „КПІ” аудиторія _____

З дисертацією можна ознайомитись у методичному кабінеті кафедри АПЕПС НТУУ „КПІ”, аудиторія .

Реферат підготовлено та представлено до розгляду „__” _____2016 р.
Робота рекомендована до захисту „__” _____ 2016 р.

Завідуючий кафедрою АПЕПС НТУУ „КПІ”,
доктор технічних наук, професор
Лук’яненко С. О.

Відповідальний за випуск магістрів
кафедри АПЕПС НТУУ «КПІ»,
кандидат технічних наук, доцент Гагарін О. О.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Експоненціальне зростання обсягу інформації є причиною все більш зростаючих труднощів пошуку необхідних документів та організації їх у вигляді структурованих за змістом сховищ. Користувачам важко знайти необхідну інформацію, традиційні механізми пошуку виявляються малоефективними.
Неструктурована інформація становить значну частину сучасних електронних документів. Системи класу Data Mining працюють зі структурованими даними, а для роботи із неструктурованим текстом необхідно використовувати засоби Text Mining. Результати аналізу засобами Text Mining можуть бути використані в автоматизованих навчальних системах для зменшення часу, який витрачається редактором при створенні статей завдяки автоматизованому виділенні головних понять і автоматичному форматуванні розмітки сторінки відповідно до виділеної суті статті. Також результати аналізів статей дозволяють автоматично створювати семантичні зв’язки між різними статтями, що суттєво покращую пошук корисної інформації. Подібні технології незамінні для вилучення знань і грають важливу роль у всій системі управління знаннями. Актуальність теми дисертації полягає у необхідності вибору та удосконалення засобів Text Mining для навчальних матеріалів з урахуванням їх особливостей та адаптації методів для обробки кириличних мов.

Зв’язок роботи з науковими програмами, планами, темами

Дисертаційна робота магістра виконувалась у НТУУ "КПІ" у відповідності з планом наукових досліджень кафедри АПЕПС.

Метою дослідження є виявлення закономірностей, нових підходів щодо автоматизованої обробки текстів, а саме виокремлення основних семантичних сутностей з навчальних статей засобами Text Mining.

Для реалізації поставленої мети були сформульовані наступні **завдання дослідження**, що визначили логіку дослідження та його структуру:

- проаналізувати структуру та функціональність автоматизованих навчальних

систем;

- виділити особливості наукових та навчальних статей;
- проаналізувати сучасні методи Text Mining та можливості їх використання з урахуванням виділених особливостей;
- розробити алгоритм виділення основних понять з текстів, що ґрунтується на комбінації методів Text Mining з урахуванням виділених особливостей текстів;
- розробити програмне забезпечення, що реалізує запропонований алгоритм.

Об'єктом дослідження є програмне забезпечення автоматизованих навчальних систем.

Предметом дослідження є методи Text-Mining для аналізу навчальних текстів.

Методи дослідження. Для досягнення мети дослідження використовується:

- статистичний метод ACSI-Matic для оцінки важливості лексем;
- дистрибутивний метод для додаткової оцінки важливості лексем;
- метод синтаксичних N-грам для об'єднання лексем в семантичні сутності.

Наукова новизна одержаних результатів. Найбільш суттєвими науковими результатами магістерської дисертації є:

- удосконалено статистичний метод виділення основних семантичних сутностей шляхом введення евристичних правил та застосування N-грам, що дозволило підвищити якість результатів;
- отримало подальшого розвитку застосування методів Text Mining для аналізу навчальних текстів з урахуванням їх специфіки

Практичне значення одержаних результатів визначається тим, що запропонований метод виділення основних семантичних сутностей на відміну від аналогів, має кращу підтримку комбінованих мов, високу швидкодію та не потребує додаткових словників. Розроблений програмний модуль може бути використаний в системах дистанційного навчання, базах знань та бібліотеках та інших системах де необхідна обробка великих обсягів текстової інформації.

Апробація результатів дисертації

Основні положення роботи доповідались і обговорювались на:

1. XIV Міжнародній науково-практичній конференції аспірантів, магістрантів і студентів «Сучасні проблеми наукового забезпечення енергетики» (м. Київ, 18-21 квітня 2016 р.).

2. III науково-практичній дистанційній конференції молодих вчених і фахівців з розробки програмного забезпечення «Сучасні аспекти розробки програмного забезпечення» (м. Київ, 15 квітня 2016 р.).

Публікації. Наукові положення дисертації опубліковані в одній роботі.

Ключові слова. *TEXT-MINING, N-ГРАМИ, АВТОМАТИЗОВАНІ НАВЧАЛЬНІ СИСТЕМИ, ОБРОБКА ТЕКСТІВ, СТЕМІНГ, СТАТИСТИЧНІ АСОЦІАЦІЇ*

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** охарактеризовано актуальність дисертаційної роботи, розписано основні питання, що розглядалися в дослідженні та напрям дослідження.

У **першому розділі** коротко охарактеризовано контекст дослідження. Зокрема, описані задачі та принципи text minig. Проаналізовано існуючі комерційні та відкриті програмні засоби і системи для аналізу тексту. Описані можливості використання засобів семантичного аналізу текстів в автоматизованих навчальних системах (АНС) та розглянута структура АНС. Сформульована проблема виокремлення семантичних сутностей в навчальному тексті. Підкреслено, що для покращення роботи екстрактивних методів необхідно враховувати додаткові правила оцінки важливості лексем.

Сформульовані завдання дослідження, а саме:

– виділити особливості наукових та навчальних статей;

– проаналізувати сучасні методи Text Mining та можливості їх використання з урахуванням виділених особливостей;

– розробити алгоритм виділення основних понять з текстів, що ґрунтується на комбінації методів Text Mining;

– розробити програмне забезпечення, що реалізує запропонований алгоритм.

У **другому розділі** проведено аналіз існуючих методів виділення семантичних сутностей та реферування текстів, а саме:

– екстрактивні методи без опори на знання, до яких відносять метод ACSI-Matic, метод Луна, метод статистичних асоціацій, метод Освальда, дистрибутивні методи. Ці методи прості в реалізації але не дають гарний результат на текстах невеликого об'єму;

– методи за опорою на знання. Недолік цих методів полягає в тому, що потрібні потужні обчислювальні ресурси для систем обробки природних мов та словники для синтаксичного розбору і генерації природно-мовних конструкцій. Крім того, для реалізації цього методу потрібні якісь онтологічні довідники, що відображають міркування здорового глузду і поняття, орієнтовані на предметну область, для прийняття рішень під час аналізу і визначення найбільш важливої інформації.

Також в другому розділі описано використання N-грам для обробки природньої мови та принципи і засоби попередньої обробки тексту, в тому числі очистки від зайвих слів, розбиття на токени, приведення слів до нормальної форми.

Далі описано принцип виділення особливостей навчальних текстів та наведено результати аналізу статей на інформаційно-навчальних веб ресурсах, таких як www.znannya.org та uk.wikipedia.org. Виділені особливості:

– Зміна регістру символів (UPPERCASE, camelCase, Firstletter).

– Виділення символами пунктуації (U -, "U", (U)).

– Вставка слів мовою, яка відрізняється від основної мови тексту статті ("система керування базами даних MySQL").

– Порядок N-грами.

– Наявність тегів (, < em>, < b>, < h1-5>, < i>, теги оформлення в документах).

– Кількість варіантів форм слова в тексті (менше – краще).

Описано удосконалені принципи оцінки важливості лексем в екстрактивних методах.

Наведено загальну структуру алгоритму виділення основних семантичних сутностей з навчального тексту (Рисунок 1).

Рисунок 1 – Схема алгоритму виокремлення семантичних сутностей

Для реалізації функції виділення основних семантичних сутностей з навчального тексту, запропоновано таку послідовність обробки тексту статті:

– Очищення тексту, вилучення стоп-слів та слів довжиною менше 3 символів. Використовується словник, що включає стоп-слова для української, російської та англійської мови.

– Токенізація тексту (розбиття на структурні складові) та виділення тегів і атрибутів. В результаті отримано дерево ієрархії токенів (слів, лексем), що дає змогу оцінити як розташування токену в тексті так і його вагу.

– Приведення слів до нормальної форма шляхом застосування процедури стемінгу на основі алгоритму Портера. В результаті отримаємо ієрархічне дерево лексем.

– Розрахунок важливості лексем методом ACSI-Matic з урахуванням ваги додаткової оцінки, що описана в пункті 2.9. В результаті отримаємо список N-грам

першого порядку з числовою оцінкою важливості та списком можливих форм.

– Застосування методу семантичних N-грам для пошуку N-грам вищих порядків, що дозволяє отримати основні семантичні сутності що складаються з 2-х та 3-х слів. Цей метод дозволяє об’єднати лексеми незалежно від словоформи.

– Вилучення надлишкових результатів шляхом відкидання N-грам нижчих порядків, що включені в N-грами більш високих порядків, та відкидання статистично незначних понять.

У **третьому розділі** описано програмну реалізацію системи для виділення основних семантичних сутностей. Програмна реалізація системи для виокремлення семантичних сутностей з навчальних текстів включає наступні етапи:

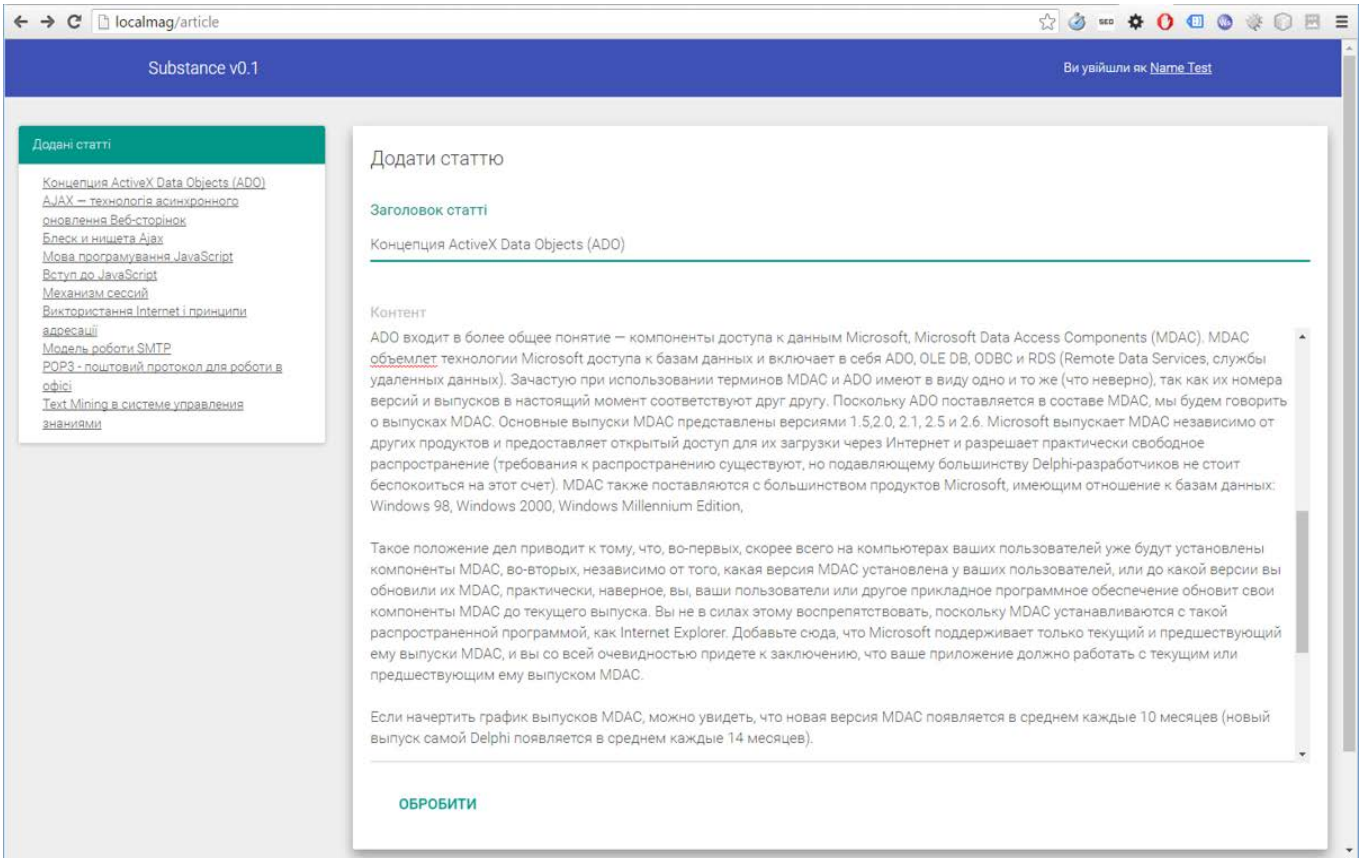
– Проектування системи. Результатом цього етапу є загальна архітектура системи, що визначає розподіл системи на програмні компоненти, концептуальна модель програмного забезпечення та функціональна декомпозиція системи, що представлені діаграмами процесів, використання та інші. Також результатом проектування є даталогічна модель представлення даних в базах даних та статичне представлення даних в системі у вигляди діаграми класів. Кінцевим результатом проектування виступає платформо орієнтована модель системи, що визначає розподіл компонентів системи на фізичних пристроях та вузлах.

– Реалізація компонентів системи згідно проекту. Результатом цього етапу є розроблені програмні модулі системи на відповідних мовах програмування та фізична структура бази даних.

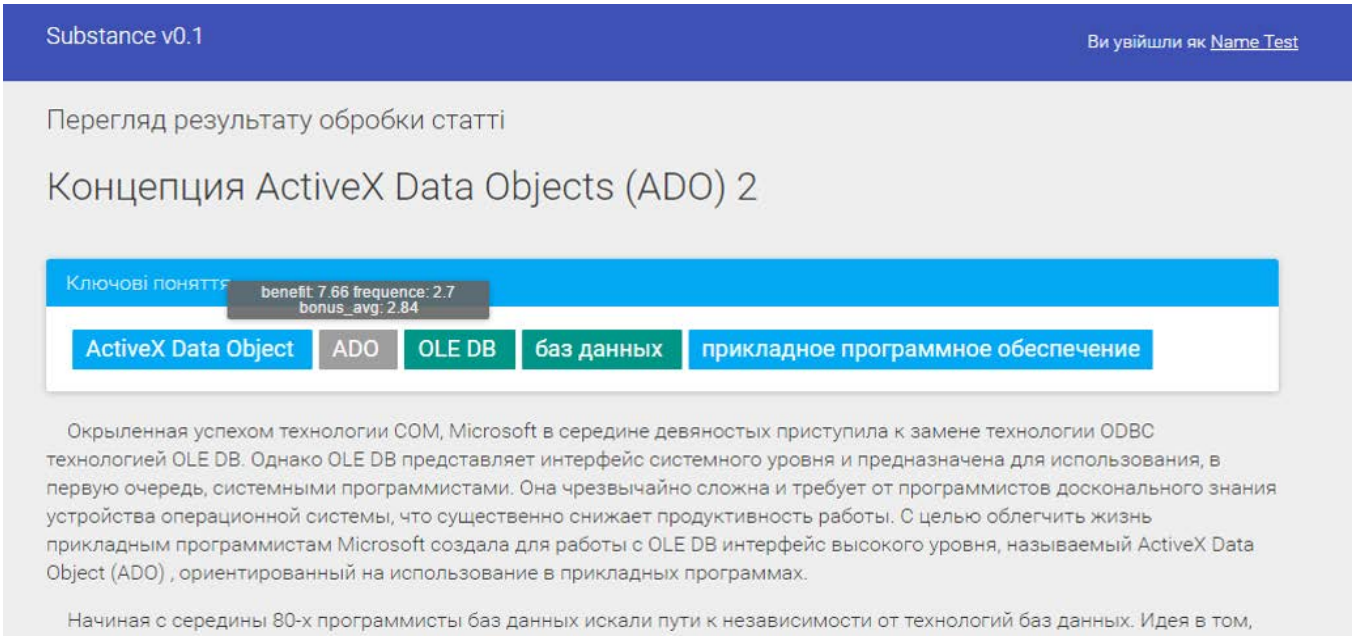
– Інтеграція системи та тестування. На етапі інтеграції програмні модулі встановлюються та налагоджуються на відповідних вузлах, проводиться комплексне тестування та інтеграційне тестування. Можлива інтеграція з компонентами «заглушками», що дозволяє тестувати окремі компоненти незалежно від інших.

– Розроблена програмна система представляє собою веб застосунок який включає модуль виділення основних семантичних сутностей, систему управляння контентом (CMS) та модуль управління користувачами.

Наведені далі скріншоти роботи системи надають уявлення про її головні функції
Додавання статей для обробки



Результати роботи програми по структуруванню текста статті .



ВИСНОВКИ

В результаті роботи над магістерською дисертацією було проведено аналіз існуючих програмних пакетів та засобів для інтелектуального аналізу текстів, що показав, що більшість існуючих рішень складно інтегрувати в автоматизовані навчальні системи на базі веб технологій через високу вартість та потребу у великій кількості обчислювальних ресурсів, а також те, що існуючі методи виділення основних семантичних сутностей не враховують специфіку направленості текстів.

На основі аналізу предметної області, та визначення мети роботи, як виявлення закономірностей, удосконалення підходів щодо автоматизованої обробки текстів, а саме виокремлення основних семантичних сутностей з навчальних статей засобами Text Mining, сформульовані завдання дослідження.

Для визначення можливостей вирішення задачі виокремлення основних семантичних сутностей в навчальному матеріалі, проведено аналіз існуючих методів, та вказані їх недоліки.

Для усунення існуючих недоліків та розробки алгоритму, що дозволяє обробити тексти навчальних матеріалів з урахуванням їх специфіки і з мінімальними витратами обчислюваних ресурсів, було удосконалено існуючі підходи, що склало наукову новизну роботи:

удосконалено статистичний метод виділення основних семантичних сутностей шляхом введення евристичних правил та застосування N-грам, що дозволило підвищити якість результатів;

отримало подальшого розвитку застосування методів Text Mining для аналізу навчальних текстів з урахуванням їх специфіки.

На основі запропонованих алгоритмів виділення основних семантичних сутностей з навчального тексту, була розроблена програмна система на основі веб технологій, що включає модуль виділення основних семантичних сутностей, який за рахунок удосконалення підходу по оцінці важливості окремих лексем, та

використання новітніх комп'ютерних технологій, по простоті інтеграції, підтримці комбінованих мов, зручності користування існуючі аналоги.

Розроблений програмний модуль та компоненти системи можуть бути використані в автоматизованих навчальних системах для зменшення часу, який витрачається редактором при створенні статей завдяки автоматизованому виділенні головних понять і автоматичному форматуванні розмітки сторінки відповідно до виділеної суті статті. Також результати аналізів статей дозволяють автоматично створювати семантичні зв'язки між різними статтями, що суттєво покращую пошук корисної інформації.

Дослідження, які були проведені під час виконання магістерської дисертації, показали, що специфіка текстів є дуже вагомим фактором при виокремленні основних понять та створенні рефератів методами без опори на знання. Актуальним є питання про можливість розробки інтелектуальних систем семантичного аналізу, що мають можливість самонавчання. Зважаючи на вищезазначене, напрямок робіт в цій області вбачається перспективним.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Нестеров, Є. Використання засобів text mining для автоматичного виділення семантичних сутностей з навчального тексту / Є. С. Нестеров, О. О. Гагарін // Сучасні проблеми наукового забезпечення енергетики. Матеріали XIV Міжнародної науково-практичної конференції аспірантів, магістрантів і студентів, присвяченої 85 річчю теплоенергетичного факультету, м. Київ, 18–21 квітня 2016 р. У 2 т. – К. : НТУУ «КПІ», 2016. – Т. 2. – С. 116.
2. Нестеров, Є. Автоматичне виділення семантичних сутностей з навчальних текстів / Є. С. Нестеров, О. О. Гагарін // Сучасні аспекти розробки програмного забезпечення: Збірник тез доповідей 3 науково-практичної дистанційної конференції молодих вчених, аспірантів та студентів, 15 квітня 2016 р. – Черкаси: видавець Чабаненко Ю.А., 2016. – С. 67.