66-12-14/2
01.03.2018

ВІДГУК

офіційного опонента, кандидата технічних наук Надутенка Максима Вікторовича на дисертаційну роботу Швороб Ірини Богданівни на тему: «Методи та засоби екстракції та аналізу слабоструктурованих текстових даних на основі документо-орієнтованого графа», представлену на здобуття наукового ступеня кандидата технічних наук за спеціальністю 10.02.21 – структурна, прикладна та математична лінгвістика

Актуальність теми дисертації

Велика кількість неструктурованих та слабкоструктурованих текстових даних та швидке збільшення їх кількості зумовлює пошук технологій для їх оптимального опрацювання та збереження, оскільки існуючі засоби та підходи в основному мають емпіричний характер, вузькоспеціалізовані та не надають можливості для впровадження нових методик та методів. Також важливим аспектом є збереження отриманих даних у реляційних базах даних, що спричинює втрату зв'язків між елементами тексту. Існуючі на сьогодні методи опрацювання слабоструктурованої інформації сильно залежать від встановленої розмітки та існуючого корпусу мови. Таким чином, аналіз слабострутурованих даних виконується напівавтоматично, оскільки для текстів з невідомою структурою необхідно визначати розмітку.

Тому, дисертація здобувача Швороб І.Б., присвячена дослідженню методів та засобів екстракції та аналізу слабкоструктурованих текстових даних на основі використання документо-орієнтованого графа, є своєчасною і актуальною.

Загальна характеристика роботи

У вступі обґрунтовано актуальність теми дисертаційної роботи, подано нерозв'язані задачі, сформульовано мету і виділено завдання дослідження, визначено об'єкт, предмет, методи дослідження, наукову новизну і практичне значення одержаних результатів, представлено загальну характеристику роботи, структуру та обсяг дисертації. Наведено відомості про впровадження результатів роботи, апробацію, особистий внесок автора, а також публікації за темою дисертації.

Перший розділ присвячено аналітичному огляду існуючих методів та засобів вирішення проблеми з екстракцією слабкоструктурованих даних. Зазначено, що задача опрацювання текстових даних вимагає застосування різних

алгоритмів аналізу та екстракції неструктурованої інформації. Проаналізовано існуючі методи екстракції, збереження, структурування, аналізу та роботи з слабкоструктурованими текстовими даними, їх переваги, області застосування, обмеження та проблеми та можливість їх застосування до аналізу слабкоструктурованих даних. Здійснено постановку задачі дослідження.

Здійснено постановку задачі, сформульовано мету та завдання дослідження методів та засобів екстракції, структурування, збереження та аналізу слабкоструктурованих природномовних текстів.

У другому розділі введено перелік концептуальних понять та визначень, аргументовано придатність NoSQL баз даних до роботи зі слабкоструктурованими даними. Здійснено порівняльну характеристику запропонованої документо-орієнтованої графової бази даних з існуючими графовою та документо-орієнтованою базами даних. Використано елементи теорії графів при роботі зі слабкоструктурованими текстами. Заслуговує на увагу розроблений дисертантом підхід до виділення різнотипних вершин та ребер документо-орієнтованого графа, що спрощує подання, збереження та доступ до отриманих даних.

Наведено приклади застосування елементів теорії графів при роботі зі слабоструктурованими даними, що дало змогу перевизначити операції над графами для аналізу слабоструктурованих текстів.

Розроблено метод перерахунку ваг ребер документо-орієнтованого графа. Здійснено формальну постановку задачі розроблення моделі подання слабоструктурованих даних на основі документо-орієнтованого графа.

У третьому розділі розроблено методи та алгоритми аналізу слабоструктурованих природномовних текстів на основі застосування текстових шаблонів, що базуються на виділених прагматичних ознаках, для екстракції даних зі слабоструктурованих природномовних текстів для побудови документо-орієнтованого графа. Розроблено метод побудови структури текстового шаблону та метод виділення прагматичних ознак для побудови текстового шаблону. Наведено алгоритм зведення текстів до єдиного вигляду. Розроблено метод первинного аналізу текстів та метод аналізу текстових блоків. Усі розроблені методи та алгоритми отримали практичну перевірку та досліджено їх характеристики.

У четвертому розділі проведено моделювання та апробацію мовно-інформаційної системи екстракції слабоструктурованих природномовних текстів та роботи з ними. Описано комплексну реалізацію практичного використання запропонованих підходів та методів в одній інтегрований системі.

Зокрема, подано архітектуру системи та її програмну реалізацію. Наведено приклади застосування розроблених методів на прикладі роботи з документо-орієнтованим графом, побудованим для системи роботи з інструкціями для медичних препаратів та для резюме найманих працівників. Продемонстровано аналіз основних результатів роботи на продемонстровано домінування розробленої мовно-інформаційної системи над аналогами.

Наукова новизна дисертаційної роботи

Метою дисертаційної роботи є розроблення є методів та засобів екстракції, структурування, збереження слабкоструктурованих текстових даних для їх подальшого аналізу. Наукова новизна одержаних результатів роботи полягає у наступному:

- вперше слабкоструктуровані природномовні тексти подано у вигляді документо-орієнтованого графа, що дало змогу використати теорію графів для встановлення зв'язків між елементами документа та знизити надлишковість результатів пошуку за запитом користувача;
- удосконалено метод екстракції даних з текстових блоків за прагматичною ознакою з метою формування документо-орієнтованого графа, який на відміну від методу на основі використання міри TF-IDF дає змогу врахувати семантику речень;
- набув подальшого розвитку метод первинного аналізу даних, який дає змогу частково структурувати природномовний текст для його подальшого аналізу та опрацювання на основі поділу тексту на блоки за прагматичними ознаками.

Виконані наукові дослідження дають змогу розв'язати наукові завдання, спрямовані на розроблення методів та засобів екстракції, збереження та аналізу слабкоструктурованих тестових даних.

Методи досліджень, використані в дисертаційній роботі

Для досягнення поставлених в дисертаційній роботі задач здобувачем використано: елементи теорії графів, статистичні методи, методи об'єктно-орієнтованого проектування.

Зв'язок дисертаційної роботи з науковими програмами, планами та темами

Обраний напрям досліджень відповідає тематикам науково-дослідних робіт «ДБ/ Інтелектуальні інформаційні технології багаторівневого управління енергоефективністю регіону» (Національний університет «Львівська

політехніка») та «Комплекс інтелектуальних інформаційних технологій інтеграції даних для обліку та аналізу підвищення кваліфікації вчителів» Національний університет «Львівська політехніка»).

Обґрунтованість і достовірність наукових результатів, висновків та рекомендацій

Ступінь обґрунтованості та достовірність результатів визначається коректним застосуванням сучасних загальноприйнятих математичних методів, використанням перевірених практикою результатів класичних досліджень. Для розв'язання задач використано повністю адекватний математичний апарат. Достовірність основних положень, обґрунтованість і адекватність результатів підтверджена результатами коректно виконаних експериментальних досліджень та практичним впровадженням розроблених методів і засобів дисертаційної роботи Швороб І.Б. в процеси діяльності низки компаній, які відображено в актах впровадження.

Наукове і практичне значення результатів, отриманих в дисертаційній роботі

Розроблено алгоритм первинного аналізу даних, який дає змогу частково структурувати природномовний текст для його подальшого опрацювання. Розроблено алгоритм перерахунку ваг ребер документо-орієнтованого графа, що дає можливість відсіювати надлишковість даних при запиті. Удосконалено алгоритм екстракції даних з текстових блоків шляхом формування документ-орієнтованого графа, який на відміну від методу на основі використання міри TF-IDF дає змогу врахувати семантику речень та на 8 % збільшити кількість збережених структурних одиниць.

На основі розробленої архітектури побудовано та впроваджено мовно-інформаційну систему екстракції слабоструктурованих природно-мовних текстів та роботи з ними

Одержані в дисертаційній роботі результати використано під час розроблення прототипів мовно-інформаційної системи та впроваджено у II травматологічному відділенні КМКЛШМД м.Львова та у ТЗоВ «To-You Sp. Z o.o.»

Публікації та апробація результатів дисертаційної роботи

За темою дисертаційної роботи опубліковано 8 наукових праць, у тому числі у тому числі 3 статтях в іноземних періодичних наукових виданнях, 3 – у фахових наукових виданнях України, 2 – у матеріалах конференцій.

Основні результати дисертації доповідалися на наукових семінарах і міжнародних науково-практичних та науково-технічних конференціях.

Відповідність дисертації встановленим вимогам ДАК України

Дисертаційна робота Швороб І.Б. «Методи та засоби екстракції та аналізу слабоструктурованих текстових даних на основі документо-орієнтованого графа» за оформленням відповідає вимогам МОН України, що пред'являються до дисертаційних робіт. Автореферат і дисертація та їх оформлення, кількість публікацій та повнота відображення результатів дисертаційних досліджень відповідають вимогам п.п.9, 11 і 12 "Порядку присудження наукових ступенів" щодо кандидатських дисертацій.

Автореферат дисертації розкриває її основні положення та висновки, є ідентичним за структурою та змістом із дисертацією.

Зауваження до дисертаційної роботи

При цілком позитивній оцінці роботи, вважаю за необхідне зробити такі зауваження до змісту і оформлення дисертаційної роботи та автореферату:

1. Матеріал підрозділу 3.2.1. Підготовка текстового документу подана описово. Доцільно було б подати блок-схему алгоритму підготовки документу до аналізу.
2. Перша частина тексту підрозділу 2.2 Визначення придатності NoSQL баз даних до роботи зі слабкоструктурованими даними має оглядовий характер, тому могла бути б перенесена у перший розділ роботи.
3. Не зовсім зрозуміло які завдання синтаксичного аналізатора виокремив автор та які саме критерії коректності даних застосував у своїй моделі слабкоструктурованих даних в розділі 1.
4. Дисертант посилається на роботи Пуппе з аналізу продуктивних правил, проте посилання в тексті на праці цих авторів відсутні.
5. Не описано призначення параметру EdgeType у розділі 2, у розділі 3 не визначено, як формується його значення. Хоча у розділі 4 наведено приклади, в яких вказаний параметр використовується.
6. В тексті дисертаційної роботи використано багато жаргонних термінів (наприклад, «дані витягуються», «токенізуватися»).
7. В тексті дисертаційної роботи є певні пунктуаційні помилки та неточності. Наприклад, на стор. 22 у фразі «HTML - hype text markup language» відсутня буква «r».

Відзначені зауваження не впливають на загальну позитивну оцінку дисертаційної роботи.

Висновки

Дисертаційна робота за змістом є закінченим науковим дослідженням, що містить нові науково-обґрунтовані результати, важливі на сучасному етапі та для перспективного розвитку національних телекомунікаційних мереж і цілком відповідає вимогам «Паспорту» спеціальності 10.02.21 – структурна, прикладна та математична лінгвістика.

Сукупність наукових положень, сформульованих та обґрунтованих в дисертаційній роботі, складає основу для розробки і впровадження методів та алгоритмів аналізу слабкоструктурованих природномовних текстів та екстракції інформації з них.

Автореферат повністю відображає зміст та основні положення дисертації.

За науковим рівнем, практичною цінністю, апробацією та публікаціями дисертаційна робота відповідає вимогам «Порядку присудження наукових ступенів», а її автор – Швороб Ірина Богданівна заслуговує присудження наукового ступеня кандидата технічних наук за спеціальністю 10.02.21 – структурна, прикладна та математична лінгвістика.

Офіційний опонент
кандидат технічних наук, завідувач відділу
інформатики Українського
мовно-інформаційного фонду НАН України М.В. Надутенко