67-72-14/1
01.03.2018

ВІДГУК

офіційного опонента доктора технічних наук, професора Лупенка Сергія Анатолійовича на дисертаційну роботу Шворoб 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

Глобалізаційні тенденції розвитку сьогодення зумовлюють необхідність створення консолідованих інформаційних продуктів, які використовують відомості з різних інформаційних джерел. Оскільки спостерігається швидке зростання кількості різноманітних за структурою джерел даних – реляційних баз даних, NoSQL-баз даних, слабоструктурованих даних тощо, то рішення з інтеграції даних вимагають постійних вдосконалень. Ще складнішим завданням є екстракція даних з природномовних текстів, оскільки відшукати структуру у такому випадку можна лише знаючи певні особливості того чи іншого тексту. На даний час є розроблені методи зі структурування даних за допомогою продукційних правил, функцій належності, однак вони вимагають зміни структури бази знань для нового джерела даних. Також такі методи сильно залежать від мови природномовного тексту, що накладає додаткові обмеження на аналіз україномовних текстів.

Отже, наукове завдання з розроблення методів та засобів екстракції, структурування, збереження та аналізу слабоструктурованих природномовних текстових даних є актуальним, а дисертація Шворoб І.Б., що присвячена дослідженню методів та засобів екстракції та аналізу слабоструктурованих текстових даних на основі документо-орієнтованого графа, є актуальною.

Загальна характеристика роботи

Дисертаційна робота складається з вступу, чотирьох розділів, висновків та додатків.

У вступі обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету і завдання дослідження, визначено об'єкт, предмет, методи дослідження, визначено 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.

У першому розділі проаналізовано поняття та методи екстракції текстової інформації зі слабоструктурованих природномовних текстів та виділено основні задачі екстракції даних. Проаналізовано найпопулярніші на сьогоднішній день системи екстракції даних зі слабоструктурованих текстів, виявлено їх обмеження та встановлено, що немає методів та засобів, які б дозволяли вирішувати усі перераховані задачі екстракції даних одночасно. Здійснено аналіз найбільш відомих синтаксичних аналізаторів та встановлено, що основною функцією синтаксичного аналізатора є визначення можливості та способу отримання аналізованих даних із початкового символу граматики. Розкрито семантику поняття «дані», визначено види даних, виділено основні проблеми в роботі зі слабоструктурованими даними.

На основі виконаного аналізу та досліджень сформульовано мету та завдання на дослідження методів та засобів екстракції, структурування, збереження та аналізу слабоструктурованих даних.

У другому розділі введено перелік базових понять та означень, що дало змогу сформувати структуру слабоструктурованого природномовного тексту на основі документо-орієнтованого графа. Проаналізовано вимоги до не реляційних баз даних. Визначено перелік проблем, нетипових для класичної реляційної моделі, які вдається вирішити за допомогою NoSQL моделей та визначено перелік вимог до нереляційної бази даних. Здійснено огляд існуючих моделей NoSQL баз даних та визначено основні аспекти вибору моделі бази даних. Введено новий тип бази даних – документо-орієнтована графова база даних, особливістю якого є можливість зберігання частин документу (документів) як вершин графа та встановлення залежностей між ними у вигляді ребер.

Продемонстровано застосування теорії графів при роботі з документо-орієнтованим графом та наведено приклади застосування операцій над такими графами на основі інструкцій до медичних препаратів та резюме найманих працівників.

Розроблено метод перерахунку ваг ребер документо-орієнтованого графа, який дає змогу формувати точніші відповіді на запитання та відсівати нерелевантну запитові користувача інформацію.

У третьому розділі розроблено метод виділення складових елементів для побудови текстового шаблону та описано текстовий шаблон з допомогою нотації Бекуса-Наура. Наведено комбінований алгоритм пошуку нечітких дублікатів в природномовних текстах, що дало змогу оптимізувати роботу розробленої системи та виключити повторний аналіз та екстракції даних текстового документа для того, щоб не формувати граф з елементів, уже наявних в базі даних. Розроблено метод формування шаблону, якого нема у базі даних шаблонів, що вирізняє пропонований підхід аналізу природномовних текстів від аналогів. Розроблено метод виділення складових елементів тексту для побудови текстового шаблону. Наведено розроблений метод поділу текстового документа на текстові блоки на основі знайденого або побудованого текстового шаблону за визначеними прагматичними ознаками. Розроблено метод виділення структурних одиниць з текстових блоків, визначення їх типу за прагматичною ознакою та формування вершин та ребер для документо-орієнтованого графа.

Четвертий розділ присвячено комплексній реалізації практичного використання запропонованих підходів. Зокрема, у розділі подано архітектуру мовно-інформаційної системи екстракції слабоструктурованих природномовних текстів та роботи з ними та розроблено відповідне програмне забезпечення. Розроблено концептуальну модель запропонованої системи. Апробовано розроблені методи для роботи зі слабоструктурованими медичними даними. Продемонстровано, як розроблені методи використано для формування системи роботи з резюме найманих працівників.

У висновках висвітлені основні положення роботи.

У додатках наведені акти про впровадження та фрагменти програмної реалізації розробленої системи.

Наукова новизна дисертаційної роботи

У роботі отримано нові наукові результати. Вважаю, що такими новими результатами, отриманими Швороб І.Б., є:

- модель документо-орієнтованого графа, що базується на застосуванні різнотипних вершин та ребер такого графа, що у свою чергу сприяє збереженню зв'язків між екстрактованими структурними одиницями, структурує отримані дані, а також дає змогу використати теорію графів для побудови процедури встановлення зв'язків між елементами документа та знизити надлишковість результатів пошуку за запитом користувача;
- метод побудови текстового шаблону та екстракції даних з текстових блоків виділених за прагматичною ознакою для структурування даних за допомогою документо-орієнтованого графа, що дає змогу врахувати семантику речень.

Ступінь обґрунтованості наукових положень, висновків і рекомендацій, сформульованих у дисертації, їхня достовірність

Автором дисертаційної роботи виконаний аналіз визначеної проблематики, здійснене комплексне теоретичне та практичне обґрунтування шляхів її розв'язання. Обґрунтованість і достовірність наукових результатів, висновків та рекомендацій, викладених в дисертаційній роботі, досягаються ретельним системним аналізом реально існуючих процесів у сфері опрацювання природномовних текстів та в об'єкті дослідження зокрема. Коректне використання методів досліджень та математичного апарату підтверджується результатами аналітичних доведень через математичні перетворення, результатами експериментальної перевірки та імітаційного моделювання, а також практичними результатами, які відображено в актах впровадження.

Практична реалізація та впровадження теоретичних результатів дисертаційної роботи у реальних системах підтверджує достовірність отриманих автором результатів.

Значення одержаних результатів для практики.

Запропонований Швороб І.Б. документо-орієнтований граф використано у реальній системі. На основі розробленої архітектури побудовано та впроваджено мовно-інформаційну систему екстракції слабоструктурованих природномовних текстів та роботи з ними.

Одержані в дисертаційній роботі результати використано під час розроблення прототипу системи та впроваджено у II травматологічному відділенні КМКЛШМД м.Львова - система аналізу інструкцій для медичних препаратів та у ТЗоВ «To-You Sp. Z o.o.» - система аналізу резюме.

Рекомендації щодо використання результатів дисертації

Наукові результати, отримані в дисертації, можуть бути використані для розв'язання практичних задач. Зокрема, це стосується синтезу структури документа за текстовими блоками та прагматичними ознаками для технічної документації, анотованих наукових звітів тощо.

Публікації та апробація результатів дисертаційної роботи

За темою дисертаційної роботи опубліковано 8 наукових праць, у тому числі три статті в іноземних періодичних наукових виданнях, три – у фахових наукових виданнях України, двох – у матеріалах конференцій. Одна наукова стаття індексується в наукометричній базі даних Scopus, дві статті у інших наукометричних базах даних.

Основні теоретичні та практичні результати дисертаційної роботи доповідались і обговорювались на науково-технічних конференціях та семінарах.

Оформлення дисертації та автореферату

Автореферат дисертації достатньо інформативний, його зміст повністю відповідає змісту дисертаційної роботи. Текст дисертації написано грамотною технічною мовою. Дисертація та автореферат викладені логічно, послідовно та коректно. Оформлення автореферату та дисертації повністю відповідає вимогам, рекомендованим Міністерством освіти і науки України.

Зауваження до дисертаційної роботи

1) Доцільно було б приділити більше уваги розробці кількісних критеріїв оцінювання точності, адекватності, ефективності розроблених у роботі методів екстракції та аналізу текстових даних.
2) Не цілком коректним є наведене на сторінці 49 означення даних як інформації, оскільки дані є лише її синтаксичним аспектом, а семантичний та прагматичний аспекти інформації не еквівалентні даним.
3) Термін «слабоструктуровані дані», а також термін «напівструктуровані дані» є надто розмиті, нечіткі, що ускладнює визначення множини текстових даних, до яких можна застосовувати розроблені у роботі моделі, методи та програмні засоби.
4) Також викликає певну засторогу і використання терміну «неструктуровані дані», оскільки абсолютно неструктурованих текстових даних не існує. Наприклад, довільні дані у тексті вже є структуровані, оскільки вони є упорядковані, слідують один за одним. Тому потрібно завжди уточняти, який саме тип структурованості у даних відсутній.
5) Термін «визначення поняття» на сторінці 49 дисертації, доцільно було замінити терміном «означення поняття».
6) Функція f2 для формування множини структурних одиниць (підрозділ 3.1) подана описово без використання строгих математичних методів задання функцій.
7) Розділ 2 містить фрагменти матеріалу, зокрема підрозділ 2.3, який має оглядовий характер, тому його було б доцільно подати у першому розділі дисертації чи додатках.
8) На сторінках 28 та 50 відсутні фрагменти речень, що ускладнює розуміння тесту.
9) У роботі присутні описки, спостерігаються порожні місця на сторінках, наприклад стор. 45, 82.

Відзначені зауваження не впливають на загальну позитивну оцінку дисертаційної роботи.

Висновки

Дисертаційна робота Швороб І.Б. «Методи та засоби екстракції та аналізу слабоструктурованих текстових даних на основі документо-орієнтованого графа» за оформленням відповідає вимогам ДАК України, що пред'являються до дисертаційних робіт. Дисертація написана сучасною науково-технічною мовою, послідовно, логічно і грамотно. Стиль викладення матеріалу забезпечує доступність його сприйняття.

Автореферат дисертації достатньо повно розкриває її зміст.

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

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 і математична лінгвістика.

Офіційний опонент:

професор кафедри комп'ютерних систем та мереж Тернопільського національного технічного університету ім. Івана Пулюя, доктор технічних наук, професор С.А. Лупенко

Підпис професора Лупенка С.А. засвідчую,

Вчений секретар Тернопільського національного технічного університету ім. Івана Пулюя, кандидат технічних наук, доцент Г.М. Крамар