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

Кваліфікаційна наукова
праця на правах рукопису

ДОРОШЕНКО АНАСТАСІЯ ЮРІЇВНА

УДК 004.89:510.635

ДИСЕРТАЦІЯ

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ФАКТОГРАФІЧНИХ ТЕКСТОВИХ РЕСУРСІВ

05.13.06 – інформаційні технології
12 – інформаційні технології

Подається на здобуття наукового ступеня
кандидата технічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____ А. Ю. Дорошенко

Науковий керівник
Шаронова Наталія Валеріївна,
доктор технічних наук, професор

Харків – 2018

АНОТАЦІЯ

Дорошенко А.Ю. Інформаційна технологія інтелектуального аналізу фактографічних текстових ресурсів. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук (доктора філософії) за спеціальністю 05.13.06 «Інформаційні технології» (122 – Комп'ютерні науки). – Національний технічний університет «Харківський політехнічний інститут», Харків, 2019.

Об'єктом дослідження є процес інтелектуального аналізу фактографічної інформації у текстових даних.

Предметом дослідження є моделі та інформаційна технологія пошуку, збору та ідентифікації фактографічної інформації в текстових мережевих ресурсах.

Дисертацію присвячено вирішенню актуальної науково-практичної задачі створення інформаційної технології інтелектуального аналізу фактографічних ресурсів для забезпечення несуперечності та актуальності результатів інтелектуального аналізу фактографічних ресурсів.

У вступі обґрунтовано актуальність теми дисертації, зазначено зв'язок роботи з науковими темами, сформульовано мету і задачі дослідження, визначено об'єкт, предмет і методи дослідження, показано наукову новизну та практичне значення отриманих результатів, наведено інформацію про практичне використання, апробацію результатів та їх висвітлення у публікаціях.

У першому розділі проведено аналіз та класифікацію задач пошуку текстової інформації, здійснено огляд проблем, переваг та недоліків існуючих методів інформаційного пошуку. Проведено аналіз досліджень в області інтелектуального тематичного аналізу текстів, розглянуто задачі класифікації та кластеризації, визначено необхідність досліджень засобів математичного моделювання інформаційної відстані, наведено постановку

задач в області інформаційних технологій обробки фактографічних даних у мережах.

Для користувачів мереж важливим є пошук необхідної інформації за запитом, що відповідає суб'єктивним критеріям, у зв'язку з чим виникає завдання інформаційного пошуку (ІП) (англ. Information Retrieval) як процесу пошуку неструктурованої інформації. Інформація, яка характеризує певний конкретний факт, фактографічну подію або їх сукупність, називається фактографічною. Для того, щоб спростити пошук інформації і зробити його релевантним, застосовується певна кількість методів ІП, але проведений огляд проблем, переваг та недоліків існуючих методів ІП показав, що при їх великої кількості проблема автоматизованого пошуку фактографічної інформації є недостатньо вирішеною. Розглянуто онтологічний підхід, який дозволяє на основі семантичного опису ресурсів знань інформаційного простору специфікувати та побудувати пошуковий образ запиту. Застосування онтології дозволяє підвищити пертинентність ІП, тобто відповідність отриманих ресурсів певній інформаційній потребі. Застосування методів інтеграції та пошуку інформації на основі онтології використовується для забезпечення підтримки управлінських рішень і є засобом представлення семантики.

Другий розділ присвячено аналізу особливостей інтелектуальної обробки фактографічної текстової інформації та обґрунтуванню вибору математичного інструментарію для моделювання процесу обробки фактів, описано базові засади та принципи моделювання лінгвістичної обробки фактографічних ресурсів, а також базові моделі інтелектуальної обробки фактографічної інформації.

Реалізація ефективного пошуку фактографічної інформації вимагає вивчення структури предметної області, знаходження її специфічних семантичних ознак, дослідження процесу пошуку релевантних джерел фахівцем в галузі. Формальне представлення фактографічної інформації

можливе лише на основі моделювання дії фахівця при аналізі повнотекстової інформації та ідентифікації її змісту.

Для цього використовується метод компараторної ідентифікації лінгвістичних об’єктів, який є ефективним засобом опису інтелектуальної діяльності людини. Теорія компараторної ідентифікації дозволяє з’ясовувати внутрішню структуру інформаційних сигналів, вигляд функції перетворення змісту інформації та вигляд предикату, який описує вибір дії фахівцем.

Ще більш абстрактним і потужним інструментом, який використовується для потреб інформатизації, у тому числі для машинного подання й обробки знань, є теорія категорій.

Теорія категорій дає можливість ясно й наочно описувати процеси формування та обробки знань. Із цією метою в роботі використовується алгебра предикатів, засобами якої побудовано предикатну інтерпретацію категорії. Наявність алгебри скінченних предикатів відкриває можливість переходу від алгоритмічного опису інформаційних процесів до опису їх у вигляді рівнянь, які задають відношення між змінними.

Усі змінні в рівнянні рівноправні, при цьому рівняння мають перевагу перед алгоритмами, оскільки дозволяють розрахувати реакцію системи навіть при неповній визначеності вхідних сигналів. Таким чином, у роботі алгебра скінченних предикатів розглядається як інструмент дослідження. Запропоновано використання предикатних категорій для формалізації фактографічної інформації. Розглянуто та побудовано відповідні реляційні моделі семантичних зв’язків елементів фактографічної інформації за допомогою алгебри предикатів.

У третьому розділі наведено еталонну модель аналізу фактографічної інформації, розроблено моделі інформаційного пошуку фактів. Проведено аналіз та класифікацію онтологій з метою використання онтологічного підходу до опису процесів інтеграції фактографічної інформації.

Розглянуто особливості інтелектуального аналізу фактографічної текстової інформації. Наведено класифікацію фактів та етапи виділення

фактів зі слабо структурованої текстової інформації. Запропоновано для опису фактів використання двох типів триплетів: «Суб'єкт→Предикат→Об'єкт» та «Предмет→Атрибут→Значення».

Це дозволяє вилучати поняття зі слабоструктурованих текстових ресурсів і описувати відношення між ними у структурованому вигляді. Запропоновано еталонну модель аналізу фактографічної інформації Здійснено моделювання інформаційного пошуку фактографічної інформації на основі математичної моделі компаратора.

Пошук факту – це пошук у семантичній мережі тексту такої підмережі, яка є ізоморфною до одного з шаблонів. Якщо підмережа знайдена, факт вважається встановленим, після чого здійснюється вилучення сутностей та їх маркування ролями, які задані у відповідних вузлах лінгвістичного опису. Інформаційна система, яка вирішує задачі пошуку та аналізу фактографічної інформації, потребує базу знань, яка відображає основні співвідношення понять у певній предметній галузі.

Відомості про ці відношення можуть бути використані при побудові тезауруса та онтології предметної галузі. За допомогою предикатних категорій множина правил виводу зберігається у вигляді ядр лінійних операторів, а сам механізм формування знань – у вигляді лінійних операторів, представлених за допомогою формул алгебри предикатів. Вирішення задач збору фактографічної інформації базується на моделях інформаційного фактографічного пошуку та екстракції даних. Моделі інформаційного пошуку фактографічної інформації задаються на базі компаратора.

Описано використання онтологій для опису процесів інтеграції фактографічної інформації. Основу предметної області складають онтології, що використовуються для опису знань з певної галузі. Узгодження онтологій є вирішенням проблеми семантичної неоднорідності, що є важливим для наступних завдань: розвиток онтологій; інтеграція схем; інтеграція каталогів; інтеграція даних; відповідь на запити тощо. Особлива увага приділяється

етапу перевірки онтології шляхом побудови семантичних дескрипторів документів та аналізу протиріч, оскільки він є критичним для всієї процедури побудови онтології та є основною відмінністю запропонованого підходу в порівнянні з відомими методами.

Четвертий розділ присвячений практичній реалізації результатів дослідження. Проведено аналіз особливостей практичної реалізації вирішення задачі екстракції фактографічних даних, розглянуто підходи та інформаційні технології вирішення задач парсингу на базі існуючих інформаційних систем. Запропоновано моделі пошуку, екстракції та обробки фактографічних даних на основі комплексу логіко-лінгвістичних моделей. Обробка текстових даних вимагає багато часу, а зберігання її результатів потребує наявності достатнього обсягу пам’яті, що може бути критичним.

Для перевірки необхідного обсягу пам’яті для зберігання важливої інформації треба оцінити один запис у базі. Для визначення необхідної кількості випробувань, результати яких застосовані для розрахунку коефіцієнтів precision, recall, forged, fallout та error, використано засоби математичної статистики та теорії ймовірності.

Розроблено еталонну архітектуру та запропоновано варіант розгортання програмної системи. Розроблено програмні компоненти серверної частини програмної системи, що дозволяє проводити екстракцію даних з торгівельних площадок на основі використання гнучкого конфігурування та предикатної моделі видобування даних. Розроблено та імплементовано програмні компоненти для збору даних на прикладі збору характеристик моделей мобільних телефонів. Проведено тестування розроблених компонентів та доведено їх працездатність для збору даних з трьох різних торгівельних площадок.

Технологія фактографічного пошуку заснована на представленні змісту тексту у формі семантичної мережі, яка містить значимі слова і словосполучення, що зв’язані різними типами синтактико-семантичних зв’язків. Елементарна семантична мережа представляє результат

синтаксичного аналізу та постсинтаксичних трансформацій дерева залежностей між словами у окремих реченнях. Повна семантична мережа тексту є сукупністю окремих семантичних мереж, які відповідають реченням.

У розділі описано застосування напівавтоматичного методу розширення базової онтології для предметної області «радіаційна безпека». Для вирішення проблеми неоднозначності слів використано словник синонімів. Наводяться результати експерименту, виконаного для поповнення онтології новими екземплярами, знайденими в спеціалізованому текстовому корпусі.

У розділі представлено оцінку ефективності та перспективи використання отриманих моделей та методів. Оцінку ефективності здійснено окремо для двох основних задач, які вирішуються у дослідженні: задачі видобування фактів з текстів та задачі видобування фактів разом з їх визначеннями. Для оцінки ефективності використані коефіцієнт точності *Precision* та коефіцієнт повноти *Recall*. Показано перспективи використання запропонованих моделей і методів ідентифікації та обробки предметних знань для індексування повнотекстових документів у задачах інтелектуального пошуку фактографічної інформації за ключовими словами, розробки інформаційної технології створення OLAP-кубів для подання багатовимірного простору знань колекції документів, визначення семантичної близькості на основі когнітивного підходу.

Таким чином, розроблена інформаційна технологія інтелектуального аналізу фактографічної інформації удосконалює та доповнює існуючий підхід обробки текстових даних і не суперечить існуючій практиці, що свідчить про її практичну цінність та ефективність використання.

Ключові слова: інформаційна технологія, фактографічна інформація, метод компараторної ідентифікації, екстракція фактів, онтологічна специфікація.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

1. Дорошенко А. Ю. Построение онтологий и фреймворк информационной системы для создания интеллектуальной системы / А. Ю. Дорошенко, Е. А. Оробинская, О. И. Король // Вісник Херсонського національного технічного університету. – Херсон : ХНТУ, 2013. – № 1 (46). – С. 196–200.

2. Дорошенко А. Ю. Применение масштабных лингвистических ресурсов для расширения онтологии предметной области (на примере области «Радиационная безопасность») / Е. А. Оробинская, Н. В. Шаронова, А. Ю. Дорошенко, Ж.-Ю. Шоша // Східно-Європейський журнал передових технологій. – 2014. – № 5/2 (71). – С. 9–14.

3. Дорошенко А. Ю. Интеллектуальные технологии идентификации фактографической информации / А. Ю. Дорошенко, Е. А. Оробинская, Аджит Пратап Сингх Гаутам // Проблеми інформаційних технологій. – Херсон : ХНТУ, 2014. – № 2 (016). – С. 103–106.

4. Дорошенко А. Ю. Розробка програмних компонентів інформаційної системи екстракції фактографічних даних з веб-ресурсів / А. Ю. Дорошенко, Н. В. Шаронова, Б. О. Єна, О. В. Янголенко // Проблеми інформаційних технологій. – Херсон : ХНТУ, 2018. – № 1 (023). – С. 27–35.

5. Дорошенко А. Ю. Розробка інформаційної технології інтелектуального аналізу фактографічної інформації / А. Ю. Дорошенко // Біоніка інтелекту. – Харків : ХНУРЕ, 2018. – № 1 (90). – С. 116–121.

6. Пат. на корисну модель 63508 Україна, МПК G06F 17/18. Цифровий гібридний медіанний фільтр / А. В. Шостак, А. Ю. Дорошенко, Ю. І. Дорошенко, М. Г. Коробков ; заявник та патентовласник Національний аерокосмічний університет «ХАІ». – № 201103302; заявл. 21.03.2011; опубл. 10.10.2011, Бюл. № 19. – 4 с.

7. Пат. на корисну модель 62818 Україна, МПК G06F 17/18. Пристрій цифрової фільтрації сигналу / А. В. Шостак, А. Ю. Дорошенко, Ю. І. Дорошенко, М. Г. Коробков, О. М. Рисований, А. В. Івашко ; заявник та

патентовласник НТУ «ХПІ». – № 201105823; заявл. 10.05.2011; опубл. 12.09.2011, Бюл. № 17. – 9 с.

8. Дорошенко А. Ю. Формальна модель природної мови як важлива частина прогресивних інформаційних технологій / А. Ю. Дорошенко, Т. О. Богдан // Proceedings of the III International Conference on Computer Science and Information Technologies. – Lviv : Institute of Computer Science and Information Technologies, 2008. – P. 394–396.

9. Дорошенко А.Ю. Использование онтологий для автоматической обработки текстов на естественном языке / А. Ю. Дорошенко, Е. А. Оробинская // Вісник Національного технічного університету «Харківський політехнічний інститут»: Тематичний вип. : Актуальні проблеми розвитку українського суспільства. – Харків : НТУ «ХПІ», 2011. – № 30. – С. 101–106.

10. Doroshenko A. Y. The problem of studying foreign languages in technical universities / A. Y. Doroshenko, A. V. Kovalyova // Integration Processes and Innovative Technologies: Achievements and Prospects of Engineering Sciences. – Kharkiv : Kharkiv National Automobile and Highway University, 2011. – P. 272–275.

11. Дорошенко А. Ю. Використання онтологій для семантичного пошуку документів / А. Ю. Дорошенко // Інтелектуальні системи та прикладна лінгвістика : тези доп. Першої Всеукраїнської науково-практ. конф. – Харків : НТУ «ХПІ», 2012. – С. 8–9.

12. Дорошенко А. Ю. Використання онтологій для семантичного пошуку документів / А. Ю. Дорошенко, Сайед Мохаммад Таухід Сіддікі, Н. В. Шаронова // Вісник Національного технічного університету «Харківський політехнічний інститут». Тематичний вип. : Актуальні проблеми розвитку українського суспільства. – Харків : НТУ «ХПІ», 2012. – № 31. – С. 95–99.

13. Дорошенко А. Ю. Системы обработки патентно-конъюнктурной информации на основе онтологий / Н. В. Шаронова, А. Ю. Дорошенко //

Інформаційні проблеми теорії акустичних, радіоелектронних і телекомунікаційних систем : тези доп. другої Міжнар. наук.-техн. конф. – Харків : НТУ «ХПІ», 2013. – С. 37–38.

14. Дорошенко А. Ю. Обработка фактографической информации для текстовых патентно-конъюнктурных данных при построении онтологий / А. Ю. Дорошенко // Экспертные оценки элементов учебного процесса : материалы XV межвуз. науч.-практ. конф. – Харьков : Изд-во НУА, 2013. – С. 29–31.

15. Дорошенко А. Ю. Система построения онтологий для обработки фактографической информации на примере текстовых патентно-конъюнктурных данных / А. Ю. Дорошенко // Актуальні проблеми прикладної лінгвістики очима наукової молоді : матеріали III регіональної наук.-практ. конф. – Харків : НАУ «ХАІ», 2013. – С. 36–37.

16. Дорошенко А. Ю. Извлечение информации из текстовых сообщений на основе правил EBNF [Електронний ресурс] / О. В. Канищева, А. Ю. Дорошенко // Прикладна лінгвістика та лінгвістичні технології Megaling-2013 : зб. наук. пр. – Режим доступу: http://megaling.ulif.org.ua/tezi-2013-rik/storinka-13.html.

17. Doroshenko A. Issues of Fact-based Information Analysis [Electronic resource] / N. Sharonova, A. Doroshenko, O. Cherednichenko // Proceedings of the International Conference on Computational Linguistics and Intelligent Systems. – 2018. – URL: http://ceur-ws.org/Vol-2136/10000011.pdf. (індексовано в Scopus).

18. Doroshenko A. Towards the Ontology-Based Approach for Factual Information Matching / N. Sharonova, A. Doroshenko, O. Cherednichenko // Інформаційні системи і технології (ICT-2018) : матеріали VII Міжнарод. наук.-техн. конф. – Харків : ХНУРЕ, 2018. – С. 230–233.

SUMMARY

Doroshenko A.Yu. Information technology of intelectual analysis of factual text resources. - Qualifying scientific work as a manuscript.

The dissertation for a candidate degree in technical sciences (PhD), speciality 05.13.06 – Information Technologies (122 - Computer Science). – National Technical University «Kharkiv Polytechnic Institute», Kharkiv, 2019.

The research object is a process of intelectual analysis of factual information in text data.

The subject of the study are models and information technology, selection and identification of factual information in text resources.

The thesis is devoted to solve the actual scientific and practical tasks and creation of information technology for intellectual analysis of factual resources to ensure consistency and relevance of the results of intellectual analysis of factual resources.

The introduction substantiates the relevance of the thesis topic, indicates the connection of work with scientific themes, and formulates the purpose and tasks of the research, defines the object, subject and methods of the research, shows the scientific novelty, and the practical significance of the obtained results, provides information on the results approbation, their testing, and publication.

In the first chapter was described an analysis and classification of search for text information and was presented an overview of the problems, advantages and disadvantages of existing methods. The analysis of the researches was carried out in the field of intellectual thematic analysis of texts, the problems of classification and clustering were considered, the necessity to build up a comprehensive list of information modeling tasks, as well as the setting of tasks in the field of information technologies of factual data processing in networks.

It is important for users of networks to search for the required information by request that answers subjective criteria, that’s why we have a task of Informational Retrieval (IR) as a process of search for unstructured information. Information

that characterizes a specific fact, a factual event, or a combination of them, is called factual.

In order to simplify the search of information and to make it relevant, was used a certain number of IR methods, but an overview of the problems, advantages and disadvantages of existing IR methods has shown that with their large amount, the problem of automated search of factual information is not sufficiently solved. An ontological approach was considered, which allows to specify and construct a search request image based on the semantic description of knowledge resources of the information space. Use of ontology can increase the pertinency of IR, which means the correspondence of the received resources to a certain information. The application of integration and search of information – based on ontologies is used to support managerial decisions and being a means of representing semantics.

The second chapter is devoted to the analysis of the peculiarities of intellectual processing of factual text information and the substantiation of the choice of mathematical tools for modeling the process of processing facts, describes the basic principles and principles of modeling linguistic processing of factual resources, as well as basic models of intellectual processing of factual information.

Implementation of effective search of factual information requires studying the structure of the subject area, finding its specific semantic features, studying the process of finding relevant sources by a specialist in the industry. Formal representation of factual information is possible only on the basis of modeling of the specialist's activity in the analysis of full-text information and identification of its contents.

For this purpose is used the method of comparative identification of linguistic objects, which are effective means of describing intellectual activity of a person. The theory of comparative identification allows us to find out the internal structure of information signals, the form of the function of transforming the content of information and the form of the predicate, which describes the choice of action by a specialist.

An even more abstract and powerful tool used for informatization, including machine presentation and processing of knowledge, is the theory of categories.

The theory of categories makes it possible to describe clearly the processes of formation and processing of knowledge. For this purpose in this work is used the algebra of predicates, with a help of the means of which the predicate interpretation of the category is constructed. The presence of algebra of finite predicates opens the possibility of a transition from algorithmic description of information processes to describing them in the form of equations that specify the relationship between variables.

All variables in the equation are equal, they are also having an advantage over algorithms, since they allow to calculate the system response even with incomplete certainty of input signals. Thus, in the work, the algebra of finite predicates is regarded as a research tool. The use of predicate categories is proposed for the formalization of factual information. The corresponding relational models of semantic connections of elements of factual information are considered and constructed with the help of predicate algebras.

The third chapter provides a reference model for analyzing of factual information, were developed models for information retrieval of facts. The analysis and classification of ontologies were carried out with the aim of using the ontological approach to the description of the processes of integration of factual information.

Features of the intellectual analysis of factual text information were considered. The classification of facts and stages of selection of facts were given from poorly structured text information. It is proposed to describe the use of two types of triplets: «Subject → Relation→ Object» and «Object → Attribute → Value».

This allows to remove the concept of weakly structured text resources and to describe the relationship between them in a structured form. There was also offered the reference model of the analysis of factual information and were carried out modeling of information search of factual information on the basis of a mathematical model of a comparator.

Searching for a fact is a search in a semantic network of the text of a subnet that is isomorphic to one of the templates. If the subnet is found, the fact is considered to be established, after which is carried out the removal of entities and their marking of the poles, which are specified in the corresponding nodes of the linguistic description. An information system that solves the problems of searching and analyzing factual information requires base of knowledge that reflects the basic relationships of concepts in a particular subject area.

Information about these relationships can be used in the construction of the thesaurus and the ontology of the subject field. By means of predicate categories, the set of output rules is stored in the form of the kernels of linear operators, and the mechanism of knowledge formation itself is represented in the form of linear operators by the formulas of predicate of algebras. The decision of the tasks of collecting factual information is based on models of informational factual search and extraction of data. Models of information search of factual information are given on the basis of a comparator.

The use of ontologies was described for the processes of integration of factual information. The basis of the subject area are used ontologies that describe the knowledge of a particular industry. Matching ontologies is a solution to the problem of semantic heterogeneity, which is important for the following tasks: the development of ontologies; integration of schemes; directory integration; data integration; response to requests, etc. Particular attention is paid to the ontology checking stage by constructing semantic document descriptors and analysis of contradictions, since it is critical for the entire ontology construction procedure and is a major difference in the proposed approach compared to known methods.

The fourth chapter is devoted to the practical implementation of the research results. The analysis of peculiarities of practical realization of the decision of the problem of extraction of factual data is carried out, approaches and information technologies of solving parsing problems on the basis of existing information systems are considered. Models of search, extraction and processing of the factual data on the basis of the complex of logical-linguistic models are offered. Processing text data

takes a lot of time, and storing its results requires a sufficient amount of memory that can be critical.

To check the required amount of memory for storing important information we need to evaluate one entry in the database. To determine the required number of tests, the results of which are used to calculate the precision, recall, forged, fallout and error, are used the means of mathematical statistics and probability theory.

In the work was developed the reference architecture and was proposed the variant of deployment of the software system. The software components of the server part of the software system are developed, which allows to make extraction of data from trading platforms based on the use of flexible configuration and predicate model of data mining. Software components for data collection are developed and implemented, on the example of collecting characteristics of mobile phone models. Testing of the developed components has been carried out and their efficiency has been proved to collect data from three different trading platforms.

The factual search technology is based on the substantive content of the text in the form of a semantic network, which contains meaningful words and phrases that are associated with different types of syntactic-cementious relationships. The Elementary Semantic Network provides syntactic analysis of postsyntactic transformations and syntactic relationships between words in the individual sentences. The full semantic network of text is a collection of individual chains that correspond to the sentence.

The chapter describes the use of the semi-automatic method for extending the basic ontology for the "radiation safety" domain. The synonym dictionary is used to solve the ambiguity of words. The results of an experiment performed to replenish the ontology with new instances that are found in a specialized text corps.

The chapter presents an assessment of the effectiveness and prospects of using the models and methods. The performance evaluation is done separately for the two main tasks that are solved in the study: the tasks of extracting facts from texts and the problems of extracting facts together with their definitions. To evaluate the efficiency, the Precision Accuracy and Recall Factor are used. The prospects of using the proposed models and methods of identification and processing of subject knowledge

were shown for indexing of full-text documents in the tasks of intellectual search of factual information by keywords, also were shown the development of information technology for the creation of OLAP cubes for the presentation of a multidimensional knowledge collection of documents, and the definition of semantic proximity based on the cognitive approach.

Thus the information technology that was developed for the intellectual analysis of factual information improves and adds the existing approach to processing of text data and does not contradict existing practice, which testifies about its practical value and efficiency.

Keywords: information technology, factual information, method of comparative identification, fact extraction, ontological specification.

REFERENCES

1. Doroshenko A. Yu. Postroenie ontologiy i freymvork informatsionnoy sistemyi dlya sozdaniya intellektualnoy sistemyi / A. Yu. Doroshenko, E. A. Orobinskaya, O. I. Korol // Visnyk Khersonskoho natsionalnoho tekhnichnoho universytetu. – Kherson : KhNTU, 2013. – № 1 (46). – S. 196–200.

2. Doroshenko A. Yu. Primenenie masshtabnyih lingvisticheskih resursov dlya rasshireniya ontologii predmetnoy oblasti (na primere oblasti «Radiatsionnaya bezopasnost») / E. A. Orobinskaya, N. V. Sharonova, A. Yu. Doroshenko, Zh.-Yu. Shosha // Skhidno-Yevropeiskyi zhurnal peredovykh tekhnolohii. – 2014. – № 5/2 (71). – S. 9–14.

3. Doroshenko A. Yu. Intellektualnyie tehnologii identifikatsii faktograficheskoy informatsii / A. Yu. Doroshenko, E. A. Orobinskaya, Adzhit Pratap Singh Gautam // Problemy informatsiinykh tekhnolohii. – Kherson : KhNTU, 2014. – № 2 (016). – S. 103–106.

4. Doroshenko A. Yu. Rozrobka prohramnykh komponentiv informatsiinoi systemy ekstraktsii faktohrafichnykh danykh z veb-resursiv / A. Yu. Doroshenko, N. V. Sharonova, B. O. Yena, O. V. Yanholenko // Problemy informatsiinykh tekhnolohii. – Kherson : KhNTU, 2018. – № 1 (023). – S. 27–35.

5. Doroshenko A. Yu. Rozrobka informatsiinoi tekhnolohii intelektualnoho analizu faktohrafichnoi informatsii / A. Yu. Doroshenko // Bionika intelektu. – Kharkiv : KhNURE, 2018. – № 1 (90). – S. 116–121.

6. Pat. na korysnu model 63508 Ukraina, MPK G06F 17/18. Tsyfrovyi hibrydnyi mediannyi filtr / A. V. Shostak, A. Yu. Doroshenko, Yu. I. Doroshenko, M. H. Korobkov ; zaiavnyk ta patentovlasnyk Natsionalnyi aerokosmichnyi universytet «KhAI». – № 201103302; zaiavl. 21.03.2011; opubl. 10.10.2011, Biul. № 19. – 4 s.

7. Pat. na korysnu model 62818 Ukraina, MPK G06F 17/18. Prystrii tsyfrovoi filtratsii syhnalu / A. V. Shostak, A. Yu. Doroshenko, Yu. I. Doroshenko, M. H. Korobkov, O. M. Rysovanyi, A. V. Ivashko ; zaiavnyk ta patentovlasnyk NTU «KhPI». – № 201105823; zaiavl. 10.05.2011; opubl. 12.09.2011, Biul. № 17. – 9 s.

8. Doroshenko A. Yu. Formalna model pryrodnoi movy yak vazhlyva chastyna prohresyvnykh informatsiinykh tekhnolohii / A. Yu. Doroshenko, T. O. Bohdan // Proceedings of the III International Conference on Computer Science and Information Technologies. – Lviv : Institute of Computer Science and Information Technologies, 2008. – P. 394–396.

9. Doroshenko A.Yu. Ispolzovanie ontologiy dlya avtomaticheskoy obrabotki tekstov na estestvennom yazyike / A. Yu. Doroshenko, E. A. Orobinskaya // Visnyk Natsionalnoho tekhnichnoho universytetu «Kharkivskyi politekhnichnyi instytut»: Tematychnyi vyp. : Aktualni problemy rozvytku ukrainskoho suspilstva. – Kharkiv : NTU «KhPI», 2011. – № 30. – S. 101–106.

10. Doroshenko A. Y. The problem of studying foreign languages in technical universities / A. Y. Doroshenko, A. V. Kovalyova // Integration Processes and Innovative Technologies: Achievements and Prospects of Engineering Sciences. – Kharkiv : Kharkiv National Automobile and Highway University, 2011. – P. 272–275.

11. Doroshenko A. Yu. Vykorystannia ontolohii dlia semantychnoho poshuku dokumentiv / A. Yu. Doroshenko // Intelektualni systemy ta prykladna linhvistyka : tezy dop. Pershoi Vseukrainskoi naukovo-prakt. konf. – Kharkiv : NTU «KhPI», 2012. – S. 8–9.

12. Doroshenko A. Yu. Vykorystannia ontolohii dlia semantychnoho poshuku dokumentiv / A. Yu. Doroshenko, Saied Mokhammad Taukhid Siddiki, N. V. Sharonova // Visnyk Natsionalnoho tekhnichnoho universytetu «Kharkivskyi politekhnichnyi instytut». Tematychnyi vyp. : Aktualni problemy rozvytku ukrainskoho suspilstva. – Kharkiv : NTU «KhPI», 2012. – № 31. – S. 95–99.

13. Doroshenko A. Yu. Sistemyi obrabotki patentno-kon'yunkturnoy informatsii na osnove ontologiy / N. V. Sharonova, A. Yu. Doroshenko // Informatsiini problemy teorii akustychnykh, radioelektronnykh i telekomunikatsiinykh system : tezy dop. druhoi Mizhnar. nauk.-tekhn. konf. – Kharkiv : NTU «KhPI», 2013. – S. 37–38.

14. Doroshenko A. Yu. Obrabotka faktograficheskoy informatsii dlya tekstovyih patentno-kon'yunkturnyih dannyih pri postroenii ontologiy / A. Yu. Doroshenko // Ekspertnyie otsenki elementov uchebnogo protsessa : materialyi XV mezhvuz. nauch.-prakt. konf. – Harkov : Izd-vo NUA, 2013. – S. 29–31.

15. Doroshenko A. Yu. Sistema postroeniya ontologiy dlya obrabotki faktograficheskoy informatsii na primere tekstovyih patentno-kon'yunkturnyih dannyih / A. Yu. Doroshenko // Aktualni problemy prykladnoi linhvistyky ochyma naukovoi molodi : materialy III rehionalnoi nauk.-prakt. konf. – Kharkiv : NAU «KhAI», 2013. – S. 36–37.

16. Doroshenko A. Yu. Izvlechenie informatsii iz tekstovyih soobscheniy na osnove pravil EBNF [Elektronnyi resurs] / O. V. Kanischeva, A. Yu. Doroshenko // Prykladna linhvistyka ta linhvistychni tekhnolohii Megaling-2013 : zb. nauk. pr. – Rezhym dostupu: http://megaling.ulif.org.ua/tezi-2013-rik/storinka-13.html.

17. Doroshenko A. Issues of Fact-based Information Analysis [Electronic resource] / N. Sharonova, A. Doroshenko, O. Cherednichenko // Proceedings of the International Conference on Computational Linguistics and Intelligent Systems. – 2018. – URL: http://ceur-ws.org/Vol-2136/10000011.pdf. (indeksovano v Scopus).

18. Doroshenko A. Towards the Ontology-Based Approach for Factual Information Matching / N. Sharonova, A. Doroshenko, O. Cherednichenko // Informatsiini systemy i tekhnolohii (IST-2018) : materialy VII Mizhnarod. nauk.-tekhn. konf. – Kharkiv : KhNURE, 2018. – S. 230–233.

ЗМІСТ