УДК 004.4'6

Аронов А. О. *Державний університет телекомунікацій, Київ*

РОЗРОБКА МЕТОДУ АВТОМАТИЗАЦІЇ ВИЯВЛЕННЯ СУПЕРЕЧЛИВОЇ ІНФОРМАЦІЇ НА ОСНОВІ АНАЛІЗУ ДАНИХ СТОРІНОК САЙТУ

Запропонований метод базується на порівнянні змісту сторінок сайту, заданих адміністратором. В якості апарату формалізації обрано представлення процедур у вигляді логічної структури алгоритму Янова, що дозволяє будувати алгоритм зверху донизу, тобто, від загальної стратегії пошуку інформації на сайті до опрацювання окремих блоків. Процес обробки даних включає процедури автоматичного виявлення протиріч і виправлення даних та автоматичного виявлення розбіжностей у змісті сторінок.

***Ключові слова**: розробка сайтів, представлення процедур, алгоритм Янова, логічна структура, аналіз даних сайту, системний адміністратор сайту, виявлення протиріч.*

Aronov A. O. *State University of Telecommunications, Kyiv*

DEVELOPMENT OF A METHOD FOR AUTOMATING DETECTION OF CONFLICTING INFORMATION BASED ON ANALYSIS OF SITE DATA

The method of automated detection of contradictory information based on the information-analytical analysis of site data is proposed. The proposed method is designed to create software that can be used by system administrators of the site. The method is based on a comparison of the content of the site pages set by the administrator, in order to identify and resolve differences. As a formalization structure, representation of procedures in the form of a logical structure of the Yanov's algorithm is chosen, which allows to build an algorithm from the top down, that is, from the general strategy of finding information on the site to the processing of individual blocks. Input data is directly the texts of the pages of the site, standard pages (pages that contains knowingly valid information), a list of search patterns of the keywords, and the result data is – updated texts of the pages of the site. A distinctive feature of the proposed method is that the data processing process includes two main procedures: the procedure for automatically detecting contradictions and correcting data; procedure for automatically detecting discrepancies in the content of pages. The process of finding and comparing of fragments of texts on the pages is performed using search patterns. Depending on the types and location of the pages being compared, a decision is made to automatically make changes or notify the site administrator of necessity to perform manual corrections. The fragments of algorithms for detecting discrepancies, contradictions and corrections of data in the text of pages of a site in the form of the Yanov's algorithm are presented.

***Keywords**: site development, representation of procedures, Yanov's algorithm, logical structure, analysis of site data, system administrators of the site, revealing contradictions.*

Аронов А. А. *Государственный университет телекоммуникаций, Киев*

РАЗРАБОТКА МЕТОДА АВТОМАТИЗАЦИИ ВЫЯВЛЕНИЯ ПРОТИВОРЕЧИВОЙ ИНФОРМАЦИИ НА ОСНОВЕ АНАЛИЗА ДАННЫХ СТРАНИЦ САЙТА

Предложенный метод основан на сравнении содержания страниц сайта, заданных администратором. В качестве аппарата формализации выбрано представление процедур в виде логической структуры алгоритма Янова, что позволяет строить алгоритм сверху вниз, то есть, от общей стратегии поиска информации на сайте к обработке отдельных блоков. Процесс обработки данных включает процедуры автоматического обнаружения противоречий и исправления данных и автоматического обнаружения расхождений в содержании страниц.

***Ключевые слова:** разработка сайтов, представление процедур, алгоритм Янова, логическая структура, анализ данных сайта, системный администратор сайта, выявление противоречий.*

1. Вступ. Аналіз невирішених задач

Достовірність інформації – це її загальна властивість, яка визначає неспотворену частку в копії документа стосовно його оригіналу. Разом з тим, за словами засновника фреймів Д. Мінскі [1], будь яка природно-мовна інформація є надлишковою, суперечливою і невизначеною одночасно. І дійсно, для отримувача інформації певна частка повідомлення є не актуальною (надлишковою), оскільки знання, які вона містить, відомі отримувачеві, певна частка – суперечливою, оскільки не збігається з картиною світу, та невизначеною, тобто такою, відносно достовірності якої отримувач не може дати однозначної відповіді. З одного боку, до змісту інформації, що міститься на сайті висуваються вимоги достовірності, актуальності, несуперечливості, цілісності. З іншого боку, з моменту створення сайту на ньому збирається великий обсяг інформації, яка в тому числі є суперечливою, призводить до проблем ефективного функціонування сайту і значно ускладнює задачу супроводження сайту з боку системного адміністратора, оскільки на сьогодні відсутні методи автоматизації виявлення суперечливої інформації в слабко структурованих природно-мовних текстах.

Розв'язання ***задачі аналізу логічної несуперечливості потоку повідомлень*** пов'язане з цілою низкою проблем, серед яких на першому місці стоїть проблема формалізації тексту, тобто перетворення довільного повідомлення до певного синтаксично однорідного представлення.

Аналіз фахових літературних джерел [2-4] показав, що сучасні підходи до розв'язання задачі аналізу повідомлень здебільшого відтворюють аналітичні операції, які виконує людина, з тією різницею, що при цьому враховується необхідність алгоритмізації та подальшої автоматизації цих операцій. Механізми сценарної синхронізації (розпізнавання знайомого ситуативного контексту), що активізуються при наявності певного соціального досвіду, забезпечують людині високу винахідливість та завадостійкість при сприйманні повідомлень.

В плані автоматизації цих процесів найбільша увага приділяється в інформаційно-аналітичних системах [2, 3]. При аналізі достовірності даних використовуються такі методи: використання примітивних мажоритарних методів; метод експертних оцінок; ранжування джерел даних; інтеграція з об'єктивними даними; залучення методів теорії ігор для аналізу оптимальних стратегій та співставлення вхідних даних про вибір певної стратегії з результатами ігрового моделювання; аналіз ціннісної орієнтації джерел; виявлення базисних процесів реального світу, що впливають на процес генерації суджень (тверджень) та інші. Але слід зазначити, що ці методи безпосередньо не розв'язують задачу виявлення протиріч, а слугують для упорядкування потоку інформації, щоб аналітику облегшити процес прийняття рішення щодо суперечливості і недостовірності інформації. Для того, щоб автоматизована інформаційна система могла автоматично добирати аргументи, вона повинна бути спроможною виділяти з повідомлень твердження, які підлягають перевірці (тобто, твердження, які містять дані, значимі для розв'язання певної задачі), виконувати перетворення сукупності даних у сукупність елементарних тверджень, які співставленні з моделлю фрагментів реального світу, на якій можливо здійснювати перевірку їх істинності та несуперечливості.

Частково проблема виявлення протиріч вирішується в системах машинного перекладу, де багатозначність і невизначеність визнається як частковий випадок протиріччя, оскільки при перекладі в тексті має бути лише перекладний еквівалент [4].

Отже, аналіз фахових джерел показав, шо розв'язання задачі щодо виявлення і прийняття рішення щодо суперечливості на масиві слабко структурованих природно-мовних текстах в плані автоматизації відсутнє навіть у постановочному плані.

Таким чином, актуальною є задача надання системному адміністратору сайту такого інструментарію, який би автоматично/автоматизовано дозволяв би виявляти суперечливу інформацію на сайті та приймати рішення щодо її подальшої долі.

Метою даної роботи є розробка методу автоматизації виявлення суперечливої інформації на основі інформаційно-аналітичного аналізу даних сторінок сайту.

2. Постановка задачі дослідження

Відповідно до поставленої в роботі мети запропоновано метод автоматизації виявлення суперечливої інформації на основі інформаційно-аналітичного аналізу даних сторінок сайту.

Метод призначений для створення програмного забезпечення, яке можуть використовувати системні адміністратори сайту.

Обмеження. Аналіз сайтів ВНЗ, зокрема, офіційного сайту Державного університету телекомунікацій показав, що до певної частини змісту інформації можливо однозначно прийняти рішення щодо її суперечливості, оскільки є визначені достовірні дані (це дані, що стосуються точних назв спеціальностей і спеціалізацій, вартості навчання за контрактною формою навчання, тощо). Для обробки таких даних передбачена процедура автоматичного виявлення протиріч та виправлення даних. Інша частина суперечливих даних на сторінках сайту пов'язана із суб'єктивним людським фактором. Так, наприклад, на різних сторінках в межах однієї кафедри можуть зустрічатися розбіжності щодо прізвищ та посад викладачів; на кафедрі можуть додати назви компаній-партнерів, а на загальну університетську сторінку «Наші партнери» – ні; буває так, що на кафедральних сторінках в різних блоках інформації вказані різні контактні номера телефонів. Для обробки таких даних передбачена процедура автоматичного виявлення розбіжностей у змісті сторінок, рішення щодо до того яка із сторінок є хибною приймає системний адміністратор.

Вхідними даними є безпосередньо тексти сторінок сайту, еталонні сторінки, список пошукових образів ключових слів.

Вихідні дані: оновлені сторінки сайту.

Процес обробки даних включає дві основні процедури:

– процедура автоматичного виявлення протиріч та виправлення даних;
– процедура автоматичного виявлення розбіжностей у змісті сторінок.

Алгоритми обробки базуються на принципах табличної обробки інформації [5, 6]. Обробка інформації представленої у вигляді таблиць має ряд переваг, а саме: наочність, відсутність рекурсій, що важливо при побудові алгоритму. Табличне представлення логічної структури алгоритму в класичній теорії алгоритмів отримало назву логічна структура алгоритму (ЛСА) Янова [2]. ЛСА Янова дозволяє наочно зберігати цілісність алгоритму при додаванні (розширенні) певних процедур. Фактично в таблиці рядок виконує всю послідовність дій відповідно до заданого об'єкта. Індивідуальність об'єкта враховується за рахунок наявності або відсутності певних умов, це дозволяє уникнути рекурсій.

3. Процедура автоматичного виявлення протиріч та виправлення даних

Як вже зазначалося, дана процедура застосовується лише для тієї частини інформації, для якої можливо однозначно прийняти рішення щодо її суперечливості, оскільки є визначені достовірні дані (це дані, що стосуються точних назв спеціальностей і спеціалізацій, вартості навчання за контрактною формою навчання). Вхідними даними процедури є еталонна сторінка сайту (тобто сторінка, яка містить завідомо достовірну інформацію щодо назв спеціальностей/спеціалізацій та вартості навчання. Вихідні дані – оновленні сторінки сайту, в яких виявлено хибну інформацію.

Алгоритм процедури виявлення протиріч та виправлення даних у тексті сторінок сайту Фрагмент алгоритму виявлення протиріч та виправлення даних у тексті сторінок сайту у формі ЛСА Янова представлений в табл. 1.

В таблиці перший стовпчик включає перелік всіх кодів еталонних сторінок сайту. Так, код 1.2.3. відповідає сторінці «Перелік спеціальностей для вступу в університет», а код 1.2.5. – «Вартість навчання». Слід зазначити, що за зміст еталонної сторінки відповідає системний адміністратор, оскільки від цього залежить достовірність роботи всієї процедури.

Фрагмент алгоритму виявлення протиріч та виправлення даних Табл. 1

Що порівнюємо	**З чим порівнюємо**	**За якими ознаками**						**Умова if**	**Дія**
Код еталон. сторінки	**Код сторінки**	L_1	L_2	L_3	L_4	L_5	L_6		
1.2.3.	1.2.2.1.2.1.8.	+	+					$T(L_i)\neq E(L_i)$	P1
		…	…	…	…	…	…	…	P1
1.2.5.	1.2.2.1.2.1.8.			+	+	+			P1
		…	…	…	…	…	…	…	P1

З метою, підвищення ефективності роботи алгоритму (зокрема швидкодії) системний адміністратор може для себе створити власний еталонний зразок, куди включає всі дані, які необхідно перевірити. Це ніяким чином не порушує роботу алгоритму, але дозволить зменшити час обробки даних, оскільки кожна сторінка, яка підлягає аналізу, перевірятиметься лише один раз за всіма параметрами. Від адміністратора сайту лише вимагається відслідковувати всі зміни щодо назв спеціальностей та вартості навчання і вчасно вносити зміни в еталонний зразок.

Другий стовпчик включає коди всіх сторінок сайту, які потенційно можуть містити інформацію подібну еталонній сторінці. Так, в таблиці 1 код сторінки 1.2.2.1.2.1.8. відповідає назві «Запрошуємо на навчання» факультету інформаційних технологій. Слід зазначити, оскільки дана сторінка є стандартизованою для всіх інститутів, факультетів та випускових кафедр університету, то розпізнавання відбувається саме за ідентифікатором сторінки, який є унікальним і містить всю інформацію щодо приналежності сторінки в ієрархічній системі.

Стовпчики 3-8 визначають ознаки (параметри), за якими здійснюється порівняння. Так, L_1– назва спеціальності, L_2 – назва спеціалізації, L_3 – вартість навчання за відповідною спеціальністю за денною формою навчання, L_4 – вартість навчання за відповідною спеціальністю за вечірньою формою навчання, L_5 – вартість навчання за відповідною спеціальністю за заочною формою навчання, L_6 – вартість навчання за відповідною спеціальністю за дистанційною формою навчання.

Слід зазначити, шо параметри L_1 і L_2 на відміну від інших можуть зустрічатися на сторінках не лише в табличній формі, але й безпосередньо у текстовій формі.

Щоб перевірити достовірність назви спеціальності чи спеціалізації у текстовій формі перевіряється контекст. Для цього вводяться пошукові образи «*спеціальн+*» та «*спеціалізаці+*», які представляють собою незмінну частину слів спеціальність та спеціалізація відповідно. Знак + означає, що слово в неповній формі і далі може йти якась послідовність букв. Далі аналізується контекст поруч (після пробілу).

Стовпчик 9 визначає умови порівняння. Якщо відповідний параметр не повністю не накладається на еталонний зразок, то інформація вважається суперечливою і алгоритм переходить до виконання процедури P1 (див. стовпчик 10 табл. 1). Процедура P1 означає, що інформація відповідно до знайденого протиріччя видаляється, а на місце відповідного параметру на сторінці сайту записуються дані з еталонного зразка.

4. Процедура автоматичного виявлення розбіжностей у текстах сторінок сайту

Процедура призначена для виявлення суперечливих даних на сторінках сайту пов’язана із суб’єктивним людським фактором. Так, наприклад, в різних сторінках в межах однієї кафедри можуть зустрічатися розбіжності щодо посад викладачів, їх прізвищ, на кафедрі можуть додати назви компаній-партнерів, а на загальну університетську сторінку «Наші партнери» – ні, буває так, що на кафедральних сторінках в різних сторінках вказані

різні контактні номера телефонів. Зазначена процедура здатна автоматично виявляти розбіжності у змісті сторінок, але рішення щодо до того яка із сторінок є хибною приймає системний адміністратор.

Структура алгоритму виявлення протиріч та виправлення даних у тексті сторінок представлена в табл. 2.

Фрагмент алгоритму виявлення розбіжностей текстових даних сайту Табл. 2

Що порівнюємо	**З чим порівнюємо**	**Параметри порівняння**				**Умова**	**Дія**
Код сторінки	**Код сторінки**	C_1	C_2	**…**	C_n		
1.2.2.1.2.2.3.1.1.	1.2.2.1.2.2.3.1.2.	+		…	…	$T_1(L_i)\neq T_2(L_i)$	P2
1.2.2.1.2.2.3.1.1.	1.2.2.1.2.2.3.1.9.	+		…	…	$T_1(L_i)\neq T_2(L_i)$	P2
1.2.2.1.2.2.3.1.1.	1.2.2.2.1.7.		+	…	…	$T_1(L_i)\neq T_2(L_i)$	P2
…	…	…	…	…	…	…	…

Особливістю даного алгоритму є те, що еталонна сторінка відсутня, тобто ми не можемо визначити достовірну інформацію, тому алгоритм порівнює зміст текстів сторінок за заданими (системним адміністратором сайту) параметрами і там, де виявлені розбіжності, повідомляє адміністратору обидва варіанти фрагмента тексту (дія P2 з табл. 2). Системний адміністратор сам приймає рішення, яка інформація є хибною і корегує її в ручному режимі.

Так в табл. 2 код сторінки 1.2.2.1.2.2.3.1.1. відповідає сторінці «Загальна інформація» кафедри комп'ютерних наук факультету інформаційних технологій навчально-наукового інституту телекомунікацій та інформатизації; код 1.2.2.1.2.2.3.1.2. – «Науково-педагогічний склад» кафедри комп'ютерних наук; 1.2.2.1.2.2.3.1.9. – «Новини та події» кафедри комп'ютерних наук; 1.2.2.2.1.7. – «Наші партнери» навчально-наукового інституту телекомунікацій та інформатизації.

Стовпчики 3-6 визначають параметри, за якими порівнюються текстові дані, їх кількість визначає системний адміністратор. Так наприклад, параметр C_1 визначає посади й прізвища педагогічного складу, а C_2 – компанії-партнери кафедри. Для пошуку необхідних даних у тексті кожний параметр має перелік ключових слів. Так параметр C_1 може мати такий набір:

C_1=:<завідувач кафедри / завідувача кафедри / завідувачем кафедри / професор / професора / професором / доцент /доцента /доцентом> тощо.

Відповідно до пошукових образів виявляються і порівнюються фрагменти текстів на сторінках (див. стовпчики 1,2 табл.2), які апріорі задає адміністратор сайту.

5. Висновки

Аналіз фахових джерел показав, шо розв’язання задачі щодо виявлення і прийняття рішення щодо суперечливості на масиві слабко структурованих природно-мовних текстах в плані автоматизації відсутнє навіть у постановочному плані. Тому актуальною є задача надання системному адміністратору сайту такого інструментарію, який би автоматично чи автоматизовано дозволяв би виявляти суперечливу інформацію на сайті та приймати рішення щодо її подальшої долі.

Запропоновано метод автоматизації виявлення суперечливої інформації на основі інформаційно-аналітичного аналізу даних сторінок сайту. Відмінною рисою запропонованого методу є те, що процес обробки даних включає дві основні процедури: процедура автоматичного виявлення протиріч та виправлення даних; процедура автоматичного виявлення розбіжностей у змісті сторінок.

Список використаної літератури

1. Мински Д. Фреймы для представления знаний / Д. Мински. – Москва: Мир, 1979. – 324 с.

2. Курносов Ю. В. Аналитика: методология, технология и организация информационно-аналитической работы / Ю. В. Курносов, П. Ю. Конотопов. – Москва: Русаки, 2004. – 550 с.

3. Юрченко А. И. Разработка автоматизированных информационно-аналитических систем в сфере науки / А. И. Юрченко // Инноватика и экспертиза. – 2011. – Выпуск 2(7). – С. 109-115.

4. Балабін В. В., Замаруєва І. В. Автоматизація когнітивного розпізнавання текстових об'єктів в умовах багатозначності і невизначеності / В. В. Балабін, І. В. Замаруєва // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка». – 2007. – №6. – С. 76-84.

5. Балашов Е. П. Информационные системы. Табличная обработка информации / Е. П. Балашов, В. Н. Негода, Д .В. Пузанков и др. – Ленинград: Энергоатомиздат,1985. – 184 с.

6. Криницкий Н. А. Программирование и алгоритмические языки / Н. А. Криницкий, Г. А. Миронов, Г. Д. Фролов. – Москва: Наука, 1979. – 509 с.

References

1. Minski D. "Frames for representation of knowledge." *Moskva: Mir* (1979): 324.

2. Kurnosov Yu. V., Konotopov P. Yu. "Analytics: methodology, technology and organization of information-analytical work." *Moskva: Rusaki* (2004): 550.

3. Yurchenko A. I. "Development of automated information and analytical systems in the field of science." *Innovatika i expertiza* 2(7) (2011): 109-115.

4. Balabin V. V., Zamaruieva I. V. "Automation of cognitive development of text objects in conditions of polysemy and uncertainty." *Collection of scientific works of the Military Institute of Kyiv National Taras Shevchenko University* 6 (2007): 76-84.

5. Balashov E. P., Negoda V. N., Puzankov D. V.and others. "Information systems. Table processing of information." *Leningrad: Energoatomizdat* (1985): 184.

6. Krinitsky N. A., Mironov G. A., Frolov G. D. "Programming and algorithmic languages." *Moskva: Nauka* (1979): 509.

Автор статті

Аронов Андрій Олексійович – аспірант, Державний університет телекомунікацій, Київ. Тел.: +380 (91) 900 08 00. E-mail: info.dut.edu.ua@gmail.com.

Author of the article

Aronov Andrii Oleksiiovych – post graduate student, State University of Telecommunications, Kyiv. Tel: +380 (91) 900 08 00. E-mail: info.dut.edu.ua@gmail.com.

Дата надходження в редакцію: 17.01.2018 р.

Рецензент: доктор технічних наук, професор В. В. Вишнівський *Державний університет телекомунікацій*