Information systems research

УДК 004.7 **doi: 10.20998/2522-9052.2018.1.05**

А. О. Аронов, В. В. Вишнівський, І. В. Замаруєва

Державний університет телекомунікацій, Київ, Україна

МЕТОД АВТОМАТИЗАЦІЇ ВИЯВЛЕННЯ ЗАСТАРІЛОЇ ІНФОРМАЦІЇ НА ОСНОВІ ІНФОРМАЦІЙНО-АНАЛІТИЧНОГО АНАЛІЗУ ДАНИХ САЙТУ

Предметом вивчення в статті являються процеси обробки даних, що розміщені на сайті для вирішення прикладного завдання автоматичного виявлення застарілої інформації. **Метою** являється підвищення ефективності роботи системного адміністратора сайту шляхом надання йому такого інструментарію, який би автоматично дозволяв виявляти застарілу інформацію на сайті та приймати рішення щодо її подальшої долі (видаляти, архівувати, переміщувати у спеціальні розділи тощо). **Завдання:** розробка методу автоматичного виявлення застарілої інформації на основі інформаційно-аналітичного аналізу даних сайту, який би надав системному адміністратору можливість подальшого автоматичного опрацювання даних сайту. Використовуваним **методом** є: інформаційно-аналітичний метод аналізу даних сайту, що представлений у вигляді моделі табличного представлення логічної структури алгоритму Янова. Отримані такі **результати.** Згідно табличної моделі процесу обробки інформації сформульовано завдання автоматичного виявлення застарілої інформації. В основу процедури автоматичного виявлення застарілої інформації покладено логічну структуру алгоритму Янова. Дана структура дозволяє наочно зберігати цілісність алгоритму при додаванні (розширенні) певних процедур. Індивідуальність об'єкта враховується за рахунок наявності або відсутності певних умов, що дозволяє уникнути рекурсій. В результаті отримано еталонні моделі (шаблони), які перетворюють текстові дані до єдиного уніфікованого представлення. Дані моделі розроблено для форматів, що характерні для інформаційних листів-повідомлень та новин. **Висновки.** Наукова новизна отриманих результатів полягає в наступному: ми розробили метод автоматичного виявлення застарілої інформації на основі інформаційно-аналітичного аналізу даних сайту, який відрізняється від існуючих тим, що для виявлення застарілої інформації аналізуються не лише часові показники часу створення/оновлення сторінок сайту, а безпосередньо зміст текстової сторінки. Для аналізу текстової інформації побудовані шаблони, які дозволяють автоматизувати процес виявлення застарілої інформації та оновлення сторінок сайту.

Ключові слова: сайт; застаріла інформація; інформаційно-аналітичний аналіз даних; логічна структура алгоритму Янова.

Вступ та постановка завдання

Разом із стрімким зростанням кількості сайтів збільшується й обсяг накопичуваної інформації на них, що приводить до проблем їх ефективного функціонування. З одного боку, до змісту інформації, що міститься на сайті висуваються вимоги достовірності, актуальності, несуперечливості та цілісності. З іншого боку, з моменту створення сайту на ньому збирається великий обсяг інформації, яка втрачає актуальність [1. 2]. Актуальність інформації — це відповідність теми інформації повідомлення потребам користувачів сайту. Таким чином, актуальність інформації є динамічною властивістю інформації, яка визначається не змістом інформації, а часом її існування.

Власне за якість інформації відповідає системний адміністратор сайту. В його обов'язки входить перевірка інформації (даних) на коректність, своєчасне оновлення інформації, видалення (або архівація) застарілої інформації. В той же час, із-за великої кількості сторінок (обсяг обліковується кількома сотнями або тисячами сторінок), часто стає неможливим прослідкувати, які сторінки оновлювалися, а які ще ні. Причина цього криється у додаванні нової інформації або ж корегуванні помилок на сторінках тим самим формально унеможливлюючи визначення застарілої інформації з використанням полів «дата останнього редагування». На подібних сторінках може бути повністю відсутні також часові маркери «дати», «числа», «року» тощо. Крім, того складність полягає ще в тому, що фіксування дати безпосередньо у змісті сторінки, яка може визначати застарілість інформації, виражається природномовними засобами і не має уніфікованого подання, що ускладнює процес автоматизації виявлення застарілої інформації [3].

Таким чином, актуальним являється **завдання** розробки методу автоматичного виявлення застарілої інформації на основі інформаційно-аналітичного аналізу даних сайту, який би надав системному адміністратору можливість подальшого автоматичного опрацювання даних сайту.

Метою роботи являється підвищення ефективності роботи системного адміністратора сайту шляхом надання йому такого інструментарію, який би автоматично дозволяв виявляти застарілу інформацію на сайті та приймати рішення щодо її подальшої долі (видаляти, архівувати, переміщувати у спеціальні рубрики тощо).

Основна частина

Відповідно до поставленої в роботі мети необхідно розробити метод автоматичного виявлення застарілої інформації на основі інформаційно-аналітичного аналізу даних сайту. Метод призначений для створення програмного забезпечення, яке можуть використовувати системні адміністратори сайту.

Обмеження: метод можна використовувати для обробки сторінок сайту, де часові характеристи-

ки повністю або обмежено неявно представлені в тексті сторінки.

Вхідними даними є безпосередньо тексти сторінок сайту та база даних еталонних моделей форматів представлення часових показників у тексті.

Вихідні дані: оновлені сторінки сайту.

Процес обробки даних базується на принципах табличної обробки інформації [4, 5]. 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 [6]. Логічна структура алгоритму Янова дозволяє наочно зберігати цілісність алгоритму при додаванні (розширенні) певних процедур. Фактично рядок в таблиці виконує всю послідовність дій відповідно до заданого об'єкта. Індивідуальність об'єкта враховується за рахунок наявності або відсутності певних умов, це дозволяє уникнути рекурсій. Структура алгоритму опрацювання застарілої інформації представлена в табл. 1.

Таблиця 1. **Фрагмент алгоритму опрацювання застарілої інформації на сайті у формі логічної структури алгоритму Янова**

Код сторінки	За якими ознаками			З чим порівнюємо			Умова if	Дія
	L_1	L_2	L_3	K_1	K_2	K_3		
1.7.	+		+	+			$T(L_3) < K_1$	$P1$
1.5.1.2.			+		+		$T(L_3) < K_2$	$P1$
1.4.5.	+			+			$T(L_1) < K_1$	$P1$
1.8.			+			+	$T(L_3) < K_3$	$P3$
1.4.6.1.		+			+		$T(L_2) < K_2$	$P2$

В табл. 1 перший стовпчик включає перелік всіх кодів сторінок сайту, які підлягають аналізу. Кожна сторінка має унікальний ключ (код) в межах сайту. Система кодування побудована на основі ієрархічного класифікатора організації даних сайту, який детально описаний в [7]. Так, код 1.7. відповідає сторінці «Новини»; 1.5.1.2. – «Склад студ. ради»; 1.4.5. – «Конференції»; 1.8. – «Фотогалерея»; 1.4.6.1. – «Аспірантура. Правила прийому». Наступні три стовпці (2-4) визначають параметри, за якими перевіряється відповідна сторінка. Так параметр L_1 означає, що пошук застарілої інформації відбувається безпосередньо за текстом сторінки. Параметр L_2 – за назвою сторінки. Параметр L_3 – за часом створення/оновлення сторінки. Стовпці 5-7 визначають з чим ми порівнюємо часові показники $L_{1\text{-}3}$. Так параметр K_1 – поточна дата; K_2 – фіксована дата (наприклад, вартість навчання, документи до вступу визначаються датою роботи приймальної комісії); K_3 – комбінована дата (наприклад, поточна дата в тексті + 1місяць, тощо). 8-й стовпчик визначає умови порівняння. 9-й стовпчик визначає дії (процедури), які необхідно виконати у разі, якщо правило до відповідної сторінки сайту спрацювало. Процедура ***P*1** означає надіслати сторінку до архіву, ***P*2** – видалити текст сторінки, ***P*3** – нагадати про необхідність оновлення (як правило, це відноситься до даних, які мають зберігатися в архівах до певної фіксованої дати. Наприклад, відомості про конференції, семінари тощо.

Слід зазначити, що часові показники L_1-L_3 не мають уніфікованого представлення, а звідси для програмної реалізації алгоритму потрібно задати певні еталонні моделі (шаблони), які б перетворювали текстові дані до єдиного уніфікованого представлення. Як вже зазначалося, ЛСА Янова дозволяє без порушення загальної логічної структури деталізувати окремі комірки табличного представлення алгоритму. Розглянемо детальніше правила уніфікації часових показників, при цьому всі часові показники приводяться до виду представлення поточної дати в комп'ютері K_1. Аналіз текстових даних сайту (зокрема, сайту Державного університету телекомунікацій) показав, що часові показники визначаються таким чином:

***G*1:** = <число> + <місяць> + <рік>
(наприклад: «**11 квітня 2018**»);

***G*2:** = <число$_1$-число$_2$> + <місяць> + <рік>
(наприклад: «**12-13 квітня 2018**»);

***G*3:** = <число$_1$-число$_2$> + <місяць>
(наприклад: «**4-8 квітня** *відбувся* ..»);

***G*4:** = <місяць>
(наприклад: «*у* **вересні** *відбувся* ..»).

Примітка: знак «+» означає, що дата не розривається контекстом.

Формати ***G*1** і ***G*2** характерні для інформаційних листів-повідомлень про конференції, семінари тощо. Особливістю представлення цих даних є те що вони можуть бути представлені, як правило, трьома мовами: українською, англійською, російською, а тому <місяць> також може містити назви 3-ма мовами. Формати ***G*3** і ***G*4**, як правило, притаманні новинам і представлені українською та російською мовами. Далі розглянемо правила уніфікації для кожного з визначених форматів текстового представлення часових показників.

Формат ***G*1:**

<{1/2/…/31}>+<{січня/января/January}>+
<201{6/7/8/9}>:=<X_1.01.201X_3>;

<{1/2/…/29}>+<{лютого/февраля/February}>+
<201{6/7/8/9}>:=<X_1.02.201X_3>;

…

<{1/2/…/31}>+<{грудня/декабря/December}>+
<201{6/7/8/9}>:=<X_1.12.201X_3>.

Таким чином, для формату ***G*1** визначається 12 шаблонів (еталонних моделей) відповідно до кількості місяців в році. В шаблоні позначення «**<…>**» означає запис на якому відбувається порівняння; фігурні дужки означають множину можли-

вих даних; знак «/» відповідає логічній операції «*або*»; позначення «:=» – операції «приймає значення»; $\boldsymbol{X_1}$ – результат операції перетинання множини чисел із першого кортежу з числом, яке зустрілося у тексті; $\boldsymbol{X_3}$ – результат операції перетинання множини чисел із третього набору даних з числом, яке зустрілося у тексті.

Так, для фрагменту текстового інформаційного листа-повідомлення: «*International Conference CoLInS 2017 Computational Linguistics and Intelligent Systems* ***21 April, 2017,*** *Kharkiv, Ukraine»,* результат накладення на шаблон буде такий вигляд:

<21>+<April>+<2017>:=< 21.04.2017>.

Формат ***G2*** також буде мати 12 шаблонів:

<{1/…/30}-{1/…/31}>+
<{січня/января/January}>+
<201{6/7/8/9}>:=<(-)$\boldsymbol{X_1}$.01.201$\boldsymbol{X_3}$>;

<{1/…/28}-{1/…/29}>+
<{лютого/февраля/February}>+
<201{6/7/8/9}>:=<(-)$\boldsymbol{X_1}$.02.201$\boldsymbol{X_3}$>;

…

<{1/…/30}-{1/…/31}>+
<{грудня/декабря/December}>+
<201{6/7/8/9}>:=<(-)$\boldsymbol{X_1}$.12.201$\boldsymbol{X_3}$>.

У даному шаблоні позначення «**(-)**$\boldsymbol{X_1}$» означає, що в якості результату операції перетинання обирається число із першого набору даних, яке розташоване після знаку «-» в тексті.

Наприклад: ***12-13 квітня 2018*** *року в Державному університеті телекомунікацій відбудеться 10-та міжнародна науково-практична конференція "Проблеми інформатизації", присвячена польоту першого космонавта світу Гагаріна Ю.А.*

Формат ***G3*** характерний для внутрішніх новин, він має також 12 шаблонів:

<{1/…/30}-{1/…/31}>+<січня>:= <(-)$\boldsymbol{X_1}$.01.$\boldsymbol{X_n}$>;

<{1/…/28}-{1/…/29}>+<лютого>:= <(-)$\boldsymbol{X_1}$.02.$\boldsymbol{X_n}$>;

…

<{1/…/30}-{1/…/31}>+<грудня>:= <(-)$\boldsymbol{X_1}$.12.$\boldsymbol{X_n}$>.

Відмінною особливістю даного формату є те, що оскільки рік відсутній або розірваний контекстом безпосередньо у тексті, тому рік визначається за поточною датою (визначеною у комп'ютері або за часом створення документу) і вставляється у результат (позначення «$\boldsymbol{X_n}$» у шаблоні.

Приклад: «*Реєстрація на вебінар обов'язкова і буде відкрита до* ***18 квітня*** *за посиланням*».

Відмінною особливістю формату ***G4*** є відсутність прямого посилання в тексті на такі часові ознаки, як: число і рік, відновлюються з поточної дати або дати створення файлу.

Даний формат як і попередній також характерний для представлення так званих внутрішніх подій і має 12 шаблонів:

<січні>:= <$\boldsymbol{XX_n}$.01.$\boldsymbol{XXXX_n}$>;

<лютому>:= < $\boldsymbol{XX_n}$.02.$\boldsymbol{XXXX_n}$>;

…

<грудні>:= < $\boldsymbol{XX_n}$.12.$\boldsymbol{XXXX_n}$>.

В шаблоні назва місяця ставиться тільки у місцевому відмінку (тобто має відповідати на питання *коли*, це потрібно для того щоб відокремити тексти, які не мають відношення до актуальності інформації в тексті.

Висновки

В статті запропоновано метод автоматичного виявлення застарілої інформації на основі інформаційно-аналітичного аналізу даних сайту, який відрізняється тим, що для виявлення застарілої інформації аналізуються не лише часові показники часу створення/оновлення сайту, а безпосередньо зміст текстової сторінки. Для аналізу текстової інформації побудовані шаблони, які дозволяють автоматизувати процес виявлення застарілої інформації та оновлення сторінок сайту.

СПИСОК ЛІТЕРАТУРИ

1. Gelenbe E. Analysis and synthesis of computer systems (2nd Edition) / E. Gelenbe, G. Pujolle // Advances in Computer Science and Engineering : Texts – Vol. 4. – 2010. – 309 p.
2. Кучук Г. А. Інформаційні технології управління інтегральними потоками даних в інформаційно-телекомунікаційних мережах систем критичного призначення / Г. А. Кучук. – Х.: ХУПС, 2013. – 264 с.
3. Гаврилова Т. А. Базы знаний интеллектуальных систем / Т. А. Гаврилова, В.Ф. Хорошевский. – С.Пб.: Питер, 2000. – 384 с.
4. Табличная обработка информации / Е. П. Балашов, В. Н. Негода, Д. В. Пузанков и др. – Л.: Энергоатомиздат, 1985. – 184 с.
5. Криницкий Н. А. Программирование и алгоритмические языки / Н. А. Криницкий, Г. А. Миронов, Г. Д. Фролов. – М.: Наука, 1979. – 509 с.
6. Аронов А.О. Розробка моделі структурно-логічного представлення даних сайту вищого навчального закладу на основі ієрархічного класифікатора / А.О. Аронов // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2018. – Вип. 59. – С. 24-28.
7. Замаруєва І.В. Автоматизація аналізу змісту природно-мовних текстів як шлях забезпечення безпеки прийняття управлінських рішень / І.В. Замаруєва, О.В. Барабаш, І.В. Пампуха // Наукові записки Українського науково-дослідного інституту зв'язку: науково-виробничий збірник. – К.: УНДІЗ, 2017. – № 3(47). – С. 33-41.

REFERENCES

1. Gelenbe, E. and Pujolle, G. (2010), "Analysis and synthesis of computer systems", *Advances in Computer Science and Engineering*, Vol. 4, 309 p.

2. Kuchuk, G.A. (2014), Information Technologies for Integrated Data Flow Control in Information and Telecommunication Networks of Critical Purposes, KhUPS, Kharkiv, 264 p.
3. Gavrilova, T.A. and Khoroshevskiy, V.F. (2000), Bazy znaniy *intellektual'nykh system* [Knowledge bases of intellectual systems], Piter, St. Petersburg, 384 p.
4. Balashov, E.P., Negoda, V.N. and Puzankov D.V. (1985), *Table processing of information*, Energoatomizdat, Leningrad, 184 p.
5. Krinitsky, N.A., Mironov, G.A. anf Frolov G.D. (1979), *Programming and algorithmic languages*, Nauka, Moscow, 509 p.
6. Aronov, A.O. (2018), "Development of the model of structural-logical representation of the data of the site of the higher educational institution on the basis of the hierarchical classifier", *Collection of scientific works of the Military Institute of the Taras Shevchenko National University of Kyiv*, VIKNU, Kyiv, No. 59, pp. 24-28.
7. Zamarueva, I.V., Barabash, O.V. and Pampukha, I.V. (2017), "Automation of the analysis of the content of natural-language texts as a way to ensure the safety of the adoption of management decisions", *Scientific notes of the Ukrainian Research Institute of Communication*, UNIIZ, Kiev, No. 3 (47), pp. 33-41.

Надійшла (received) 23.02.2018
Прийнята до друку (accepted for publication) 11.04.2018

Метод автоматизации обнаружения устаревшей информации на основе информационно-аналитического анализа данных сайта

А.А. Аронов, В.В. Вишневский, И.В. Замаруева

Предметом изучения в статье являются процессы обработки данных, размещенные на сайте для решения прикладной задачи автоматического обнаружения устаревшей информации. **Целью** является повышение эффективности работы системного администратора сайта путем предоставления ему такого инструментария, который бы автоматически позволял выявлять устаревшую информацию на сайте и принимать решение о ее дальнейшей судьбе (удалять, архивировать, перемещать в специальные разделы и т.д.). **Задача**: разработка метода автоматического обнаружения устаревшей информации на основе информационно-аналитического анализа данных сайта, который бы предоставил системному администратору возможность дальнейшей автоматической обработки данных сайта. Используемым **методом** является: информационно-аналитический метод анализа данных сайта, представленый в виде модели табличного представления логической структуры алгоритма Янова. Получены следующие **результаты**. Согласно табличной модели процесса обработки информации сформулирована задача автоматического обнаружения устаревшей информации. В основу процедуры автоматического обнаружения устаревшей информации положена логическая структура алгоритма Янова. Данная структура позволяет наглядно сохранять целостность алгоритма при добавлении (расширении) определенных процедур. Индивидуальность объекта учитывается за счет наличия или отсутствия определенных условиях, что позволяет избежать рекурсий. В результате получены эталонные модели (шаблоны), которые превращают текстовые данные к единому унифицированному представлению. Данные модели разработаны для форматов, которые характерны для информационных писем-сообщений и новостей. **Выводы**. Научная новизна полученных результатов заключается в следующем: мы разработали метод автоматического обнаружения устаревшей информации на основе информационно-аналитического анализа данных сайта, который отличается от существующих тем, что для выявления устаревшей информации анализируются не только временные показатели времени создания / обновления страниц сайта, а непосредственно содержание текстовой страницы. Для анализа текстовой информации построены шаблоны, которые позволяют автоматизировать процесс выявления устаревшей информации и обновления страниц сайта.

Ключевые слова: сайт; устаревшая информация; информационно-аналитический анализ данных; логическая структура алгоритма Янова.

Method for automation of detecting detailed information on the basis of information and analytical analysis of the site data

A. Aronov, V. Vyshnivsky, I. Zamarueva

The **subject matter** of the article is the data processing processes posted on the site for solving the applied problem of automatic detection of outdated information. The **goal** is to increase the efficiency of the work of system administrator site by providing him with a tool that automatically allows to identify outdated information on the site and decide its future fate (delete, archive, move to special sections, etc.). The **tasks** to be solved are: to develop a method for automatically detecting outdated information based on information analysis of site data that would provide the system administrator with the ability to further automatically process site data. The **method** used is: an information-analytical method for analyzing site data, presented in the form of a table representation model of the logical structure of the Yanov's algorithm. The following **results** were obtained. According to the tabular model of information processing, the task of automatic detection of outdated information was formulated. The basis of the procedure for automatic detection of outdated information is the logical structure of the Yanov's algorithm. This structure allows visually to preserve the integrity of the algorithm when adding (expanding) certain procedures. The individuality of the object is accounted for by the presence or absence of certain conditions, which avoids recursion. As a result, we obtained a reference models (templates), which convert text data to a single unified representation. These models are designed for formats that are typical for letters-messages and news. **Conclusions**. The scientific novelty of the results obtained is as follows: we have developed a method for automatically detecting outdated information based on information analysis of site data that differs from existing ones in that not only time indicators for creating / updating pages of a site are analyzed, but the all content of text page. For the analysis of text information templates were build, which allow to automate the process of revealing outdated information and updating the pages of the site.

Keywords: web-site; outdated information; information analytical analysis of data; logical structure of Yanov's algorithm.