УДК 510.23

А.О. Осідач, аспірант
Національний університет «Львівська політехніка», м. Львів, Україна

МАТЕМАТИЧНА МОДЕЛЬ ЕЛЕКТРОННОГО ДОКУМЕНТА

А.О. Осидач, аспирант
Национальный университет «Львовская политехника», г. Львов Украина

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ЭЛЕКТРОННОГО ДОКУМЕНТА

Andrii Osidach, PhD student
National University "Lviv Polytechnic", Lviv, Ukraine

MATHEMATICAL MODEL OF ELECTRONIC DOCUMENT

Запропоновано математичну модель електронного документа, яка ґрунтується на застосуванні логічних областей, що дозволяє розробляти методи оброблення різнокласових електронних документів у сучасних системах електронного документообігу. Закінчена модель документа складається з двох частин: фізичної і логічної структур. Фізична структура групує фізичні об'єкти в документі; логічна структура документа відображає його логічну організацію. Між фізичною і логічною структурами немає однозначної відповідності, але їх окремі елементи можуть знаходитися у прямій залежності один від одного. На основі цього було проведено моделювання структури електронних документів, а також формалізовано їх елементи для застосування цих моделей до постійно змінюваної корпоративної документації.

***Ключові слова**: математична модель, електронний документ, електронний документообіг, структура електронного документа.*

Предложена математическая модель электронного документа, основанная на применении логических областей, которая позволяет разрабатывать методы обработки разноклассовых электронных документов в современных системах электронного документооборота. Законченная модель документа состоит из двух частей: физической и логической структур. Физическая структура группирует физические объекты в документе; логическая структура документа отражает его логическую организацию. Между физической и логической структурами не существует однозначного соответствия, но их отдельные элементы могут находиться в прямой зависимости друг от друга. На основе этого было проведено моделирование структуры электронных документов, а также формализовано их элементы для применения этих моделей в постоянно меняющейся корпоративной документации.

***Ключевые слова**: математическая модель, электронный документ, электронный документооборот, структура электронного документа.*

In the article the mathematical model of an electronic document based on the use of logical areas, which allows us to develop methods for processing electronic documents different class, in modern systems of electronic document. Completed model of the document consists of two parts: the physical and logical structures. The physical structure includes physical objects in the document; the logical structure of the document reflects its logical organization. Between the physical and logical structures there is no-one correspondence, but some elements may be in direct proportion to each other. On the basis of this article was to simulate the structure of electronic documents, as well as formalized their elements for the application of these models in the ever-changing corporate documentation.

***Key words**: mathematical model, electronic document, electronic document management, electronic document structure.*

Аналіз останніх досліджень і публікацій. Відповідно до отриманого зі стандартів [1] визначення, електронний документ є сукупністю даних у пам'яті обчислювальної системи, призначений для сприйняття людиною за допомогою відповідних програмних і апаратних засобів. Електронний документ може включати текстову, графічну і звукову інформацію, мати нелінійну структуру; різні користувачі можуть переглядати його в різній формі і змінювати його.

Електронний документ може бути розглянутий у вигляді сукупності двох структур: фізичної і логічної [2]. Між фізичною і логічною структурами немає однозначної відповідності, але їх окремі елементи можуть знаходитися у прямій залежності один від одного.

Згідно з мірою структурованості всі корпоративні електронні документи можуть бути розділені таким чином:

– клас документів зі строгою фіксованою структурою S_F. До цього класу відносяться документи, що мають фіксовану, жорстко задану структуру: чеки, бланки, кредитні карти і тому подібне. Документи цього класу можуть бути легко збережені за допомогою традиційних реляційних і об'єктно орієнтованих баз даних;

– клас неструктурованих документів S_N. Документи, що належать до цього класу, є повністю неструктуровані. Прикладом таких електронних документів можуть служити документи мультимедіа і відеодані;

– клас напівструктурованих документів S_P. Це найбільш великий клас документів, до якого відноситься найбільший об'єм даних. До напівструктурованих відносяться такі документи, які можна виділити в деяку структуру, але ця структура недостатньо чітка для зберігання цих документів у традиційних системах.

Крім цього, додатково кваліфікують:

– частково структуровані документи S_S. До цього класу відносяться документи, велика частина якого має правильну структуру і може зберігатися у традиційних СУБД, а інша, неструктурована частина документів, деяким чином пов'язується зі структурованими даними;

– документи з неявною структурою S_I. До цього дуже великого класу відносяться такі документи, які хоча і мають деяку досить строгу структуру, але ця структура не чітко виражена. Вони поділяються на:

– клас документів з обмеженою областю змін S_V. Електронні документи представленого класу можуть мати структуру, що змінюється, проте кількість цих варіацій є обмеженою. До цього класу можна віднести деякі види фінансової або внутрішньої офісної документації;

– клас документів зі змінюваною структурою S_C. Електронні документи, що відносяться до цього класу, мають постійну структуру, що змінюється; крім того, сам склад цих документів може постійно змінюватися. Проте в кожен момент часу усі ці документи мають строгу логічну структуру, яка може бути описана певним чином. До цього класу можуть бути віднесені більшість корпоративних документів управління – документація системи менеджменту якості, технологічна документація і т. ін.

Постановка проблеми. Нині є важливою проблема розмітки логічної структури електронних документів у класі документів S_C. Наприклад, у CALS-технологіях (Continuous Acquisition and Life-cycle Support) рекомендується використовувати стандартизовані інтерактивні електронні технологічні керівництва (ІЕТК), які є організованими в базу даних набору документів у форматі SGML [3]. Введення стандартів для структуризації документів забезпечує можливість передачі даних між різними організаціями, централізоване управління даними і можливість автоматизації процесу розроблення структурованих документів.

Таким чином, важливою і перспективною розробкою у сфері створення систем документообігу є завдання відображення структури електронного документа за допомогою стандартизованого формату представлення даних, розпізнавання логічної структури документів з метою збереження її в цьому форматі й об'єднання безлічі структурованих документів у базу даних з потужною мовою запитів.

У зв'язку з цим необхідно провести моделювання структури електронних документів, а також формалізувати їх елементи з метою застосування цих моделей до постійно змінюваної корпоративної документації класу S_C.

Метою статті є розроблення математичної моделі електронного документа, що дозволяє розробляти методи оброблення різнокласових електронних документів у системах електронного документообігу.

Виклад основного матеріалу. В основу стандартних мов опису документів SGML і ODA покладено уявлення про ієрархічну структуру тексту. Це твердження використовується для опису як фізичної, так і логічної структури документів. Дослідження природи документів різного типу показали, що незважаючи на обмеження, які накладає ця умова на допустимі описи структури документів, існують різні способи для подолання труд-

нощів, що виникають у зв'язку з цим [4; 5]. Таким чином, структура практично будь-якого документа може бути або представлена в ієрархічному виді, або розбита на безліч ієрархічних підструктур.

Дослідники, що працюють у сфері розпізнавання структури документів, передусім звертають увагу на побудову моделі структури документа. У [6] пропонується модель документа, що допускає відображення як логічної, так і фізичної структури. Представлена модель документа заснована на проміжних вузлах, що визначають елементи логічної структури, які поставлені в однозначну відповідність елементам фізичної структури. Таким чином, сукупність вузлів логічної структури, фізичної структури і взаємодії між ними дозволяє передати повний опис структури документа.

Модель документа, яка не робить відмінностей між фізичною і логічною структурами, представлена в [6]. Елементи логічної і фізичної структур представлено в цій моделі еквівалентними і розглядаються в сукупності. В цьому випадку загальна структура документа може не мати ієрархічного представлення; тут ієрархічними є тільки підмножини загальної структури, якими є окремо взяті фізична і логічна структури. Така неієрархічна модель дещо утрудняє задачі маніпуляції з документом.

У [7] модель документа представлено у вигляді сукупності полігональних областей, а також набору відображень, що визначають зв'язки цих багатокутних областей з фізичними і логічними атрибутами відповідних областей.

Цікавий підхід до побудови моделі документа представлено в [8]. Логічна структура тут визначається за допомогою геометричних взаємозв'язків між сегментами тексту, тобто великою мірою незалежно від тексту. Проблема отримання логічної структури документа з властивої йому фізичної структури тоді може бути розділена на дві стадії: отримання сегментів документа, що представляють ієрархічну послідовність логічних елементів, і класифікація вузлів ієрархії згідно з типами елементів логічної структури, які вони представляють.

Модель документа, що використовується в [9] і названа дескриптором документа, також здатна описати як логічну, так і фізичну структуру документів. Логічна структура тут представлена відповідно до правил виведення регулярних граматик. Фізична структура є атрибутами, які пов'язані з нетермінальними символами граматики логічної структури. Адекватне поширення запропонованого підходу на опис логічної структури за допомогою контекстів *n*-вільних граматик було розглянуто в роботі [10].

Нарешті, статистична модель структури документа, заснована на застосуванні методу *n*-грамів до ієрархічних структур, представлена в [2]. Тут фізична і логічна деревовидні структури представлено ймовірностями контексту окремо взятих вузлів дерева подібно до вірогідності *n*-грамів. У цьому випадку проведено поширення лінійної моделі природної мови, названої *n*-грамами [11], на ієрархічну структуру документа.

Усі розглянуті вище моделі документів, за винятком останньої, дозволяють описувати тільки документи, що відносяться до класу S_S і частково документи класу S_V. Статистична модель дозволяє описати документи класу S_C, проте не дозволяє явної побудови граматики розглянутого класу документів.

Виходячи з цього, нами запропоновано математичну модель електронного документа, засновану на застосуванні логічних областей, що дозволяє розробляти методи оброблення різнокласових електронних документів у системах електронного документообігу.

Позначимо через M – безліч всіляких логічних міток заданого документа D.

Визначення 1. Припустимо, що $m \in M$ – мітка, що визначає тип логічного об'єкта й O – область, обмежена межею логічною об'єкта. Тоді пара (m, O) називається логічною областю заданого об'єкта.

Межа логічного об'єкта задається за допомогою відповідних тегів логічної розмітки.

Визначення 2. Логічні області (m_1, O_1) *і* (m_2, O_2) називаються рівними, тобто:

$$(m_1, O_1) = (m_2, O_2), \quad (1)$$

якщо $m_1 = m_2$ і $O_1 = O_2$.

Визначення 3. Логічна область (m_1, O_1) називається вкладеною в логічну область (m_2, O_2), тобто:

$$(m_1, O_1) \prec (m_2, O_2), \quad (2)$$

якщо $O_1 \prec O_2$, причому $O_1 = O_2$ тоді і тільки тоді, коли $(m_1, O_1) = (m_2, O_2)$.

Нехай Z – безліч усіх логічних областей документа D. Визначимо відношення $\prec z$ таким чином:

$$(m_1, O_1) \prec z\,(m_2, O_2) \Leftrightarrow (m_1, O_1) \prec (m_2, O_2) \text{ и } (m_1, O_1), (m_2, O_2) \in Z. \quad (3)$$

Таким чином, структура документа може бути відображена у вигляді вкладених логічних областей (рис. 1).

Рис. 1. Приклад зображення логічної структури документа у вигляді вкладених логічних областей
Джерело: розроблено автором.

Для обґрунтування моделі сформулюємо і доведемо такі теореми.

Теорема 1. Відношення $\prec z$ на множині Z є порядком.

Доказ. Згідно з визначенням [3] для доведення теореми необхідно перевірити для відношення $\prec z$ виконання умов рефлексивності, транзитивності й антисиметричності.

Рефлексивність. Умова рефлективності має вигляд $(m_1, O_1) \prec z\,(m_2, O_2)$ для усіх $(m, O) \in Z$, оскільки в умові задана приналежність логічної області (m, O) множині Z, та залишається перевірити правильність нерівності $(m_1, O_1) \prec (m_2, O_2)$. Згідно з визначенням [3], з рівності $O_1 = O_2$ повинна слідувати рівність $(m_1, O_1) = (m_2, O_2)$. Ця рівність правильна, за визначенням [2], тоді і тільки тоді, коли рівні логічні мітки, тобто $m_1 = m_2$. Отже, умова рефлексивності виконана.

Транзитивність. Умова транзитивності має такий вигляд: якщо $(m_1, O_1) \prec z\,(m_2, O_2)$ і $(m_2, O_2) \prec z(m_3, O_3)$, то $(m_1, O_1) \prec z\,(m_3, O_3)$. З перших двох нерівностей витікає, що $(m_1, O_1), (m_2, O_2) \text{ і } (m_3, O_3) \in Z$, тому залишається довести, що якщо $(m_1, O_1) \prec z\,(m_2, O_2)$ і $(m_2, O_2) \prec z\,(m_3, O_3)$, то $(m_1, O_1) \prec z\,(m_3, O_3)$. З $(m_1, O_1) \prec z\,(m_2, O_2)$ витікає, що $O_1 \subseteq O_2$, причому $O_1 = O_2 \Leftrightarrow m_1 = m_2$, а з $(m_2, O_2) \prec z\,(m_3, O_3)$ витікає, що $O_2 \subseteq O_3$, причому $O_2 = O_3 \Leftrightarrow m_2 = m_3$. Згідно з властивістю транзитивності для множин отримаємо, що $O_1 \subseteq O_3$, причому $O_1 = O_2 = O_3 \Leftrightarrow m_1 = m_2 = m_3$. Отже, $(m_1, O_1) \prec (m_3, O_3)$, $(m_1, O_1) \prec z\,(m_3, O_3)$.

Антисиметричність. Умова антисиметричності визначається таким чином: якщо $(m_1, O_1) \prec z\,(m_2, O_2)$ і $(m_2, O_2) \prec z\,(m_1, O_1)$, то $(m_1, O_1) = (m_2, O_2)$. З будь-якого з перших двох нерівностей отримуємо, що $(m_1, O_1), (m_2, O_2) \in Z$, тому перевіряємо умову, що залишилася, що якщо $(m_1, O_1) \prec z\,(m_2, O_2)$ і $(m_2, O_2) \prec z\,(m_1, O_1)$, то $(m_1, O_1) = (m_2, O_2)$. З перших двох умов витікає, що $O_1 \subseteq O_2$ і $O_2 \subseteq O_1$. Згідно з властивістю антисиметрич-

ності для множин, позначимо, що $O_1 = O_2$. За визначенням (3) це можливо тоді і тільки тоді, коли $(m_1, O_1) = (m_2, O_2)$.

Таким чином, відношення $\prec z$ – порядок на множині Z.

З теореми 1 витікає, що якщо множина Z – «не порожня», то вона є частково впорядкована.

Теорема 2. Частково впорядкована множина Z є ґратами.

Доказ. Згідно з визначенням [12] ґратами є частково впорядкована множина, в якій будь-яка пара елементів має точну нижню і верхню грань.

Нехай (o_1,o_2), $o_1,o_2 \in Z$ – пара логічних областей з частково впорядкованої множини Z. Нехай також $x = inf\ (o_1,o_2)$ – точна нижня грань $y = sup\ (o_1,o_2)$ – точна верхня грань пари (o_1,o_2).

Розглянемо побудову точної нижньої грані. Можливі ті, що наслідують випадок:

1. Логічні області o_1 і o_2 порівнянні між собою, тобто або $o_1 \prec z\ o_2$ або $o_2 \prec z\ o_1$. У першому випадку отримуємо $x = o_1$, а в другому – $x = o_2$.

2. Логічні області o_1 і o_2 не порівнянні між собою, але є така максимальна логічна область o_3, що $o_3 \neq \otimes$, де $\otimes$ означає порожню область, і правильні нерівності $o_3 \prec z\ o_1$ і $o_3 \prec z\ o_2$. Тоді $x = o_3$.

3. Якщо жоден з перших двох випадків не застосуємо, то $x \neq \otimes$.

Побудова точної верхньої грані виробляється аналогічно:

1. Якщо $o_1 \prec z\ o_2$ то $y = o_2$ і якщо $o_2 \prec z\ o_1$ то $y = o_1$.

2. Якщо існує така мінімальна логічна область $o_3 \neq D$, де D – увесь документ, що $o_1 \prec z\ o_3$ і $o_2 \prec z\ o_3$, то $y = o_3$.

3. Якщо не застосовні перші два випадки, то $y = D$.

Таким чином, множина Z є ґратами.

Слід також зазначити, що Z є ієрархічними ґратами. Цей факт виходить з властивостей відношення $\prec$, що не допускає ніяких інших перетинів логічних областей, крім повного вкладення.

Таким чином, структура документа може бути відображена у вигляді дерева (рис. 2).

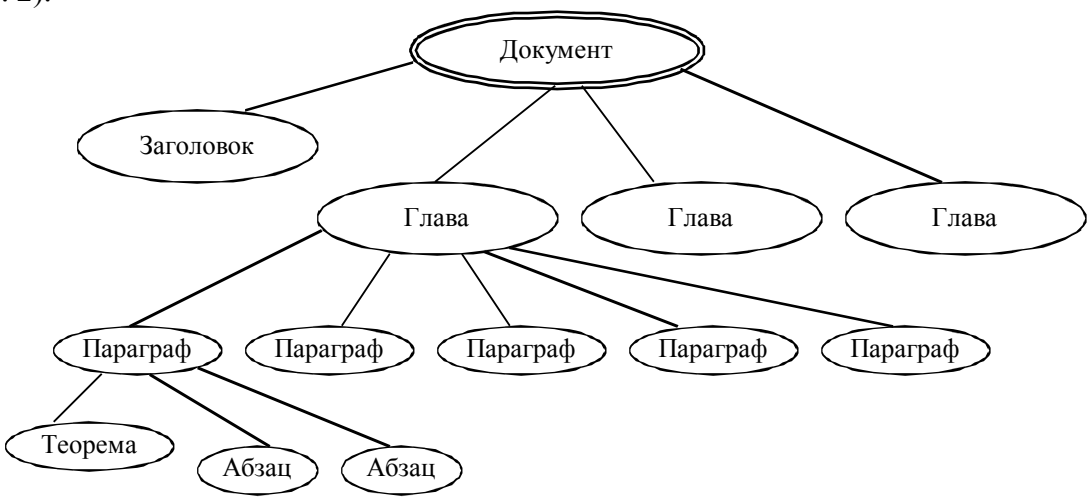

Рис. 2. Приклад зображення логічної структури документа у вигляді дерева структури
Джерело: розроблено автором.

Побудовані ієрархічні грати Z пов'язана з такими властивостями [7]:

1. Позначимо через T множину усіх можливих у документі D атрибутів, що відносяться як до області фізичних типів (малюнок, таблиця, формула), так і до області формату (тип, розмір і зображення шрифту, вирівнювання тексту). Тоді відображення $P: Z \rightarrow T$ пов'язує логічні області документа з їх фізичними ознаками.

2. Нехай I – алфавіт символів, прийнятих в документі D. Тоді I' – безліч усіх визначених над цим алфавітом рядків. Відображення $C: Z \rightarrow I'$ дозволяє з'єднати логічні області документа з їх змістом.

3. Позначивши через W безліч роздільників, можливих у документі D, отримуємо відображення $R: Z \times Z \rightarrow W$, яке визначає роздільник між парами логічних областей.

Таким чином, сукупність фізичної структури документа $S^F = (T)$ і логічної структури документа $S^L = (C,W)$ визначає структуру $S = (S^F, S^L)$ заданого документа D.

Висновки і пропозиції. Одним з найбільш важливих етапів системного дослідження складної системи документообігу є розроблення математичних моделей. Щоб описати набір правил, характерних для електронних документів, необхідно розробити математичну модель документа, яку іноді також називають універсальним документом [2].

Зазвичай розрізняють два типи структур електронного документа [6]: фізична структура, яка групує фізичні об'єкти в документі, і логічна структура документа, яка відображає його логічну організацію. Таким чином, закінчена модель документа складається з двох частин: фізична структура і логічна структура.

Виходячи з цього, у статті запропонована математична модель електронного документа, яка ґрунтується на застосуванні логічних областей, що дозволяє розробляти методи оброблення різнокласових електронних документів у сучасних системах електронного документообігу.

Список використаних джерел

1. *ДСТУ 4163-2003*. Вимоги до оформлювання документів. – [Чинний від 2003-09-01]. – К., 2003. – 46 с.

2. *Смирнова Г. Н.* Учебное пособие по дисциплине «Электронные системы управления документооборотом» / Г. Н. Смирнова. – М. : Московский международный институт эконометрики, информатики, финансов и права, 2003. – 168 с.

3. *Структура* та компоненти системи електронного документообігу [Електронний ресурс]. – Режим доступу : http://nauch.com.ua/geografiya/20401/index.html?page=2 .

4. *Ткачук Г. І.* Використання електронної системи документообігу у ВНЗ / Г. І. Ткачук, С. А. Постова // Магістратура в умовах євроінтеграційних процесів вищої школи. – Житомир : ЖДУ, 2014. – С. 254.

5. *Удосконалення* корпоративних інформаційних систем [Електронний ресурс]. – Режим доступу : http://ukrbukva.net/page,8,69376–Sovershenstvovanie–korporativnyh–informacionnyh–sistem.html.

6. *Електронний* документообіг: сучасні тенденції та проблеми провадження [Електронний ресурс]. – Режим доступу : http://www.rusnauka.com/34_VPEK_2012/Philologia/7_121024.doc.htm.

7. *Losee R. M.* Learning syntactic rules and tags with genetic algorithms for information retrieval and filtering: an empirical basis for grammatical rules. University of North Carolina, USA, 2005. – 20 p.

8. *Summers K. M.* Automatic discovery of logical document structure. Phd thesis, Cornell University. Cornell, USA. 2008. – 196 p.

9. *Azokly A. S.* Tune approche uniforme pour la reconnaissance de la structure physique de documents composites fondee sur l'analyse des cspaces. Phd thesis, l'lnstitut d'lnformatique, University de Fribourg, Suisse, 2005. – 155 p.

10. *Azokly A. S.* Tune approche uniforme pour la reconnaissance de la structure physique de documents composites fondee sur l'analyse des cspaces. Phd thesis, l'lnstitut d'lnformatique, University de Fribourg, Suisse, 2009. – 155 p.

11. *Srihari S.N.,* Lam S.W., Govindaraju V.,, Srihari R.K., Hull J J. Document image understanding. Center of Excellence for Document Analysis and Recognition, State University of New York at Buffalo, Buffalo, USA, 2008.

12. *Биркгоф Г.* Теория решеток : пер. с англ. / Г. Биркгоф. – М. : Наука, 1984. – 432 с.