DOI: https://doi.org/10.26642/ten-2023-1(91)-178-183
УДК 004.02

І.О. Черняк, аспірант
Державний університет «Житомирська політехніка»

Аналіз алгоритмів та математичних моделей для автоматизації електронного документообігу

(Представлено: PhD Граф М.С.)

У роботі детально досліджуються можливості використання на покращення математичних моделей та алгоритмів, які використовуються для автоматизації та покращення електронного документообігу. Також у цій статті досліджується потенціал алгоритмів і математичних моделей в автоматизації документообігу. Розглянуто роль математичних моделей в автоматизації документообігу та можливості вдосконалення цих моделей. У досліджених наукових роботах простежується велика перспектива в покращенні електронного документообігу та математичної моделі для автоматизації документообігу. Проте в більшості опрацьованих робіт не розглядаються можливі ризики і проблеми при імплементації покращень від математичної моделі для електронного документообігу. Також було висунуто припущення про ідеальну математичну модель сьогодні для електронної документації.

Ключові слова: *математична модель; електронна документація; документообіг; електронний документообіг; автоматизація.*

Актуальність теми. Алгоритми та математичні моделі для автоматизації документообігу – це використання алгоритмів і математичних моделей для автоматизації різних аспектів документообігу, таких як вилучення документів, класифікація, маршрутизація, контроль версій, зберігання та пошук, а також прогнозування [1]. Це може допомогти підвищити ефективність і точність управління документами, а також знизити витрати і поліпшити дотримання нормативних вимог. Потенціал алгоритмів і математичних моделей для автоматизації документообігу є дуже перспективним. Алгоритми можуть допомогти автоматизувати процес управління документами, зменшивши кількість необхідної ручної роботи. Вони також можуть допомогти підвищити точність та ефективність управління документами. Математичні моделі можуть допомогти оптимізувати процес управління документами, надаючи спосіб аналізу та прогнозування поведінки складних систем.

Організації розглядають документообіг як важливий аспект. Автоматизувавши документообіг, організація може заощадити час, ресурси та підвищити його ефективність і точність. У цій статті досліджується потенціал алгоритмів і математичних моделей в автоматизації документообігу. У ній висвітлюються переваги та труднощі, пов’язані з автоматизацією документообігу, а також те, як алгоритми та математичні моделі можуть вирішити ці проблеми. Наприклад, алгоритми можуть мінімізувати ручну роботу і підвищити точність, автоматично витягуючи з документів таку інформацію, як імена, дати та адреси. Це значно економить час і зусилля порівняно з ручним введенням даних.

Аналіз останніх досліджень та публікацій, на які спирається автор. Тема статті досліджувалася у різних сферах вітчизняними вченими, серед яких О.Матвієнко [2], О.Січова [3], Н.Лиско [4]. Теоретичні аспекти впровадження електронного документообігу в установах галузі освіти вивчали Н.Задорожна, В.Петрушко [1]. Також багато зарубіжних вчених досліджували тематику електронної документації та документообігу, її впровадження в органах державної влади. У цій категорії більшість досліджень пропонують необхідність інтеграції компонентів системи електронного документообігу (СЕД) між собою. Також досліджувалися фактори щодо впровадження СЕД в уряді. Серед них: підтримка вищого керівництва [10], планування впровадження [6], якість даних [8] та співпраця [6]. Ці фактори є одними з найпоширеніших. Також Баходір Мумінов і Адільбек Даулетов розробили математичну та інформаційну модель для керування електронними документами в системах електронного документообігу на основі набору даних і подій. Вони також пропонують ієрархічний погляд на структуру управління та задачу оптимізації для пошуку інформації [10].

Мета статті є провести аналіз існуючих досліджень та на їх основі визначити основні складові в сучасних математичних моделях для автоматизованого документообігу, а також знайти можливі проблеми та способи для покращення або модернізації існуючих моделей для автоматизованого документообігу.

Викладення основного матеріалу. Автоматизація документообігу має багато переваг, зокрема підвищення ефективності, точності та відповідності нормативним вимогам. Автоматизація документообігу також може допомогти зменшити витрати, усуваючи потребу в паперових документах і місцях для їх зберігання. Крім того, автоматизація документообігу може допомогти поліпшити обслуговування клієнтів, надаючи їм швидший доступ до інформації.

Можна виокремити такі основні переваги, як:

1. Підвищення ефективності, оскільки автоматизований документообіг скорочує час і зусилля, необхідні для обробки, організації та пошуку документів, що призводить до підвищення продуктивності та ефективності;
2. Краща організація, через те, що автоматизовані системи дозволяють ефективно класифікувати та індексувати документи, що призводить до покращення організації та полегшення пошуку документів;
3. Підвищена точність автоматизованих систем, які зменшують ризик людських помилок і забезпечують послідовну, точну та актуальну інформацію;
4. Покращена співпраця, через те що автоматизовані системи полегшують спільну роботу, забезпечуючи безпечний доступ до документів у режимі реального часу з будь-якого місця і часу;
5. Краще прийняття рішень, оскільки автоматизовані системи надають дані та інформацію, які можуть бути використані для прийняття рішень, що дозволяє організаціям працювати більш ефективно;
6. Відповідність та безпека. Автоматизовані системи можуть забезпечити дотримання нормативних вимог і захистити конфіденційну інформацію шляхом контролю доступу, моніторингу використання та забезпечення належного зберігання і знищення документів;
7. Економія коштів, зменшення витрат, пов'язаних з ручним управлінням документами, такими як друк, зберігання та пошук, за рахунок оптимізації процесів та зменшення відходів.

Математичні моделі відіграють вирішальну роль в автоматизації документообігу, надаючи необхідні для цього алгоритми та методи. Вони використовуються для представлення документів у структурованому форматі, що дозволяє маніпулювати ними та проводити аналіз. Однією з математичних моделей для представлення документів є модель векторного простору, яка показує кожен документ як вектор числових значень. У цій моделі кожен документ представлений як набір термінів (або слів) і пов'язаних з ними ваг, які відображають важливість термінів у документі. У моделі векторного простору документ представлений у вигляді вектора високої розмірності, де кожен вимір відповідає терміну, а його значення відображає вагу цього терміна в документі. Потім вектори можна порівняти за допомогою математичних методів, таких як косинусна подібність, щоб визначити схожість між документами. Наприклад, розглянемо два документи: «Документ А» і «Документ Б», з термінами «дані» й «аналіз». Якщо термін «дані» частіше використовується в документі А, він матиме більшу вагу у векторі, що представляє документ А, тоді як термін «аналіз» може мати більшу вагу у векторі, що представляє «Б».

Використовуючи моделі векторного простору, документи можна легко порівнювати та аналізувати, що дає змогу ефективно здійснювати пошук і класифікацію документів на основі їхнього змісту. Крім того, модель векторного простору є гнучкою і може бути розширена для включення додаткових функцій, таких як частота терміна, обернена до частоти документа (TF-IDF) і n-грами, щоб покращити представлення документів. Ось простий приклад того, як модель векторного простору можна реалізувати на Python за допомогою бібліотеки scikit-learn:

```
from sklearn.feature_extraction.text import TfidfVectorizer
# Define a list of documents
documents = [
    "Data analysis is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, drawing conclusions, and supporting decision-making.",
    "Data visualization is the graphic representation of data. It involves creating visualizations, like charts and graphs, to make the data easier to understand.", "Lorem Ipsum is simply dummy text of the printing and typesetting industry."
]

# Create an instance of the TfidfVectorizer
vectorizer = TfidfVectorizer()
# Fit the vectorizer to the list of documents and transform the documents into a sparse matrix
X = vectorizer.fit_transform(documents)

# The sparse matrix X represents the vector space model of the documents, where each row is a document and each column is a term
print(X.shape)  # Output: (3, 69)

# The terms (or words) in the documents can be obtained from the `vocabulary_` attribute of the vectorizer
vocabulary = vectorizer.vocabulary_
# The weights of the terms in the documents can be obtained from the matrix X
weights = X.toarray()
```

У цьому прикладі TfidfVectorizer використовується для перетворення текстових документів у модель векторного простору, де кожен документ представлено у вигляді розрідженої матриці (X) числових

значень. Метод fit_transform використовується для припасування векторизатора до списку документів і перетворення документів у розріджену матрицю. Атрибут vocabulary_ векторизатора містить терміни (або слова) в документах, а ваги термінів у документах можна отримати з матриці X.

Індексація та пошук документів. Математичні моделі, такі як моделі векторного простору та імовірнісні моделі пошуку, використовуються для індексування та пошуку релевантних документів на основі запитів користувачів. Моделі індексування та пошуку використовуються в системах управління документами для ефективного зберігання та пошуку документів на основі запитів користувачів. Дві найпоширеніші моделі індексування та пошуку документів – це моделі векторного простору та моделі імовірнісного пошуку.

Моделі векторного простору. У цих моделях кожен документ представляється у вигляді вектора де кожен вимір відповідає терміну, а його значення – вазі цього терміна в документі. Потім вектори можна порівняти, щоб визначити схожість між документами. Коли користувач надсилає запит, він також має вигляд вектора, а документи, які найбільше схожі на вектор запиту, виводяться як результати.

Імовірнісні моделі пошуку. Імовірнісні моделі пошуку базуються на ідеї, що кожен документ асоціюється з набором ключових слів і ймовірністю релевантності. Коли користувач надсилає запит, ймовірність релевантності для кожного документа обчислюється на основі ключових слів запиту та ймовірності ключових слів у документах. Документи з найвищою ймовірністю релевантності відображаються в результатах пошуку.

Як моделі векторного простору, так і моделі імовірнісного пошуку мають свої переваги та недоліки, і вибір моделі залежить від конкретних вимог системи управління документами. Наприклад, моделі векторного простору є простими і швидкими, але вони можуть не враховувати контекстні зв'язки між термінами, тоді як імовірнісні моделі пошуку є більш складними і можуть давати кращі результати, але вони можуть бути повільними і складними у впровадженні [13]. Ось приклад реалізації моделі векторного простору для індексування та пошуку документів на Python за допомогою бібліотеки scikit-learn:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
# Define a list of documents
documents = [
    "Data analysis is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, drawing conclusions, and supporting decision-making.",
    "Data visualization is the graphic representation of data. It involves creating visualizations, like charts and graphs, to make the data easier to understand.",
    "Lorem Ipsum is simply dummy text of the printing and typesetting industry."
]
# Create an instance of the TfidfVectorizer
vectorizer = TfidfVectorizer()
# Fit the vectorizer to the list of documents and transform the documents into a sparse matrix
X = vectorizer.fit_transform(documents)

# The sparse matrix X represents the vector space model of the documents, where each row is a document and each column is a term
print(X.shape)  # Output: (3, 69)

# Calculate the cosine similarity between each document and all other documents
similarities = cosine_similarity(X)
# The similarities matrix contains the cosine similarity scores between each pair of documents
print(similarities)

# Example of a user query
query = "process of data analysis"
# Represent the query as a vector using the same vectorizer
query_vector = vectorizer.transform([query])
# Calculate the cosine similarity between the query and each document
query_similarities = cosine_similarity(query_vector, X)

# The query_similarities array contains the cosine similarity scores between the query and each document
print(query_similarities)

# The documents with the highest similarity scores are the most relevant to the query
relevant_documents = [documents[i] for i in query_similarities.argsort()[-2:][::-1]]
print(relevant_documents)
```

Кластеризація та класифікація документів. Математичні моделі, такі як кластеризація за методом k-середніх та дерева рішень, використовуються для автоматичного розподілу документів на значущі групи.

Визначайте теми та взаємозв'язки. Математичні моделі, такі як латентний розподіл Діріхле (LDA) і графові моделі, використовуються для визначення основних тем, що є у колекції документів, і для створення репрезентативного резюме документів.

Вимірювання схожості текстів. Косинусні математичні моделі, наприклад, використовуються для вимірювання подібності між документами та групування схожих документів разом.

Забезпечення конфіденційності та безпеки. Математичні моделі, такі як алгоритми шифрування та системи контролю доступу, використовуються для забезпечення конфіденційності та цілісності конфіденційної інформації.

Інтеграція з іншими системами. Математичні моделі використовуються для інтеграції з іншими системами, такими як системи управління контентом та інструменти аналізу даних, для покращення співпраці та прийняття рішень.

В цілому математичні моделі створюють основу для автоматизації управління документами, надаючи алгоритми і методи, необхідні для ефективної організації, класифікації, пошуку та аналізу великих колекцій документів.

Математичні моделі можна використовувати для автоматизації різних аспектів документообігу, таких як ідентифікація та класифікація документів, маршрутизація документів для затвердження та управління версіями документів. Моделі також можна використовувати для оптимізації зберігання та пошуку документів, а також для прогнозування того, як документи будуть використовуватися в майбутньому.

Автоматизація документообігу може запропонувати багато переваг, враховуючи підвищення ефективності та точності, зниження витрат і підвищення гнучкості. Це також може допомогти організаціям відповідати нормативним вимогам і покращити дотримання внутрішніх політик.

Перше, що є зараз популярним, – це впровадження алгоритмів машинного навчання: наприклад, алгоритми ML для автоматичної класифікації, кластеризації та категоризації документів на основі їхнього змісту, зменшуючи потребу в ручному введенні. Також доволі часто можна побачити роботи, які направлені на обробку природної мови (NLP Natural Language Processing) та застосування методів NLP для вилучення важливої інформації, наприклад, дат, імен, сутностей і зв'язків з неструктурованих документів. Важливим також є включення контекстної інформації, яка б враховувала контекст, в якому створюються та використовуються документи, щоб краще зрозуміти їх значення та важливість для покращення автоматизації роботи з документами. Використання онтології для визначення зв'язків між поняттями, сутностями та категоріями, які можна використовувати для покращення пошуку та категоризації документів. Інтегрування з існуючими системами, такими як CRM, ERP і сховища даних, щоб обмінюватися інформацією та покращувати співпрацю. І як головне можна виокремити алгоритми для пошуку, оскільки вдосконалені алгоритми пошуку можна використати, щоб краще розуміти наміри користувачів і знаходити релевантні документи.

Ідеальна математична модель для електронного документообігу на теперішній момент на основі проаналізованих досліджень, ймовірно, містила б такі функції:

– представлення документів (документи подаються як математичні об'єкти, що дозволяє маніпулювати ними та аналізувати);

– індексування та пошук (ефективна система індексації та пошуку, така як модель векторного простору або імовірнісна модель пошуку, використовується для швидкого пошуку відповідних документів на основі запитів користувачів);

– кластеризація та класифікація (неструктуровані документи автоматично кластеризуються і класифікуються за категоріями, що полегшує управління та аналіз великих колекцій);

– тематичне моделювання (алгоритми тематичного моделювання, такі як Latent Dirichlet Allocation (LDA), використовуються для визначення основних тем, наявних у колекції документів, і створення репрезентативного резюме документів);

– подібність текстів (алгоритми схожості тексту використовуються для виявлення схожих документів і групування їх у кластери, зменшуючи дублювання і покращуючи організацію);

– конфіденційність і безпека (модель містить заходи конфіденційності та безпеки, такі як шифрування даних і контроль доступу, щоб забезпечити конфіденційність і цілісність конфіденційної інформації);

– інтеграція з іншими системами (модель легко інтегрується з іншими системами, такими як системи управління контентом, системи документообігу та інструменти аналізу даних, для покращення співпраці та прийняття рішень).

Висновки та перспективи подальших досліджень. Незважаючи на численні переваги автоматизації документообігу, існують певні виклики, які необхідно враховувати. Однією з них є потенційна можливість помилок в алгоритмах, які використовуються для автоматизації процесу. Ці виклики містять потребу в спеціалізованих навичках і знаннях, ризик помилок, а також потенційні перебої в бізнес-операціях. Іншою

проблемою є вартість впровадження та обслуговування автоматизованої системи. Нарешті, існує ризик того, що автоматизовані системи не зможуть впоратися з усіма типами документів або ситуацій.

Потенціал алгоритмів і математичних моделей для автоматизації документообігу є дуже перспективним. Алгоритми можуть допомогти автоматизувати процес управління документами, зменшивши кількість необхідної ручної роботи. Вони також можуть допомогти підвищити точність та ефективність управління документами. Математичні моделі можуть допомогти оптимізувати процес управління документами, надаючи спосіб аналізу та прогнозування поведінки складних систем.

Існує низка проблем, які необхідно вирішити для того, щоб повністю реалізувати потенціал алгоритмів і математичних моделей для автоматизації документообігу. Однією з них є потреба в додаткових дослідженнях того, як ці інструменти можуть бути ефективно використані в реальних сценаріях. Іншим викликом є необхідність розробки більш зручних інтерфейсів для цих інструментів, щоб ними могло користуватися ширше коло людей.

Використання алгоритмів і математичних моделей для автоматизації документообігу може запропонувати багато переваг, враховуючи підвищення ефективності та точності. Однак автоматизація документообігу пов'язана з певними проблемами, такими як необхідність спеціальних знань і можливість помилок. Незважаючи на ці виклики, потенціал алгоритмів і математичних моделей для автоматизації документообігу є значним, і подальші дослідження в цій галузі є виправданими.

Список використаної літератури:

1. *Задорожна Н.* Інформаційна система менеджменту наукових досліджень у НАПН України / *Н.Задорожна, В.Петрушко, С.Тукало* // Інформаційні технології в освіті [Електронний ресурс]. – Режим доступу : https://lib.iitta.gov.ua/926/.
2. *Матвієнко О.* Основи організації електронного документообігу : навч. посіб. / *О.Матвієнко, М.Цивін*. – К., 2008. – 112 с.
3. *Січова О.* Основні аспекти впровадження електронного документообігу в Україні / *О.Січова* // Наукові праці Національної бібліотеки України ім. В.І. Вернадського. – К., 2006. – Вип. 16. – С. 323–331.
4. *Лиско Н.* Державне регулювання у сфері електронного документообігу в Україні / *Н.Лиско* // Вісник соціально-економічних досліджень. – 2013. – Вип. 1. – С. 230–235 [Електронний ресурс]. – Режим доступу : http://nbuv.gov.ua/UJRN/Vsed_2013_1_37.
5. Про електронні документи та електронний документообіг : Закон України [Електронний ресурс]. – Режим доступу : https://zakon.rada.gov.ua/laws/show/851-15?%20lang=ua#Text.
6. *Willis A.* Corporate Governance and Management of Information and Records / *A.Willis* // Records Management Journal. – 2005. – № 15. – P. 86–97.
7. *Wong T.Y.* Web-based Document Management Systems in the Construction Industry / *T.Y. Wong, H.K. Sar*. – 2012 [Electronic resource]. – Access mode : https://www.fig.net/pub/fig2012/papers/ts01c/TS01C_wong_5393.pdf.
8. *Singh R.* An Important Data Quality Tools for Data Warehouse: Case Study / *R.Singh*. – 2015. – № 2 (7) [Electronic resource]. – Access mode : http://www.jetir.org/papers/JETIR1507029.pdf.
9. *Muminov B.B.* Mathematical and Information Model of Electronic Document Management System / *B.B. Muminov*. – 2021 [Electronic resource]. – Access mode : https://ieeexplore.ieee.org/document/9670326.
10. *Yaacob R.A.* Electronic Records Management in Malaysia: A Case Study in one Government Agency / *R.A. Yaacob, R.Mapong Sabai*. – 2011.
11. *Haidera Ahmad S.* Opportunities and Challenges in Implementing Electronic Document Management Systems / *S.Haidera Ahmad*. – 2015.
12. *Ralph S.H.* Electronic Document Management: Challenges and Opportunities for Information Systems Managers / *S.H. Ralph* // MIS Quarterly. – 1995. – P. 29–49.
13. *Manning K.D.* Introduction to information retrieval / *K.D. Manning, P.Raghavan, H.Schütze* // Williams. – 2011 [Electronic resource]. – Access mode : http://om.univ.kiev.ua/users_upload/15/upload/file/pr_lecture_02.pdf.
14. *Blahušiaková M.D.* Automation and Digitalization of Business Processes – New Challenges Arising, Inter Alia, from the COVID-19 Pandemic / *M.D. Blahušiaková.* – 2022.
15. *Bunawan A.A.* The Challenges in Preserving the Electronic Records Metadata / *A.A. Bunawan, S.Nordin* // International journal of information systems and engineering. – 2015. – Vol. 1 (1).

References:

1. Zadorozhna, N., Petrushko, V. and Tukalo, S. «Informatsiina systema menedzhmentu naukovykh doslidzhen u NAPN Ukrainy», *Informatsiini tekhnolohii v osviti*, [Online], available at: https://lib.iitta.gov.ua/926/
2. Matviienko, O. and Tsyvin, M. (2008), *Osnovy orhanizatsii elektronnoho dokumentoobihu*, navch. posib., K., 112 p.
3. Sichova, O. (2006), «Osnovni aspekty vprovadzhennia elektronnoho dokumentoobihu v Ukraini», *Naukovi pratsi Natsionalnoi biblioteky Ukrainy im. V.I. Vernadskoho*, K., Issue 16, pp. 323–331.
4. Lysko, N. (2013) «Derzhavne rehuliuvannia u sferi elektronnoho dokumentoobihu v Ukraini», *Visnyk sotsialno-ekonomichnykh doslidzhen*, Issue 1, pp. 230–235, [Online], available at: http://nbuv.gov.ua/UJRN/Vsed_2013_1_37
5. *Pro elektronni dokumenty ta elektronnyi dokumentoobih*, Zakon Ukrainy, [Online], available at: https://zakon.rada.gov.ua/laws/show/851-15?%20lang=ua#Text

6. Willis, A. (2005), «Corporate Governance and Management of Information and Records», *Records Management Journal*, No. 15, pp. 86–97.
7. Wong, T.Y. and Sar, H.K. (2012), «Web-based Document Management Systems in the Construction Industry», [Online], available at: https://www.fig.net/pub/fig2012/papers/ts01c/TS01C_wong_5393.pdf
8. Singh, R. (2015), «An Important Data Quality Tools for Data Warehouse: Case Study», No. 2 (7), [Online], available at: http://www.jetir.org/papers/JETIR1507029.pdf
9. Muminov, B.B. (2021), «Mathematical and Information Model of Electronic Document Management System», [Online], available at: https://ieeexplore.ieee.org/document/9670326
10. Yaacob, R.A. and Mapong Sabai, R. (2011), «Electronic Records Management in Malaysia: A Case Study in one Government Agency».
11. Haidera Ahmad, S. (2015), «Opportunities and Challenges in Implementing Electronic Document Management Systems».
12. Ralph, S.H. (1995), «Electronic Document Management: Challenges and Opportunities for Information Systems Managers», *MIS Quarterly*, pp. 29–49.
13. Manning, K.D., Raghavan, P. and Schütze, H. (2011), «Introduction to information retrieval», *Williams*, [Online], available at: http://om.univ.kiev.ua/users_upload/15/upload/file/pr_lecture_02.pdf
14. Blahušiaková, M.D. (2022), «Automation and Digitalization of Business Processes – New Challenges Arising, Inter Alia, from the COVID-19 Pandemic».
15. Bunawan, A.A. and S.Nordin (2015), «The Challenges in Preserving the Electronic Records Metadata», *International journal of information systems and engineering*, Vol. 1 (1).

Черняк Ілля Олександрович – аспірант факультету інформаційно-комп'ютерних технологій Державного університету «Житомирська політехніка».

https://orcid.org/0000-0003-0029-7679.

Наукові інтереси:

– інтеграція процесів у розподілених системах та хмарних технологіях.

Cherniak I.O.

Analysis of algorithms and mathematical models for automation of electronic document management

The paper examines in detail the possibilities of using mathematical models and algorithms used to automate and improve electronic document management. This article also explores the potential of algorithms and mathematical models in document management automation. The role of mathematical models in document flow automation and the possibilities of improving these models are considered. The researched scientific papers show great promise in improving electronic document management and mathematical models for document management automation. However, most of the studies do not consider possible risks and problems in implementing improvements from the mathematical model for electronic document management. Also, an assumption was made about the ideal mathematical model for electronic documentation at the moment.

Keywords: mathematical model; electronic documentation; document flow; electronic document management; automation.

Стаття надійшла до редакції 01.05.2023.