УДК 004.891

Сергій Васильович Лєнков
Володимир Миколайович Джулій
Ігор Володимирович Муляр
Леся Володимирівна Охрамович

АЛГОРИТМ ПОШУКУ СЕМАНТИЧНО ПОДІБНИХ ДОКУМЕНТІВ

Постановка проблеми. Аналіз останніх досліджень і публікацій

Найважливішою складовою інфраструктури суспільства є сукупність людино-машинних інформаційних систем, в яких інформація виступає одним з головних ресурсів його життєдіяльності. В сучасних умовах виробництво не може існувати і розвиватися без високоефективної системи управління, що базується на автоматизованій інформаційній технології. Автоматизована інформаційна технологія тісно пов'язана з інформаційною системою, яка є для неї основним середовищем. В даний час в умовах глобалізації спостерігається тенденція укрупнення промислового виробництва, створення фінансово-промислових груп, холдингів. У сучасних умовах керівників підприємств цікавить:

агрегація даних (а не велика кількість конкретних значень);

динаміка, перспективи, тенденції (а не статистика);

корпоративні рішення (а не рішення для підрозділів);

мінімальні витрати на пошук необхідної інформації;

повнота і несуперечність інформації;

аналітичні зрізи для підтримки прийняття рішень.

Становлення інформаційного суспільства, зміна економічних умов, розвиток сучасних комп'ютерних технологій – фактори, які призвели до зміни умов управління підприємствами і, як наслідок, пред'явили нові вимоги до автоматизованих інформаційних систем і технологій обробки інформаційних ресурсів.

Основними з них є:

підвищення якості управління за рахунок більш оперативного і повного використання інформації про хід виробничого процесу, про матеріальні, фінансові, енергетичні потоки і витрати, про запаси сировини і матеріалів;

визначення та ефективне використання комплексних показників у системах управлінського та бухгалтерського обліку, що поліпшують інформаційне забезпечення оперативного управління;

наявність комплексної системи управління фінансовим станом підприємства, об'єднаної з інформаційними базами даних;

наявність корпоративної мережі, побудованої на архітектурі клієнт-сервер, як основної інформаційної магістралі підприємства;

наявність єдиного інформаційного простору всього підприємства, до складу якого входять фактографічні бази даних, бази документів, бази прецедентів і об'єднуючий їх компонент – предметно-орієнтоване сховище даних, що дозволяє використовувати всю накопичену інформацію для процесу прийняття управлінських рішень. В сучасних автоматизованих системах управління підприємством циркулює великий обсяг різнорідної інформації. В останні роки спостерігається тенденція до скорочення зростання обсягу структурованих даних і зростанню обсягу частково структурованих і не структурованих даних. Основна мета інформаційної системи: організація обробки, зберігання та передачі інформації. Інформаційні системи, в яких зберігання і обробка інформації здійснюється за допомогою обчислювальної техніки, називаються автоматизованими інформаційними системами. Інформаційні системи є основним засобом, інструментарієм вирішення завдань та інформаційного забезпечення. Інформаційне забезпечення – це сукупність процесів збору, обробки, зберігання, аналізу та видачі інформації, необхідної для забезпечення управлінської діяльності та технологічних процесів.

Формулювання мети статті. Виклад основного матеріалу

Інформація є сполучною ланкою між різними видами інтелектуальної та матеріальної діяльності колективів людей, між управлінням і виробництвом. Обсяг інформації, на відміну від інших видів ресурсів, не зменшується з часом, а навпаки, постійно збільшується, створюючи умови для накопичення досвіду, сприяючи виробленню обґрунтованих управлінських рішень. Керівництву середньої та вищої ланки холдингів, фінансово-

промислових груп для прийняття якісних управлінських рішень необхідно мати оперативний доступ до зацікавленої їх інформації. Проведений аналіз показав, що на пошук необхідної інформації йде до 20 % робочого часу; більшості користувачам складно сформулювати запит, що точно відображає його інформаційну потребу, це призводить до отримання нерелевантних документів; в інформаційно-довідкових системах недостатньо представлений механізм зворотного зв'язку з користувачем.

У зв'язку з цим важливе значення має організація ефективних механізмів пошуку в інформаційному фонді автоматизованої системи. Наявність в рамках автоматизованої системи інформаційно-довідкової підсистеми дає можливість отримувати оперативний доступ до достовірної інформації, необхідної для прийняття рішень і дозволяє підвищити ефективність управління.

Традиційно в інформаційних системах обробки даних виділяють наступні функціональні підсистеми: техніко-економічного планування, технічної підготовки виробництва, оперативного управління виробництвом, матеріально-технічного постачання, бухгалтерського обліку, збуту і реалізації продукції, управління кадрами, управління якістю продукції, такі підсистеми що оптимізують роботу підприємства, як формування портфеля замовлень, оптимальний розподіл капіталовкладень, управління оснащенням, оптимальний розподіл випуску продукції між підприємствами об'єднання, оперативно-календарне планування запуску-випуску, відвантаження продукції та інші.

Структуру інформаційної системи, що включає в себе інформаційну підсистему обробки даних представлено на рисунку 1.

Рис. 1. Структура інформаційної системи, що включає в себе інформаційну підсистему обробки даних

Особливості підсистем, що забезпечують її функціонування:

1) Підсистема інформаційного забезпечення – включає в себе інформаційні масиви (документи, запити, метадані), а також засоби і способи їх опису, подання та класифікації.

2) Підсистема лінгвістичного забезпечення – логіко-семантичний апарат, що складається з інформаційно-пошукової мови, методик індексування, критерію видачі, положень та інструкцій передмашинної та машинної обробки і пошуку інформації.

3) Підсистема математичного та програмного забезпечення – алгоритми та програмні засоби, що реалізують всі функції інформаційно-довідкової підсистеми, з допомогою комп'ютера.

4) Підсистема технічного забезпечення – технічні засоби, забезпечують зберігання, пошук і передачу інформації.

При розробці інформаційної системи обробки даних слід враховувати наступні особливості:

1) Використання критеріїв релевантності, що відрізняються від критеріїв, які використовуються в глобальній мережі Інтернет (PageRank, індексу цитованості).

2) Підвищені вимоги до точності і повноти результатів, що видаються.

3) Здійснення пошуку з урахуванням прав доступу співробітників.

Також бажано передбачити: тезауруси, що використовують прийняті на підприємстві термінологію і скорочення; засоби уточнення запиту; різні пошукові фільтри; персоналізацію пошуку.

В інформаційній системі обробки даних виділимо наступні підсистеми: підсистема попередньої обробки документів / запитів; підсистема індексування документів; підсистема аналізу запиту / документа-зразка; підсистема

побудови кластерів асоціативно пов'язаних пошукових термінів документа; підсистема реалізації діалогового режиму взаємодії з користувачем; підсистема «Тезаурус»; підсистема пошуку; підсистема формування результатів пошуку.

Структурна схема пропонованої інформаційної системи обробки даних представлена на рисунку 2. Розглянемо детальніше функції зазначених підсистем.

У **підсистемі попередньої обробки документів / запитів** здійснюються наступні операції: визначення мови тексту; лексичний аналіз; видалення стоп-слів; перетворення слів до початкової форми; приведення регістру.

Рис. 2. Структурна схема інформаційної системи обробки даних

Підсистема індексування призначена для виразу змісту документа на інформаційно-пошуковій мові. Результатом індексування є пошукові зразки документів. Пошуковий зразок документа являє собою «текст, що складається з лексичних одиниць інформаційно-пошукової мови, виражає зміст документа або інформаційного запиту і призначений для реалізації інформаційного пошуку».

У **підсистемі аналізу запиту / документа-зразка** здійснюється визначення інформаційної потреби користувача, формування пошукового образу запиту, завдання обмежень пошуку.

Інформаційну систему обробки даних пропонується доповнити підсистемою побудови кластерів асоціативно пов'язаних пошукових термінів документа. Ця підсистема призначена для побудови візуального представлення основного змісту документа у вигляді графа, вершинами якого є пошукові терміни, а ребра відображають їх асоціативний зв'язок. Ефективність пошуку неструктурованої інформації залежить від багатьох факторів, серед яких важливе місце займає точність відображення пошукової потреби користувача в запиті. Оскільки формулювання запиту є одним з ключових і складних аспектів інформаційного пошуку, в системі передбачається використання наступних методів уточнення запиту:

глобальних методів, які передбачають розширення запиту або нове формулювання запиту за допомогою тезауруса;

зворотного зв'язку за релевантністю;

пошуку семантично схожих документів, тобто документів, семантично схожих із заданим документом-зразком.

При пошуку семантично схожих документів необхідно вказати один релевантний документ, а при використанні зворотного зв'язку за релевантністю – кілька релевантних документів. Оскільки документ середнього розміру може охоплювати не одну тематику, причому не всі тематики документа можуть цікавити користувача, то використання підсистеми побудови кластерів асоціативно пов'язаних пошукових термінів документа допоможе у вирішенні завдання коректного відображення пошукової потреби користувача.

В інформаційній системі обробки даних введений діалоговий режим взаємодії з користувачем, особливістю якого є використання візуалізації графа, що відображає взаємозв'язки

між термінами інформаційного масиву. У поданні документа, сформованому підсистемою побудови кластерів асоціативно пов'язаних пошукових термінів документа, користувач зможе видалити пошукові терміни або кластери пошукових термінів або розширити запит термінами, асоціативно пов'язаними з термінами документа. Даний підхід дозволить користувачеві правильно підібрати набір пошукових термінів. Використання графа для представлення взаємозв'язків між термінами дозволяє застосувати алгоритми обходу графа в глибину і в ширину для виявлення семантично близьких термінів і уточнення запиту.

Підсистема **«Тезаурус»** призначена для створення, введення і коригування словників. Під інформаційно-пошуковим тезаурусом розуміється «нормативний словник дескрипторної інформаційно-пошукової мови із зафіксованими в ньому парадигматичними відношенями лексичних одиниць. Парадигматичні відношення вказують спільність або протиставлення значень і використання лексичних одиниць». Спочатку в пошукових системах тезауруси використовувалися для індексування документів. У сучасних системах тезауруси застосовуються для уточнення запиту. Процес побудови словників складається з наступних етапів: вибір лексичних одиниць; визначення їх морфологічних, синтаксичних і семантичних характеристик; розташування лексичних одиниць у певному порядку. Перед побудовою необхідно визначити типи відношень (частина-ціле, рід-вид і т.д.), ступінь деталізації словників. Розрізняють три способу побудови словників: апріорний, апостеріорний, динамічний. При апріорному способі проводиться вибір лексичних словників з різних термінологічних ресурсів по заданій тематиці (довідників, енциклопедій). Даний спосіб вимагає великих інтелектуальних витрат, його неможливо автоматизувати. При апостеріорному підході формування лексики здійснюється з представницької вибірки майбутнього масиву документів. У цьому випадку можлива автоматизація процесу побудови словника, але потрібно багато витрат на складання вибірки документів. Динамічний спосіб -процес накопичення лексики і побудови словників поєднаних з процесом експлуатації інформаційно-пошукової системи. Цей спосіб є найбільш перспективним. Одна з його великих переваг полягає в тому, що всі процеси побудови словників можна організувати в режимі діалогового зворотного зв'язку з користувачами системи, підвищуючи тим самим якість словників.

У **підсистемі пошуку** здійснюється пошук в інформаційному масиві документів, що задовольняють запиту. Розрізняють дві схеми організації інформаційних масивів: пряму і інвертовану. Пряма схема передбачає, що кожен запис масиву містить пошуковий зразок конкретного документа. У цьому випадку пошук документів зводиться до повного перегляду масиву. У інвертованому файлі кожен запис зберігає інформацію про всі документи, в пошукових зразках яких зустрічається заданий термін. Терміни впорядковані, що дозволяє використовувати бінарний пошуковий метод і значно прискорити швидкість пошуку. Більшість документальних інформаційно-пошукових систем в даний час використовують інвертовану схему організації інформаційного масиву.

У **підсистемі формування результатів пошуку** відбувається обчислення міри релевантності документів запиту та видача результатів. Як правило, результати пошуку відображаються у вигляді списку. Для більш зручного перегляду результати пошуку можуть групуватися на основі метаданих або піддаватися кластерному аналізу. Ранжування документів може виконуватися за датою створення / оновлення документа, за ступенем важливості. Виділяють два типи релевантності: формальну і змістову. Формальна релевантність відображає ступінь близькості пошукового зразка документа пошуковому припису, на підставі застосовуваного в інформаційно-пошуковій системі критерію видачі, а змістовна релевантність – ступінь близькості документа запиту. Для користувача первинною інформацією є не формулювання запиту, а саме інформаційна потреба, яка послужила причиною для пошуку, тобто користувач оцінює відповідність документа своїй інформаційній потребі. Мірою якості пошуку, що визначає наскільки добре результат пошуку задовольняє інформаційну потребу користувача, є партинентність. Інструменти підвищення партинентності, крім можливостей уточнення формулювання запитів, передбачають використання вагових критеріїв, що дозволяє ранжувати знайдені документи і видавати для перегляду користувачеві найбільш вагомі документи або обмежуватися видачею не більш ніж заданого числа вагомих документів. У сучасних системах проблемам релевантності, а особливо партинентності, приділяється все більше уваги.

Пошук документів, семантично схожих на заданий документ-зразок, можна розглядати як спрощений варіант зворотного зв'язку за релевантністю. При використанні функції пошуку семантично схожих документів, користувачеві для складання запиту необхідно вказати документ, що більшою мірою відображає його інформаційну потребу. Пошук документів, семантично схожих на документ-зразок, сприяє вирішенню проблеми коректного відображення його інформаційних потреб. Аналіз методів семантично схожих документів показав, що часто схожість між документами обчислюється на підставі критеріїв, визначених розробниками системи, і, як правило, не відомих користувачеві. Крім того, він не має можливості впливати на механізм пошуку схожих документів. Тому актуальним завданням є підвищення якості даного виду пошуку.

Основні кроки пошуку семантично схожих документів:

виконується лексичний, морфологічний аналіз, нормалізація термінів, видалення стоп-слів, виявлення значущих двослівних термінів;

здійснюється побудова списку термінів, що зустрічаються в документі, і обчислення частоти їх появи;

формуються кластери асоціативно пов'язаних значущих пошукових термінів документа-зразка. Метою даного кроку є побудова кластерів термінів, що відображають основний зміст документа;

дозволяє виконати уточнення інформаційної потреби користувача. Побудовані на попередньому кроці кластери термінів візуалізуються і виводяться на екран. Користувач має право видалити кластери або терміни, які виходять за рамки його пошукових інтересів. Можна уточнити запит за рахунок додавання асоціативно пов'язаних пошукових термінів, які не містяться в документі.

Пропонований алгоритм побудови пошукового зразка документа заснований на статистичному підході до аналізу текстів. Схема розробленого алгоритму зображена на рис. 3. Розглянемо його основні кроки.

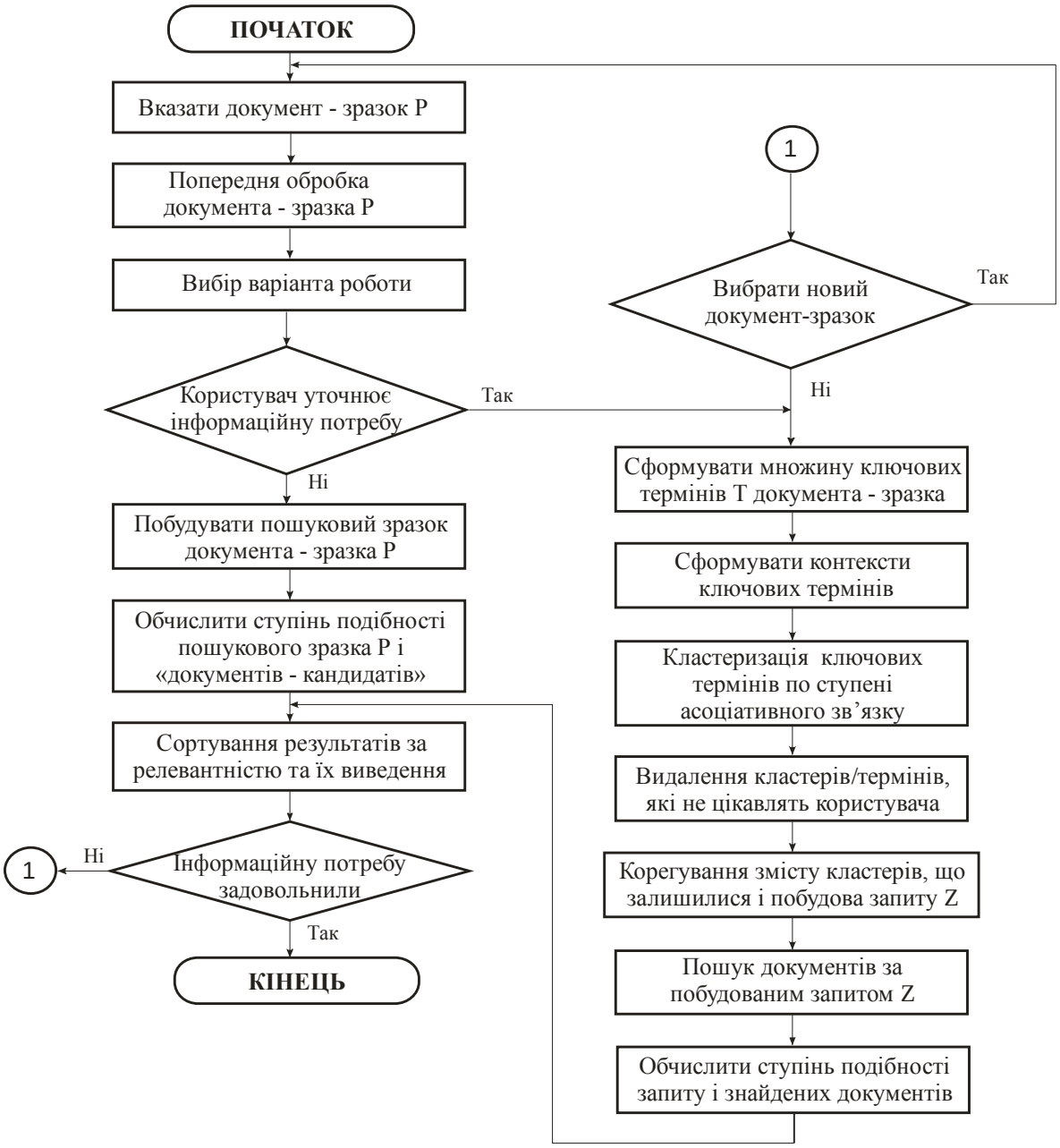

Рис. 3. Алгоритм пошуку семантично подібних документів

1. Виконати попередню обробку документа: лексичний аналіз, в ході якого виконується видалення розмітки, тобто елементів форматування, видалення пунктуації цифр, математичних формул; приведення регістра – перетворення всіх символів до верхнього або

нижнього регістру; видалення стоп-слів, тобто слів, які є допоміжними і несуть мало інформації про зміст документа. Зазвичай заздалегідь складаються списки таких слів, і в процесі попередньої обробки вони видаляються з тексту; морфологічний аналіз, полягає в перетворенні кожного слова до його початкової форми, яка виключає відмінювання слова, множинні форми, особливості усного мовлення і т.п.

2. Сформувати список пошукових термінів, що зустрічаються в документі.

3. Для всіх пошукових термінів, що містяться в списку, обчислити ступінь приналежності документу.

4. Сформувати матрицю, яка відображає частоту появи пари термінів t_i, t_j в одному контексті в документі d.

5. Для всіх пошукових термінів отримати контекст. Обчислити функцію приналежності.

Висновки

На основі проведеного дослідження запропоновано структуру інформаційної системи обробки даних, особливість якої полягає в принципах функціонування підсистем, націлених на підвищення партинентності і релевантності пошуку неструктурованої інформації, а саме: підсистеми діалогового режиму взаємодії з користувачем, підсистеми пошуку семантично подібних документів і підсистеми формування кластерів асоціативно зв'язаних значущих термінів документа.

Запропоновано метод та розроблено алгоритм пошуку семантично подібних документів, особливістю якого є можливість уточнення інформаційної потреби користувача та побудови більш точного пошукового запиту.

Література

1. Гагарина Л.Г. Разработка и эксплуатация автоматизированных информационных систем / Л.Г. Гагарина, Д.В. Киселев, Е.Л.Федотова // учеб. пособие. М.: ИД «Форум»: Инфа-М, 2007. – 384 с. **2. Гайдамакин Н.А.** Автоматизированные информационные системы, базы и банки данных / Н.А. Гайдамакин // Москва «Гелиос АРВ», – 2002. – 368 с. **3. Муляр І.В.** Метод пошуку текстових та графічних об'єктів в середовищі інформаційного забезпечення процесу діагностування/ І.В.Муляр // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2005. – №2. – С.94-97. **4. Муляр І.В.**Гіпертекстова модель представлення інформації в базах діагностичних даних / І.В. Муляр, В.М. Джулій // Вісник ТУП. – 2001. – №1. – С. 189-191.

В статье проведен анализ и предложена структура информационной системы обработки данных, алгоритм поиска семантически подобных документов.

На основе результатов исследования определена структура информационной системы обработки данных, особенность которой заключается в принципах функционирования подсистем, нацеленных на повышение партинентности и релевантности поиска неструктурированной информации, а именно: подсистемы диалогового режима взаимодействия с пользователем, подсистемы поиска семантически подобных документов и подсистемы формирования кластеров ассоциативно связанных значимых терминов документа.

Предложен метод и разработан алгоритм поиска семантически подобных документов, особенностью которого является возможность уточнения информационной потребности пользователя и построения более точного поискового запроса.

Ключевые слова: информационная система, поисковая системы, структурирована информация.

This paper analyzes the structure and proposed information system data processing algorithm for finding semantically similar documents.

Based on the results of the study determined the structure of the information processing systems, a feature which is based on the principles of functioning of subsystems aimed at improving search relevance partynentnosti and unstructured information, namely the subsystem dialog mode user interaction subsystem finding semantically similar documents and subsystems forming clusters associated air knitted important terms of the document.

The method and algorithm of finding semantically similar documents, a feature which is the ability to clarify the information needs of users and build a more precise search.

Key words: information system, search engine, unstructured information.