УДК 004.923:65.011.56
© **Я. Меденець[1], І. Стрепко[2], О. Тимченко[1], 2011**

МЕТОДИ КЛАСТЕРИЗАЦІЇ ДАНИХ У ТЕХНОЛОГІЧНИХ ПРОЦЕСАХ ПОЛІГРАФІЧНОГО ВИРОБНИЦТВА

Описані основні напрями вдосконалення автоматизації друкарських процесів. Проаналізовані варіанти декомпозиції. Проведений опис кластеризації параметрів у друкарському процесі. Запропоновані основи кластерного аналізу на методі k-середніх.

Basic directions of automation printing processes improvement have beed described. Decouplig variants analised. Printing process clasterisation parameters description conducted. Claster analysis foundations based on k-means method offered.

1. ВСТУП

Основними напрямами вдосконалення автоматизації друкарських процесів останніми роками є: розробка комплексних систем управління агрегатами друкарської машини і параметрами друкарського процесу; розробка систем автоматизованого налаштування фарбового апарату друкарської машини; створення систем збору і відображення параметрів друкарського процесу, а також управління машиною із застосуванням технологічного контролера (ТК) і персонального комп'ютера (ПК).

Продуктивність і якість друку постійно зростають. Значною мірою це зв'язано із застосуванням комп'ютерних систем управління, комп'ютеризованих електроприводів змінного струму. Світова практика показує, що при помірному збільшенні обсягів виробництва друкарських виробів спостерігається істотне підвищення їх якості та продуктивності машин, що свідчить про підвищення попиту і збільшені вимоги споживачів до якості друку.

Це є наслідком вдосконалення технологічних агрегатів машин, оснащення їх високодинамічними регульованими електроприводами, комп'ютерними засобами автоматизації і датчиками, що забезпечують контроль і регулювання електромагнітних, механічних і технологічних змінних, діагностику стану устаткування.

Метою роботи є вдосконалення процесу управління роботою рулонної ротаційної машини для підвищення якості виготовлення поліграфічної продукції.

[2] Українська академія друкарства

2. РУЛОННІ ДРУКАРСЬКІ МАШИНИ. КЛАСТЕРИЗАЦІЯ ПАРАМЕТРІВ

Процес вдосконалення технологічних агрегатів, електроприводів і засобів автоматизації є взаємозв'язаним процесом. Так, поява на ринку надійних і високодинамічних частотнорегульованих електроприводів змінного струму привела до постановки завдань підвищення швидкості машин в 2-2,5 рази, продуктивність сучасних машин досягає P=60 - 80 тис. відтисків на годину. Рулонні друкарські машини мають швидкість обертання друкарської пари до r=(1000-1500)об/хв при швидкості руху паперової стрічки V=8-15 м/с. Число фарб, що наносяться на задруковуваний матеріал, в машині досягає значень N=30-60. Несуміщення кольорів має бути не більше Δs =(0,05-0,1)мм. Значення товщини шару фарби на растрових елементах відбитків при офсетному друці дорівнюють Δt =(0,9-1,25)мкм.

З метою підвищення точності підтримки заданої швидкості V і натягу F паперового полотна між секціями машин, зростають вимоги до динаміки систем управління швидкістю і співвідношення швидкостей секцій. Підвищення швидкостей машин, безумовно, спричиняє за собою і необхідність підвищення якості систем управління технологічними змінними процесу виробництва друку, основним завданням якого є підтримка із заданою точністю основних технічних показників несуміщення кольорів і товщини шару фарби.

Вирішення завдань регулювання технологічних змінних можливо лише на основі використання спеціальних алгоритмів управління, що враховують взаємозв'язки змінних, транспортні запізнювання у виконавчих механізмах, прямих і перехресних каналах об'єкту управління.

Актуальним завданням є підвищення швидкодії виконавчих механізмів в каналах регулювання несуміщення кольорів Δs і товщини шару фарби Δt . У поліграфічних машинах явно виокремилась тенденція на заміну механічних засобів синхронізації друкарських циліндрів, регулювання натягу паперової стрічки, подачі фарби на електромеханічні засоби. Зі всього різноманіття друкарських машин найбільш складні і продуктивні — рулонні ротаційні друкарські машини (РРМ).

Під час друкарського процесу на основну консоль панелі управління багатофарбової РРМ надходять параметри від різноманітних давачів: параметри швидкості руху полотна, натягу, зволоження, подачі фарби, суміщення кольорів, подібні.

На якість друку також впливають й інші параметри, у даному випадку їх можна віднести до другорядних: якість фарби (відповідає за параметри насичення, яскравості) та зволожуючого розчину, темпера-

тура в приміщенні, значення відносної вологості, граматура паперу та ін. Причому лише перелік цих параметрів може сягати сотень.

3. СТРУКТУРА ЗБОРУ І ВІДОБРАЖЕННЯ ДАНИХ

Для будь-якої поліграфічної машини механічні показники не є вирішальними у друкарському процесі. Дані показники повинні постійно піддаватись електронній обробці різноманітними контролерами, а інформація передаватись на головний пульт управління в узагальненому вигляді.

Розпаралелювання програмних засобів керування у друкарському процесі зводиться до процесу декомпозиції отриманих результатів на незалежні, що не вимагають послідовного виконання і можуть бути виконані, відповідно, різними (контролерами) процесорами системи паралельно.

Розглянемо можливі варіанти декомпозиції: проста декомпозиція, функціональна і декомпозиція даних.

Тривіальна декомпозиція застосовується в тому випадку, коли різні копії лінійного коду програми керування можуть виконуватися незалежно одна від одної і не залежать від результатів, що отримані у процесі виконання інших копій коду. У цьому випадку основна програма, одержавши різні початкові параметри, може бути запущена на різних процесорах для отримування даних на різних ланках друкарського процесу.

При функціональній декомпозиції вихідна задача розбивається на ряд послідовних задач, що можуть бути виконані на різних вузлах незалежно від проміжних результатів, але строго послідовно. Кожна з цих задач може бути виконана на окремому вузлі. Загальний час обробки всього масиву елементів даних помітно зменшується за рахунок того, що відбувається обробка відразу кількох послідовних елементів масиву даних. Дані обробляються в режимі конвеєра.

На відміну від функціональної декомпозиції, коли між процесорами розподіляються різні задачі, декомпозиція даних припускає виконання на кожному процесорі однієї і тієї ж задачі, але над різними наборами даних. Частини даних спочатку розподілено між процесорами, що обробляють їх, після чого результати складаються в одному місці. Дані повинні бути розподілені так, щоб обсяг роботи для кожного процесора був приблизно однаковий, тобто декомпозиція має бути збалансованою. У випадку дисбалансу ефективність роботи може бути знижена.

У деяких виробничих процесах кількість параметрів може досягати тисяч, а частота, з якими вони оновлюються, може доходити до 10 ГГц. Зрозуміло, що процесор, який використовується на пульті ке-

рування, не зможе впоратись з такою кількістю інформації, а крім того, її ще потрібно опрацювати, перевірити, надати зворотні команди для виконання та здійснити запис у архів.

Таким чином, для ефективного опрацювання великої кількості різнорідних даних, виникає необхідність кластеризації параметрів та декомпозиції оброблюваних даних.

У системах оброблення інформації, кластером є:

– описувач абстрактного типу даних;
– підмножина об'єктів з певними наборами ознак.

Обчислювальний кластер, як і будь-яка система для паралельних обчислень, є ефективним, коли обчислювальна задача, яку необхідно вирішити, принципово не може бути вирішена за допомогою комп'ютерів широкого вжитку (наприклад, персональних комп'ютерів), або вирішення задачі за допомогою поширених систем вимагає тривалого часу.

До таких задач належать:

– задачі, що «не вміщуються» в оперативну пам'ять (вимагають десятки гігабайт і більше);
– обчислення, що вимагають значної кількості операцій і відповідно тривалого часу (дні, тижні, місяці);
– коли потрібно обрахувати велику кількість задач (десятки, сотні) за короткий проміжок часу.

Зауважимо, якщо задача ефективно вирішується за допомогою поширених систем, то використання кластеру може бути неефективним. Задачі базуються на обробці великих масивів даних, структурованих у регулярні структури (матриці або послідовності).

У випадку, коли оброблювана структура даних може бути розбита на регулярний (непересічний) масив підструктур (областей), задача може бути розподілена між процесорами і вирішена в паралельному режимі. Це дозволить скоротити час рішення задачі, або поставити задачу для більшого масиву даних (наприклад зробити різницеву сітку більш дрібною).

Виробничий цикл складається з багатьох процесів, кожний з яких обробляється своїм процесором й має свій адресний простір. Причому безпосередній доступ до пам'яті іншого процесу неможливий, а обмін даними між процесами відбувається за допомогою операцій приймання й посилки повідомлень.

Процес, який повинен одержати дані, викликає операцію *прийняти повідомлення*, і вказує, від якого саме процесу він повинен одержати дані, а процес, який повинен передати дані іншому, викликає операцію *надіслати повідомлення* і вказує, якому саме процесу потрібно передати ці дані.

Загалом високопродуктивний виробничий процес складних систем включає:

– завантаження параметрів;

– паралельне опрацювання інформації на окремих вузлах кластерної системи;

– отримання і об’єднання результатів окремих вузлів;

представлення результатів кластеризації та її візуалізація.

Основні етапи функціонування інформаційної технології обробки кластеризованих даних показані на рис. 1.

Базові параметри потрапляють на модуль завантаження параметрів роботи, де порівнюються з біжучими параметрами. Завантаження відбувається усіма паралельними процесами одночасно. Відбувається обчислення та перевірка, чи дані параметри підпадають під прийнятні критерії.

У модулі представлення результатів на основі опрацювання усієї інформації отримуємо загальний результат, що показує повну характеристику виробничого процесу. Дані виводяться на екран панелі управління та фіксуються.

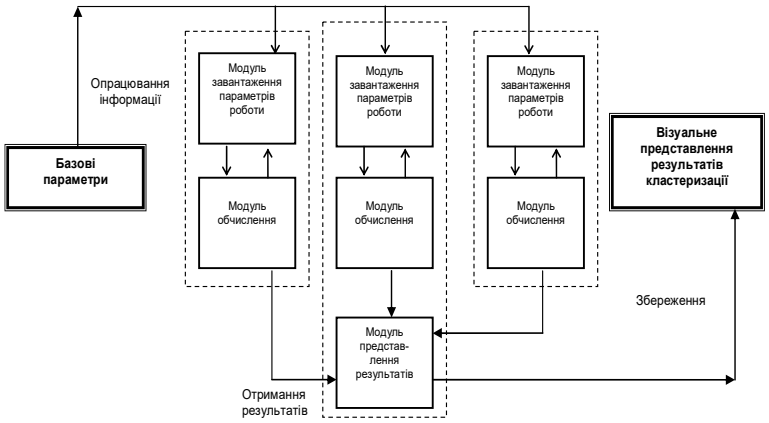

Рис. 1. Основні модулі інформаційної технології кластеризації технологічного процесу

На прикладі РРМ можна відобразити одним елементом (центороїдом) наступні параметри, що формують кластери даних:

– швидкість машини – відноситься до кластеру *швидкості*;

– вузли машини (секції подачі рулонів паперу, друкарські пари, фальцапарат) – кластер *вузлів*;

– якість фарби (насичення, яскравість) – кластер якості *відбитку*…

Основна ідея кластеризації – виділити з елементів основні, а також центр кластера, який повинен задовільняти обмеженням івимогам технологічного процесу.

Опишемо параметри друкарської машини такими функціональними блоками:

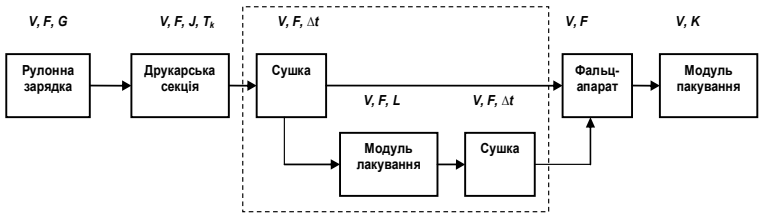

Рис. 2. Функціональна схема багато фарбової друкарської машини.

Продуктивність друкарської описуємо наступною залежністю:

$$Pr = < \forall \ Vi, Fi, Gg, Jj, \Delta t, Ll, Kk \ > , \quad (1)$$

де V – швидкість паперового полотна,
F – натяг стрічки,
i = 1..N, N – кількість фарбових апаратів,
G – якість паперу (граматура), g – ширина рулону,
J – якість фарби (яскравість, насиченість, густина),
j= 1..S, S – кількість фарб,
Tk – точність роботи,
k = 1..k, k – точність роботи окремих елементів,
Δt – зміна температури на ділянках сушки,
Ll – якість лаку,
l=0..1, l – лакування,
Km – блок пакування, m – задається оператором.

Як видно, параметри натягу та швидкості фігурують на усіх ділянках друкарської машини. Це є базові параметри, їх можна узагальнити, розрахувавши бажані показники при оптимальних параметрах.

Додатковими параметрами також виступають грама тура паперу та ширина полотна. Зрозуміло, що при різній ширині параметри натягу не будуть однаковими. Друкарська секція додатково вносить свої корективи у параметри швидкості та натягу, крім того, враховуємо також кількість друкарських секцій та фарбових апаратів. Саме тут потрібно досягнути максимальної точності, оскільки мова іде про накладання кольорів. Несуміщення кольорів відразу вкаже на неякісно налагоджений технологічний процес.

У випадку друкарської машини, де передбачений процес сушіння та лакування, потрібно також враховувати зміну температури на ділянках сушки. Аналогічно, визначальними параметрами у фальцапараті є швидкість та натяг паперового полотна.

4. МАТЕМАТИЧНА МОДЕЛЬ

Завдання кластеризації найчастіше вирішується методами математичної статистики. Завдяки їх використанню з'являється можливість автоматизації процесу кластеризації за рахунок електронно-обчислювальних засобів. Для адаптивних систем, реалізованих на програмному рівні, завдання кластеризації вирішуються лише за допомогою апарату математичної статистики.

Формальна постановка задачі кластеризації виглядає наступним чином: нехай X – множина об'єктів, Y – множина номерів кластерів.

При модернізації систем керування, а також розробці нових систем множина Y може бути невідомою. Задана функція відстані між об'єктами $p(x,x')$. Маємо кінцеву вибірку об'єктів:

$$X^m = \{x_1, ..., x_m\} \subset X . \quad (2)$$

Потрібно розбити вибірку на підмножини (кластери), що перетинаються, так, щоб кожен кластер складався з об'єктів, близьких до p, а об'єкти різних кластерів істотно відрізнялись.

При цьому кожному об'єкту $x_i \in X^m$ приписується номер кластера y_i.

Алгоритм кластеризації – функція $a: X \to Y$, що будь-якому об'єкту $x \in X$ ставить у відповідність номер кластера $y \in Y$. Множина Y у деяких випадках відома завчасно, однак найчастіше ставиться завдання визначення оптимального числа кластерів з точки зору того або іншого критерію якості кластеризації.

Кластеризація відрізняється від класифікації тим, що мітки вихідних об'єктів y_i не задані за замовчуванням, а також може навіть бути невідомою сама множина Y. Остання обставина вводить наступні корегування при використанні кластерного аналізу:

– відсутність можливості швидкого визначення класів досліджуваних об'єктів;

– відсутність можливості застосування стандартизованих методик;

– відсутність побудови таксономій.

Однак, статистична кластеризація володіє і рядом переваг:

– можливість задання завчасно відомого класу об'єктів згідно початкових характеристик;

– можливість кластеризації великої кількості об'єктів у стислі терміни;

– можливість уведення у досліджувані сукупності стандартизованих індикаторів.

Статистичні алгоритми засновані на припущенні, що кластери описуються деяким сімейством імовірнісних розподілень, а саме завдання кластеризації зводиться до розділу по кінцевій вибірці.

У основі кластерного аналізу лежить дві гіпотези баєсівського підходу по розділенню розподілень.

Гіпотеза 1. Об'єкти вибірки X^l з'являються випадково і незалежно, відповідно імовірному розподіленню, що являє собою сукупність розподілень

$$p(x) = \sum_{y \in Y} w_y p_y(x), \quad \sum_{y \in Y} w_y = 1, \tag{3}$$

де $p_y(x)$ – функція щільності розподілення кластеру y,

w_y – невідома апріорна імовірність появи об'єктів з кластеру y.

Конкретизуючи вид розподілення $p_y(x)$, частіше усього беруть сферичні гаусівські щільності. Це звичайна практика – уявляти кластери у вигляді еліпсоїдів обертання.

Гіпотеза 2. Об'єкти описуються n числовими ознаками $f_1(x),\ldots, f_n(x)$, $X=R^n$. Кожен кластер $y \in Y$ описується n-мірною гаусівською щільністю $p_y(x)=N(x;\mu_y,\sum_y)$ з центром $\mu_y=(\mu_{y1},\ldots,\mu_{yn})$ і діагональною коваріаційною матрицею $\sum_y=diag(\sigma^2_{y1},\ldots,\sigma^2_{yn})$.

При цих уявленнях задача кластиризації співпадає з задачею розділення ймовірностей розподілень і для її вирішення можна застосувати ЕМ-алгоритм. На Е-кроці, згідно формули Байеса, вираховують приховані змінні g_{iy}. Значення g_{iy} рівне імовірності того, що об'єкт $x_i \in X^l$ належить кластеру $y \in Y$. На М-кроці уточнюються параметри кожного кластеру $\mu_y,\sum_y$, при цьому використовуються приховані змінні g_{iy}.

Однак, ЕМ-алгоритм володіє недоліком, що у результаті кластеризації кожен об'єкт $x_i \in X^l$ приписується кожному кластеру з визначеною імовірністю.

Найчастіше з практичною метою використовується метод ***k-середніх***, що являється спрощенням ЕМ-алгоритму. Головною відмінністю є те, що у ЕМ-алгоритмі кожен об'єкт x_i розподіляється усіма кластерами з імовірністю $g_{iy} = P\{y_i = y\}$.

У алгоритмі *k-середніх* кожен об'єкт приписується лише одному кластеру, тому форма кластерів не формалізується. Існує спрощений варіант ЕМ, у якому форма кластерів також не буде налаштовуватись – для цього достатньо зафіксувати коваріційні матриці $\sum_y y \in Y$.

Можливий також варіант *k-середніх*, за яким будуть визначатись дисперсії кластерів вздовж координатних осей.

Загальний алгоритм кластеризації з використанням *k-середніх* можна зобразити наступним виглядом:

1) сформувати початкове наближення центрів усіх кластерів $y \in Y$. У якості центрів можна взяти найбільш віддалені один від одного об'єкти вибірки μ_y ;

2) присвоїти кожен об'єкт вибірки найближчому центру:

$y_i := argmin \underset{y \in Y}{p} (x_i, \mu_y), i = 1, ..., l;$

3) обчислити нове положення центрів:

$$\mu_{yi} := \frac{\sum_{i=1}^{l} [y_i = y] f_i(x_i)}{\sum_{i=1}^{l} [y_i = y]};$$
$$y \in Y, i = 1, ..., n$$

4) продовжувати обчислення поки y_i не перестануть змінюватись.

Алгоритм *k-середніх* найбільш повно відповідає вимогам простоти використання й зручності інтерпретації результатів кластеризації у випадку його використання у адаптивних системах. Велика кількість програмно-апаратних продуктів, що використовують алгоритм *k-середніх*, дозволяє реалізувати системи різного типу складності.

В процесі аналізу поліграфічного виробництва кожен параметр повинен бути пов'язаним з *n*-мірним вектором. Даний вектор містить у собі оцінки параметрів.

Отже, кожен об'єкт описується вектором:

$$x_i = \{x_{il}, ...x_{in}\}, \; x_i \in X^l, \qquad (4)$$

де X^l – сукупність векторів, що характеризує деякий кластер чи визначений напрямок (вектор).

Для проведення кластерного аналізу сукупності стратифікуючи векторів необхідне задання першочергових коректних умов кластеризації. Для виявлення цих умов потрібне першочергове сортування. Суть даної процедури заключається у тому, щоб із множини X^l виділити необхідну кількість стійких груп векторів, у котрих в подальшому будуть виділятись вектори, що мають середньостатистичний для даної даної групи векторів набір параметрів. Кожен такий вектор повинен представляти собою центр *n*-го кластеру x_{in} при подальшому проведенні кластеризації.

При проведенні ряду процедур кластеризації технологічного процесу поліграфічного виробництва формуються стійкі вектори – центри x_{in}. При заданні першочергової похибки можна чітко виділити ці стійкі вектори, які в подальшому будуть використовуватись в якості індикаторів при наступних процедурах кластеризації. Основною метою цих

індикаторів є пришвидшення процесу кластеризації за рахунок того, що у якості центрів кластерів будуть вибрані саме ці вектори.

Корегування індикаторів можна виконувати по бажанню або необхідності у випадках суттєвої зміни програми. Самі індикатори будуть представляти собою набори параметрів, характерні визначеним об'єктам.

Зрозуміло, що вказані операції неможливо проводити без програмно-апаратних засобів автоматизації. Для написання коректного програмного коду потрібно використовувати наступні операції:

– провести першочергове сортування, для чого з усієї сукупності виділити необхідне (виходячи з заданих умов) кількості кластерів;

– виділити вектори – індикатори. Для цього необхідно опрацювати декілька вибірок вихідних даних, використовуючи алгоритм *k-середніх*. У результаті отримуємо декілька стійких векторів x_{in}, що являються індикаторами певних параметрів. Чим більша кількість опрацьованих вибірок, тим точнішими будуть індикатори;

– використовувати індикатори x_{in} у якості центрів кластерів y_i при подальших операціях кластеризації.

Рішення задачі кластеризації принципово неоднозначне. Цьому є декілька причин:

– не існує однозначно найкращого критерія якості кластеризації;

– число кластерів, як правило, невідоме завчасно і встановлюється у відповідності з деяким суб'єктивним критерієм;

– результат кластеризації суттєво залежить від метрики, вибір котрої, як правило, також суб'єктивний і визначається експертом.

Під час друкарського процесу найбільш суттєвими є параметрами є швидкість V та натяг F. Дані параметри доцільно віднести до *кластеру швидкості* та *кластеру натягу* відповідно. Після цього формуємо стійкі стратифікуючи вектори – центроїди, які будуть основою даних кластерів. Отже, кожен параметр технологічного процесу може бути представлений у якості вектора, компонентами якого є характеристиками параметрів, що необхідні для його оцінки.

5. ВИСНОВКИ

Розподілення технологічного процесу на кластери дозволить розвантажити центральний процесор методом виключення з-під його розрахунків значної кількості інформації. Розподіленими системами легше управляти та вносити доповнення у вже існуючий процес. Крім того, за допомогою кластеру, ми отримуємо узагальнені параметри, що відображаються в загальній системі.

Управління локальними елементами відображається на основі диференціювання (оцінювання) параметрів взаємозв'язаних кластерів.

Відображується не значна кількість параметрів, а лише параметри, що впливають на якість друкарського процесу.

У результаті використання методів кластеризації отримуємо підвищення якості поліграфічної продукції: обслуговуючий персонал РРМ лише вносить мінімальні корективи, оператор звільнений від аналізу ідентичних параметрів; основні ж операції контролюються автоматикою, - локалізація обчислень відбувається у функціональних вузлах РРМ. Це суттєво економить час налаштування, а також дозволяє отримати більшу кількість поліграфічної продукції протягом одиниці часу.

1. Веретенников В. І. Управління проектами / В. І. Веретенников, Л. М. Тарасенко, Г. І. Гевлик. — К. : ЦНЛ, 2006. — 280 с. 2. Самарин Ю. Н. Автоматизация управления полиграфическим предприятием / Ю. Н. Самарин, П. К. Иванов // КомпьюАрт. — 2006. — № 8. — С. 40–47. 3. Булычев Ю.Г. Бурлай И.В. Системный подход к моделированию сложных динамических систем в задачах оптимизации с прогнозирующей моделью // АиТ. – М.: Наука, 1996. – № 3. – С. 34-47. 4. Яковлев Ю.П. Контролінг на базі інформаційних технологій. – К.: ЦНЛ, 2006. – 318с.