ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1999
Seria: INFORMATYKA z. 36 Nr kol. 1414

Janusz ŚWIERZOWICZ
Politechnika Rzeszowska, Zakład Informatyki

ANALIZA MOŻLIWOŚCI WYKORZYSTANIA TECHNIK DRĄŻENIA DANYCH

Streszczenie. Wzrost zastosowania baz danych w wielu dziedzinach wywołuje potrzebę tworzenia narzędzi nowej generacji dla automatycznej analizy danych. Narzędzia te są przedmiotem badań pozyskiwania wiedzy z baz danych, a szczególnie drążenia danych. W artykule przedstawiono problemy związane z eksploatacją systemu informacyjnego, wspierającego klasyfikację dorobku naukowego nauczycieli akademickich. Przedstawiono ogólną ideę, jak te problemy mogą być rozwiązywane wykorzystując metody drążenia danych. Opisano komponenty systemu oraz podano przykłady realizacji algorytmów. Zaprezentowano projekt modułów i przykłady zapewniające automatyczną analizę danych i wykrywania nieprawidłowości dotyczących klasyfikacji gromadzonej bibliografii pracowników.

ANALISYS OF USING DATA MINING TECHNIQUES

Summary. A huge increase of database applications in business and scientific domain has creating a need for a new generation of tools for automated data analysis. Knowledge discovery in databases and data mining aim at these tools. The paper deals with utilization issues of a computer-aided assessment of scientific achievements. General idea of using data mining methods in solving of these issues is depicted. Components of the system are described. Project of software modules in DBMS that using data mining techniques for automated database analysis is out-lined. These modules are used for verification of scientific achievements category.

1. Wprowadzenie

W ostatniej dekadzie obserwujemy gwałtowny wzrost zdolności systemów informacyjnych do generowania i gromadzenia danych. Postępy w gromadzeniu danych naukowych [1,4,7-10], komputeryzacja wielu dziedzin handlu (rozpowszechnienie kodów paskowych dla więk-

szości produktów) [2], administracji [2, 5], ochrony zdrowia [6] powodują, że mamy do czynienia z nadmiarem danych operacyjnych.

Z drugiej strony obserwujemy postępy w technologii przechowywania danych. Postęp ten dotyczy:

- sprzętu (dyski magnetyczne, CD-ROM, DVD) i jest wyrażony wzrostem pojemności, szybkości dostępu, obniżką ceny nośnika;
- metod dostępu do danych – przez rozpowszechnienie standardów (ODBC, Universal Data Access, CORBA);
- metod integracji systemów zarządzania bazami danych (systemy rozproszone, systemy wielobazowe);
- zastosowania hurtowni danych (*data warehousing*) i bieżącego przetwarzania analitycznego (*on-line analytical processing OLAP*).

Duża ilość danych przekracza możliwości tradycyjnych, ręcznych metod analizy danych i zastosowanie arkuszy kalkulacyjnych lub kwerend staje się nieskuteczne.

Istnieje zatem istotna potrzeba stosowania technik i narzędzi nowej generacji, zdolnych do inteligentnego wspomagania użytkownika w analizie danych dla pozyskiwania cennych informacji. Te narzędzia są przedmiotem rozwijającej się dziedziny badań *pozyskiwania wiedzy z baz danych - PWBD* (ang. *Knowledge Discovery in Databases KDD*) [3,5]. W dalszej części artykułu zdefiniowano proces PWBD i podstawowe metody drążenia danych oraz podano przykład zastosowania algorytmu PWBD w aplikacji bazodanowej.

Termin *pozyskiwanie wiedzy z baz danych* pojawił się jako hasło w 1989 r., do opisania złożonego procesu, ze szczególnym naciskiem na metody drążenia danych. Termin *drążenie danych - data mining* był szeroko używany przez statystyków, analityków danych i społeczność związaną z informacyjnymi systemami zarządzania, natomiast PWBD przez naukowców zajmujących się sztuczną inteligencją.

PWBD odnosi się do całkowitego procesu pozyskiwania użytecznej wiedzy z danych, podczas gdy drążenie danych odnosi się do stosowania algorytmów dla wydobywania wzorców z danych bez dodatkowych elementów procesu PWBD (takich jak włączanie wstępnej wiedzy i właściwej interpretacji wyników). Te dodatkowe kroki są niezbędne do uzyskania użytecznej wiedzy z danych. Czysto mechaniczne zastosowanie metod drążenia danych, krytykowane w literaturze statystycznej jako *fishing, dredging*, może dostarczać błędnych wzorców.

Odnajdywanie i interpretacja wzorców na podstawie danych odnoszą się zazwyczaj do interaktywnego i iteracyjnego procesu PWBD. Proces ten obejmuje powtarzalne zastosowanie specyficznych metod lub algorytmów drążenia danych oraz do interpretacji wzorów generowanych przez te algorytmy. Artykuł koncentruje się na elementach procesu pozyskiwania wie-

dzy łącznie ze wstępnym przetwarzaniem danych oraz interpretacją odkrytych wzorców jako wiedzy o szczególnych metodach i algorytmach drążenia danych ukierunkowanych na wydobywanie wzorców z danych operacyjnych.

PWBD jest dziedziną zainteresowań naukowców zajmujących się: statystyką, bazami danych, sztuczną inteligencją, pozyskiwaniem wiedzy dla systemów ekspertowych oraz wizualizacją danych. Obejmuje metody, algorytmy i techniki z różnorodnych dziedzin. Zasadniczym celem jest pozyskiwanie wiedzy z danych w kontekście dużych baz. Pokrewną dziedziną są hurtownie danych, odwołujące się do popularnego trendu w systemach informacyjnych, dotyczącego gromadzenia, czyszczenia i przekształcania danych transakcyjnych w dane zorientowane na realizację procesu podejmowania decyzji. Popularnym podejściem dla analizy danych w hurtowni jest bieżące przetwarzanie analityczne. Narzędzia OLAP zapewniają wielowymiarową analizę danych, rozszerzając wykorzystanie SQL. Zarówno narzędzia pozyskiwania wiedzy z baz danych, jak i OLAP są przykładami narzędzi nowej generacji dla inteligentnego wydobywania informacji i zarządzania.

2. Algorytmy drążenia danych w procesie pozyskiwania wiedzy z baz danych

Drążenie danych odnosi się do wydobywania wzorców i modeli z danych gromadzonych w bazach, przy czym w procesie PWBD algorytmy drążenia danych realizowane są iteracyjnie i interaktywnie.

2.1. Etapy procesu pozyskiwania wiedzy z baz danych

Proces pozyskiwania wiedzy z baz danych jest procesem wieloetapowym [2, 3, 5], w którym występują:

- *Analiza* dziedziny aplikacji, związanej z nią wstępnej wiedzy i celów użytkownika.
- *Selekcja* danych dla procesu PWBD, skupienie uwagi na podzbiorze próbek danych.
- *Czyszczenie i wstępne przetwarzanie danych* - usuwanie szumu lub odległych wartości, strategia obsługi brakujących pól danych.
- *Transformacja danych* – polegająca na redukcji i projekcji danych, odnajdywaniu użytecznych cech reprezentujących dane, w zależności od celu zadania.
- *Drążenie danych* obejmujące:
  - *Wybór zadania drążenia danych*: określenie celu procesu PWBD, np. klasyfikacja, regresja, grupowanie itd.

- *Wybór algorytmu(ów) drążenia danych*: wybór metody użytej do szukania wzorców w danych. Zawiera decyzje, jakie modele i jakie ich parametry mogą być odpowiednie i odpowiadają szczególnym metodom drążenia danych z ogólnym kryterium procesu PWBD (np. użytkownik może być bardziej zainteresowany w zrozumieniu modelu niż w jego predykcyjnych właściwościach).
- *Realizacja algorytmu drążenia danych*: szukanie wzorców w szczególnej reprezentacyjnej postaci lub zestawu takich reprezentacji: reguły lub drzewa klasyfikacji, regresja, grupowanie itd. Użytkownik może istotnie wspomóc metody drążenia danych przez poprawne przeprowadzenie poprzednich kroków.

- *Interpretacja odkrytych wzorców*, możliwy powrót do dowolnego z poprzednich punktów celem dalszej iteracji.
- *Konsolidacja pozyskanej wiedzy*: włączenie wiedzy w system prezentacji. W tym punkcie jest także zawarte sprawdzenie i rozwiązywanie potencjalnych konfliktów z uprzednio pozyskaną wiedzą.

Proces PWBD może wymagać istotnych iteracji i może zawierać pętle sprzężenia zwrotnego pomiędzy dowolnymi etapami. Większość prac z dziedziny PWBD koncentrowało się na etapie drążenia danych. Jednakże, dla skutecznego wdrożenia procesu PWBD, w praktyce istotne są również inne kroki [2].

Drążenie danych jest etapem w procesie PWBD, zawierającym algorytmy, za pomocą których, przy akceptowalnych ograniczeniach wydajności obliczeń, uzyskujemy wzorce opisujące zbiór danych. Drążenie danych dotyczy dopasowania modeli do obserwowanych danych lub do określania wzorców na podstawie obserwowanych danych.

2.2. Cele drążenia danych

Podstawowe cele drążenia danych sprowadzają się do predykcji i opisu. Predykcja umożliwia określenie przyszłych lub nieznanych wartości poprzez użycie pewnych zmiennych lub pól w bazie dla przewidywania innych interesujących nas wielkości. Opis koncentruje się na odnajdywaniu łatwo interpretowalnych wzorców opisujących dane.

2.3. Zadania drążenia danych

Wspomniane cele drążenia danych można najczęściej zrealizować rozwiązując zadania:

- *Klasyfikacji*, określającej funkcje przypisującą dane do jednej z predefiniowanych klas.
- *Regresji* (*estymacji, predykcji*), określającej funkcję odwzorowującą dane na przewidywaną wartość typu rzeczywistego.

- *Grupowania* wg podobieństwa oraz grupowania w homogeniczne podgrupy, w którym szuka się skończonego zestawu podgrup (klastrów), aby opisać dane. Grupowanie nie zależy od predefiniowanych klas.
- *Streszczania* – dotyczy metod znajdywania zwartego opisu zestawu danych.

2.4. Komponenty algorytmów drążenia danych

W każdym algorytmie drążenia danych można wyróżnić trzy główne komponenty: reprezentację modelu, oszacowanie modelu i metodę przeszukiwania.

Reprezentacja modelu to język służący do opisania odkrywanych wzorców. Jeżeli reprezentacja jest zbyt ograniczona, wówczas żadna ilość czasu treningowego lub przykładów nie będzie w stanie stworzyć dokładnego modelu danych. Tak więc ważne jest, żeby analityk danych w pełni pojmował reprezentacyjne założenia, właściwe dla określonego algorytmu. Należy stwierdzić, że większa reprezentacyjna moc modelu powoduje redukcję predykcyjnej dokładności ukrytych danych. Dodatkowo, poszukiwanie staje się bardziej złożone i interpretacja modelu jest zazwyczaj trudniejsza.

Oszacowanie modelu określa, jak dobrze wzorzec (model i jego parametry) spełnia kryteria procesu pozyskiwania wiedzy. Przy ocenie modelu mogą być brane pod uwagę zarówno kryteria logiczne, jak i statystyczne.

Metoda przeszukiwania składa się z dwóch komponentów: szukania parametrów i szukania postaci modelu. Algorytm szukania parametrów przeszukuje parametry optymalizujące kryteria oceny modelu dla zestawu danych ustalonej reprezentacji modelu. Szukanie postaci modelu występuje w pętli sprzężenia zwrotnego nad metodą szukania parametrów. Dla każdej określonej postaci modelu uruchamiana jest natychmiast metoda przeszukiwania parametrów, aby ocenić jego jakość. Implementacje metod przeszukiwania postaci modelu mają tendencje do stosowania heurystycznych technik przeszukiwań [3].

2.5. Przykłady metod drążenia danych

Istnieje duża różnorodność metod drążenia danych. Do najpopularniejszych metod należą:

- metody bazujące na przykładach *Example-based Methods*, w których stosuje się reprezentatywne przykłady z baz danych, aby aproksymować model. Do tej grupy należą algorytmy klasyfikacji najbliższego sąsiada – *nearest-neighbor classification*, algorytmy wnioskowania w oparciu o przypadki – *case-based reasoning* i algorytmy wnioskowania z pamięcią *Memory-Based Reasoning MBR* [11]. Dla metod tych wymagana jest dobrze zdefiniowana miara odległości między punktami w przestrzeni da-

nych i funkcja kombinacji uwzględniające wartości najbliższych sąsiadów w celu dokonania predykcji;

- metody nieliniowej regresji składające się z rodziny metod dla predykcji dopasowującej liniowe i nieliniowe kombinacje funkcji podstawowych dla kombinacji zmiennych wejściowych. Przykładami metod należących do tej grupy są metody wykorzystujące sieci neuronowe oraz metody adaptacyjnych splajnów;
- metody drzew decyzyjnych (dla modelowania predykcyjnego, klasyfikacji, streszczeń). Przykładami są metody *Classification And Regresion Trees CART* oraz *CHi-squared Automatic InDuction CHAID*.

W następnym rozdziale przedstawiono zagadnienia związane z eksploatacją systemu informacyjnego, wspierającego klasyfikację dorobku naukowego nauczycieli akademickich. Przedstawiono ogólną ideę, jak te problemy mogą być rozwiązywane z wykorzystaniem metody drążenia danych.

3. Opis systemu informatycznego dla klasyfikacji dorobku naukowego nauczycieli akademickich

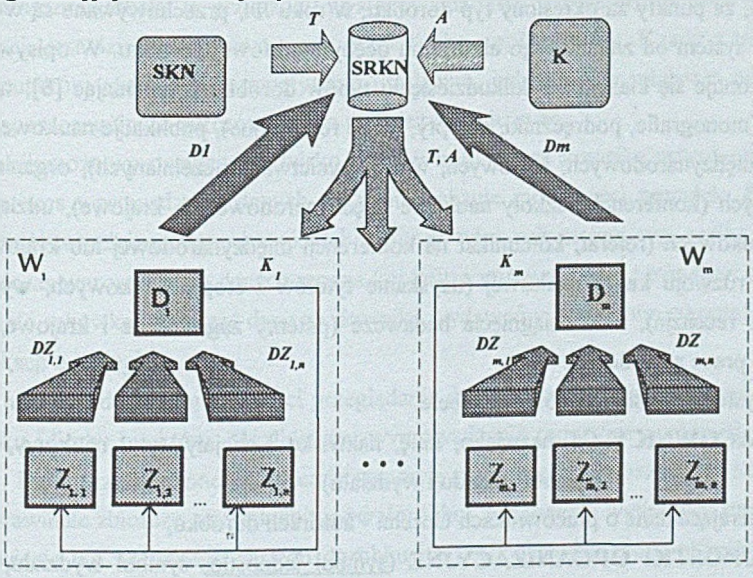
Jednym z czynników branych pod uwagę przy dotowaniu bieżących badań naukowych jest ocena dotychczasowej działalności naukowo-badawczej poszczególnych pracowników uczelni oraz zespołów badawczych. Jest to problem słabo strukturalizowany, trudno go ująć w sposób ilościowy. Najprostszym sposobem oceny dotychczasowego dorobku jest sklasyfikowanie dokonań pracownika lub zespołu, przyjmując za dany typ dorobku określoną liczbę punktów, a następnie ich zliczenie. Liczba punktów powinna odzwierciedlać wagę, wpływ dorobku pracownika na rozwój określonej dyscypliny naukowej. Omawiany w artykule komputerowy system klasyfikacji dorobku naukowego nauczycieli akademickich powstał jako narzędzie wspomagające ocenę ich dokonań [8-10]. W dalszej części artykułu przedstawiono strukturę przepływu informacji w systemie, stosowane modele danych, związki między tabelami oraz omówiono programy komputerowe wchodzące w skład systemu.

3.1. Struktura przepływu informacji w systemie

Informacje o dorobku naukowym i organizacyjnym pracownika, zebrane w określonym przedziale czasu, są jednym z elementów jego indywidualnej oceny oraz jednostki organizacyjnej Uczelni. Na rys. 1 przedstawiono strukturę przepływu informacji w systemie klasyfikacji dorobku naukowego.

Kryteria oceny dorobku, ustalone przez Senacką Komisję Nauki (SKN), są jednolite w skali całej Uczelni. Kryteria te w postaci tabeli punktów za określony rodzaj działalności są przekazywane z samodzielnej Sekcji Rozwoju Kadry Naukowej (SRKN) do poszczególnych jednostek organizacyjnych. Do sekcji dostarczana jest informacja o miejscu zatrudnienia pracowników - autorów dorobku z systemu kadrowego. Dorobek pracowników, wpisywany w zakładzie lub w katedrze, weryfikowany jest merytorycznie przez Wydziałową Komisję Nauki w dziekanacie i po korekcie gromadzony jest w SRKN. Tam dokonuje się ostatecznej korekty wprowadzonego dorobku i na jego podstawie określone zostają udziały dorobku jednostek organizacyjnych w dorobku całej Uczelni.

Zadaniem systemu jest zbieranie informacji o dorobku w bazie danych w celu szybkiego wyszukiwania informacji wg przyjętych kryteriów, tworzenia wykazów, zestawień zbiorczych oraz dostarczania władzom Uczelni informacji ilościowych o dokonaniach poszczególnych jednostek organizacyjnych w zadanym przedziale czasu.



Rys. 1. Schemat przepływu informacji w systemie klasyfikacji dorobku naukowego
Fig. 1. Structure of the information system for classification of scientific achievements

Na rys. 1 oznaczono: T – kryteria oceny dorobku, SKN – Senacka Komisja Nauki, SRKN – Samodzielna Sekcja Rozwoju Kadry Naukowej, K – system kadrowy, A – informacja o miejscu zatrudnienia pracowników – autorów dorobku, $Z_{i,j}$ – i-ty zakład j-tego wydziału, W_j – j-ty wydział, D_j – dziekanat j-tego wydziału, $DZ_{i,j}$ – dorobek pracowników spisywany w i-tym zakładzie i j-tym wydziale, K_j – weryfikacja merytoryczna i korekta dorobku dokonana przez wydziałową komisję nauki w j-tym dziekanacie.

3.2. Modele danych w systemie

Spośród technik organizowania informacji, zwanych modelami danych, możemy wyróżnić model relacyjnej bazy danych. Dane przechowywane są w tabelach składających się z wierszy i kolumn, atrybuty klucza głównego zostały podkreślone. W głównej tabeli systemu

DOROBEKN (ID, autorzy, tytuł, wydawca, tom, zeszyt, strony, rok, typ, zasięg, punkty, udziały, data_zap, data_kor, ...)

pamiętane są: identyfikator pozycji, nazwiska i inicjały autorów dorobku, tytuł, wydawca, tom, zeszyt, strony, rok publikacji, typ dorobku, rodzaj zasięgu dorobku, przyznane punkty za dorobek, procentowe udziały współautorów oraz automatycznie wstawiana jest data zapisu lub korekty danych oraz pewne atrybuty robocze.

Inną ważną tabelą jest tabela **PUNKT*nn*** z atrybutami

PUNKT*nn* (symbol, opis, punkty, ...).

Fakt, że punkty za określony typ dorobku, w roku *nn*, przechowywane są w tabeli, uniezależnia system od zmienianego algorytmu oceny punktowej dorobku. W opisywanym systemie dokonuje się klasyfikacji kilkudziesięciu typów dorobku rozróżniając [8]: wydawnictwa zwarte (monografie, podręczniki, skrypty, prace redakcyjne), publikacje naukowe (w czasopismach międzynarodowych, krajowych, w wydawnictwach uczelnianych), organizacje imprez naukowych (konferencje, szkoły naukowe międzynarodowe lub krajowe), udziały w imprezach naukowych (referat, komunikat na konferencji międzynarodowej lub krajowej), aktywność w rozwoju kadry naukowej (uzyskanie tytułów i stopni naukowych, wypromowanie doktora, recenzje), inne osiągnięcia badawcze (patenty zagraniczne i krajowe, wdrożenia, twórcze prace zawodowe).

W systemie występują również tabele:

PRACOWNICY (id, nazwisko, imię, nazwisko i inicjały, tytuł naukowy, stanowisko, symbol zakładu i wydziału)

zawierające dane o pracownikach Uczelni - autorach dorobku,

JEDNOSTKI ORGANIZACYJNE (symbol_jednostki, symbol wydziału, nazwa jednostki, suma punktów, udział),

EWIDENCJA TRANSMISJI DANYCH (symbol_jednostki, symbol wydziału, nazwa jednostki, nazwisko i imię osoby odpowiedzialnej, telefon, data importu i eksportu danych),

WYDAWNICTWA (symbol_wydawnictwa, nazwa wydawnictwa),

oraz kilka tabel roboczych, np. struktury katalogów, tabela przenoszonych pracowników, zestawień zbiorczych typów dorobku, ewidencji plików archiwalnych.

Bazy danych obsługują programy Dorobek Naukowy DNZ i Administrator Systemu DN_AS zainstalowane w zakładach oraz w Sekcji Rozwoju Kadry Naukowej.

3.3. Opis działania systemu

System wspomagający zarządzanie Uczelnią w zakresie oceny dorobku naukowego wykorzystuje programy DNZ i DN_AS. Umożliwia prowadzenie ewidencji dorobku naukowego, oblicza punkty za dorobek, tworzy na życzenie użytkownika wykazy publikacji oraz generuje różne zestawienia o dorobku. Programy są zabezpieczone przed niepowołanym dostępem oraz przed błędnymi działaniami użytkowników.

Źródłem informacji w systemie są dane o dorobku wprowadzone przez użytkowników w zakładach lub w katedrach. Każda pozycja wprowadzana jest tylko raz przez pierwszego autora. Poszczególnym pozycjom dorobku użytkownik przydziela określoną liczbę punktów, zgodnie z dostarczoną tabelą. W przypadku występowania kilku współautorów ilość punktów za daną pozycję jest dzielona automatycznie przez liczbę współautorów. Każdy z nich uzyskuje część punktów zgodnie z automatycznie ustalanym, jednakowym udziałem procentowym; udziały poszczególnych autorów można również zróżnicować.

Użytkownik może wprowadzać nazwisko i inicjały autorów ze słownika pracowników, co przyspiesza proces wprowadzania danych oraz minimalizuje możliwości pomyłek. Program bada poprawność wypełnienia danych w polach: autorzy, tytuł, rok, strony, typ, zasięg, punkty i udziały. Podczas wprowadzania danych można kopiować grupy nazwisk i inicjałów autorów ze schowka i do schowka, wybierać dane ze słownika wydawców, sprzężono również wybrane typy publikacji z zasięgiem.

Program ma rozbudowane możliwości przeglądania danych uporządkowanych wg autora, wydawcy, roku i liczby punktów. Możliwe jest wyprowadzanie raportów w różnych kodach polskich liter. Dane przetworzone w systemie stanowią podstawę do tworzenia wielu zestawień (np. zestawienie zbiorcze wszystkich typów dorobku, zestawienie wybranych typów dorobku, zestawienie wg wybranego zasięgu dorobku itp.). Użytkownik może przeglądać lub uzyskać wydruki dorobku naukowego dla katedry, zakładu lub wydziału oraz dla wybranego autora w określonym przedziale czasu zarówno z punktacją, jak i bez punktacji.

Program DN_AS stanowi uzupełnienie funkcjonującego w Sekcji Rozwoju Kadry Naukowej programu DNZ. Program ten jest wykorzystywany okresowo w przypadku zmian w strukturze uczelni (np. tworzenie nowego wydziału, łączenie wydziałów, tworzenie nowych zakładów, łączenie zakładów, przenoszenie zakładów z wydziału na wydział), przenoszenia pracowników pomiędzy wydziałami, zmian w regulaminie przyznawania punktów za dorobek naukowy i organizacyjny. Program umożliwia wówczas przekazanie niezbędnych parametrów dla prawidłowego funkcjonowania programu DNZ. Inną funkcją programu jest przygotowy-

wanie baz danych dla zakładów, import baz z zakładów oraz dołączenia danych z zakładów do istniejących struktur danych z wydziału na dysku twardym w SRKN. Program prowadzi również ewidencję przyjętych i wysłanych danych z zakładów oraz osób odpowiedzialnych za ich przygotowanie. Dane z dysku twardego są okresowo archiwizowane na dyskietkach w celu zabezpieczenia przed awarią systemu lub przypadkowym wykasowaniem. Program na życzenie użytkownika wykonuje kopie archiwalne oraz prowadzi ewidencję archiwizowanych zbiorów.

3.4. Wnioski z eksploatacji systemu

Opisywany system przyczynia się do zwiększenia efektywności zarządzania Uczelni w zakresie oceny dorobku naukowego. Ważnym problemem do rozwiązania w istniejącym systemie jest usprawnienie procesu weryfikacji istniejących danych, wyeliminowanie dublowania danych już raz wprowadzonych przez innych współautorów, wykrywania pozycji pod nieco zmienionym tytułem, wykrywania pozycji zbliżonych tematycznie. Problemy te rozwiązano wykorzystując algorytm drążenia danych oparty na metodzie wnioskowania z pamięcią MBR [11]. Rozwiązanie przedstawiono w następnym rozdziale.

4. Przykład wykorzystania metody drążenia danych w systemie klasyfikacji dorobku naukowego

W prezentowanym przykładzie przedstawiono wybrane etapy procesu PWBD wykorzystując metodę MBR.

- *Analiza dziedziny aplikacji:* ***Zadaniem algorytmu jest wykrycie w bazie danych identycznych lub podobnych pozycji dorobku naukowego.***
- *Selekcja danych dla procesu:* Docelowym zestawem danych będzie tabela ***DorobekN*** zawierająca kilkaset rekordów.
- *Czyszczenie i wstępne przetwarzanie danych:* – rozdzielanie omyłkowo złączonych wyrazów spacjami, korekta wyrazów.
- *Transformacja danych*:
  - Utworzenie tabeli ***WYRAZY (IDWyraz, Wyraz, IDDN)*** z polami: ***IDWyraz*** – identyfikator wyrazu, ***Wyraz*** – wyraz występujący w polu ***DorobekN.Tytuł, IDDN*** – identyfikator pozycji dorobku w tabeli ***DorobekN***.
  - Utworzenie tabeli ***WYJĄTKI*** i przepisanie do niej wyrazów z tabeli ***WYRAZY***, nieistotnych z punktu widzenia rozróżniania tekstu, np. spójników, przedimków itp.
  - Usunięcie wyjątków z tabeli ***WYRAZY***.

– Utworzenie tabeli *Częstości wyrazów* (*Wyraz, Liczba wystąpień, Częstość, Waga*), gdzie:

Wyraz – szukany wyraz,

Częstość = 1/*liczba wystąpień* (1)

Waga = – *Ln(Częstość)/Ln(2)* (2)

- *Drążenie danych* obejmujące:
  - *Wybór zadania drążenia danych*: grupowanie podobnych tytułów dorobku.
  - *Wybór algorytmu(ów) drążenia danych*: Metoda MBR

    W metodzie MBR należy:
    - Wybrać treningowy zestaw danych

      Wybrany zestaw liczy kilkaset rekordów, pole tytułu każdego rekordu zawiera kilka (kilkanaście) wyrazów.
    - Określić funkcję odległości **d(A,B)** między punktami A i B w przestrzeni danych. Funkcja ta określa miarę podobieństwa między dwoma pozycjami powinna spełniać następujące własności:

      $d(A,B) \geq 0$ (3)

      $d(A,A)=0$ (4)

      $d(A,B)=d(B,A)$ (5)

      $d(A,B) \geq d(A,C)+d(C,B)$ (6)

      Dla wyznaczenia funkcji **d** wprowadzono funkcję $\delta(A,B)$

      $\delta(A,B)=\Sigma$**wag wyrazów wspólnych A,B,** (7)

      tak więc

      $d(A,B)=1-\delta(A,B)/\delta(A,A)$ (8)
    - Utworzyć tabele **δAB(ID_A,ID_B,SumaWaga)**

      Oraz **dAB(ID_A,ID_B,d(A,B))**
  - *Realizacja algorytmu drążenie danych*:

    Przykładowe wyniki realizacji algorytmu przedstawiono w tabeli 1. Wyświetlono wartości uporządkowane pod względem odległości.
- *Interpretacja odkrytych wzorców:*
  - Pozycje (27,648) oraz (162, 691) mają odległość 0 ⇒ prace mają taki sam tytuł (z punktu widzenia algorytmu) ⇒ należy porównać ich treść.
  - Pozycje (46,53) mają odległość bliską 0, prace mają taki sam tytuł ⇒ należy porównać ich treść.
  - Pozycje (306,371), (112,731), (169,694) wykazują pewne podobieństwo⇒należą do tej samej tematyki.

- Pozycje (149,803) mają wspólny wyraz często występujący w zestawie danych⇒należą do różnej tematyki.

Tabela 1

Przykładowe wyniki realizacji algorytmu MBR

ID_A	ID_B	Odległość	A.TYTUŁ	B.TYTUŁ
27	648	0	Metoda wyznaczania luzów międzyzębnych w falowej przekładni zębatej	Komputerowa metoda wyznaczania luzów międzyzębnych w falowej przekładni zębatej.
162	691	0	Opis ruchu 3-kołowego mobilnego robota	Opis ruchu mobilnego robota kołowego
46	53	9,3811E-02	The relation between grain size and surface roughness afterplastic deformation of M85 brass sheets	The relation between grain size and surface roughness afterplastic deformation of M85 brass sheet
306	371	0,12206454623	On some transformations of the Caratheodory class of functions	On some subclass of Caratheodory functions and corresponding class of starlike functions
112	731	0,14963587030	Badania własności kompozytu Ti-SiC	Wytwarzanie i własności kompozytu Ti-SiC
169	694	0,30702998551	Sterowanie krzepkie ruchem nadążnym mobilnego robota kołowego	Algorytm sterowania ruchem nadążnym mobilnego robota kołowego
172	690	0,45398922506	Adaptacyjne sterowanie w ruchu nadążnym mobilnego robota kołowego	Identyfikacja dynamicznych równań ruchu mobilnego robota kołowego
149	803	0,93536247486	Computation by large Particle Method on Modern Supercomputer.Numerical Methods and Computational Mech. in Science end	The methods for spheroidal graphite iron producing

5. Uwagi końcowe i wnioski

Prezentowany system, rozpatrywany jako narzędzie dla ilościowego oszacowania dorobku naukowego nauczycieli akademickich, wspomaga władze uczelni w zakresie prowadzenia polityki dotyczącej rozwoju kadry naukowej. Wyposażenie go w mechanizmy drążenia danych usprawnia analizę gromadzonych danych.

LITERATURA

1. Baryłko M., Łoś B., Markiewicz A.: Baza danych wyników naukowych pracowników Uniwersytetu Gdańskiego, Materiały konferencji INFOBAZY'97, Gdańsk, s.310-312, 1997.
2. Brachmann R. J., Tej A.: The process of Knowledge Discovery in Databases. In Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. – editors, Advances in Knowledge Discovery and Data Mining, pp.37-53, AAAI Press/The MIT Press, Menlo Park, CA, 1996.
3. Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R.: From data mining to knowledge discovery: An overview. In Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. – editors, Advances in Knowledge Discovery and Data Mining, pp1-34, AAAI Press/The MIT Press, Menlo Park, CA, 1996.
4. Magyar G.: Quality control and enhancement of R+D activity in the higher education in Hungary, Information Systems Development '98, Bled, Slovenia, 1998.
5. Mannila H.: Methods and problems in data mining, Proceedings of International Conference on Database Theory, Delphi, Greece 1997, Springer –Verlag.
6. Matheus C. J., Piatetsky-Shapiro G.: Selecting and reporting what is interesting. In Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. – editors, Advances in Knowledge Discovery and Data Mining, pp495-515, AAAI Press/The MIT Press, Menlo Park, CA, 1996.
7. Plichta S., Jasińska-Suwada S.: MAGIC jako system do budowy oprogramowania wspomagającego pracę wyższej uczelni. Materiały konferencji INFOBAZY'97, Gdańsk, str.108-115., 1997.
8. Świerzowicz J.: A Management Information System for Classification of Scientific Achievements. Evolution and Challenges in System Development, pp.735-740, Kluwer Academic/Plenum Publishers, New York 1999 (przyjęto do druku).
9. Świerzowicz J.: Komputerowy system klasyfikacji dorobku naukowego nauczycieli akademickich, Materiały II Krajowej Konferencji „Komputerowe wspomaganie badań naukowych", WTN, Wrocław 1995.
10. Świerzowicz J.: An information system for classification of scientific achievements", The annual Fifth International Conference, UKRSOFT-95, Ukraine, Lwow 1995.
11. Waltz D., Brij M., Linoff G.: Classifying News Stories using Memory Based Reasoning, Proceedings, SIGIR'92, ACM Press, 1992.

Recenzent: Dr inż. Katarzyna Stąpor

Wpłynęło do Redakcji 9 kwietnia 1999 r.

Abstract

A huge increase of database applications in business and scientific domain has creating a need for a new generation of tools for automated data analysis. Knowledge discovery in databases and data mining aim at these tools.

The paper deals with utilization issues of a computer-aided assessment of scientific achievements. General idea of using data mining methods in solving of these issues is depicted. Components of the system: data models, data structures, software and data flow are described. Input-output procedures: data acquisition, data preparation, and data retrieval are discussed. Project of software modules in DBMS that using data mining techniques for automated and intelligent operational database analysis is outlined. These modules are used for verification of scientific achievements category. Finally, basic features of the system: fault tolerance, fault detection, flexibility and users' friendliness are delineated.

The system presented above supports University authorities in the area of staff development policy making. The system provides facilities for effective assessment of scientific achievements and may be considered as an assessment support tool for quantitative measuring of scientific achievements. Equipping the system with data mining mechanism further facilitates data analysis