http://dx.doi.org/10.18778/7969-549-2.02

Grzegorz Bryda
Uniwersytet Jagielloński

CAQDAS, Data Mining i odkrywanie wiedzy w danych jakościowych

Streszczenie. Celem artykułu jest refleksja metodologiczna nad procesem rozwoju wspomaganej komputerowo analizy danych jakościowych (CAQDAS) od tradycyjnej analizy jakościowej (Qualitative Analysis) opartej przede wszystkim na teorii ugruntowanej, poprzez analizę treści (Qualitative Content Analysis), w kierunku wykorzystania w socjologii jakościowej czy naukach społecznych zaawansowanych metod eksploracji danych i odkrywania wiedzy (Data Mining, DM and Knowledge Discovery in Datasets, KDD). Rozwój technologii informatycznych w zakresie gromadzenia i przetwarzania informacji oraz algorytmów i technik analitycznych doprowadził do sytuacji, w której wykorzystywanie ich osiągnięć na gruncie socjologii jakościowej i nauk społecznych staje się naturalnym procesem rozwoju CAQDAS. Obecnie wykorzystywanie CAQDAS w obszarze socjologii jakościowej jest na tyle powszechne, że nie budzi zdziwienia, że coraz więcej badaczy, także w Polsce, sięga po oprogramowanie komputerowe w analizie danych jakościowych. Specyfika CAQDAS uczy swoistego rygoryzmu metodologicznego, dokładności i precyzji w procesie analizy danych jakościowych, co pozytywnie odbija się na jakości prowadzonych analiz i badań. Jednakże analiza danych jakościowych wykorzystująca metodologię Data Mining to *novum* na gruncie socjologii jakościowej. Wiąże się to nie tylko z rozwojem nowych algorytmów czy technik analitycznych, ale także ze zmianami w podejściu do komputerowej analizy danych jakościowych, wzbogacaniem programów o możliwości pogłębionej analizy treści i struktury lingwistycznej dokumentów tekstowych. W obszarze CAQDAS towarzyszy temu zjawisku obserwowany od kilku lat zwrot metodologiczny w kierunku paradygmatu *mixed-methods* w naukach społecznych, a w szczególności w badaniach jakościowych. Jego konsekwencją jest implementacja wielowymiarowych technik statystycznej analizy danych, technik eksploracji danych tekstowych (Text Mining), a także algorytmów z dziedziny inteligencji komputerowej czy przetwarzania języka naturalnego w programach do wspomaganej komputerowo analizy danych jakościowych (QDA Miner, Qualrus czy T-Lab). Zdecydowana większość tych rozwiązań ma swe korzenie właśnie w dynamicznie rozwijającej się od kilkunastu lat metodologii Data Mining. Jeśli oprogramowanie CAQDAS wykorzystuje się najczęściej do pracy z mniejszymi zbiorami danych jakościowych, to Data Mining pozwala na prowadzenie analiz, w których wielkość zbioru danych jest w zasadzie nieograniczona. Celem tego artykułu jest przybliżenie środowisku badaczy jakościowych w Polsce metodologii Data Mining i odkrywania wiedzy w danych, a tym samym zachęcenie do eksperymentowania z nowymi podejściami w obszarze CAQDAS. W artykule staram się także ukazać relacje pomiędzy CAQDAS i teorią ugruntowaną a Data Mining i procesem odkrywania wiedzy w danych na gruncie socjologii jakościowej i szerzej – nauk społecznych.

Słowa kluczowe: analiza danych jakościowych, teoria ugruntowana, Data Mining, odkrywanie wiedzy w danych, CAQDAS, metody mieszane (*mixed-methods*).

Wstęp. Komputerowa analiza danych jakościowych

W ciągu ostatnich kilkunastu lat w naukach humanistycznych i społecznych coraz bardziej odczuwalny jest wpływ nowych technologii informatycznych na sposób prowadzenia badań, proces analizy danych i teoretyzowania. Wpływ ten wiąże się bezpośrednio z ideą szeroko rozumianej digitalizacji nauk humanistycznych i społecznych określanej jako Digital Humanities, Digital Social Sciences. Digital Humanities jest dziedziną nauki, prowadzenia analiz i badań, nauczania, która powstała na styku informatyki i dyscyplin humanistycznych. Skupia się na badaniu wpływu elektronicznych form zapisu danych tekstowych na rozwój tych dyscyplin oraz na tym, co te dyscypliny oraz nauki humanistyczne wnoszą do rozwoju wiedzy informatycznej. Za początek digitalizacji nauk humanistycznych uznaje się pionierską pracę z końca lat 40. XX w. *Index Thomisticus*[1] włoskiego jezuity Roberto Brusa. Wsparcie ze strony firmy IBM pozwoliło mu na wykorzystanie ówczesnych komputerów do archiwizacji oraz analizy lingwistycznej i literackiej dzieł św. Tomasza z Akwinu oraz powiązanych z nim autorów. Idea elektronicznego kodowania tekstów pisanych, zapoczątkowana przez Brusa, rozwijała się w kierunku stworzenia standardowego schematu kodowania humanistycznych tekstów elektronicznych i stałą się podstawą wdrożenia osiągnięć z zakresu informatyki w obszarze humanistyki. W konsekwencji w 1987 r. uruchomiono projekt Text Encoding Initiative, którego celem było opracowanie standardów digitalizacji tekstów humanistycznych. W 1994 r. opublikowano pierwszą wersję wytycznych w tym zakresie[2]. Od drugiej połowy lat 90. XX w. zaczęły pojawiać się elektroniczne archiwa danych tekstowych i graficznych, na początku w Stanach Zjednoczonych, później zaś w Europie. Digitalizacja tekstów w naukach humanistycznych nie szła w parze z możliwościami komputerowej analizy dużych zbiorów danych tekstowych. Te dopiero pojawiły się wraz z rozwojem algorytmów drążenia danych (Data Mining) i większymi zasobami obliczeniowymi współczesnych komputerów.

Digitalizacja w polu nauk społecznych, w tym w socjologii, miała odmienny charakter. Zainteresowanie technologiami informatycznymi skupiało się na możliwościach wykorzystania komputerów w obszarze analiz danych i badań empirycznych[3]. Udokumentowane zastosowanie programów komputerowych w analizie danych ilościowych w naukach społecznych datuje się na drugą połowę lat

[1] Zob. strona projektowa Index Thomisticus, www.corpusthomisticum.org/it/.

[2] Zob. strona projektowa The TEI Guidelines for Electronic Text Encoding and Inter Change, www.tei-c.org/Guidelines/.

[3] Charakterystykę wzajemnego wpływu i kształtowania się relacji między oprogramowaniem do wspomaganej komputerowo analizy danych jakościowych a procesem badawczym można znaleźć w artykule Brydy (2014).

60. XX w. (Brent, Anderson 1990; Tesch 1990). W tym czasie powstały funkcjonujące do dziś programy do statystycznej analizy danych ilościowych SPSS (obecnie IBM Statistics) czy Statistica. Początkowo były to narzędzia o ograniczonej funkcjonalności, jednakże wraz z rozwojem technologii informatycznych deweloperzy wzbogacali je o nowe algorytmy i techniki analityczne. Idea wspomaganej komputerowo analizy danych jakościowych ma również długą tradycję w naukach społecznych. Pierwsze udokumentowane zastosowanie komputerów w analizie danych jakościowych odnosi się do publikacji z 1966 r. *The General Inquirer: A Computer Approach to Content Analysis* autorstwa Philipa J. Stone'a, Dextera C. Dunphyego, Marshalla S. Smitha i Daniel M. Ogilvie pokazujące możliwości wykorzystania komputerów do analizy treści, np. danych antropologicznych (etnograficznych), ale także konieczność nowego spojrzenia na sposób definiowana analizy treści[4]. Oczywiście powszechność tego typu rozwiązań była ograniczona ze względu na brak łatwego dostępu do komputerów i oprogramowania analitycznego, które trzeba było tworzyć na potrzeby konkretnych projektów badawczych realizowanych przez humanistów i przedstawicieli nauk społecznych[5].

Dopiero w latach 80. XX w. na szerszą skalę zaczęły powstawać programy do wspomaganej komputerowo analizy danych jakościowych (CAQDAS, ang. *Computer Asssisted Qualitative Data Analysis Software*). CAQDAS rozwijano dla komputerów na platformie IBM PC w Stanach Zjednoczonych, Niemczech, Wielkiej Brytanii, Danii, Holandii i Australii. Jednakże wraz z pojawieniem się pierwszych programów – takich jak Text Base Alpha, Ethno, Qualpro, TAP czy The Ethnograph (Tesch 1990; Drass 1989; Fischer 1994) – wykorzystanie komputerów w analizie danych jakościowych budziło szereg kontrowersji wśród badaczy jakościowych. Na przełomie lat 80. i 90. XX w. w wielu publikacjach naukowych w socjologii, dotyczących wspomaganej komputerowo analizy danych, przewijała się debata na temat możliwości oraz pozytywnych i negatywnych skutków zastosowania oprogramowania w badaniach jakościowych (Conrad, Reinharz 1984; Richards, Richards 1989; Richards, Richards 1991; Seidel 1991; Kelle 1995). Punktem zwrotnym w rozwoju oprogramowania do analizy danych jakościowych było powołanie do życia, w 1994 r. na University of Surrey, CAQDAS Networking

[4] General Inquirer to system analizy danych tekstowych rozwijany od lat 60. XX w. przy wsparciu USA National Science Foundation and Research Grant Councils of Great Britain and Australia. Do połowy 1990 r. rozwijany był na dużych komputerach typu mainframe IBM obsługujących język programowania PL/1, następnie przy wsparciu Gallup Organization został przeprogramowany przez Philipa Stone'a w języku TrueBasic, a później ponownie napisany w języku Java przez Vanja Buvaca. System nie jest rozwijany komercyjnie.

[5] Obecnie system General Inquirer umożliwia analizy treści w języku angielskim z wykorzystaniem słowników „Harvard" i „Lasswell" oraz słowników rozwijanych przez użytkowników. Zob. strona projektu General Inquirer, www.wjh.harvard.edu/~inquirer/homecat.htm; strona projektowa Laswell Value Dictionary, www.wjh.harvard.edu/~inquirer/lasswell.htm.

Project, którego celem stała się integracja środowiska badaczy jakościowych przez: dostarczanie informacji, organizowanie szkoleń z zakresu wykorzystania programów do komputerowej analizy danych jakościowych, tworzenie platformy dla debaty dotyczącej kwestii analitycznych, metodologicznych i epistemologicznych wynikających z korzystania z oprogramowania CAQDAS oraz prowadzenie badań socjologicznych dotyczących ich zastosowań[6].

W ciągu ostatnich dwóch dekad, wraz z rozwojem technologii informatycznych na masową skalę, zaczęto szerzej korzystać z programów CAQDAS w badaniach jakościowych wykorzystujących technikę indywidualnych i grupowych wywiadów socjologicznych oraz analizę treści dokumentów tekstowych (Berelson 1952; Krippendorf 1986; Becker, Gordon, LeBailly 1984; Gerson 1984; Brent 1984; Pfaffenberger 1988). Pierwsze programy CAQDAS były pisane przez badaczy-entuzjastów, którzy nie tylko sami realizowali badania terenowe czy prowadzili analizy, lecz także posiadali umiejętności programowania lub znali kogoś, kto je posiadał. Wielu rozwijało programy niezależnie od siebie, często pozostając nieświadomymi faktu, że inni również pracują nad tego typu narzędziami analitycznymi. Programy rozwijano w zgodzie z indywidualnym podejściem badaczy do procesu analizy i dominującą ówcześnie metodologią badań jakościowych. Największy wpływ na rozwój oprogramowania CAQDAS miały metodologia teorii ugruntowanej i analizy treści (zob. Berelson 1952; Bong 2002; Glaser, Strauss 2009). Obecnie pierwotne różnice między programami CAQDAS zacierają się ze względu na postępującą ich komercjalizację oraz podobieństwo oferowanych funkcjonalności. Towarzyszy temu implementacja nowych technik i algorytmów analitycznych z zakresu pogłębionej eksploracji danych jakościowych, w tym danych tekstowych. Wiąże się to ze zmianami w podejściu do komputerowej analizy danych jakościowych, wzbogacaniem jej o analizę treści i struktury lingwistycznej dokumentów tekstowych. W obszarze CAQDAS towarzyszy temu zwrot metodologiczny w kierunku paradygmatu *mixed-methods* w naukach społecznych, a w szczególności w badaniach jakościowych (Tashakkori, Teddlie 2003). Jego wyrazem jest proces przechodzenia od tradycyjnej analizy danych jakościowych (Qualitative Analysis), przez Qualitative Content Analysis, w kierunku pogłębionej eksploracji danych jakościowych Text Mining wykorzystującej techniki statystyczne i algorytmy z dziedziny inteligencji komputerowej[7] czy przetwarzania języka

[6] Zob. strona projektowa The CAQDAS Networking Project, www.surrey.ac.uk/sociology/research/researchcentres/CAQDAS/about/.

[7] Sztuczna inteligencja (Artificial Intelligence, AI) to dziedzina badań naukowych informatyki na styku z neurologią, psychologią i kognitywistyką, obejmująca logikę rozmytą, obliczenia ewolucyjne, sieci neuronowe itp. Zajmuje się tworzeniem modeli zachowań inteligentnych oraz programów komputerowych symulujących te zachowania. Termin wymyślił amerykański informatyk John McCarthy. Inteligencja komputerowa (Computational Intelligence, CI) to dziedzina nauki zaj-

naturalnego[8]. Text Mining ma swe korzenie w rozwijającej się od kilkunastu lat metodologii Data Mining. Celem tego artykułu jest przybliżenie metodologii Data Mining środowisku badaczy jakościowych w Polsce oraz refleksja nad możliwościami wykorzystania eksploracji danych i odkrywania wiedzy w obszarze socjologii jakościowej oraz wspomaganej komputerowo analizy danych jakościowych.

Data Mining. Eksploracja i odkrywanie wiedzy w danych

Od kilkunastu lat można zaobserwować zarówno gwałtowny wzrost liczby informacji gromadzonych w formie elektronicznej, jak i rozwój technologii pozyskiwania, zapisu danych oraz ich magazynowania w postaci dużych baz danych: repozytoriów, hurtowni, archiwów statystycznych, sondażowych czy dokumentów tekstowych. Można je spotkać w każdym obszarze życia codziennego, począwszy od baz danych dotyczących transakcji bankowych, informacji z kas fiskalnych, rejestrów użycia kart kredytowych, zestawień rozmów telefonicznych, przez statystyki urzędowe, archiwa danych statystycznych i sondażowych, aż po rejestry medyczne, biologiczne itp. Zjawisku temu towarzyszy rozwój technologii informatycznych w zakresie przetwarzania i statystycznej analizy danych, algorytmów lingwistyki komputerowej czy sztucznej inteligencji. Wiąże się to z rozwojem metodologii w zakresie technik i algorytmów analitycznych służących modelowaniu procesów lub zjawisk społecznych. Kluczowe znaczenie odgrywa w tym rozwoju eksploracja danych (ang. *Data Mining*) określana także jako: drążenie danych, pozyskiwanie wiedzy, wydobywanie danych, ekstrakcja danych. Data Mining to podstawowy etap procesu odkrywania wiedzy w bazach danych (ang. *Knowledge Discovery in Databases*, KDD)[9]. Logika KDD zawiera się w sekwencji następujących etapów: zrozumienia danych, wyboru danych do analizy, wstępnego przetworzenia danych, przekształcenia danych do analizy, przeprowadzenia

mująca się rozwiązywaniem problemów, które nie są efektywnie algorytmizowalne za pomocą obliczeń. CI wykorzystuje metody matematyczne z wielu dziedzin, korzysta z inspiracji biologicznych, biocybernetycznych, psychologicznych, statystycznych, matematycznych, logicznych, informatycznych, inżynierskich i innych, jeśli mogą się one przydać do rozwiązywania efektywnie niealgorytmizowalnych problemów. W skład CI wchodzą: sieci neuronowe, logika rozmyta, algorytmy genetyczne i programowanie ewolucyjne, metody uczenia maszynowego, rozpoznawania obiektów (*pattern recognition*), metody statystyki wielowymiarowej, metody optymalizacji, metody modelowania niepewności – probabilistyczne, posybilistyczne itp.

[8] Charakterystyka Text Mining została przedstawiona w artykule znajdującym w tej publikacji (Bryda, Tomanek 2014).

[9] Termin ten zrodził się w obszarze badań nad sztuczną inteligencją. Data Mining jest przede wszystkim wykorzystywany w biznesie, stąd ostatnim etapem metodologii KDD jest zazwyczaj implementacja i integracja modeli analitycznych z systemami bazodanowymi.

eksploracji w celu odkrycia struktury wzorców i zależności, konstruowania modeli analitycznych, oceny stopnia dopasowania modeli do danych, a następnie oceny i interpretacji wyników pod kątem uzyskanej wiedzy. Nie ma jednoznacznej, ogólnie przyjętej definicji eksploracji danych. Większość istniejących definicji zwraca jednak uwagę na trzy rzeczy: analizę dużych zbiorów danych (w szczególności danych zastanych), poszukiwanie struktury zależności w danych i podsumowań oraz wizualizacje jako formę reprezentacji wyników.

Dynamika KDD w różnych obszarach nauki oraz rozwój zaawansowanych technik i algorytmów drążenia danych doprowadziły do sytuacji, w której idea odkrywania wiedzy staje się możliwa do zastosowania na gruncie socjologii analitycznej, w tym socjologii jakościowej. Staje się to możliwe ponieważ rozwój oprogramowania do wspomaganej komputerowo analizy danych jakościowych (CAQDAS) idzie w kierunku metod mieszanych, a więc równoczesnego wykorzystywania w procesie analizy danych ilościowych i jakościowych[10]. Są to dane ustrukturyzowane (statystyki urzędowe, dane z badań sondażowych, dane pomiarowe itp.), częściowo ustrukturyzowane zbiory danych tekstowych (dane z Internetu, ze stron WWW, publikacji elektronicznych) oraz dane nieustrukturyzowane (luźne dokumenty, książki, artykuły, zapiski, notatki, transkrypcje wywiadów) czy też inne rodzaje danych z badań jakościowych (np. zdjęcia, rysunki, filmy). Integracja tych danych w procesie analitycznym stanowi bogactwo informacji i źródło wiedzy o życiu społecznym. Wymaga także odpowiednich technik analitycznych, zdolnych nie tylko do ich przetworzenia, wydobycia zawartych informacji, lecz przede wszystkim ujęcia w struktury interpretowalnej wiedzy. Obecne na rynku programy do wspomaganej komputerowo analizy danych jakościowych pozwalają tylko w pewnym stopniu na tego typu analizy. Istnieje możliwość „inteligentnego uczenia się" wzorców kodowania danych (Qualrus)[11] czy automatycznego kodowania treści dokumentów tekstowych w oparciu o model klasyfikacyjny skonstruowany na bazie analizy słownikowej istniejącego zbioru danych tekstowych (QDA Miner)[12]. Rozwiązania te wykorzystują techniki i algorytmy analityczne właśnie z obszaru Data i Text Mining, a także przetwarzania języka naturalnego (NLP)[13]. Zanim przejdę do refleksji nad możliwościami zastosowania Data Mining w procesie eksploracji

[10] Doskonałym przykładem są tu metody mieszane (mixed methods).

[11] Zob. strona producenta oprogramowania: www.ideaworks.com/download/qualrus/QualrusManual.pdf.

[12] Zob. strona producenta oprogramowania: http://provalisresearch.com/Documents/QDAMiner40.pdf.

[13] Przetwarzanie języka naturalnego (Natural Language Processing, NLP) to dział informatyki, w skład którego wchodzi teoria gramatyk i języków formalnych oraz reprezentacja wiedzy zawartej w tekstach. Analiza języka naturalnego dotyczy przetwarzania komputerowego tekstów zapisanych w języku naturalnym w celu wydobywania z nich informacji, reguł i prawidłowości, wzorców.

danych i odkrywania wiedzy w obszarze wspomaganej komputerowo analizy danych jakościowych, chciałbym krótko scharakteryzować proces drążenia danych i stojącą u jego podstaw metodologię drążenia danych CRISP.

Czym jest Data Mining?

Data Mining, eksploracja, drążenie danych to proces analityczny, którego celem jest odkrywanie wiedzy, czyli uogólnionych reguł i prawidłowości w ustrukturyzowanych i nieustrukturyzowanych danych w oparciu o metody statystyczne, techniki i algorytmy sztucznej inteligencji. Wiedza ta nie wynika wprost z danych. Jest konsekwencją określonej struktury relacji między analizowanymi danymi, wynikiem tego, iż to takie, a nie inne dane znalazły się w bazie. Cel eksploracji nie ma ścisłego związku ze sposobem pozyskiwania danych. Może ona dotyczyć zarówno danych zgromadzonych w systemach bazodanowych, jak i danych pozyskiwanych w toku badań empirycznych. Najczęściej odnosi się do danych zastanych. Nie jest to reguła, ale cecha odróżniająca Data Mining od statystyki czy badań socjologicznych, w których dane są zbierane, aby odpowiedzieć na określone pytania badawcze. Dlatego drążenie danych często nazywane jest wtórną analizą danych. Data Mining ma związek z wielkością wolumenu danych[14], mocą obliczeniową komputera czy wykorzystaniem zaawansowanych technik statystycznych i algorytmów sztucznej inteligencji do znajdowania ukrytych dla człowieka, ze względu na jego ograniczone możliwości czasowe i percepcyjne, związków przyczynowo-skutkowych, prawidłowości czy podsumowań zawartych w danych, które są zrozumiałe i mają moc wyjaśniającą. Zależności te stanowią formę reprezentacji wiedzy zawartej w danych. W procesie eksploracji specyfikuje się cechy badanego zjawiska tak, aby móc je ująć, w formalne reguły, strukturę relacji, modele[15] lub wzorce. Eksploracja i modelowanie danych są więc tworzeniem wyidealizowanej, ale użytecznej repliki realnego świata. W przypadku nauk społecznych modelowanie dotyczy ukazania takiej reprezentacji relacji między

[14] Jeśli wolumen jest stosunkowo niewielki, to możemy skorzystać z tradycyjnej, statystycznej eksploracji danych lub jeśli mamy do czynienia z danymi jakościowymi z algorytmów analitycznych dostępnych w programach CAQDAS. Kiedy jednak liczba danych rośnie, stajemy przed nowymi problemami. Niektóre z nich dotyczą sposobu przechowywania danych, ich jakości, standaryzacji zapisu, występowania braków danych itp. Inne odnoszą się do sposobu wyznaczania danych do analizy, badania regularności, dynamiki zjawisk czy procesów społecznych, konstruowania i walidacji modeli analitycznych, weryfikacji tego, czy nie są przypadkowym odzwierciedleniem jakiejś wewnętrznej rzeczywistości zbioru danych.

[15] Model jest uproszczoną reprezentacją realnego procesu społecznego. Służy do redukcji złożoności relacji pomiędzy danymi. Model dostarcza odpowiedzi na pytania: jak coś działa, jakie są mechanizmy działania, jakie są prawidłowości, jakie są relacje.

zmiennymi, która zgodnie z założeniem izomorfizmu odzwierciedla relacje między zmiennymi opisującymi własności świata danego procesu społecznego czy rzeczywistości.

Data Mining to pojęcie, pod którym kryją się różne techniki analityczne służące odkrywaniu wiedzy w danych. Błędnym jest przekonanie, jakoby proces drążenia danych polegał na analizie olbrzymich ilości danych przez inteligentne algorytmy, które same, bez udziału człowieka, odnajdują prawidłowości czy relacje. Data Mining to proces interaktywny i iteracyjny. Odkrycie związków między danymi wymaga użycia nie tylko zaawansowanych technologii, lecz przede wszystkim wiedzy eksperckiej, znajomości danych i umiejętności analitycznych badacza. Pozornie nieistotne wzory czy struktury relacji zawarte w danych, odkryte przy pomocy metod i technik eksploracji, dzięki doświadczeniu i wiedzy badacza mogą stać się cennymi informacjami. Stąd rzetelne drążenie danych wymaga wiedzy z zakresu problematyki, która jest przedmiotem Data Mining, umiejętności rozumienia danych oraz interpretacji związków między nimi. W procesie drążenia danych wykorzystywane są różnorodne metody i techniki poszukiwania związków między zmiennymi. Wiele z nich określa się mianem algorytmów „uczących się" (machine learning) lub „modelujących". Należą do nich m.in. metody statystyczne (analiza regresji, analiza wielowymiarowa, algorytmy klasyfikacyjne i taksonomiczne, drzewa decyzyjne), sieci neuronowe, metody ewolucyjne, logika rozmyta czy zbiory przybliżone. Wywodzą się ze statystyki matematycznej, uczenia maszynowego czy badań nad sztuczną inteligencją. W praktyce wykorzystuje się także różne modele przetwarzania danych, tj. streszczanie, poszukiwanie asocjacji, analizę funkcjonalną, klasyfikację czy grupowanie. W analizach typu Data Mining z reguły nie stawia się hipotez *a priori*. „Hipotezy" powstają w drodze eksploracji danych jako efekty identyfikacji systematycznych relacji pomiędzy zmiennymi w sytuacji, gdy natura tych relacji nie jest z góry określona. Drążenie danych utożsamia się więc zazwyczaj z podejściem indukcyjnym do odkrywania wiedzy. Data Mining może jednak czasami przyjąć logikę dedukcyjną w procesie analizy danych. Techniki i algorytmy analityczne mogą być wykorzystywane wówczas jako sposób weryfikacji modeli powstałych wcześniej na etapie eksploracji danych lub istniejących i wymagających empirycznego sprawdzenia. W trakcie eksploracji danych znajdowana jest często bardzo duża liczba wzorców. W większości przypadków są to wzorce znane, mało interesujące dla analityka. Problemem jest identyfikacja wzorców, które mają charakter nieznanej, odkrywanej wiedzy. Ocena ich wartości leży po stronie badacza. Poza miarami dopasowania modeli do danych czy ich użytecznością nie ma bowiem żadnych kryteriów obiektywnej oceny ich wartości. Wzorce są bowiem zdeterminowane poprzez zestawy cech czy danych oraz są efektem zastosowania określonych technik czy algorytmów analitycznych. W praktyce duża różnorodność

technik i algorytmów analitycznych nie ułatwia wyboru tych, które są najtrafniejsze w odniesieniu do analizowanych zagadnień, dlatego powszechnie uznaje się, że Data Mining jest procesem interakcyjnym i iteracyjnym, w którym istotną rolę odgrywa badacz, jego wiedza, umiejętności i doświadczenie, a nie zaimplementowane w danym programie techniki czy algorytmy analityczne.

Data Mining różni się od statystycznej, eksploracyjnej analizy danych ilościowych (Exploratory Data Analysis, EDA). Różnica ta dotyczy celu i podejścia do analizy. Eksploracyjna analiza danych jest podejściem analitycznym służącym odkrywaniu struktury zależności między analizowanymi zmiennymi. W tym celu wykorzystuje się głównie proste techniki wizualizacji zależności w danych oraz metody statystyki opisowej. Techniki wizualizacji stosowane w EDA zapewniają wgląd w analizowane dane, pozwalają odkrywać ich strukturę, znajdować przypadki odstające i nieprawidłowości. Siła wizualizacji w EDA opiera się na wykorzystywaniu, posiadanych przez każdego człowieka, naturalnych zdolności rozpoznawania wzorców czy regularności. Dzięki wizualizacji analityk zyskuje właściwy dystans do danych, pozostaje otwarty na wyłaniające się wzorce czy struktury zależności, skupiając się na ich zrozumieniu. Nie oznacza to jednak niemożności wykorzystywania innych technik analizy niż statystyki opisowe czy wizualizacje. Drążenie danych jest raczej ukierunkowane na praktyczne zastosowania niż na zrozumienie istoty analizowanego zjawiska czy wykrywanie konkretnych związków pomiędzy rozważanymi zmiennymi. Data Mining ma bowiem silny związek z biznesem, w którym techniki i algorytmy analityczne wykorzystuje się do znajdowania rozwiązań pozwalających na dokonywanie użytecznych prognoz lub przewidywań. W procesie eksploracji danych i odkrywania wiedzy w polu biznesu wykorzystuje się często metody tzw. „czarnej skrzynki", techniki statystyczne: statystyki opisowe, tabele kontyngencji, analizę czynnikową, dyskryminacyjną, hierarchiczną analizę skupień itp. albo zaawansowane techniki analizy tj. sieci neuronowe czy drzewa klasyfikacyjne umożliwiające generowanie prognoz, niepozwalające jednak na identyfikowanie natury zależności pomiędzy zmiennymi, na których opierają się prognozy. W polu nauki Data Mining znajduje zastosowanie w odkrywaniu struktur wiedzy zawartej w danych przez konstruowanie algorytmów, funkcji aproksymacyjnych, reguł indukcyjnych, tworzenie typologii, klasyfikacji lub generowanie struktur wielowymiarowych zależności między zmiennymi[16]. Współcześnie procesy eksploracji danych znajdują szereg

[16] Podstawą eksploracyjnej analizy danych jakościowych jest poszukiwanie podobieństw między danymi, kodami, fragmentami tekstu czy dokumentami. W tym celu wykorzystuje się mechanizmy przeszukiwania tekstu czy zakodowanych fragmentów, dokonuje porównań w tabelach kontyngencji, macierzach typu: kod–kod, kod–dokument czy słowo–dokument. Podobnie jak w tradycyjnej eksploracyjnej analizie danych istotną rolę w tym procesie odgrywają podstawowe wizualizacje struktury zależności między analizowanymi danymi jakościowymi.

zastosowań w analizie danych o ruchu internetowym (analiza logów), rozpoznawaniu sygnałów obrazu, mowy, pisma, sensu wyrazów i zdań, struktur chemicznych, stanu zdrowia człowieka, wspomaganiu diagnostyki medycznej, biologii i badaniach genetycznych, analizie operacji bankowych, prognozowaniu wskaźników ekonomicznych, pogody, plam na Słońcu, aż po zagadnienia z zakresu kogniwistyki, doświadczeń psychologicznych, analizy sposobu rozumowania i kategoryzacji, poruszania się i planowania itp. Ważną rolę w tym zakresie pełni także wykorzystanie metodologii Data Mining w rozwijaniu systemów eksperckich czy systemów uczących się[17]. Zastosowanie tej metodologii w naukach społecznych, a w szczególności w socjologii jakościowej czy wspomaganej komputerowo analizie danych jakościowych (CAQDAS), jest naturalnym procesem rozwoju tej dziedziny nauki.

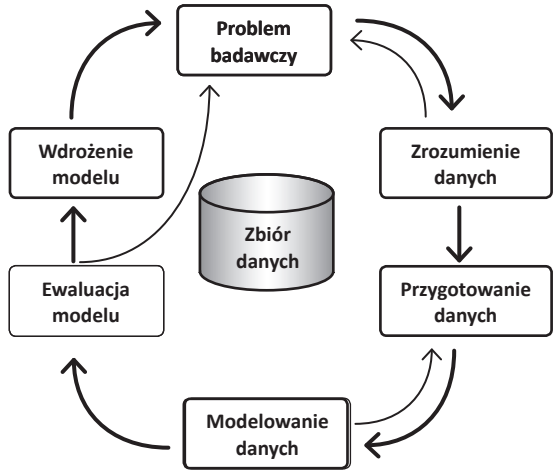
Metodologia Data Mining

Proces eksploracji, drążenia danych, przeprowadza się zazwyczaj w oparciu o tzw. metodologię CRISP-DM (ang. *Cross Industry Standard Process for Data Mining*)[18]. Metodologię tę opisuje się jako model hierarchiczny, składający się z zestawów zadań opisywanych na czterech poziomach abstrakcji od najbardziej ogólnego do konkretnego: na poziomie faz, zadań ogólnych, zadań szczegółowych oraz procedur analitycznych dotyczących bezpośrednio procesu drążenia. Każdy z nich

[17] Uczenie się oznacza autonomiczne zmiany w systemie mające na celu polepszenie jakości jego działania, dokonujące się na podstawie obserwacji otaczającego świata lub analizy układu danych. Zmiana ta polega na zdobyciu lub udoskonaleniu przez system wiedzy lub umiejętności, zapamiętaniu tej wiedzy lub umiejętności i wykorzystaniu jej do wykonania stawianych mu zadań. Rodzaj uczenia się, z którym będziemy mieć do czynienia, zależy od postaci i sposobu dostarczania systemowi uczącemu się jego obserwacji i doświadczeń, mechanizmu generowania jakiejś wiedzy na ich podstawie oraz sposobu wykorzystania tej wiedzy. Jeżeli zadaniem systemu uczącego się miałoby być odpowiednie zakwalifikowanie obiektu do danej kategorii, pokazalibyśmy mu szereg prawidłowo zakwalifikowanych obiektów do kategorii i na tej podstawie system uzyskałby wiedzę potrzebną do przypisania dowolnego obiektu do odpowiedniej kategorii. Systemy uczące się to systemy wykorzystujące techniki i algorytmy Data Mining do poprawy jakości działania przez zdobywanie nowych doświadczeń, które są następnie przekształcane w reprezentację wiedzy i dzięki możliwości samodzielnego wnioskowania wykorzystywane w interakcji tych systemów ze środowiskiem. Uczenie może odbywać się pod nadzorem analityka lub bez nadzoru. Systemy uczące się znajdują zastosowanie w automatyce, ekonomii, systemach wspomagania decyzji, symulacjach komputerowych, zagadnieniach optymalizacyjnych, diagnostyce technicznej i medycznej (zob. Cichosz 2007).

[18] Zob. Chapman i in. (2000) CRISP – DM 1.0. Step-by-step data mining guide, ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf.

składa się z pewnej liczby ogólnych zadań odnoszących się do sytuacji, jakie występują w trakcie eksploracji danych. Zgodnie z zasadami CRISP-DM proces drążenia danych zachodzi zawsze w określonych warunkach, co oznacza, że modele analityczne wymagają kontekstualizacji. Rozróżnia się cztery wymiary drążenia danych: obszar zastosowania – konkretna dziedzina, w której przebiega projekt drążenia danych; typ problemu drążenia danych – opisuje rodzaj zagadnienia, którego dotyczy dany projekt analityczny; kontekst techniczny – obejmujący różne techniczne wyzwania pojawiające się zwykle podczas realizacji projektu oraz narzędzia i techniki – wykorzystywane w procesie eksploracji danych. Metodologię procesu drążenia danych, po pewnych modyfikacjach, przedstawia ilustr. 1.



Ilustr. 1. Proces drążenia danych według metodologii CRISP-DM

Źródło: opracowanie własne na podstawie dokumentacji metodologii CRISP-DM

Proces drążenia danych składa się z sześciu etapów. Ich kolejność nie jest jednak „sztywna". Często bowiem niezbędny jest powrót do etapów poprzednich. Wynik każdego z etapów decyduje o tym, który etap lub jakie zadanie zostanie wykonane w następnej kolejności. Strzałki wskazują najważniejsze i najczęściej pojawiające się zależności między etapami. Odzwierciedlają również cykliczny i procesualny charakter drążenia danych. Etap pierwszy dotyczy zrozumienia celów projektu analitycznego i przetworzenia tej wiedzy w definicję problemu badawczego oraz stworzenia planu działań. Etap drugi ma na celu zaznajomienie się z danymi, rozpoznanie problemów z ich jakością, dotarcie do pierwszych spostrzeżeń, podsumowań i odkrycie interesujących grup obserwacji pozwalających na skonstruowanie hipotez o wiedzy zawartej w danych. Na etap ten składa się opis danych, ich eksploracja i jakościowa weryfikacja pod kątem wykorzystania

w procesie drążenia i modelowania. Musimy odpowiedzieć sobie na pytania: czy dane są kompletne, czy reprezentują wszystkie sytuacje, czy zawierają błędy, czy występują braki danych, jeżeli tak, to jakie są to braki, gdzie i jak mamy z nimi do czynienia, jak sobie z nimi radzić itp. Etap trzeci odnosi się do przygotowania danych do budowania modeli analitycznych. Na tym etapie dokonujemy wyboru danych, które mają być uwzględnione w analizie i modelowaniu. Do analizy wybiera się „wyczyszczone dane". Istotne znaczenie na tym etapie mają również operacje służące tzw. tworzeniu danych, a więc kodowanie i przekształcenia, dodawanie nowych czy rekodowanie już istniejących danych. Przygotowanie danych to także integracja danych ilościowych i jakościowych. W procesie integracji zawiera się agregacja danych tj. łączenie informacji pochodzących z różnych zbiorów danych. Ostatnie trzy etapy to modelowanie danych w oparciu o wybrane algorytmy analityczne, ocena dopasowania modelu do danych i wdrożenie, jeśli projekt ma na celu użyteczność praktyczną modelu. Data Mining nie kończy się w momencie skonstruowania modelu analitycznego, ewaluacji czy jego aplikacji. Efekty drążenia danych oraz to, czego nauczymy się w trakcie eksploracji danych i budowania modeli analitycznych, przynoszą nowe pytania i problemy badawcze. Doświadczenie w realizacji projektów Data Mining w sferze biznesu i praktyka analityczna uczy, że tworzenie modeli analitycznych wspomagających wyjaśnianie naukowe wymaga poświęcenia ok. 80% czasu na trzy pierwsze etapy: zrozumienie uwarunkowań problemu badawczego w procesie eksploracji danych, zrozumienie danych (ich treści i relacji między nimi), a także odpowiednie ich przygotowanie do modelowania. Te trzy etapy decydują o jakości procesu drążenia danych i wyniku modelowania.

Z metodologicznego punktu widzenia w procesie drążenia danych i budowania modeli analitycznych dzięki zastosowaniu odpowiedniego oprogramowania analitycznego badacz społeczny otrzymuje dwojakiego rodzaju wsparcie: pasywne i aktywne. Wsparcie pasywne ma miejsce wtedy, gdy sformułował on wstępnie hipotezę badawczą lub rozpoczął poszukiwanie prawidłowości występujących w danych i wykorzystuje swoje doświadczenie oraz wiedzę w procesie ich analizy. Wykorzystując odpowiednie oprogramowanie, badacz może przeglądać dane, dokonywać na nich operacji, przedstawiać je w formie graficznej – w postaci tabel i różnego rodzaju wykresów. Może także wyliczać różne statystyki i testować postawione hipotezy, badając zaobserwowane związki między danymi. W przypadku technik aktywnego wsparcia użytkownik również jest inicjatorem procesu Data Mining i decyduje o jego przebiegu. Jednak rola, jaką w procesie analizy odgrywa komputer, jest znacznie ważniejsza niż poprzednio. To on samodzielnie identyfikuje prawidłowości i związki, jakie tkwią w danych. Jeśli chcemy zbudować model analityczny zjawiska lub procesu społecznego, możemy skorzystać z bogatego zestawu technik aktywnego wsparcia – uczących się i modelujących.

Wiele technik modelowania opiera się na statystyce i sztucznej inteligencji. Techniki te można podzielić ze względu na podejście do modelowania na: predykcyjne, grupujące (segmentacyjne) i asocjacyjne. W modelowaniu predykcyjnym, nazywanym także nadzorowaną techniką uczenia się, dane są stosowane do przewidywania wartości zmiennych wynikowych. Najczęściej stosowane to: sieci neuronowe, drzewa decyzyjne, regresja liniowa i regresja logistyczna. Metody grupowania, segmentacji, nazywane nienadzorowanymi technikami uczenia się, nie mają z góry założonej zmiennej wynikowej. Celem technik segmentacji jest próba wydzielenia zbiorów składających się z jednostek o podobnych wartościach zmiennych źródłowych. Najczęściej stosowane to: sieci Kohonena, metoda *k*-średnich i metoda dwustopniowa. Techniki asocjacyjne nazywane są uogólnionym modelowaniem predykcyjnym. Zmienne ze zbioru danych są wykorzystywane jako zmienne źródłowe i wynikowe jednocześnie. Reguły asocjacyjne usiłują powiązać określone skutki z zestawem przyczyn.

Metody i techniki eksploracji danych

Różnorodność metod i technik eksploracji danych wywodzących się z różnych dyscyplin badawczych utrudnia identyfikację tych, które wydają się najbardziej odpowiednie w zakresie analizy danych. Można je, jak wskazuje Tadeusz Morzy (2013), sklasyfikować ze względu na: charakterystykę (metody opisowe i metody predykcyjne); cel eksploracji (odkrywanie asocjacji, klasyfikacja i predykcja, grupowanie, analiza sekwencji i przebiegów czasowych, eksploracja: tekstu i danych semistrukturalnych, WWW, grafów i sieci społecznościowych, danych multimedialnych i przestrzennych itp.); typy eksplorowanych danych (płaskie pliki danych, relacyjne bazy danych, a także wraz z rozwojem narzędzi do generowania i przechowywania danych oraz technologii eksploracji dane multimedialne – zdjęcia, filmy, muzyka, tekstowe i semistrukturalne, przestrzenne – mapy, grafy, struktury chemiczne, sieci społecznościowe itd.) oraz typy odkrywanych wzorców. Najpopularniejszym i najczęściej stosowanym sposobem ich rozróżnienia jest klasyfikacja ze względu na cel analityczny samej eksploracji danych. W praktyce proces drążenia danych odnosi się do trzech rodzajów działań: opisu, przewidywania oraz odkrywania wzorców i reguł w zbiorze danych. Opis danych pozwala na charakterystykę zależności i ich graficzną reprezentację. Ma na celu poszukiwanie wzorców i trendów znajdujących się w analizowanym zbiorze danych. Wzorce te muszą być przejrzyste dla badacza tak, aby można je było intuicyjnie i sensownie interpretować. Efektem opisu danych są modele analityczne: grupujące obserwacje ze względu na podobieństwo cech czy opisujące związki między zmiennymi.

Grupowanie (ang. *clustering*) oznacza łączenie ze sobą obiektów, obserwacji czy respondentów w klasy. Grupa jest zbiorem obserwacji podobnych do siebie nawzajem pod względem określonych cech, a niepodobnych do innych grup. W grupowaniu nie chodzi o oszacowanie czy przewidywanie wartości zmiennej, lecz o podzielenie zbioru danych na homogeniczne podgrupy lub grupy. W modelach grupujących podobieństwo obserwacji wewnątrz grup jest maksymalizowane, a podobieństwo do obserwacji spoza grupy – minimalizowane. W modelu opartym na grupowaniu nie występuje zmienna zależna. Grupowanie opiera się tylko na wybranych zmiennych niezależnych. Przykładem grupowania są różne segmentacje psychograficzne lub socjodemograficzne stylu życia, jak również segmentacje stosowane w naukach biologicznych, np. grupowanie ekspresji genów pod względem podobnych zachowań. Grupowanie stanowi krok wstępny do procesu eksploracji danych z grupami wynikowymi używanymi jako dane wejściowe w technikach modelowania lub jako kategorie zmiennej zależnej w budowaniu modeli klasyfikacyjnych. Do popularnych technik eksploracji danych w tym zakresie zalicza się algorytm *k*-średnich, dwustopniowe grupowanie, analizę głównych składowych, hierarchiczną analizę skupień, skalowanie wielowymiarowe, analizę korespondencji czy samoorganizujące się sieci Kohonena. Algorytmy te można także stosować w odniesieniu do tradycyjnych danych jakościowych, konstruując np. koszyki semantyczne (zestawy synonimów, antonimów czy słów kluczowych) w analizie treści.

Kolejny rodzaj działań w zakresie eksploracji danych odnosi się do budowania modeli analitycznych, których celem jest klasyfikacja lub przewidywanie wartości danej zmiennej na podstawie innych zmiennych niezależnych. Proces podziału rzeczy, zachowań, obiektów, słów czy obrazów na klasy, grupy, kategorie to jedna z podstawowych czynności poznawczych człowieka ułatwiająca poruszanie się w złożonym świecie życia codziennego. Dystynkcja jest ściśle związana z myśleniem, postrzeganiem, uczeniem się i działaniem. Stąd w wymiarze epistemologicznym klasyfikacja nie tylko umożliwia zrozumienie rzeczywistości na drodze redukcji entropii w bazie danych, ale i stanowi narzędzie odkrywania zależności między danymi czy konstruowania teorii naukowych. Podstawą klasyfikacji jest uzyskanie jednorodnego zbioru danych, w odniesieniu do którego łatwiej wyróżnić cechy systematyczne. Dotyczy to redukcji dużej liczby obiektów do kilku kluczowych kategorii w celu ujawnienia struktur w istniejących danych. W zależności od rodzaju danych klasyfikację dzieli się na wzorcową (znana jest częściowa charakterystyka klas) oraz bezwzorcową, zwaną taksonomią (celem jest dopiero odkrycie struktury klas). W eksploracji danych klasyfikację wzorcową określa się również jako uczenie z nauczycielem (nadzorowane), zaś bezwzorcową jako uczenie bez nauczyciela (nienadzorowane). Uczenie nadzorowane polega na analizie zbioru danych liczbowych lub tekstowych, których

przynależności do klas są znane (np. poprzez istniejące słowniki) i konstruowaniu modeli dla każdej z klas, opierając się na charakterystyce posiadanych danych. Wynikiem tej klasyfikacji są drzewa decyzyjne lub zbiór reguł decyzyjnych, które są wykorzystane zarówno w celu lepszego zrozumienia własności każdej, wyróżnionej klasy, jak i zgodnie z danym modelem określenia przynależności klasowej nowych obiektów. Uczenie nienadzorowane dotyczy sytuacji, gdy nie ma informacji o cechach istniejącego wzorca, a wybrane techniki analityczne wspomagają znalezienie reguł klasyfikacyjnych jedynie na podstawie dostępnych danych. W celu zwiększenia precyzji klasyfikacji stosuje się podział zbioru danych na uczący i testowy. W pierwszym kroku buduje się model analityczny na zbiorze uczącym, a później weryfikuje się jego skuteczność na zbiorze testowym, porównując wyniki przed i po wprowadzeniu nowych danych do modelu. W ten sposób powstaje model, który jest rozwijany i udoskonalany wraz z pojawianiem się nowych danych. Przewidywanie (predykcja) jest podobne do klasyfikacji, ale jego wynik odnosi się do przyszłości. Model predykcyjny budowany jest na danych historycznych lub teraźniejszych, a jego wartością jest przewidywanie wystąpienia określonych zdarzeń, słów czy wartości zmiennych w analizowanym zbiorze danych. W budowaniu modelu klasyfikacyjnego czy predykcyjnego podstawową zasadą jest triangulacja technik analitycznych, której celem jest wybór najlepiej dopasowanego do danych modelu. Do najczęściej stosowanych technik klasyfikacji danych w obszarze Data Mining zalicza się: regresję liniową, regresję logistyczną, analizę dyskryminacyjną, drzewa decyzyjne (C5.0, CART, CHAID, QUEST), sieci neuronowe czy algorytmy genetyczne.

Ostatni rodzaj działań w obszarze Data Mining to odkrywanie wzorców i reguł indukcyjnych. Odkrywanie reguł to proces szukania złożonych zależności asocjacyjnych lub korelacji pomiędzy cechami w obrębie zestawu analizowanych danych[19]. Reguły asocjacyjne przybierają postać „Jeżeli poprzednik, to następnik". Na przykład reguła przedstawiona w taki sposób X => Y jest interpretowana jako sytuacja, w której elementy spełniające X, spełniają również Y. Oprócz tego dla każdej reguły stosuje się

[19] Zależności między zmiennymi, obiektami lub zdarzeniami mają charakter indukcyjny. Potocznie poprzez pojęcie indukcji rozumiemy przechodzenie od wielu drobnych faktów do jednego prawa ogólnego, opisującego je wszystkie. Uczenie się na podstawie wnioskowania indukcyjnego oznacza wygenerowanie na podstawie analizy danych empirycznych hipotezy indukcyjnej stanowiącej ogólny obraz dotyczący relacji zawartych w danych. Uzyskana w ten sposób wiedza (hipoteza indukcyjna) może być później stosowana do wnioskowania dedukcyjnego. W Data Mining mamy więc dwa rodzaje wnioskowania: indukcyjne (służące odkrywaniu wiedzy) i dedukcyjne (służące jej weryfikowaniu). W praktyce mamy zwykle do czynienia z danymi zorganizowanymi w rekordach opisanych przez odpowiednio dobrany zestaw atrybutów. Wnioskowanie indukcyjne polega na odnalezieniu zależności między tymi atrybutami, a wnioskowanie dedukcyjne polega na zastosowaniu znalezionych hipotez do sprawdzania poprawności lub przewidywania przyszłych wartości nowych rekordów lub atrybutów.

miarę wsparcia (pokrycia) oraz miarę dokładności (ufności). Załóżmy, że w analizowanym zbiorze danych we wczesnej analizie eksploracyjnej dokonaliśmy klasyfikacji obserwacji pod względem postaw obywatelskich oraz wyróżniliśmy regułę „jeśli zaufanie do innych (X), to działanie na rzecz społeczności lokalnej (Y)". Wsparcie dla tej reguły X => Y będzie liczone jako stosunek liczby obserwacji zawierających X i Y wobec całkowitej liczby obserwacji w badanej grupie osób (prawdopodobieństwo zajścia X i Y). Natomiast ufność dla danej reguły X => Y jest miarą dokładności reguły, określoną jako stosunek liczby obserwacji zawierających jednocześnie X i Y, do liczby obserwacji zawierających tylko X. Badacz może więc preferować reguły, które mają duże wsparcie lub dużą ufność. W praktyce kierujemy się jednak uzyskaniem wysokich wartości dla obu tych miar. Mocne reguły asocjacyjne to te, dla których wsparcie i ufność są wyższe niż przyjęte w modelu wartości graniczne. Aby dana reguła miała sens, istotna jest weryfikacja częstości jej występowania w zbiorze danych. Do najczęściej używanych algorytmów indukcyjnych w drążeniu danych zalicza się algorytmy asocjacyjne APRIORI, GRI, CARMA[20]. Ich zaletą jest to, że odnoszą się zarówno do zmiennych ilościowych, jak i jakościowych. W procesie drążenia danych reguły indukcyjne znajdują bardzo szerokie zastosowanie. Począwszy od wykrywania fałszerstw lub nadużyć w transakcjach finansowych, bankowych, poprzez analizę danych sprzedaży bezpośredniej przy podejmowaniu decyzji marketingowych, aż do poszukiwania związków między zachowaniami podmiotów i instytucji w obszarze bezpieczeństwa państwa – przeciwdziałanie przestępczości zorganizowanej. W nauce astronomia wykorzystuje reguły indukcyjne do poszukiwań nieznanych gwiazd lub galaktyk, a medycyna, biologia czy genetyka molekularna – do znajdowania powiązań między genami na poziomie cząsteczkowym[21].

CAQDAS. Od eksploracji do odkrywania wiedzy w danych jakościowych

Po scharakteryzowaniu metodologii oraz technik drążenia danych „w obszarze big data" chciałbym powrócić do wskazanej we wstępie tego artykułu relacji między Data/Text Mining a wspomaganą komputerowo analizą danych

[20] Wyszukiwanie reguł asocjacyjnych w danych jest jednym z podstawowych zagadnień w ramach odkrywania wiedzy. Zagadnienie odkrywania reguł asocjacyjnych zostało po raz pierwszy przedstawione przez Agrawala i in. (1993, 1994). Z roku 1994 pochodzi wspomniany powyżej algorytm Apriori.

[21] W dziedzinie Data Mining mogą być prowadzone także zaawansowane analizy czasowe uwzględniające zmienność zjawisk. Polegają one na badaniu obszernych zestawów danych w perspektywie czasowej celem odnajdywania prawidłowości, poszukiwania podobnych sekwencji i wydobywania z nich wzorców, badania okresów wystąpień tychże sekwencji, a także czynników mających wpływ na ich wystąpienie oraz odstępstw od znalezionych reguł.

jakościowych, poszukując odpowiedzi na następujące pytania: Co metodologia Data Mining oferuje socjologii jakościowej? Jaki jest związek między Data Mining, a CAQDAS? Jakie są możliwości wykorzystywania algorytmów i technik analitycznych Data Mining w środowisku CAQDAS? Czy uprawianie analizy danych jakościowych przy wsparciu Data Mining ma sens?

Jeśli przyjmiemy, że podstawowym celem socjologii jest dostarczanie ugruntowanej empirycznie oraz dającej się zweryfikować wiedzy o zjawiskach, mechanizmach procesach społecznych, to odpowiedź na te pytania jest jednoznaczna. Wiedza socjologiczna służy zrozumieniu, wyjaśnianiu oraz przewidywaniu, a więc wykorzystanie nowych podejść metodologicznych czy technik analitycznych jest jak najbardziej uzasadnione. Wraz z rozwojem socjologii jakościowej rozwija się jej metodologia, techniki prowadzenia badań empirycznych, procedury analizy danych oraz sposoby wnioskowania. Pojawienie się w programach CAQDAS zaawansowanych algorytmów i technik drążenia danych[22] jest konsekwencją lawinowego wzrostu elektronicznych baz, repozytoriów czy hurtowni danych tekstowych, możliwości ich archiwizowania oraz kompleksowego przetwarzania i analizy. W obszarze CAQDAS towarzyszy temu zjawisku obserwowany od kilku lat zwrot metodologiczny w kierunku paradygmatu *mixed-methods* w naukach społecznych, a w szczególności badaniach jakościowych (Tashakkori, Teddlie 2003). Jego wyrazem jest implementacja wielowymiarowych technik statystycznych, algorytmów czy innych funkcjonalności z dziedziny inteligencji komputerowej i przetwarzania języka naturalnego w specjalistycznych programach do wspomaganej komputerowo analizy danych jakościowych tj. QDA Miner, Qualrus czy T-Lab[23]. W programach tych znajdujemy szereg zaawansowanych rozwiązań wspomagających tradycyjną analizę danych tekstowych tj.: analiza tematyczna, korespondencji, asocjacji semantycznych, kontekstualna, leksykalna, a także umożliwiających modelowanie semantyczne, klasyfikację czy kategoryzację. Wszystkie te rozwiązania opierają się właśnie na dynamicznie rozwijającej się metodologii eksploracji i odkrywania wiedzy w danych.

CAQDAS a proces ewolucji analizy danych jakościowych

Rozwój CAQDAS w kierunku wykorzystania zaawansowanych metod eksploracji i odkrywania wiedzy w danych (głównie tekstowych) jest możliwy nie tylko dzięki zastosowaniu nowych technologii informatycznych, lecz przede wszystkim

[22] Mam tu na myśli przede wszystkim algorytmy i techniki analityczne z obszaru Text Data Mining (krótko Text Mining). Charakterystyka tego podejścia została przedstawiona w artykule Bryda, Tomanek (2014).

[23] Zob. strona producenta oprogramowania T-lab, http://tlab.it/en/presentation.php.

dzięki ewolucji metodologii i technik analizy danych jakościowych, w szczególności w obszarze danych tekstowych. Ewolucyjny charakter tych zmian odzwierciedlają cztery strategie prowadzenia wspomaganej komputerowo analizy danych jakościowych:

1. Metody tradycyjne (Qualitative Data Analysis) (zob. Lewins, Silver 2007)[24]:
   – etnografia/netnografia/etnometodologia,
   – analiza ramowa/konwersacyjna/narracyjna/dyskursu,
   – teoria ugruntowana,
   – studia przypadków,
2. Analiza zawartości/treści (zob. Krippendorf 2004; Schreier 2012);
   – ilościowa analiza zawartości/treści (Quantitative Content Analysis),
   – jakościowa Analiza zawartości/treści (Qualitative Content Analysis),
3. Metody mieszane (Mixed Methods) (zob. Johnson, Onwuegbuzie 2004);
4. Metody i techniki Data/Text Mining (w tym algorytmy przetwarzania języka naturalnego) (zob. Fayyad, Piatetsky-Shapiro, Smyth 1996; Han, Kamber 2006).

Rdzeń współczesnej analizy danych jakościowych stanowi nieodłącznie teoria ugruntowana, której procedury były od początku implementowane w programach CAQDAS. Wytyczyła ona nie tylko wzorce przeprowadzania analiz jakościowych (zob. Fielding, Lee 1993; Fielding, Lee 1998; Fielding 2012; Bryda 2014), ale jej założenia metodologiczne stały u podstaw rozwoju wielu obecnych funkcjonalności narzędzi CAQDAS. Jeśli jednak dokonamy analizy pojawiania się nowych funkcjonalności w programach CAQDAS na przestrzeni ostatnich kilkunastu lat, to zobaczymy, że rozwój wspomaganej komputerowo analizy danych jakościowych w kierunku Data czy Text Mining nie byłby możliwy bez rozwoju ilościowej i jakościowej analizy treści oraz metod mieszanych (*mixed methods*). Szczególne znaczenie ma tu ilościowa i jakościowa analiza zawartości/treści, której procedury analityczne odnajdujemy w metodologii eksploracji danych tekstowych i odkrywaniu wiedzy (Data i Text Mining). Od połowy XX w. analiza treści jest definiowana jako systematyczna technika opisu danych tekstowych i redukcji semantycznej ich znaczenia dokonywanej w procesie kodowania (Berelson 1952; Weber 1990; Krippendorf 2004). W przeciwieństwie do tego Holsti przedstawia szeroką definicję analizy zawartości jako dowolnej techniki wnioskowania opartej na obiektywnych i systematycznych identyfikacjach określonych cech przekazów zawartych w danych tekstowych (Holsti 1969: 4). Nie ogranicza tym samym analizy treści tylko do dziedziny analizy tekstu, ale wskazuje, że podejście to może być stosowane z powodzeniem

[24] Zob. także strona projektu Online QDA, http://onlineqda.hud.ac.uk/Intro_QDA/what_is_qda.php.

w innych dziedzinach, np. analizie obrazów, jednak z zastrzeżeniem dotyczącym jej stosowalności tylko do danych, które są trwałe w naturze. Analiza treści występuje w dwóch odmianach ilościowej (Quantitative Content Analysis) i jakościowej (Qualtitative Content Analysis), określanej również jako analiza tematyczna (Thematic Analysis) (Saldana 2013; Guest, MacQueen, Namey 2012). Pierwsza bywa określana jako podejście typu *concept-driven* wykorzystujące słowniki klasyfikacyjne w procesie analizy. Druga zaś nazywana jest jako podejście typu *data-driven*, co przybliża ją do logiki teorii ugruntowanej. Ilościowa analiza treści sformułowana przez Berelsona jest techniką służącą do obiektywnego, systematycznego i ilościowego opisu jawnej zawartości komunikatów (dokumentów tekstowych). Definicja ta jest kłopotliwa dla badacza jakościowego, ponieważ ogranicza się tylko do jawnej zawartości przekazów, bez uwzględniania ukrytej treści sensu, intencji czy społecznych reakcji, które tekst może wywoływać. Kładzie nacisk na podejście kwantytatywne w analizie danych tekstowych. Traktowaniu analizy treści jako metody ilościowej sprzeciwił się Kracauer (1952). Podkreślając jakościowy charakter tej metody, wskazuje, że analiza treści powinna sięgać w strukturę sensu wypowiedzi, dokumentu tekstowego. Znaczenie jest bowiem często skomplikowane, holistyczne i zależne od kontekstu, nie zawsze oczywiste i jasne na pierwszy rzut oka. Czasami konieczne jest, aby przeczytać tekst bardziej szczegółowo, by określić jego znaczenie. Niektóre aspekty znaczenia mogą pojawić się tylko raz w tekście. Nie oznacza to jednak, że są one mniej ważne niż te wymieniane częściej. Akcentując jakościowy charakter analizy treści, Kracauer zwraca uwagę, że nie powinna ograniczać się ona jedynie do analizowania tego, co widoczne czy zliczania częstotliwości słów[25]. Współcześnie, niezależnie od powyższych kontrowersji odnośnie sposobu rozumienia analizy treści, zgodnie z logiką metod mieszanych i logiką rozwoju narzędzi CAQDAS, algorytmów i technik analitycznych oraz technik przetwarzania języka naturalnego zawartość dokumentów tekstowych może być analizowana na dwóch poziomach: opisowym (analiza tego, co zostało powiedziane) i interpretacyjnym (analiza tego, jak to zostało powiedziane). Innymi słowy ilościowa analiza treści powinna poprzedzać analizę jakościową

[25] Podejście to podjął George (1959) w analizie wojennej propagandy. Jego zdaniem analiza treści propagandy wymaga analizowania strategii, co zwykle przejawia się w pojedynczych wystąpieniach pewnej frazy lub słowa w całym tekście (a nie w częstotliwości tych wystąpień). W rzeczywistości stosował określenie nie-częstotliwości fraz czy słów kluczowych jako wskaźnika jakościowej odmiany analizy treści. Określenie to opisuje wymiary niekwantytatywne, niestatystyczne w analizie treści, podkreślając obecność lub nieobecność pewnych wartości cechy lub zespołu cech wskaźnikowych w trakcie wnioskowania opartego na danej hipotezie (George 1959: 8). Podobnie jak Kracauer czy później Holsti (1969) podkreśla wartość jakościowej analizy treści opartej na eksploracji i odkrywaniu znaczeń.

w procesie odkrywania wiedzy (sensu) ukrytej w treści dokumentów tekstowych. Można więc przyjąć, że proces rozwoju wspomaganej komputerowo analizy danych jakościowych od tradycyjnej analizy (Qualitative Data Analysis) opartej przede wszystkim na teorii ugruntowanej w kierunku eksploracji danych i odkrywania wiedzy nie byłby możliwy bez osiągnięć, jakie niesie ze sobą analiza treści. Data Mining w analizie danych jakościowych oferuje nie tylko nowe spojrzenie na proces analityczny, lecz przede wszystkim umożliwia modelowanie procesów społecznych dzięki nieograniczonej liczbie apppendowanych zbiorów danych tekstowych. Data czy szerzej Text Mining to w zasadzie analiza treści w nieograniczonej skali wykorzystująca nowe algorytmy i techniki analityczne oraz metody uczenia maszynowego. Stąd wypracowane przez nie procedury metodologiczne stoją u podstaw rozwoju algorytmów i technik analitycznych Data, Text Mining w programach CAQDAS (zob. Ho Yu, Jannasch-Pennell, DiGangi 2011).

Data Mining w procesie analizy danych jakościowych

Rozwój zaawansowanych metod analitycznych – takich jak Data czy szerzej Text Mining – jest z pewnością dużym krokiem naprzód. Należy jednak pamiętać, że nie ma jednej metody, która rozwiązuje wszystkie lub chociaż większość problemów analitycznych lub badawczych, co doskonale widać na gruncie socjologii jakościowej. Porównanie metodologii Data czy Text Mining oraz metodologii badań jakościowych w socjologii pozwala uznać te podejścia za logicznie kompatybilne, zarówno ze względu na komplementarność w podejściu do danych, jak i na etapy bądź procedury analityczne. Widoczne jest podobieństwo pomiędzy logiką metodologii teorii ugruntowanej a logiką metodologii procesu eksploracji danych i odkrywania wiedzy. Konstruowanie czy rozwijanie teorii ugruntowanej przebiega – podobnie jak w metodologii Data Mining – od danych do modelu koncepcyjnego, wyjaśniającego. W teorii ugruntowanej odkrywanie wiedzy w analizowanych danych to etap generowania teorii substancjalnej. Model analityczny, który powstaje na tym etapie, odnosi się jednak do jakiejś kategorii centralnej, podobnie jak teoria. W przypadku metodologii Data Mining „znika" pojęcie kategorii centralnej. Nacisk jest tu raczej położony na rekonstrukcję struktury głębokiej zawartej w danych jakościowych, której przejawami mogą być np. odkryte w trakcie analizy reguły indukcyjne, wzorce zależności, modele procesów społecznych. Budowanie modeli analitycznych w oparciu o metody i techniki Data Mining wymaga od badaczy społecznych logiki i systematycznych działań, ciągłego sprawdzania etapów pośrednich, bycia świadomym na każdym etapie procesu generowania „teorii z danych". Zgodnie z logiką i metodologią drążenia

danych (CRISP DM) wspomagany komputerowo proces eksploracji i odkrywania wiedzy w danych jakościowych/tekstowych da się opisać przez wyodrębnienie takich etapów, jak[26]:

1. Zrozumienie celu analitycznego i problemu eksploracji danych:
   – wstępny wybór metod/-y i technik/-i eksploracji danych;
   – przygotowanie danych do analizy i modelowania;
   – ocena jakości danych jakościowych;
   – integracja danych z różnych źródeł (baza danych);
   – transformacje i przekształcania, redukcja wolumenu danych;
   – preprocessing (tekstowy, lingwistyczny) i wstępne przetwarzanie danych;
   – proste i złożone wyszukiwanie treści w dokumentach;
   – kodowanie danych i linkowanie treści w dokumentach;
2. Analiza danych (ilościowa i jakościowa):
   – analiza frekwencyjna występowania słów kluczowych i fraz;
   – eksploracja (drążenie) danych/analizy tabelaryczne i opisowe;
   – implementacja, budowanie i rozwijanie słowników semantycznych, klasyfikacyjnych;
   – podział zbioru danych na uczące i testowe;
3. Modelowanie, predykcja i odkrywanie wiedzy:
   – klasyfikacja z wykorzystaniem słowników/konstruowanie typologii;
   – modelowanie z wykorzystaniem technik statystycznych, algorytmów indukcyjnych lub sieci neuronowych;
   – diagnostyka różnych modeli poznawczych (dopasowanie do danych);
   – wizualizacja zależności, reguł czy odkrytych wzorców/konstruowanie map kognitywnych;
4. Ewaluacja modeli poznawczych, zależności, reguł czy odkrytych wzorców:
   – walidacja modeli na zbiorach/z danych testowych;
   – interpretacja wyników analizy i oceny wartości odkrytej wiedzy.

W tak ujętym procesie analizy danych jakościowych punktem wyjścia, zgodnie z metodologią badań społecznych, jest problem badawczy. Następne kroki to integracja i przygotowanie danych do analizy. Kolejny etap to budowanie na zbiorze danych uczących modeli analitycznych oraz ich walidacja na danych testowych pod kątem poprawności metodologicznej, analitycznej i interpretacyjnej w odniesieniu do wyjściowego problemu badawczego. Gotowe modele analityczne można weryfikować także poprzez „zasilanie modeli analitycznych" nowymi danymi zastanymi lub wywołanymi w trakcie kolejnych badań terenowych. Proces ten ma charakter iteracyjny i opiera się na ciągłej interakcji między

[26] Proces odkrywania wiedzy w bazach danych (Knowledge Discovery in Databases – KDD) składa się z pięciu etapów: selekcji, preprocessingu, transformacji, data mining oraz interpretacji/ewaluacji.

danymi, na bazie których powstał model wyjaśniający, a danymi pojawiającymi się podczas kolejnych cykli badań. Skonstruowany indukcyjnie model analityczny podlega więc ciągłej weryfikacji w zależności od nowych danych lub pojawiających się nowych cech, które nie występowały wcześniej w modelu. Iteracyjność tego procesu można również odnieść do ciągłej interakcji między wnioskowaniem indukcyjnym i hipotetyczno-dedukcyjnym stosowanym w podejściu *mixed methods*. Punktem wyjścia jest rozumowanie indukcyjne w analizie oparte na algorytmach i technikach analitycznych służących eksploracji danych. W procesie tym generowane są hipotezy pobocze, które mogą być z kolei testowane w kolejnych cyklach badań i analiz socjologicznych. Rozumowanie dedukcyjne pojawia się wtedy, gdy gotowe są modele analityczne. W socjologii jakościowej proces ten sprzyja ciągłej weryfikacji rozproszonej wiedzy, ale przede wszystkim pozwala na swoiste novum w obszarze wspomaganej komputerowo analizy danych – eksperymentowanie z aposteriorycznymi modelami rzeczywistości społecznej, wygenerowanymi w toku analizy danych jakościowych/tekstowych.

W dziedzinie Data Mining funkcjonuje kilka prawd i nieporozumień dotyczących analizy danych z wykorzystaniem tej metodologii. Po pierwsze uważa się, że proces drążenia danych wymaga sztucznej inteligencji. I tak, i nie. Pomimo że wiele programów używa sieci neuronowych, które uznawane są za jedną z metod sztucznej inteligencji, to ich użytkownik, badacz społeczny, nie musi znać szczegółów funkcjonowania metod sztucznej inteligencji. Wyniki ich stosowania są całkowicie przezroczyste. Ważne jest, że badacz posiada pewien model do wykorzystania, model wyjaśniający relacje w danych. Po drugie uważa się, że proces eksploracji danych jest bardzo trudny. Skomplikowane są jedynie algorytmy analityczne, a coraz nowsze narzędzia, z odpowiednim interfejsem użytkownika, pozwalają na łatwe ich użycie. Po trzecie uznaje się, co jest faktem, że proces przygotowania danych do analiz Data Mining jest bardzo złożony. Szacuje się, że zajmuje ok. 80% całkowitego czasu samej analizy. Dane muszą być przeczyszczone, zintegrowane oraz dobrze zorganizowane tak, aby można było uzyskać właściwy model wyjściowy. Po czwarte zastosowanie metod i technik Data Mining wymaga posiadania hurtowni danych. Faktem jest, że Data Mining pozwala pracować bardziej inteligentnie z danymi składowanymi w hurtowniach danych. Okazuje się jednak, że wiele programów służących do eksploracji danych daje zadowalające wyniki w systemach bez hurtowni danych. Nawet wtedy, gdy mamy do czynienia z ogromną liczbą danych, to cel badań określa zakres wykorzystywanych danych lub ich próbkę. Tylko nieliczne programy wymagają używania wszystkich posiadanych danych. Najpierw należy się upewnić, co chcemy osiągnąć, zanim zaczniemy czasochłonny proces eksploracji danych. Kolejnym nieporozumieniem jest kwestia posiadanego sprzętu i mocy obliczeniowej. Chociaż sprzęt ma istotne znaczenie dla szybkości wykonywanych analiz, to wiele aplikacji

Data Mining można uruchomić na dobrze wyposażonym komputerze stacjonarnym lub laptopie. Warto w tym miejscu zwrócić również uwagę na to, że nie ma żadnych automatycznych narzędzi Data Mining, które na poczekaniu i w mechaniczny sposób rozwiązują problemy badawcze czy generują wzorce lub modele analityczne. Data Mining to proces ciągłej analizy danych, wymagający stałego zaangażowania i kontroli badacza jakościowego. Nawet jeśli w modelowaniu danych jakościowych (tekstowych) użyjemy różnych nienadzorowanych technik uczenia, to na każdym etapie procesu drążenia danych wymagana jest obecność człowieka. Analityk – badacz jakościowy musi kontrolować sposób przygotowanego modelu wyjaśniającego, szczególnie wtedy, gdy pojawiają się nowe dane. Musi pamiętać, że proces eksploracji i odkrywania wiedzy w danych, dzięki zastosowaniu zaawansowanych algorytmów i technik analitycznych oraz wsparcia oprogramowania komputerowego, pozwala budować modele analityczne i znajdować wzorce, reguły czy zależności. Wyłącznie jednak w gestii badacza – analityka pozostaje identyfikacja przyczyn ich występowania, interpretacja czy teoretyzowanie.

Zakończenie. Odkrywanie wiedzy w socjologii i naukach społecznych

Od kilkunastu lat w obszarze socjologii jakościowej i nauk społecznych wzrasta zainteresowanie zaawansowanymi, nowatorskimi metodami i technikami analizy danych w odniesieniu do różnego rodzaju danych tekstowych: wywiadów (swobodnych, pogłębionych, biograficznych, narracyjnych, zogniskowanych wywiadów grupowych itp.), zapisów obserwacji, materiałów prasowych, literackich, blogów, forów czy danych hipertekstowych. Zjawisku temu towarzyszy dynamiczny rozwój oprogramowania do wspomaganej komputerowo analizy danych jakościowych, w tym właśnie danych tekstowych (CAQDAS). Wzrasta liczba użytkowników programów CAQDAS w naukach społecznych, a także humanistycznych czy medycznych, pojawiają się nowe funkcjonalności w programach będące odpowiedzią na dynamiczny rozwój nowych technologii, algorytmów i technik analitycznych. W praktyce analizy i badania jakościowe nie ograniczają się „do kilku lub kilkunastu wywiadów", lecz dzięki rozwojowi nowych technologii zbierania, przechowywania i przetwarzania danych – tak jak w przypadku metodologii Data Mining – umożliwiają prowadzenie wielowymiarowych analiz na dużych zbiorach danych jakościowych (nie tylko tekstowych). Ponadto w większości przypadków dostępne na rynku programy CAQDAS umożliwiają łączenie w procesie analizy danych ilościowych i jakościowych zgodnie z logiką *mixed methods* (metod mieszanych). Niestety wciąż brakuje paradygmatu analityczno-badawczego, spójnego schematu pojęć i definicji, procedur analitycznych zdolnych

do ich przetworzenia, wydobycia zawartych w tych danych informacji, ujęcia ich w struktury interpretowalnej wiedzy naukowej. Nie ma wystandaryzowanych oraz sprawdzonych reguł i procedur metodologicznych prowadzenia analizy danych jakościowych w naukach społecznych. Nie ma także takich narzędzi analitycznych, jak słowniki semantyczne czy klasyfikacyjne, które można by efektywnie wykorzystywać w procesie wspomaganej komputerowo analizy danych jakościowych. Metodologia drążenia danych stanowi moim zdaniem istotny krok w kierunku wypracowania ram paradygmatycznych dla wspomaganego komputerowo procesu odkrywania wiedzy w danych jakościowych. Należy jednak pamiętać, że eksploracja danych jakościowych (ustrukturyzowanych i nieustrukturyzowanych) i odkrywanie wiedzy jest złożonym procesem integracji różnych źródeł danych, selekcji i transformacji danych, eksploracji, ekstrakcji wiedzy, wizualizacji związków, testowania modeli analitycznych oraz interpretacji uzyskanych wyników. W przypadku danych ustrukturyzowanych wszystkie te operacje są obecnie wspierane przez wyspecjalizowane, komercyjne oprogramowanie analityczne tj.: IBM Modeller (dawniej SPSS Clementine), Statistica Data Miner czy SAS Enterprise Miner, współpracujące zwykle z systemami bazodanowymi[27]. W przypadku nieustrukturyzowanych danych jakościowych możliwości kompleksowej eksploracji i odkrywania wiedzy przy wsparciu programów CAQDAS wykorzystujących algorytmy i techniki Data czy Text Mining są w praktyce nieograniczone. Do znanych w środowisku badaczy jakościowych i osób zajmujących się analizą treści programów, które zawierają te rozwiązania, należą QDA Miner, Qualrus, T-lab bądź RapidMiner[28]. Zgodnie z logiką teorii ugruntowanej wiedza o rzeczywistości społecznej tkwi w danych empirycznych (w szczególności danych jakościowych), a krokiem jej poznania jest ich kompleksowa analiza i zrozumienie struktury zależności między nimi. Wiąże się to nierozerwalnie z procesem eksploracji i odkrywania wiedzy w danych poprzez twórczą rekonstrukcję ich relacji i pogłębioną analizę, generowanie i testowanie hipotez, a także modelowanie zależności między nimi czy modelowanie procesów społecznych przy użyciu zaawansowanych, wielowymiarowych technik i algorytmów analitycznych, uczenia maszynowego czy metod sztucznej inteligencji. Zarówno podejście „od danych empirycznych" (bottom – up), niezależnie od stosowanych metod badawczych: ilościowych czy jakościowych, jak i generowanie wiedzy z danych jest zgodne z podejściem analitycznym dominującym od dawna w tradycji socjologii jakościowej. Data Mining pokazuje, że analiza danych jakościowych – na etapie wstępnym – jest przede wszystkim procesem indukcyjnym. Jeśli przyjąć, że indukcja jest podstawowym sposobem wnioskowania w metodologii badań jakościowych, to większość

[27] Zob. strona Data i Text Mining Community, www.kdnuggets.com/software/text.html.

[28] Zob. strona producenta oprogramowania Rapid Miner, http://rapidminer.com/solutions/.

technik eksploracji korzysta z mechanizmów indukcyjnych w procesie odkrywania wiedzy w danych. Konstruowanie modeli analitycznych zachowań, działań czy procesów społecznych dokonuje się zawsze w odniesieniu do określonego układu danych empirycznych, jakimi dysponuje badacz lub jakie zebrał w trakcie badań terenowych. W tym sensie model analityczny jest jedynie pewną pozakontekstową aproksymacją (przybliżeniem) rzeczywistego stanu rzeczy lub procesu społecznego. Dlatego też proces eksploracji i odkrywanie prawidłowości w zbiorach danych jakościowych możemy śmiało nazwać modelowaniem, uczeniem się na podstawie danych. Wiedza, reguły czy wzorce odkryte podczas analizy danych są w pewnym sensie uogólnieniem ich struktury semantycznej. Analiza danych jakościowych wychodzi wtedy poza wymiar opisu czy tworzenia typologii. Dzięki temu istotą analizy jakościowej staje się rozwijanie teorii normatywnej zgodnie z przedstawioną w artykule logiką oraz metodologią eksploracji i odkrywania wiedzy w danych.

Bibliografia

Agrawal Rakesh, Imieliński Tomasz, Swami Arun (1993), *Mining Association Rules Between Sets of Items in Large Databases*, "ACM SIGMOD Record", 22 (2), s. 207–216.

Agrawal Rakesh, Srikant Ramakrishnan (1994), *Fast Algorithms for Mining Association Rules in Large Databases*. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago, s. 487–499.

Becker Howard S., Gordon Andrew C., LeBailly Robert K. (1984), *Fieldwork with the Computer: Criteria for Assessing Systems*, "Qualitative Sociology", 7 (1–2), s. 16–33.

Berelson Bernard (1952), *Content Analysis in Communication Research*, Free Press, Glencoe, IL.

Bong Sharon A. (2002), *Debunking Myths in Qualitative Data Analysis*, Forum Qualitative Sozialforschung, vol. 3, no. 2; www.qualitative-research.net/index.php/fqs/article/view/849 [dostęp: 1.06.2014].

Brent Edward E. (1984), *Qualitative Computing: Approaches and Issues*, "Qualitative Sociology", 7 (1/2), s. 36–60.

Brent Edward E., Anderson Ronald E. (1990), *Computer Applications in the Social Sciences*, Temple University Press, Philadelphia.

Bryda Grzegorz (2014), *CAQDAS a badania jakościowe w praktyce*, „Przegląd Socjologii Jakościowej", t. 10, nr 2, s. 12–38; www.przegladsocjologiijakosciowej.org/Volume26/PSJ_10_2_Bryda.pdf [dostęp: 1.06.2014].

Bryda Grzegorz, Tomanek Krzysztof (2014), *Od CAQDAS do Text Miningu. Nowe techniki w analizie danych jakościowych*, [w:] Jakub Niedbalski (red.), *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Chapman Pete, Clinton Julian, Kerber Randy, Khabaza Thomas, Reinartz Thomas, Shearer Colin, Wirth Rüdiger (1999, 2000), *CRISP – DM 1.0. Step-by-step Data Mining Guide*, ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf [dostęp: 27.05.2014].

Cichosz Paweł (2007), *Systemy uczące się*, Wydawnictwa Naukowo-Techniczne, Warszawa.
Conrad Peter, Reinharz Shulamit (1984), *Computers and Qualitative Data*, Human Sciences Press, New York, NY.
Dey Ian (1993), *Qualitative Data Analysis: A User-Friendly Guide for Social Scientists*, Routledge, London–New York, NY.
Drass Kriss A. (1989), *Text Analysis and Text-Analysis Software: A Comparison of Assumptions*, [w:] Grant Blank, James L. McCartney, Edward E. Brent (eds), *New in Technology in Sociology: Practical Applications in Research and Work*, Transaction Publishers, New Brunswick, NJ.
Fayyad Usama M. (1996), *Advances in Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, Calif.
Fayyad Usama, Piatetsky-Shapiro Gregory, Smyth Padhraic (1996), *From Data Mining to Knowledge Discovery in Databases*, AI Magazine 17 (3), Menlo Park, California, s. 37–54; www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf [dostęp: 01.06.2014].
Fielding Nigel G. (2012), *The Diverse Worlds and Research Practices of Qualitative Software*, Forum Qualitative Sozialforschung, vol. 13, no. 2; www.qualitative-research.net/index.php/fqs/article/view/1845/3369 [dostęp: 1.06.2014].
Fielding Nigel G., Lee Raymond M. (1998), *Computer Analysis and Qualitative Research*, Sage, London.
Fielding Nigel G., Lee Raymond M. (1993), *Using Computers in Qualitative Research*, Sage, London.
Fischer Michael D. (1994), *Applications in Computing for Social Anthropologists*, Routledge, London.
George, Alexander L. (1959), *Propaganda Analysis: A Study of Inferences Made From Nazi Propaganda in World War II*, Evanston III Row, Peterson.
Gerson Elihu (1984), *Qualitative research and the computer*, "Qualitative Sociology", 7 (1/2), s. 61–74.
Gibbs Graham (2011), *Analiza danych jakościowych*, przeł. Maja Brzozowska-Brywczyńska, Wydawnictwo Naukowe PWN, Warszawa.
Glaser Barney G., Strauss Anselm Leonard (2009), *Odkrywanie teorii ugruntowanej: strategie badania jakościowego*, przeł. Marek Gorzko, Zakład Wydawniczy Nomos, Kraków.
Guest Greg, MacQueen Kathleen M., Namey Emily E. (2012), *Applied Thematic Analysis*, Sage Publications, Los Angeles.
Han Jiawei, Kamber Micheline (2006), *Data Mining: Concepts and Techniques*, Elsevier, Amsterdam.
Hand D., Mannila H., Smyth P. (2005), *Eksploracja danych*, WNT, Warszawa.
Ho Yu Chong, Jannasch-Pennell Angel, DiGangi Samuel (2011), *Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability*, The Qualitative Report, vol. 16, no. 3, May, s. 730–744; www.nova.edu/ssss/QR/QR16-3/yu.pdf [dostęp: 1.06.2014].
Holsti Ole R. (1969), *Content Analysis for the Social Sciences and the Humanities*, Addison-Wesley, Reading, MA.
Johnson R. Burke, Onwuegbuzie Anthony J. (2004), *Mixed methods Research: A Research Paradigm Whose Time Has Come*, "Educational Researcher", 33 (7), s. 14–26.
Kelle Udo (ed.), (1995), *Computer-aided Qualitative Data Analysis: Theory, Methods and Practice*, Sage Publications, London.
Kracauer Siegfried (1952), *The Challenge of Qualitative Content Analysis*, "Public Opinion Quarterly", 16, s. 631–642.
Krippendorf Klaus (2004), *Content Analysis. An Introduction to its Methodology*, Sage, Thousand Oaks, CA.

Krippendorff Klaus (1986), *Information Theory Structural Models for Qualitative Data*, Sage Publications, Beverly Hills, California.
Larose Daniel T. (2008), *Metody i modele eksploracji danych*, PWN, Warszawa.
Larose Daniel T. (2006), *Odkrywanie wiedzy z danych: wprowadzenie do eksploracji danych*, PWN, Warszawa.
Lewins Ann, Silver Christina (2007), *Using Software in Qualitative Research: A Step-by-Step Guide*, Sage Publications, London.
Lofland John (2009), *Analiza układów społecznych. Przewodnik metodologiczny po badaniach jakościowych*, przeł. Elżbieta Hałas, Anna Rosińska-Kordasiewicz, Sylwia Urbańska, Monika Żychlińska i in., Wydawnictwo Naukowe Scholar, Warszawa.
Miles Matthew B., Huberman Michael A. (2000), *Analiza danych jakościowych*, przeł. Stanisław Zabielski, Trans Humana, Białystok.
Morzy Tadeusz (2013), *Eksploracja danych: metody i algorytmy*, Wydawnictwo Naukowe PWN, Warszawa.
Pfaffenberger Bryan (1988), *Microcomputer Applications in Qualitative Research*, Sage, Newbury Park, CA.
Piatetsky-Shapiro Gregory, Frawley William (1991), *Knowledge Discovery in Databases*, AAAI Press, Menlo Park, Calif.
Richards Lyn, Richards Tom (1989), *The Impact of Computer Techniques for Qualitative Analysis*, Technical Report, no. 6/89, Department of Computer Science, La Trobe University.
Richards Lyn, Richards Tom (1991), *The Transformation of Qualitative Method: Computational Paradigms and Research Processes*, [w:] Nigel G. Fielding, Raymond M. Lee (eds), *Using Computers in Qualitative Research*, Sage, London.
Seidel John (1991), *Method and Madness in the Application of Computer Technology to Qualitative Data Analysis*, [w:] Nigel G. Fielding, Raymond M. Lee (eds), *Using Computers in Qualitative Research*, Sage, London.
Saldaña Johnny (2013), *The Coding Manual for Qualitative Researchers*, 2 ed., Sage, London.
Schreier Margrit (2012), *Qualitative Content Analysis in Practice*, Sage Publications, London.
Silvana di G., Davidson J. (2008), *Qualitative Research Design for Software Users*, Open University Press, Milton Keynes.
Stone Philip J., Dunphy Dexter C., Smith Marshall S., Ogilvie Daniel M. (1966), *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press, Cambridge, Massachusetts.
Tashakkori Abbas, Teddlie Charles (2003), *Handbook of Mixed methodsMixed methods in Social & Behavioral Research*, Sage Publications, Thousand Oaks, Calif.
Tesch Renata (1990), *Qualitative Research: Analysis Types and Software Tools*, Falmer Press, London and Philadelphia.
Weber Robert P. (1990), *Basic content analysis*, Sage, Newbury Park, CA.
Wiedemann Gregor (2013), *Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences*, Forum Qualitative Sozialforschung, vol. 14, no. 2; www.qualitative-research.net/index.php/fqs/article/view/1949 [dostęp: 1.06.2014].

Data i Text Mining Community, www.kdnuggets.com/software/text.html.
The CAQDAS Networking Project, www.surrey.ac.uk/sociology/research/researchcentres/caqdas/about/.
General Inquirer, www.wjh.harvard.edu/~inquirer/homecat.htm.
Index Thomisticus, www.corpusthomisticum.org/it/.
Laswell Value Dictionary, www.wjh.harvard.edu/~inquirer/lasswell.htm.
Online QDA, www.onlineqda.hud.ac.uk/Intro_QDA/what_is_qda.php.

The TEI Guidelines for Electronic Text Encoding and Inter Change, www.tei-c.org/Guidelines/.
www.ideaworks.com/download/qualrus/QualrusManual.pdf.
www.provalisresearch.com/Documents/QDAMiner40.pdf.

Oprogramowanie

QDA Miner – http://provalisresearch.com/products/.
Qualrus – www.qualrus.com/.
T-Lab – http://tlab.it/en/presentation.php.
RapidMiner – http://rapidminer.com/solutions/.

CAQDAS, Data Mining and Knowledge Discovery in Qualitative Data

Summary. The aim of this article is methodological reflection over the process of the development of the computer-assisted qualitative data analysis (CAQDAS) from traditional qualitative analysis based primarily on grounded theory procedures by qualitative content analysis techniques, towards the use of advanced methods and techniques of Data Mining and Knowledge Discovery in Datasets, in the field of qualitative sociology and social sciences. This process is accompanied by expansion of information technology in the qualitative sociology, the evolution of ways of collection and processing informations, and erosion of the new algorithms and analytical techniques, what has led to a situation in which the usage of their achievements in the field of qualitative sociology and social science is a natural process of development of CAQDAS. Currently the use of CAQDAS in the area of qualitative sociology is rather common and that it is not surprising that more and more researchers, including Poland, reaches for the computer software in qualitative data analysis. CAQDAS teaches methodological rigor, accuracy and precision in qualitative data analysis, what positively affects the quality of the analyzes and research. However, the analysis of qualitative data using the methodology of Data Mining is a novelty in the field of qualitative sociology. This involves not only the use of new algorithms and analytical techniques, but also with changes in the approach to computer-aided analysis of qualitative data, adding new functionalities such as the possibility the analysis of the content and structure of the linguistic text documents. The changes in the area of CAQDAS are accompanied by observed for several years methodological return towards mixed-methods in the sociology and social sciences, particularly in qualitative research and data analysis. Its consequence is the implementation of multivariate statistical techniques, textual data mining techniques, algorithms of computer intelligence and natural language processing into programs for computer-assisted qualitative data analysis (QDA Miner, Qualrus or T-Lab). The majority of these solutions have its roots in the booming Data Mining methodology. Generally CAQDAS software are mostly used to work with smaller data sets, but Data Mining allows to conduct analyzes in which the size of the data set is basically unlimited. The purpose of this article is to present to the community of qualitative researches in Poland the methodology of Data Mining and procedures of knowledge discovery in data, and thus encourage them to experiment with new approaches in the area of CAQDAS. I also try to show the relationship between CAQDAS and grounded theory, Data Mining and process of knowledge discovery in data in the field of qualitative sociology, and broadly in the social sciences.

Keywords: qualitative data analysis, grounded theory, knowledge discovery in data, CAQDAS, Data Mining, mixed-methods.