Metody i techniki odkrywania wiedzy

NARZĘDZIA CAQDAS W PROCESIE ANALIZY DANYCH JAKOŚCIOWYCH

pod redakcją
Jakuba Niedbalskiego

WYDAWNICTWO UNIWERSYTETU ŁÓDZKIEGO

Metody i techniki odkrywania wiedzy

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Metody i techniki odkrywania wiedzy

NARZĘDZIA CAQDAS W PROCESIE ANALIZY DANYCH JAKOŚCIOWYCH

pod redakcją
Jakuba Niedbalskiego

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

ŁÓDŹ 2014

Jakub Niedbalski – Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny, Instytut Socjologii
Katedra Socjologii Organizacji i Zarządzania, 90-214 Łódź, ul. Rewolucji 1905 r. nr 41/43
e-mail: jakub.niedbalski@gmail.com

RECENZENT
Marian Niezgoda

REDAKTOR WYDAWNICTWA UŁ
Dorota Stępień

SKŁAD I ŁAMANIE
AGENT PR

PROJEKT OKŁADKI
Łukasz Orzechowski

Zdjęcie na okładce: © momius – Fotolia.com

Publikacja dofinansowana z funduszy Rektora Uniwersytetu Łódzkiego
oraz Dziekana Wydziału Ekonomiczno-Socjologicznego UŁ

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I.W.06685.14.0.K

Ark. wyd. 19,0; ark. druk. 18,375

ISBN 978-83-7969-549-2 (wersja papierowa)
ISBN 978-83-7969-550-8 (wersja online)

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. (42) 665 58 63, faks (42) 665 58 62

Spis treści

Wprowadzenie do komputerowej analizy danych jakościowych

Rozwój nowoczesnych technologii komputerowych powoduje, że coraz bardziej widoczny staje się wpływ innowacyjnych narzędzi takich jak specjalistyczne oprogramowanie na proces badawczy zarówno w naukach ścisłych, jak i humanistycznych (Niedbalski 2013b). Na przestrzeni ostatniej dekady możemy obserwować niezwykle dynamiczny rozwój oprogramowania komputerowego wspomagającego analizę danych jakościowych, a lista dostępnych programów staje się coraz dłuższa (zob. Lewins, Silver 2004). Pierwsze tworzone były przez samych badaczy, obecnie za kolejnymi wersjami stoją całe zespoły projektowe składające się z naukowców oraz zaplecza informatycznego, a wiele czołowych programów stało się już dobrze rozpoznawalną marką, znaną na całym świecie. Jednocześnie kolejne wersje programów wzbogacane są o nowe funkcje poszerzające możliwości wykonywanej za ich pomocą analizy danych (Niedbalski 2013b).

Niniejsza publikacja jest inspirowana aktualnymi trendami w naukach społecznych i humanistycznych, które już od kilkudziesięciu lat są prężnie rozwijane w czołowych ośrodkach naukowych za granicą. W Polsce również mamy do czynienia z rosnącym zainteresowaniem zarówno świata akademickiego, jak i podmiotów rynkowych z możliwościami oprogramowania CAQDAS (komputerowego wspomagania analizy danych jakościowych) w projektowaniu i prowadzeniu badań jakościowych. Nieustanie wzrasta liczba badaczy, naukowców, ale także praktyków zaciekawionych prowadzeniem badań jakościowych, poszukujących przy tym narzędzi, które mogłyby wspomóc proces analityczny. Osoby zainteresowane metodami ilościowymi mogą czerpać z bogatej literatury prezentującej takie programy komputerowe, jak SPSS czy Statistica. Na rynku wydawniczym nie ma jednak zbyt wielu tego typu opracowań, odnoszących się do programów CAQDAS. W rodzimej literaturze metodologicznej występują jedynie pojedyncze opracowania odnoszące się do tej tematyki (Trutkowski 1999; Bieliński, Iwińska, Kordasiewicz 2007; Niedbalski, Ślęzak 2012; Brosz 2012; Niedbalski 2013a, 2014). Jednocześnie w naszym kraju istnieją badacze wykorzystujący i specjalizujący się w rozmaitych CAQDAS, a nawet tworzący polskie programy do analizy danych jakościowych.

Prezentowana książka ma szansę stać się publikacją, która zaprezentuje możliwości i sposób wykorzystania programów CAQDAS w badaniach opartych na metodach jakościowych, uzupełniając w ten sposób literaturę przedmiotu dostępną na polskim rynku.

Publikacja, którą oddajemy w ręce czytelników, jest zbiorem artykułów badaczy posiadających przeważnie wieloletnie doświadczenie w stosowaniu nowoczesnych narzędzi wspomagających proces badawczy, takich jak specjalistyczne oprogramowanie komputerowe. Powstanie niniejszej grupy było podyktowane obserwowanym od dłuższego czasu zapotrzebowaniem środowiska naukowego, w którym pojawia się coraz więcej osób zarówno korzystających z oprogramowania komputerowego, jak i zainteresowanych jego wdrożeniem w planowanych oraz realizowanych przez siebie przedsięwzięciach badawczych, ale które jak dotychczas nie miały okazji do wymiany doświadczeń oraz poszukiwania fachowej wiedzy w tym zakresie. Proponowana pozycja ma za zadanie przybliżyć nowe spojrzenie na metodologię badań jakościowych i przyczynić się do rozpropagowania idei stosowania nowych technologii w naukach społecznych i humanistycznych.

Książka zawiera teksty przygotowane przez badaczy i praktyków, dla których praca w środowisku oprogramowania komputerowego jest codziennością. Dzięki temu otrzymujemy bardzo rzetelną wiedzę opartą na wieloletnim doświadczeniu poszczególnych autorów, którzy w danym zakresie reprezentują wiedzę ekspercką. Zbiór ten zawiera i pokazuje w sposób przekrojowy, ale też systematyczny, korzystanie z różnych programów w ramach prowadzenia badań opartych na rozmaitych metodach i z wykorzystaniem wielu narzędzi badawczych. W ten sposób zyskujemy szerokie spektrum możliwości wykorzystania obecnie istniejących, popularnych programów z rodziny CAQDA, a zarazem możemy przyjrzeć się różnym ich zastosowaniom. Prezentowana książka powinna więc zaspokoić oczekiwania zarówno niedoświadczonych jeszcze użytkowników oprogramowania, którzy pragną zasięgnąć nieco informacji na temat jego zastosowania, jak i wytrawnych badaczy, którzy dzięki niej mogą nieco zrewidować swój warsztat badawczy, a być może odnaleźć świeży powiew inspiracji.

Wśród wielu zagadnień poruszanych przez autorów warto zwrócić uwagę na tak istotne kwestie, jak: podejmowanie dyskusji nad zgodnością zasad, na jakich funkcjonuje oprogramowanie CAQDA z regułami oraz procedurami metodologii badań jakościowej; wskazanie możliwości zastosowania oprogramowania CAQDA w realizacji projektów badawczych opartych na różnych metodach jakościowych i w ramach różnych podejść analitycznych; zaprezentowanie zgodności „architektury oprogramowania" z procedurami wybranych metod badawczych; przedstawienie wpływu nowych technologii na przebieg procesu badawczego; a także wytyczenie kierunków rozwoju, w jakich powinien podążać proces implementowania nowoczesnych rozwiązań technologicznych

w proces realizacji projektów badawczych opartych na metodach jakościowych oraz ukazanie przyszłości metod jakościowych w kontekście zastosowania oprogramowania CAQDA.

Książkę rozpoczyna niezwykle interesujący artykuł **Grzegorza Brydy**, w którym wraz z autorem możemy prześledzić proces 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 szerzej, w naukach społecznych zaawansowanych metod eksploracji danych i odkrywania wiedzy (Data Mining, DM and Knowledge Discovery in Datasets, KDD). Celem artykułu jest przybliżenie metodologii Data Mining i odkrywania wiedzy w danych przez badaczy jakościowych w Polsce, a tym samym zachęcenie do eksperymentowania z nowymi podejściami w obszarze CAQDAS.

Kamil Brzeziński zapoznaje z kolei czytelników z badaniami dotyczącymi motywów podjęcia decyzji o zamieszkaniu na „osiedlu grodzonym", dostrzeganych przez ich mieszkańców zalet i wad takich kompleksów, a także wewnętrznych relacji sąsiedzkich. Prezentowane badania stanowią tło dla sposobu i charakterystyki wykorzystania programu QDA Miner, który posłużył autorowi do przeprowadzenia analizy danych i realizacji wspomnianego problemu badawczego.

Dzięki artykułowi **Izabeli Ślęzak** i **Jakuba Niedbalskiego** mamy natomiast wgląd w to, jak poszczególne opcje programu NVivo mogą zostać wykorzystane, aby stanowiły skuteczny środek do wsparcia analizy danych prowadzonej zgodnie z procedurami metodologii teorii ugruntowanej. Autorzy pokazują, w jaki sposób określony program należący do rodziny CAQDA może sprostać wymaganiom badacza stosującego wybraną metodę badawczą. Nie stronią również od uwag nad rozwiązaniami, które zostały zaimplementowane do opisywanego narzędzia, odnosząc się w ten sposób krytycznie do jego wewnętrznej architektury i niektórych funkcji programu.

Na przykładzie określonego projektu badawczego **Jakub Niedbalski** stara się przybliżyć, jak realizować badania zgodnie z procedurami metodologii teorii ugruntowanej, korzystając z dostępnych funkcji trzech bezpłatnych programów komputerowych Audacity, WeftQDA oraz CmapTools. Artykuł ma charakter poglądowy i edukacyjny, pozwalający zapoznać się z możliwościami narzędzi CAQDA oraz ich faktycznym zastosowaniem w realizacji projektów badawczych opartych na wskazanej metodzie badawczej.

Z kolei **Artur Piszek** opisuje narzędzie Qualify, które dzięki nowatorskiemu zastosowaniu pozwala zwiększyć użyteczność oprogramowania Evernote o możliwość wykonywania za jego pomocą jakościowej analizy treści. Autor prezentuje najważniejsze informacje dotyczące wspomnianego narzędzia, zapoznając czytelnika krok po kroku ze sposobami wykorzystania jego poszczególnych funkcji.

Z podobną inicjatywą mamy do czynienia w przypadku artykułu **Jerzego Żurko**, który od kilku lat z powodzeniem stosuje program Socjolog, biorąc jednocześnie czynny udział w pracach nad jego udoskonalaniem. Wspomniana aplikacja jest dobrym przykładem efektywnej współpracy badaczy reprezentujących nauki humanistyczne oraz profesjonalnych informatyków, którzy potrafili wspólnymi siłami stworzyć od podstaw ciekawe i co ważne – rodzime oprogramowanie.

Krzysztof Tomanek w swoim tekście poświęconym autorskiej koncepcji analizy treści polegającej na klasyfikacji wypowiedzi lub tekstów opartej na metodologii stosowania algorytmów zapożyczonych z obszaru machine learning (ML) akcentuje natomiast dwie różnice wobec podejścia ML w stosunku do własnych koncepcji metodologicznych. Po pierwsze proponuje budowę słowników tematycznych, które składają się ze słów i fraz kluczowych (podobnie jak ML), ale które wzbogacone o reguły semantyczne i pragmatyczne (inaczej niż w ML) identyfikują dodatkowe, specyficzne dla wypowiedzi cechy. Po drugie proponuje wyposażenie słowników klasyfikacyjnych w reguły rządzące logiką analizowanych wypowiedzi.

Ten sam autor – **Krzysztof Tomanek** – w artykule *Jak nauczyć metodę samodzielności? O uczących się metodach analizy treści* wprowadza czytelników w niezwykle interesujące zagadnienie zaawansowanych statystycznie systemów znajdujących zastosowanie w jakościowych analizach danych tekstowych. Opisuje w nim podstawowe, dostępne w wybranych programach CAQDAS (ze szczególnym uwzględnieniem programu Qualrus), techniki wspierające opracowanie materiałów tekstowych, takie jak automatyczne i półautomatyczne metody kodowania.

W kolejnym artykule **Grzegorz Bryda** i **Krzysztof Tomanek** podejmują refleksję metodologiczną nad procesem rozwoju klasycznych analiz jakościowych w obszarze nauk społecznych, a szczególnie w socjologii, która charakteryzuje się przechodzeniem od „stylu" CAQDAS w kierunku Text Miningu.

Celem następnego artykułu – napisanego również przez **Grzegorza Brydę**, **Krzysztofa Tomanka** – jest prezentacja strategii stosowanych w analizie danych tekstowych. Autorzy pokazują jak budować narzędzia służące do analizy dużych zbiorów danych tekstowych, wskazując przy tym, że w ramach analiz treści stosować można metody inspirowane podejściem zgodnym z teorią ugruntowaną, analizą z zastosowaniem reguł leksykalnych, metod statystycznych oraz podejściem specyficznym dla logiki falsyfikacjonizmu.

Tekst napisany przez **Jacka Burskiego** odsłania zaś kolejny, aplikacyjny aspekt zastosowania programu komputerowego QDA Miner służącego do wsparcia analiz danych jakościowych. Główne zadanie, jakie stawia sobie autor tekstu, dotyczy konsekwencji użycia techniki komputerowej do skomplikowanych analiz jakościowych, a także jej ewentualnego wpływu na wyniki procesu badawczego.

Jacek Burski stara się w ten sposób wykazać, iż pomimo zastosowania zaawansowanych narzędzi komputerowych intuicja badacza oraz jego zdolności analityczne i syntetyczne zawsze powinny odgrywać główną rolę.

W ostatnim artykule niniejszej książki **Kamil Głowacki** prezentuje pakiet narzędzi służących organizacji i zarządzaniu wiedzą gromadzoną oraz wytwarzaną przez badacza. Jest to także zestaw narzędzi wspomagających badacza w procesie koordynowania całego przedsięwzięcia badawczego. Z całą pewnością wśród opisywanych przez autora programów każdy znajdzie ten, który będzie najlepiej spełniał jego własne wymagania, biorąc pod uwagę rodzaj, przedmiot oraz zakres prowadzonych przez siebie badań.

Wszystkie teksty zawarte w publikacji stanowią istotny wkład w zrozumienie specyfiki oraz istoty rozmaitych kontekstów i uwarunkowań związanych ze stosowaniem oprogramowania komputerowego wspomagającego analizę danych jakościowych. Książka ma szansę przyczynić się do lepszego poznania tej dynamicznie rozwijającej się tematyki oraz może wzbudzić refleksję nad aktualnym stanem wiedzy dotyczącej oprogramowania CAQDA.

Jakub Niedbalski

Bibliografia

Bieliński Jacek, Iwańska Katarzyna, Rosińska-Kordasiewicz Anna (2007), *Analiza danych jakościowych przy użyciu programów komputerowych*, „ASK. Społeczeństwo. Badania. Metody", nr 16, s. 89–114.

Brosz Maciej (2012), *Komputerowe wspomaganie badań jakościowych. Zastosowanie pakietu NVivo w analizie materiałów nieustrukturyzowanych*, „Przegląd Socjologii Jakościowej", t. 8, nr 1, s. 98–125; www.przegladsocjologiijakosciowej.org [dostęp: 20.11.2012].

Lewins Ann, Silver Christina (2004), *Choosing CAQDAS Software. CAQDAS Networking Project*, University of Surrey, Guildford.

Niedbalski Jakub, Ślęzak Izabela (2012), *Analiza danych jakościowych przy użyciu programu NVivo a zastosowanie procedur metodologii teorii ugruntowanej*, „Przegląd Socjologii Jakościowej", t. 8, nr 1, s. 126–165; www.przegladsocjologiijakosciowej.org [dostęp: 20.11.2013].

Niedbalski Jakub (2013a), *Odkrywanie CAQDAS. Wybrane bezpłatne programy komputerowe wspomagające analizę danych jakościowych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Niedbalski Jakub (2013b) *CAQDAS – oprogramowanie do komputerowego wspomagania analizy danych jakościowych. Historia ewolucja i przyszłość*, „Przegląd Socjologiczny", t. LXII/1, s. 153–166.

Niedbalski Jakub (2014), *Komputerowe wspomaganie analizy danych jakościowych. Zastosowanie oprogramowania NVivo i Atlas.ti w projektach badawczych opartych na metodologii teorii ugruntowanej*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Trutkowski Cezary (1999), *Analiza treści wspomagana komputerowo. Badanie społecznych reprezentacji polityki*, „ASK. Społeczeństwo. Badania. Metody", nr 8, s. 113–133.

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.