http://dx.doi.org/10.18778/7969-549-2.10

Grzegorz Bryda
Krzysztof Tomanek
Uniwersytet Jagielloński

Od CAQDAS do Text Miningu
Nowe techniki w analizie danych jakościowych

Streszczenie. Celem artykułu jest refleksja metodologiczna nad procesem rozwoju komputerowo wspomaganej analizy danych jakościowych (CAQDAS), który zmierza w kierunku metod eksploracji danych tekstowych służących odkrywaniu wiedzy (Knowledge Discovery in Text Databases, Text Mining). W rozważaniach tych skupiamy się na naukach społecznych, szczególnie w socjologii jakościowej. Zastosowanie wspomaganej komputerowo analizy danych jakościowych w obszarze socjologii jakościowej stało się już poniekąd faktem. Środowisko badaczy jakościowych w Polsce coraz częściej sięga po oprogramowanie CAQDAS w projektach badawczych. Praca z różnorodnymi programami CAQDAS prowadzi do wzrostu świadomości metodologicznej, co przekłada się na większą dokładność i precyzję w procesie analizy danych jakościowych. Jednakże analiza danych jakościowych wykorzystująca metodologię, algorytmy i techniki Text Mining to swoiste *novum* na gruncie socjologii jakościowej. Text Mining (TM) to zestaw technik, w które wyposażone są programy przeznaczone do automatycznego lub semiautomatycznego wydobywania informacji z danych tekstowych. Text Mining polega na wykorzystaniu oprogramowania komputerowego do znajdowania ukrytych dla człowieka, z uwagi na ograniczone możliwości percepcyjne i czasowe, prawidłowości zawartych w danych tekstowych. Jeśli algorytmy analityczne CAQDAS wykorzystuje się w pracy z mniejszymi zbiorami danych jakościowych, to techniki Text Mining pozwalają na prowadzenie analiz, w których wielkość zbioru danych jest w zasadzie nieograniczona. W artykule staramy się ukazać proces rozwoju algorytmów analitycznych CAQDAS w kierunku Text Mining. Staramy się także znaleźć odpowiedź na pytanie, czy te podejścia są względem siebie konkurencyjne czy raczej komplementarne?

Słowa kluczowe: odkrywanie wiedzy w danych, CAQDAS, Data Mining, Text Mining, teoria ugruntowana, przetwarzanie języka naturalnego (NLP), odkrywanie wiedzy w danych tekstowych (KDT).

Wstęp – eksploracja i odkrywanie wiedzy w danych tekstowych

Z metodologicznego punktu widzenia badacz może korzystać z danych pochodzących z własnych lub istniejących badań empirycznych (dane wywołane, pierwotne) lub z już istniejących źródeł danych empirycznych (dane

zastane, wtórne). W przypadku tradycyjnej analizy typu Data Mining najczęściej wykorzystuje się dane zastane, zgromadzone w systemowych bazach danych, hurtowniach czy repozytoriach. Zalicza się do nich zarówno dane statystyczne i demograficzne, transakcyjne, sprzedażowe, rejestry, oficjalne sprawozdania urzędowe, dokumentacje techniczne, ewidencje, ankiety personalne pochodzące z różnego rodzaju instytucji, kroniki, spisy ludności, księgi parafialne i inne informacje archiwalne, kwerendy biblioteczne, jak i wszelkiego rodzaju dane tekstowe: dane ze stron internetowych, zarchiwizowane dane z badań jakościowych, dokumenty osobiste, tj. blogi, listy, dzienniki, pamiętniki, autobiografie, transkrypcje wywiadów, zapiski obserwacji, genealogie – opisujące i rejestrujące wydarzenia z punktu widzenia ich uczestników itp. W naukach społecznych posługujemy się danymi ze spisów ludności prowadzonych w celach administracyjnych lub publicznych, aby na przykład zbadać strukturę gospodarstw domowych, rozkład dochodów i wydatków, wzorce imigracji i migracji, zmiany w strukturze rodziny, mobilność społeczną czy cechy obszarów wiejskich, miejskich i metropolii. Dane zbierane przez ośrodki badania opinii, instytuty naukowe czy organizacje pozarządowe są wykorzystywane do analizy zmian opinii publicznej, postaw politycznych lub aktywności społecznej. Jednocześnie wraz z rozwojem nowych mediów i technologii informatycznych jako badacze dysponujemy coraz większą liczbą danych jakościowych dotyczących bogactwa życia społecznego. Jednak złożoność i wielowymiarowość tych informacji wymaga określonej metodologii oraz dysponowania odpowiednimi technikami i narzędziami, zdolnymi do przetworzenia dużej liczby danych tekstowych. Do takich należy rozwijająca się w ostatnich latach dziedzina odkrywania wiedzy w bazach danych (KDD, Knowledge Discovery in Databases) i metodologia eksploracji, drążenia danych (Data Mining)[1]. Jej rozwinięcie stanowi odkrywanie wiedzy w bazach dokumentów elektronicznych, wykorzystujące możliwości przetwarzania języka naturalnego w procesie analizy tekstu oraz zaawansowane techniki i algorytmy eksploracji danych tekstowych (Text Mining). Konsekwencją tego procesu jest przechodzenie w obrębie środowiska CAQDAS od tradycyjnych analiz w stylu Qualitative Analysis, poprzez Qualitative Content Analysis w kierunku Text Mining. Celem tego artykułu jest przybliżenie problematyki Text Mining oraz refleksja metodologiczna nad możliwościami jej wykorzystywania w obszarze wspomaganej komputerowo analizy danych jakościowych (CAQDAS).

[1] O procesie rozwoju komputerowo wspomaganej analizy danych jakościowych (CAQDAS) w kierunku metod eksploracji danych i odkrywania wiedzy w danych (Data Mining) w obszarze nauk społecznych, a szczególnie socjologii jakościowej zob. Bryda (2014).

Co to jest Text Mining?

Text Mining lub szerzej odkrywanie wiedzy z danych tekstowych (KDT, Knowledge Discovery in Texts) to dziedzina metod i technik eksploracji danych, która łączy w sobie zaawansowane algorytmy i techniki Data Mining oraz logikę analizy treści tekstowych (Hearst 1999). Podejście to ma charakter interdyscyplinarny. W ramach KDT wykorzystujemy bowiem wiedzę z zakresu:

a) drążenia danych (DM, Data Mining),
b) uczenia maszynowego (ML, Machine Learning),
c) przetwarzania języka naturalnego (NLP, Natural Language Processing),
d) metod wyszukiwania i ekstrakcji informacji (Information Retrieval and Extraction),
e) tłumaczenia maszynowego (Machine Translation),
f) statystyki matematycznej,
g) lingwistyki komputerowej,
h) informatyki.

Wykorzystanie wskazanych rozwiązań, podobnie jak w przypadku Data Mining, umożliwia znajdowanie w analizowanych dokumentach tekstowych nieznanych wcześniej zależności, reguł, sekwencji czy wzorców. Reprezentacja tak odkrytej wiedzy polega na tworzeniu opisów świata empirycznego lub jego stanów za pomocą metod i technik przetwarzania oraz analizy danych, a zwłaszcza procedur wnioskowania[2]. Techniki Text Mining wydobywają ukrytą w tekstach wiedzę w oparciu o analizę języka naturalnego, który stanowi dla nas strukturę danych. Mimo wspólnego rdzenia metodologicznego Text Mining i Data Mining różnią się właśnie co do pierwotnej struktury zbioru danych. Data Mining przystosowany jest do analizy danych o określonej strukturze, gdzie wartości analizowanych zmiennych wyrażone są na tradycyjnych skalach pomiarowych. Text Mining polega na przetwarzaniu oraz analizie nieustrukturyzowanych lub częściowo ustrukturyzowanych danych tekstowych, np. zapisy wypowiedzi na forach internetowych, wiadomości poczty elektronicznej, artykuły prasowe, odpowiedzi na otwarte pytania ankietowe, opisy dolegliwości podawanych przez pacjentów, komentarze do sesji giełdowych i zdarzeń dotyczących spółek, życiorysy, listy motywacyjne, teksty reklamacji konsumenckich itp. Text Mining zapożycza z metodologii Data Mining: eksploracyjne podejście do procesu analizy danych, algorytmy oraz techniki statystycznej analizy wielowymiarowej, techniki klasyfikacji i grupowania, metody wizualizacji i interpretacji uzyskanych wyników. Nie byłoby jednak drążenia danych tekstowych bez rozwoju metod i narzędzi informatycznych, a w szczególności osiągnięć lingwistyki komputerowej.

[2] W obszarze eksploracji danych zagadnienie reprezentacji wiedzy wiąże się z rozwojem sztucznej inteligencji i jej zastosowaniami w życiu codziennym (Rutkowski 2011).

Metody eksploracji danych tekstowych

W procesie eksploracji dokumentów tekstowych i odkrywania wiedzy w danych tekstowych wykorzystuje się metody:

a) wyszukiwania tekstu (Information Retrieval, IR),
b) ekstrakcji informacji (Information Extraction, IE),
c) przetwarzania języka naturalnego (Natural Language Processing, NLP).

Information Retrieval jest procesem wyszukiwania i lokalizowania w bazie danych tekstowych informacji będących efektem zapytania ze strony użytkownika. Wyszukiwania mogą być oparte na istniejących metadanych określonego dokumentu, jego pełnym tekście lub na podstawie indeksowania treści dokumentów.

Systemy IR nie informują o zawartości treściowej danego dokumentu, lecz o fakcie, że poszukiwana informacja występuje w tym dokumencie[3]. Obecnie wykorzystuje się dwie główne metody indeksowania i wyszukiwania dokumentów, bazujące na algorytmach Boole'a oraz rankingach (Feldman, Sanger 2006). Zgodnie z modelem Boole'a algorytm na podstawie połączonych operatorami logicznymi słów dokonuje podziału zbioru danych tekstowych na dwie części: dopasowaną i niedopasowaną do zapytania wyszukującego[4]. System rankingowy do oceny podobieństwa treści zapytania z treścią dokumentów tekstowych wykorzystuje najczęściej model wektorowy (ang. Vector Space Model; Salton, Wong, Yang 1975) lub probabilistyczny (van Rijsbergen 1979), a następnie określa kolejność dopasowania dokumentów do zapytania wyszukującego[5]. Zaletą Information Retrieval jest niezależność od danego systemu wiedzy czy języka zapytań użytkownika.

Information Extraction to proces identyfikacji i ekstrakcji treści w dokumentach pisanych w języku naturalnym na podstawie wygenerowanych analitycznie lub predefiniowanych wzorców wiedzy. Wzorce te bazują na rozwiązaniach zbliżonych do NLP oraz wykorzystują tzw. dziedzinowe generatory wiedzy (wzorce treści), dla których podstawą jest leksykalna analiza tekstu. Systemy ekstrakcji informacji nie wyszukują dokumentów, ale zawarte w nich treści. W systemach informatycznych

[3] Nie jest naszym celem omawianie w tym artykule systemów wyszukiwania informacji opartych o zbiory słów kluczowych do reprezentacji dokumentów i definiowania zapytań. Mimo istotnych wad i ograniczeń, podejście to jest bardzo popularne i szeroko stosowane w wielu praktycznych systemach wyszukiwania informacji z uwagi na swoją efektywność i prostotę. Zainteresowanych tą tematyką odsyłamy do podstawowej literatury przedmiotowej (Manning, Raghavan, Schutze 2008).

[4] W modelu boolowskim wyszukiwanie treści polega na łączeniu słów za pomocą operatorów logicznych: AND, OR, NOT („i", „lub", „nie"). Nazwa pochodzi od matematyka George'a Boole'a, pioniera logiki matematycznej zwanej niegdyś logiką symboliczną.

[5] Modele wektorowy i probabilistyczny wyszukiwania informacji uwzględniają stopień relewancji dokumentu i zapytania użytkownika w procesie generowania odpowiedzi.

gromadzących dane tekstowe Information Retrieval oraz Information Extraction występują zazwyczaj jako rozwiązania współzależne. W programach CAQDAS metody te znajdują odzwierciedlenie w procedurach wyszukiwania i kodowania treści.

Model przestrzeni wektorowej

Większość systemów wyszukiwania informacji i eksploracji baz danych tekstowych opiera się na prostych technikach dopasowania i zliczania częstości występowania słów (Key Word Search) i fraz (Key Phrase Search) kluczowych, opisujących zbiory dokumentów. Miarą oceny wyszukiwania słów i fraz kluczowych stosowaną w tych systemach są precyzja (ang. precision), tj. odsetek poprawnie wyszukanych dokumentów w odniesieniu do zapytania i zwrot (ang. recall, określany czasami jako kompletność) czy odsetek relewantnych wyszukanych dokumentów[6] (Berry 2004: 162; Feldman, Sanger 2006; Hotho, Nurnberger, Paaß 2005; Manning, Raghavan, Schütze 2008).

$$precision = \frac{|\{Relevant\} \cap \{Retrived\}|}{|\{Retrived\}|} \qquad recal = \frac{|\{Relevant\} \cap \{Retrived\}|}{|\{Relevant\}|}$$

Z punktu widzenia reprezentowania i odkrywania wiedzy zawartej w danych tekstowych podejście to jest niewystarczające. W analizach Text Mining formalnym sposobem reprezentacji dokumentów tekstowych jest model przestrzeni wektorowej (Vector Space Model, VSM; Manning, Raghavan, Schütze 2008). W modelu tym każdy dokument jest reprezentowany przez wektor należący do tak zwanej przestrzeni cech. Wektorowa reprezentacja dokumentów pozwala na wykonywanie matematycznych przekształceń, które uznaje się za odzwierciedlenie operacji na rzeczywistych dokumentach. Model przestrzeni wektorowej umożliwia także wyszukiwanie dokumentów zawierających określone słowa, których wektory są odpowiednio bliskie wektorowi zapytania (wektor zapytania jest przyporządkowany wektorowi określającemu dany dokument). Model ten jest zwykle przedstawiany w postaci macierzy dokumenty – słowa/zbiory słów (tzw. reprezentacja unigramowa, bag of words), gdzie wartością jest liczba wystąpień *i*-tego słowa w *j*-tym dokumencie. Kolejność wystąpień wyrazów w tekście nie jest uwzględniania. Macierz może również odzwierciedlać relację między

[6] Używa się również innych wskaźników oceny poprawności wyszukiwania dokumentów tekstowych. Na przykład (1) prawdopodobieństwo znalezienia dokumentu nierelewantnego wśród wyszukanych dokumentów, fall-out = {{irrelevant}{retrieved}}/{{irrelevant}}, (2) średnia ważona precyzji i zwrotu, F = (1 +) * precision*recall/(*precision + recall), która przyjmuje najczęściej wartości 1, 0,5 oraz 2.

dokumentami a *n*-wyrazowymi ciągami wyrazów (reprezentacja *n*-gramowa) lub dokumentami a pojęciami, ideami czy faktami (reprezentacja pojęciowa), przy czym konieczna jest identyfikacja pojęć w dokumentach[7] (ilustr. 1).

$$X_{ij} = \begin{bmatrix} & T_1 & T_2 & \dots & T_n \\ D_1 & w_{11} & w_{21} & \dots & w_{n1} \\ D_2 & w_{12} & w_{22} & \dots & w_{n2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{1n} & w_{2n} & \dots & w_{nn} \end{bmatrix}$$

Ilustr. 1. Model macierzy dokumenty – słowa/zbiory słów

Objaśnienia: T – wyrazy/*n*-wyrazowe fragmenty tekstu/pojęcia, idee, fakty;
D – dokumenty tekstowe

Źródło: opracowanie własne

W modelu reprezentacji wektorowej przyjmuje się założenie, że dokument tekstowy jest reprezentowany jako wektor częstości występowania słów kluczowych, a zbiór dokumentów można przedstawić za pomocą macierzy (*Term_Frequency_Matrix*), której elementy reprezentują liczbę wystąpień danego słowa kluczowego w danym dokumencie.

Tabela 1. Fragment przykładowej macierzy TFM w programie Wordstat opartej na pomiarze podobieństwa dokumenty/słowa

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
D1	1									
D2	0,011	1								
D3	0,004	0	1							
D4	0,006	0,004	0	1						
D5	0,019	0,027	0,001	0,013	1					
D6	0,026	0,02	0,003	0,019	0,138	1				
D7	0,013	0,008	0,004	0,007	0,015	0,014	1			
D8	0,013	0,005	0,005	0,012	0,014	0,018	0,005	1		
D9	0,081	0,005	0,005	0,009	0,008	0,016	0,006	0,013	1	
D10	0,02	0,011	0,004	0,012	0,018	0,026	0,009	0,014	0,015	…
D11	0	0,003	0	0,001	0,001	0,001	0,002	0,002	0	…

[7] Pojęcie jest wtedy reprezentowane jako struktura złożona semantycznie (lista, drzewo itp.).

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
D12	0,376	0,014	0,004	0,01	0,038	0,053	0,016	0,017	0,105	…
D13	0,012	0,016	0,001	0,008	0,027	0,023	0,013	0,007	0,009	…
D14	0,001	0,001	0,004	0,001	0	0	0	0	0,001	…
D15	0,004	0,003	0	0,003	0,003	0,002	0,004	0,003	0,003	…
D16	0,003	0	0	0	0,001	0,001	0	0,001	0,003	…
D17	0,021	0,006	0,003	0,007	0,011	0,016	0,003	0,008	0,016	…
D18	0,024	0,027	0,002	0,009	0,042	0,054	0,014	0,013	0,012	…
D19	0,028	0,014	0,002	0,008	0,038	0,031	0,03	0,009	0,009	…

Źródło: opracowanie własne.

Każdy element macierzy jest wagą słowa w dokumencie. W najprostszej reprezentacji boolowskiej wagi słów w wektorze dokumentu przyjmują wartości: 0 (nie występuje) lub 1 (występuje). W pełnej reprezentacji wagi te odpowiadają częstości występowania słów w dokumentach. Każdy element wektora oznacza słowo (lub grupę słów) w zbiorze dokumentów, a wielkość wektora jest określona przez liczbę słów (lub grupy wyrazów) kompletnego zbioru dokumentów[8]. Zaletą modelu wektorowego w stosunku do reprezentacji opartej na zbiorze słów kluczowych jest możliwość zdefiniowania miary odległości pomiędzy dokumentami a zapytaniem. Podobne treściowo dokumenty winny cechować się podobną częstością występowania tych samych słów kluczowych. Reprezentacja wektorowa dokumentów pozwala interpretować każdy dokument jako punkt w wielowymiarowej przestrzeni, której wymiary odpowiadają słowom kluczowym. Stąd do oceny odległości pomiędzy dokumentami czy dokumentami a zapytaniem można stosować miary odległości w przestrzeni euklidesowej. W systemach wyszukiwania informacji używa się również specyficznych miar, takich jak: odległość kosinusowa[9] czy miara odległości słów (Manning, Raghavan, Schütze 2008). Jak już wcześniej wspominaliśmy, istotnym elementem reprezentacji wektorowej jest możliwość określania podobieństwa

[8] W praktyce proces wyszukiwania i eksploracji danych tekstowych z wykorzystaniem modelu wektorowego można podzielić na trzy etapy: indeksowanie słów kluczowych w dokumentach, ważenie indeksowanych słów pod kątem wyszukiwania dokumentów i rangowanie dokumentów według przyjętej miary podobieństwa.

[9] Jeśli dwa dokumenty leżą blisko siebie w przestrzeni słów kluczowych, to prawdopodobnie zawierają podobne treści (są do siebie podobne). Odległość kosinusowa dwóch dokumentów d1 i d2 jest zdefiniowana jako znormalizowany iloczyn skalarni wektorów d1 i d2 i reprezentuje kosinus kąta pomiędzy dwoma wektorami reprezentującymi dokumenty. Dwa dokumenty d1 i d2 leżą blisko siebie w przestrzeni wektorowej (dotyczą tej samej problematyki), gdy kosinus kąta między nimi dąży do 1. Jeżeli wartość kosinusa kąta jest bliska 0, oznacza to, że dokumenty są do siebie niepodobne.

dokumentów względem siebie. Wspomniany powyżej sposób konstruowania macierzy dokumentów opartej na prostym ważeniu dokumentów według częstości występowania słów kluczowych „preferuje" dokumenty, w których istnieje większe prawdopodobieństwo wystąpienia danego słowa w zapytaniu. Ich zdolność różnicowania (dyskryminacji) dokumentów tekstowych jest mała. W praktyce dokumenty tekstowe lepiej opisują te słowa, których częstość występowania jest mniejsza. Stąd w modelu macierzy TFM przyjęto schemat nadawania wag dokumentom, który uwzględnia siłę dyskryminacyjną słów kluczowych. Schemat ten nosi nazwę TF-IDF, gdzie TF to waga częstości słów (term frequency), a IDF waga odwrotna częstości dokumentu (inverse document frequency). Waga TF jest liczbą wystąpień słowa w dokumencie, zaś waga IDF logarytmem (dziesiętnym lub naturalnym) ilorazu łącznej liczby dokumentów do liczby dokumentów zawierających dane słowo kluczowe. TF-IDF słowa kluczowego w dokumencie jest iloczynem obu wag. Miara TF-IDF posiada większą moc dyskryminacji dokumentów tekstowych niż klasyczny system wagowy prostej macierzy TFM (Berry 2004; Manning, Raghavan, Schütze 2008).

Z punktu widzenia eksploracji i odkrywania wiedzy w danych tekstowych (Text Mining) modele reprezentacji powinny dążyć do maksymalnego zachowania i odtwarzania zawartości semantycznej dokumentu oraz efektywnego wyszukiwania informacji, zwracając ocenę ich podobieństwa do treści zdefiniowanej w zapytaniu użytkownika. W modelach reprezentacji dokumentów testowych opartych na zbiorze słów kluczowych problem ten dotyczy kwestii synonimiczności (wyrażania tej samej treści za pomocą słów bliskoznacznych) i polisemiczności (występowania różnych znaczeń danego słowa w różnych kontekstach), sposobu definiowania słów kluczowych (liczba pojedyncza czy mnoga) czy odmiany słów w niektórych językach. W modelach reprezentacji wektorowej zagadnienie to wiąże się ze zdolnością grupowania dokumentów tekstowych opartych na miarach dyskryminacji i podobieństwa. Podejścia te jednak nie do końca rozwiązują problem podobieństwa semantycznego dokumentów. Najbardziej obiecujące rozwiązania w tym zakresie oferują techniki przetwarzania języka naturalnego, które próbują wprost modelować i „wydobywać" zawartość semantyczną dokumentów tekstowych.

Przetwarzanie języka naturalnego

Rozwój metod i technik analitycznych Text Mining wiąże się przede wszystkim z możliwością wykorzystania algorytmów służących przetwarzaniu języka naturalnego na strukturę języka formalnego, rozumianego przez komputer. Przetwarzanie języka naturalnego jest dziedziną sztucznej inteligencji zajmującą się automatyzacją procesu analizy, tłumaczenia i generowania informacji w języku

naturalnym[10]. Składa się z teorii gramatyk i języków formalnych oraz reprezentacji wiedzy zawartej w tekstach. W praktyce NLP odnosi się do przetwarzania danych tekstowych oraz rozpoznawania i generowania mowy. Przetwarzanie języka naturalnego znajduje zastosowanie m.in. w takich obszarach życia codziennego, jak: programy i urządzenia przeznaczone dla osób niepełnosprawnych, sterowanie urządzeniami za pomocą głosu, wspomaganie nauki języków obcych, automatyczne tłumaczenie tekstów pomiędzy językami, automatyczne generowanie streszczeń tekstów, robotyka itp. W procesie przetwarzania języków naturalnych systemy i algorytmy informatyczne próbują dokonać „zrozumienia" kontekstu semantycznego analizowanego tekstu. W metodzie tej nie oblicza się podobieństwa słów czy dokumentów względem siebie, ale w analizowanych dokumentach oznacza się poszczególne części mowy (analiza składniowa, gramatyczna) oraz poszukuje się znaczenia danego wyrażenia w kontekście (analiza semantyczna). Pomimo że z jednej strony metody przetwarzania języka naturalnego pozwalają dzięki kontekstualizacji na lepsze dopasowanie i odwzorowanie treści danego zapytania wyszukującego do dokumentów tekstowych, to z drugiej – kontekstualizacja stanowi ich wadę. Ogranicza bowiem transferowalność modelu analitycznego poza system językowy, w którym dokonywana jest analiza zapytań. Nie jest to zadanie niemożliwe, ale bardzo złożone, czasochłonne, wymagające dużego nakładu pracy szczególnie, gdy w grę wchodzi adaptacja określonych rozwiązań analitycznych na inne języki naturalne[11]. Dlatego też podstawą analizy eksploracyjnej dokumentów tekstowych opartej na przetwarzaniu języka naturalnego jest wykorzystywanie istniejących słowników danego języka lub budowanie nowych w oparciu o analizowane teksty i zestawy słów. Logika analizy danych tekstowych wymaga zrozumienia technik w ramach Text Mining oraz podstaw, jakie dla zastosowania tych technik dają procedury NLP. Scharakteryzowanie kluczowych etapów procesu analitycznego, w którym wykorzystujemy TM i NLP, pozwoli nam na pokazanie kolejnych jego etapów, jakie możemy realizować w środowisku CAQDAS.

[10] Język naturalny to język stosowany przez ludzi w codziennej komunikacji interpersonalnej do wytwarzania i przekazywania określonych treści. Język naturalny powstaje poprzez formalne, świadome i ścisłe zdefiniowane wszystkich reguł, jakie nim rządzą. Na język formalny składa się system znaków używanych w procesie poznania rzeczywistości społecznej. Jego przeciwieństwem są języki formalne stanowiące etap pośredni między człowiekiem a maszyną i zapewniające skuteczną komunikację między nimi. Są zapisywane z użyciem przystępnych dla człowieka symboli, liter i wyrazów, a jednocześnie w pełni precyzyjne, co umożliwia ich automatyczne przetwarzanie przez komputer (Feldman, Sanger 2007; Hotho, Nurnberger, Paaß 2005).

[11] Przykład takiego rozwiązania można znaleźć w: Tomanek, Bryda (2014).

Proces eksploracji i odkrywania wiedzy w dokumentach tekstowych wymaga mocy obliczeniowej komputerów[12] oraz przygotowania danych do dalszej analizy. Punktem wyjścia jest wstępna obróbka pliku tekstowego (ang. Text Preprocessing), podczas której dane tekstowe zapisane w różnych formatach są importowane do pojedynczego zbioru, łatwego do późniejszego odczytywania i dokonywania przekształceń. Każdy „surowy" dokument tekstowy, który zostanie poddany analizie danych musi być przekształcony w odpowiednią formę. W tym celu konieczne jest przetworzenie struktury danych tekstowych na taką, która jest bardziej odpowiednia na kolejnych etapach procesu analizy danych. Wiele podejść próbuje w tym zakresie wykorzystać wprost strukturę składniową i semantyczną danego dokumentu tekstowego. W metodologii Text Mining przyjmuje się założenie, że każdy dokument tekstowy jest reprezentowany przez zestaw występujących w nim słów (ang. bag of words), opisujących jego strukturę syntaktyczną i semantyczną. Przetwarzanie dokumentu w oparciu o język naturalny tak, aby możliwa była jego dalsza analiza komputerowa wiąże się z redukcją syntaktyczną dokumentu tekstowego, której celem jest wykluczenie z procesu analizy nieistotnych składników tekstu. Etapy tego procesu przedstawia ilustr. 2.

Ilustr. 2. Etapy obróbki dokumentów tekstowych w ramach Text Mining

Źródło: opracowanie własne

Pierwszym krokiem jest **tokenizacja** dokumentu tekstowego, czyli podział tekstu wejściowego na zdania, słowa, tokeny, czyli znaki interpunkcyjne i nietekstowe (przecinki, kropki itp.[13]). Proces ten jest uzależniony od języka, w jakim

[12] Dane są analizowane w postaci tekstowej, a nie liczbowej, co w przypadku dużych zbiorów dokumentów tekstowych wymaga odpowiedniej mocy obliczeniowej komputerów. Dopiero po przetworzeniu dane są analizowane w formie macierzy liczbowej.

[13] Tokenami mogą być również inne znaczniki tekstu mające status wyrażeń regularnych.

został napisany dany tekst i obszaru tematycznego, do którego się odnosi[14]. Tokenizacja jest zazwyczaj procesem automatycznym zależnym jedynie od języka formalnego, w jakim napisany został program do eksploracji danych tekstowych. W wyniku tokenizacji tworzony jest zbiór słów występujących we wszystkich analizowanych dokumentach tekstowych, znaki interpunkcyjne są usuwane z dalszej analizy, a inne separatory nietekstowe zastępowane tzw. pojedynczymi białymi znakami (white spaces[15]). Zestawy słów (bag of words, BOW) uzyskane w efekcie połączenia wszystkich dokumentów tekstowych „tworzą" słownik klasyfikacyjny, który jest poddawany dalszej obróbce przy rozwijaniu modelu analitycznego. Jednocześnie w celu zmniejszenia rozmiaru słownika, a tym samym wymiarowości opisu zbioru analizowanych dokumentów tekstowych dokonuje się dalszej redukcji zestawu słów i fraz opisujących te dokumenty[16].

Proces tworzenia słownika klasyfikacyjnego i modelu analitycznego wymaga na etapie wstępnym „zubożenia" naturalnego języka tekstu poddawanego procedurom analitycznym Text Mining. Nie jest to rezygnacja z tradycyjnego „jakościowego wglądu w dane", poszukiwania sensu zawartego w analizowanych wypowiedziach czy dokumentach tekstowych, lecz raczej specyficzna redukcja analityczna tekstu (jego struktury syntaktycznej i gramatycznej), celem późniejszej rekonstrukcji zawartej w nim semantyki. Analiza danych tekstowych, zgodnie z logiką i metodologią drążenia danych, ma charakter iteracyjny, stąd w praktyce weryfikacja kontekstu semantycznego danych tekstowych dokonuje się w ciągłym procesie dekonstruowania i rekonstruowania (odkrywania) struktur znaczeniowych poprzez stosowanie różnych procedur analitycznych[17]. Tradycyjne, hermeneutyczne rozumienie w analizie danych jakościowych (analiza semantyczna języka naturalnego dokonywana przez badacza) zostaje wsparte przez

[14] W przypadku języka polskiego przetwarzanie języka naturalnego jest utrudnione ze względu na jego zróżnicowanie i bogactwo form fleksyjnych. Im większa fleksyjność języka naturalnego, tym więcej czasu potrzeba na redukcje występującej w nim odmiany wyrazów i sprowadzenie ich do formy podstawowej, nadającej się do zastosowania w Text Miningu.

[15] Whitespace to język programowania stworzony przez Edwina Bradiego i Chrisa Morrisa, który do zapisu instrukcji wykorzystuje tylko tzw. „białe znaki", czyli spacje, tabulatory i znaki nowej linii, a wszelki tekst jest jedynie komentarzem: http://compsoc.dur.ac.uk/whitespace/.

[16] W praktyce istnieje szereg innych metod redukcji wymiarowości tekstów, jednakże ze względu na zakres tego artykułu scharakteryzowaliśmy te najczęściej wykorzystywane w podejściu Text Mining. Należy jednak pamiętać, że ich przydatność w eksploracji danych tekstowych jest zależna od celu analitycznego i badawczego.

[17] W analizie CAQDAS wspartej algorytmami i technikami Text Mining poszukuje nie samych słów kluczowych czy fraz, lecz ich kontekstu semantycznego, tj. reprezentacji w określonych sekwencjach zdaniowych języka naturalnego. W trakcie analizy Text Mining badacz poszukuje znaczeń słów w analizowanych strukturach języka, a nie tylko słów czy fraz, poszukuje ich semantycznej, a nie statystycznej reprezentacji przy założeniu swoistego izomorfizmu lingwistycznego między językiem naturalnym i formalnym.

technologie informatyczne. Technologie te umożliwiają zapis formalny języka naturalnego, a co za tym idzie – jego analizę syntaktyczno-semantyczną wykorzystującą zaawansowane algorytmy i techniki analityczne. W analizie formalnej języka naturalnego do rozwijania słowników klasyfikacyjnych czy budowy reguł i modeli analitycznych wykorzystywane są przede wszystkim słowa czy frazy, ponieważ to relacje między nimi, tak jak w języku naturalnym, stanowią reprezentację wiedzy zawartej w zbiorze analizowanych tekstów.

Języki naturalne zawierają pod względem gramatycznym szereg wyrazów pomocniczych, które z punktu widzenia analiz Text Mining nie niosą ze sobą istotnych informacji o treści dokumentu tekstowego. Są to najczęściej spójniki lub wyrażenia funkcyjne. Stąd kolejnym krokiem jest ich eliminacja poprzez tworzenie tzw. **stop listy** (ang. stop words) w celu dalszej redukcji syntaktycznej danych tekstowych. W wielu programach komputerowych takie listy są już zaimplementowane. Lista taka nie jest zamknięta i badacz może dodawać do niej kolejne wyrazy, które uznaje za nieistotne z punktu widzenia analizy danych tekstowych. Stop lista jest ściśle powiązana z językiem dokumentów tekstowych. Pomijanie wyrazów funkcyjnych jest prostym sposobem redukcji szumu informacyjnego i poprawy jakości reprezentacji tekstów, stosowanym od dawna w niemal wszystkich aplikacjach z dziedziny pozyskiwania informacji i eksploracji danych tekstowych. Stop listy mogą być tworzone: (a) ręcznie na podstawie wiedzy i doświadczenia, (b) automatycznie na podstawie frekwencji ciągów znaków występujących w bazie tekstów, (c) automatycznie z nadzorowaniem, gdy analityk weryfikuje działanie automatyczne.

Kolejnym etapem przetwarzania tekstu jest **lematyzacja**, czyli analiza morfologiczna słownika i sprowadzenie podobnych form leksykalnych słów do jednej formy podstawowej (lematu, lemma[18]). W języku mówionym różnorodność fleksyjna wyrazów jest niezbędna do zbudowania poprawnego syntaktycznie zdania. Wskazuje również na funkcję danego wyrazu w zdaniu. Jednak z punktu widzenia poprawnej reprezentacji dokumentu tekstowego podobne pod względem leksykalnym słowa zawierają tę samą informację, a więc powinny być rozpoznane jako wystąpienie tego samego leksemu. Proces lematyzacji jest szczególnie ważny dla języków o bogatej fleksji, tj. język polski. W procesie eksploracji dokumentów tekstowych, w celu obliczenia podobieństwa pomiędzy dwoma dokumentami nie jest konieczne znalezienie poprawnej formy leksykalnej wyrazu, wystarczy

[18] Lemma to kanoniczna, najprostsza forma leksemu używana do jego reprezentacji słownikowej. Lemma może być reprezentowana przez jeden wyraz tekstowy. Ma szczególne znaczenie w językach ze złożonym systemem odmiany, np. polskim. Natomiast **leksem** to wyraz, abstrakcyjna jednostka systemu słownikowego języka, na którą składa się znaczenie leksykalne oraz spełniane przez nią funkcje gramatyczne. W informatyce (języki programowania) **leksem** to podstawowa jednostka leksykalna tekstu kodu źródłowego. Odnosi się zarówno do kompilatorów, jak i interpreterów.

jego **rdzeń** (ang. stem), czyli taka jego cząstka, która umożliwi identyfikowanie leksemu. Proces sprowadzania wyrazu do jego rdzenia jest określany jako **stemming**[19]. W praktyce sprowadzanie wyrazu do rdzenia oznacza usuwanie form przedrostkowych, przyrostkowych czy deklinacyjnych charakterystycznych dla danego wyrazu, słowa, przy zachowaniu jego znaczenia. W wyniku stemmingu otrzymujemy rdzeń klasyfikacyjny dla słów zawartych w słowniku. Uwzględnienie kontekstu semantycznego w procesie eksploracji danych tekstowych zwiększa rzetelność analiz Text Mining i poprawia jakość grupowania tekstów. Jak już wcześniej zaznaczaliśmy, nie wszystkie wyrazy występujące w tekście winny się pojawić w wektorowej reprezentacji dokumentu, ponieważ duża część z nich nie zawiera żadnych istotnych informacji. Stąd oprócz wspomnianych powyżej etapów przygotowania danych tekstowych do analizy eksploracyjnej, w zależności od celu badawczego, wykorzystuje się również metodę selekcji wyrazów ze względu na określone części mowy. Inne znaczenie w tekście mają bowiem rzeczowniki, czasowniki czy przymiotniki. Przykładowo rzeczowniki opisują określone obiekty, zdarzenia, fakty, a także wskazują na tematykę. Natomiast przymiotniki wskazują na cechy i właściwości rzeczowników. Podobnie można pokusić się o redukcję danych tekstowych ze względu na występującą w dokumentach sekwencję wyrazów odnoszących się do pojedynczych bytów, takich jak osoby, instytucje czy organizacje. Sekwencje wyrazów ułatwiają bowiem poprawną reprezentację słów, szczególnie gdy ich znaczenie jest wąskie. Nieuwzględnienie w analizie tego typu sytuacji może prowadzić do błędnej interpretacji semantycznej dokumentów tekstowych. Aby jeszcze bardziej zmniejszyć liczbę wyrazów w słowniku, można wykorzystać algorytmy indeksowania lub wyszukiwania słów kluczowych. Prostym sposobem weryfikacji słów kluczowych jest wykorzystanie ich entropii[20]. Odzwierciedla ona zakres dopasowania danego słowa kluczowego do różnych dokumentów tekstowych. Jeśli słowa występują w wielu dokumentach, to wskaźnik entropii będzie niski. Entropia może być postrzegana jako wskaźnik znaczenia (wagi) słowa w danej dziedzinie lub kontekście analizy

[19] Stemming to sprowadzenie grupy wyrazów do ich wspólnego rdzenia, postaci podstawowej, umożliwiającej traktowanie ich wszystkich jak to samo słowo. Przykład stemmingu, tej procedury analitycznej, można znaleźć w artykule: Tomanek, Bryda (2014). Klasycznym przykładem tej procedury jest algorytm Portera, który znajduje rdzenie dla słów angielskich (Porter 1980). Tego typu algorytmy tworzone są zwykle w oparciu o reguły specjalnie skonstruowane dla konkretnego języka.

[20] Entropia to miara w teorii informacji wyrażająca średnią liczbę informacji, jaka przypada na pojedynczą wiadomość. Można ją interpretować jako niepewność wystąpienia danego zdarzenia elementarnego w następnej chwili. Jeżeli zdarzenie występuje z prawdopodobieństwem równym 1, to jego entropia wynosi 0, gdyż z góry wiadomo, co się stanie – nie ma niepewności. W analizach Text Mining entropia kładzie nacisk na rzadkie słowa, występujące tylko w kilku dokumentach z całego zbioru. Otrzymują one największą wagę w zbiorze dokumentów.

danych tekstowych. Oprócz wspomnianych powyżej metod przygotowania danych tekstowych do Text Mining, w celu lepszej reprezentacji wiedzy w dokumentach tekstowych, stosuje się rozwiązania z zakresu lingwistycznej obróbki danych (ang. Linguistic Preprocessing). Do rozwiązań tych należą m.in.:

1. **Part-of-speech tagging** (POS) – tagowanie części mowy według rodzaju danego języka,
2. **Text chunking** – podział tekstu na mniejsze jednostki analityczne w celu późniejszego grupowania sąsiadujących słów, wyrazów w zdaniu tak, aby łatwiej było znaleźć frazy,
3. **Word Sense Disambiguation** – kategoryzacja semantyczna celem ustalenia znaczenia pojedynczych słów czy fraz w dokumentach, pod kątem ich lepszej reprezentacji w przestrzeni wektorowej,
4. **Parsing** – analiza składniowa dokumentu tekstowego, jego struktury gramatycznej i zgodności z gramatyką języka naturalnego[21].

W analizie Text Mining w przetwarzaniu języka naturalnego wykorzystuje się w równym stopniu metody formalne i statystyczne. Metody formalne służą do opisu języka w wymiarze: fonologicznym (rozpoznawanie i generowanie mowy), leksykalnym (tokenizacja tekstu, identyfikacja części mowy, tagowanie słów), morfologicznym (rozpoznawanie sufiksów, prefiksów i form fleksyjnych, analiza wyrażeń, stemming/lematyzacja itp.), syntaktycznym (analiza gramatyczna zdań, poszukiwanie schematów lingwistycznych w danych tekstowych), semantycznym (reprezentacja wiedzy, spójność semantyczna słów, wzbogacanie reprezentacji wiedzy przez synonimy, homonimy), pragmatycznym (weryfikacja sensów, interpretowanie intencji, analiza wyrażeń metaforycznych, przekształcanie informacji w wiedzę) i dyskursywnym (analiza kontekstu narracyjnego tekstu, wypowiedzi). Natomiast metody statystyczne koncentrują się na wyszukiwaniu regularności cechujących dane teksty i języki naturalne. Odnoszą się one – podobnie jak w klasycznej analizie treści – do ilościowego podejścia do jakościowej analizy tekstu. W metodologii Text Mining do metod tych zalicza się przedstawioną w poprzednim rozdziale wektorową reprezentację dokumentów oraz analizę ukrytych grup semantycznych dokumentów. Znaczenie technologii opartych na przetwarzaniu języka naturalnego wynika z tego, że uwalniają one użytkownika od problemów ekstrakcji wiedzy i interpretacji istotnej informacji znajdującej się w tekstach pisanych w języku naturalnym. Dzięki wsparciu informatycznemu reguły przetwarzania i analizy języka naturalnego pozwalają na:

[21] W kontekście analizy Text Mining zdanie jako jednostka semantyczna w procesie komunikacji wymaga ustalenia jego struktury gramatycznej i zgodności z gramatyką danego języka naturalnego. Większość współczesnych parserów jest przynajmniej częściowo oparta na analizie statystycznej korpusu języka, co pozwala na zgromadzenie informacji o częstości występowania poszczególnych wyrazów i fraz w różnych kontekstach.

a) rozumienie i badanie struktury języków,
b) trafną kategoryzację dokumentów tekstowych,
c) tworzenie słowników,
d) automatyczne generowanie wypowiedzi, streszczeń,
e) wyszukiwanie dokumentów, fragmentów tekstu,
f) grupowanie i klasyfikację dokumentów,
g) automatyczne przetwarzanie treści dokumentów WWW,
h) odkrywanie nowych elementów ontologii (pojęć, klas, atrybutów, relacji, twierdzeń),
i) automatyczne wyszukiwanie elementów wiedzy, reprezentację i odkrywanie wiedzy zawartej w danych tekstowych.

CAQDAS, Text Mining i odkrywanie wiedzy w danych jakościowych

Obecnie eksploracja danych tekstowych to przede wszystkim wyszukiwanie podobnych dokumentów tekstowych na podstawie zapytania lub wyszukiwanie podobnych dokumentów w oparciu o przykładowe dokumenty, a także klasyfikacja, grupowanie, kategoryzacja dokumentów, rankingi ważności dokumentów czy analiza zależności pomiędzy dokumentami (np. analiza cytowań, plagiatów). Na gruncie socjologii jakościowej zarówno eksploracja, jak i zagadnienie odkrywania wiedzy w danych tekstowych nie ma utrwalonej tradycji. W pewnym sensie problematyka ta jest obecna w metodologii teorii ugruntowanej oraz analizy treści. Jak pisze Krzysztof Konecki: „metodologia [teorii ugruntowanej] poprzez swoją elastyczność umożliwia zatem utrzymanie w trakcie badań tzw. 'kontekstu odkrycia' (serendipity), tj. dzięki jej procedurom posiadamy zdolność poszukiwania i odkrywania zjawisk, których na początku badań nie szukaliśmy" (Konecki 2000; Glaser, Strauss 2009). Serendipity jest umiejętnością, której można się nauczyć w praktyce, stosując określone techniki badawcze i procedury analizy danych. Jest to immanentna cecha teorii ugruntowanej, której klarowne procedury pozwalają na procesualne odkrywanie struktury zjawisk społecznych. Nie oznacza to jednak, że przy użyciu innych metod nie jest to możliwe. Serendipity może odnosić się zarówno do implementowanej coraz częściej w programach CAQDAS metodologii mixed methods, jak i dowolnej metodologii drążenia danych tekstowych (Data Mining, Text Mining czy Web Mining). Kontekst odkrycia może być również konsekwencją rozumowania dedukcyjnego (wnioskowanie logiczne, od ogółu do szczegółu), indukcyjnego (wnioskowanie z prawdziwości racji, od szczegółu do ogółu), abdukcyjnego (wnioskowanie o prawdopodobnych przyczynach na podstawie znajomości skutku, wyjaśnianie tego, co wiadome, odwrotność dedukcji) czy nawet w szczególnych warunkach

heurystycznego (wnioskowanie bez ścisłych reguł, na podstawie skojarzeń lub analogii z czymś znanym uprzednio). W związku z tym pojawia się pytanie, czy ideę serendipity znaną z metodologii badań jakościowych można odnieść bezpośrednio do analiz Text Mining, łącząc ją z odkrywaniem wiedzy w danych tekstowych? Jak w związku z tym rozumieć zagadnienie odkrywania wiedzy w danych, a w szczególności danych jakościowych?

W tradycyjnym rozumieniu serendipity odnosi się do sekwencji działań badacza-analityka, w trakcie których „odkrywa" on, niejako przez przypadek, zależności, prawidłowości czy własności badanego zjawiska lub procesu społecznego. Odkrywanie wiedzy dokonuje się w sposób naturalny, poza badaczem, ale i także przy współudziale badacza, w zgodzie z logiką i procedurami danego paradygmatu metodologicznego. Stąd serendipity w socjologii jakościowej jawi się jako naturalny kontekst odkrywania wiedzy w procesie badań terenowych i analizy danych wspomaganej lub nie oprogramowaniem CAQDAS[22]. W tym sensie serendipity jest wypadkową wyobraźni socjologicznej, wiedzy, doświadczenia, warsztatu terenowego i umiejętności analitycznych badacza. W odniesieniu do analiz Text Mining odkrywanie wiedzy ma raczej charakter czysto analityczny i wiąże się z eksperymentowaniem z posiadanymi danymi. Kontekst odkrycia jest raczej konstruowany eksperymentalnie w procesie eksploracji danych, przy wykorzystaniu różnych mniej lub bardziej zaawansowanych technik czy algorytmów analitycznych. W tym sensie odkrywanie wiedzy w danych jakościowych (Text Mining, drążenie danych jakościowych) to interaktywny i iteracyjny proces poszukiwania nowych (nieoczekiwanych) konfiguracji i regularności tkwiących w danych tekstowych. Jego celem jest przechodzenie z poziomu „surowych" danych terenowych do generowania wzorców, prawidłowości, które stanowią fundament dalszej analizy i rozwijania wiedzy teoretycznej. Nie ma specjalnie znaczenia, jakie to są dane: wywiady indywidualne, grupowe, dokumenty pisane, nagrania audio, pamiętniki, życiorysy, dzienniki itp. Ważne jest, aby dane surowe zostały przetworzone i zapisane w odpowiednich formatach jako dokumenty tekstowe, które dają się później przekształcać zgodnie z procedurami analitycznymi Text Mining czy przetwarzania języka naturalnego.

Z metodologicznego punktu widzenia należy rozróżnić eksplorację i odkrywanie wiedzy w danych jakościowych. Celem eksploracji, najogólniej mówiąc, jest analiza danych tekstowych wykorzystująca techniki Text Mining i algorytmy NLP dla lepszego zrozumienia sieci relacji ukrytych w tych danych. Automatyczna eksploracja danych jakościowych korzysta z technik Text Mining i algorytmów

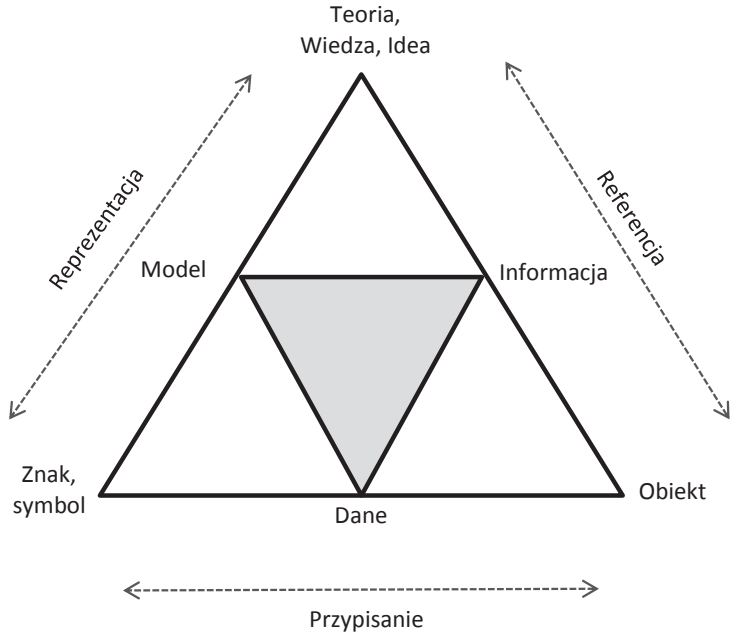
[22] W socjologii jakościowej, antropologii społecznej czy etnografii odkrycia mają charakter rzeczowy i prowadzą przede wszystkim do głębszego i lepszego rozumienia badanych zjawisk.

przetwarzania języka naturalnego. Dzięki temu otwiera nowe możliwości w zakresie interakcji badacza-analityka z danymi tekstowymi. Między innymi daje możliwość budowania różnorodnych modeli reprezentacji wiedzy. Dlatego modele eksploracji danych powinny być przejrzyste tak, aby były w stanie opisywać wzorce, które można intuicyjne zinterpretować i wyjaśnić. Istotą procesu eksploracji jest dobór odpowiednich technik czy algorytmów analitycznych, uwzględniających kontekst znaczeniowy danych jakościowych (tekstowych). Przykładowo algorytmy reguł indukcyjnych plus przetwarzanie języka naturalnego obejmujące formalizację tekstu (przedstawienie wypowiedzi stworzonej w języku naturalnym w języku formalnym) oraz implementacja jej wyniku w programach komputerowych odzwierciedlają eksperymentalny kontekst odkrywania wiedzy w danych tekstowych. Nie pozwalają jednak na zbyt duży poziom uniwersalizacji tej wiedzy, ponieważ NLP z natury ma charakter dziedzinowy, określony przez kontekst semantyczny i strukturę lingwistyczną danego języka naturalnego[23]. Stąd niezwykle ważny jest dobór metody eksploracji do celu analizy i rodzaju danych jakościowych (artykuły prasowe, blogi, wywiady indywidualne, wywiady grupowe itp.), uwzględniający ich kontekst znaczeniowy. Eksploracja danych tekstowych obejmuje m.in. następujące rodzaje zadań analitycznych: opis i charakterystyka danych, odkrywanie reguł asocjacyjnych, klasyfikację, grupowanie (analiza skupień, *k*-średnich, dwustopniowe grupowanie), predykcję, analizy statystyczne (regresja, dyskryminacja), odkrywanie wzorców sekwencji, poszukiwanie odchyleń, anomalii, tradycyjne wyszukiwanie oraz ekstrakcję treści dokumentów tekstowych i WWW itp.[24] Odkrywanie wiedzy w danych jakościowych ma charakter ogólniejszy i odnosi się do całego procesu transformacji surowych danych tekstowych we wzorce czy reguły semantyczne. Wiąże się nie tylko z rozumieniem języka danych i treści dokumentów, lecz także z umiejętnością ich wielowymiarowej analizy, syntezy wiedzy, nadawania sensu czy interpretacji. Dane tekstowe są jakościowymi reprezentacjami obiektów empirycznych świata społeczno-kulturowego: wypowiedzi, faktów, zdarzeń itp. Przez odniesienie do tego, co reprezentują, zawierają znaczenie, a więc są nośnikami określonych informacji. Treści informacyjne organizowane są w procesie analizy i interpretacji w struktury wiedzy. Interpretacja danych wymaga wiedzy o opisywanym świecie

[23] Próbka takich reguł jest zawarta w artykule: Tomanek, Bryda (2014). Budując reguły, posługujemy się procedurami i założeniami przetwarzania języka naturalnego. Na przykładzie programu QDA Miner pokazujemy, jak zbudować reguły słownikowe, które w ramach języka polskiego będą dawały dość trafne wyniki analityczne. Budując reguły, posługujemy się procedurami i założeniami przetwarzania języka naturalnego.

[24] Wiedza może przyjmować wiele postaci: wartości miar statystycznych, opisy charakterystyczne/dyskryminujące, reguły asocjacyjne, drzewa i reguły klasyfikacyjne, funkcje i równania, klauzule logiczne, skupienia i ich opis, taksonomia (hierarchia), trendy i zależności czasowe.

i o języku, w którym dane są zapisane. Wiedza pełni aktywną rolę w procesie interpretacji danych, nadawania im znaczenia (sensu). Relacje między danymi, informacją i wiedzą można opisać podobnie, jak relacje między obiektem, symbolem i ideą w trójkącie semiotycznym Ogdena i Richardsa (1923). Z kolei dane, informacja i wiedza to trzy wierzchołki trójkąta określanego jako trójkąt epistemiczny, odzwierciadlający charakter reprezentowania i odkrywania wiedzy w danych tekstowych. Jeśli nałożymy na siebie te trójkąty, odwracając jednocześnie podstawę trójkąta epistemicznego, tak by była ona bliższa szczytowi trójkąta semiotycznego, otrzymamy diagram przedstawiający relacje pomiędzy strukturami poznawczymi (wiedza) i strukturami językowymi (lingwistyka, NLP), występującymi w analizach Text Mining. Relacje te przedstawia ilustr. 3. W konsekwencji każda odkrywana struktura epistemiczna w procesie analizy danych tekstowych czerpie ze struktury semiotycznej języka naturalnego. Rozróżnienie danych, informacji i wiedzy jest więc nie tylko istotne dla określenia relacji między nimi, lecz przede wszystkim dla zrozumienia roli wspomaganej komputerowo analizy danych (CAQDAS), analiz typu Text Mining czy przetwarzania języka naturalnego w procesie odkrywania wiedzy w danych jakościowych (ilustr. 3).



Ilustr. 3. Relacje między strukturami epistemicznymi (poznawczymi) i semiotycznymi w procesie odkrywania wiedzy w danych tekstowych

Źródło: opracowanie własne na podstawie idei trójkąta semiotycznego

Wykorzystywanie wiedzy o składni języka, semantyce i logice powiązań pomiędzy elementami wypowiedzi to atrakcyjny obszar w eksploracji oraz analizie danych tekstowych w programach CAQDAS. W programach tych istnieje wiele algorytmów czy technik analitycznych wydobywania informacji z danych tekstowych. Każda metoda analizy tekstu ma jednak swoje mocne i słabe strony. Większość narzędzi CAQDAS istniejących obecnie na rynku opiera się zwykle na jednym podejściu do analizy tekstu, u podstaw którego zazwyczaj stoi określony paradygmat metodologiczny (metody mieszane, teoria ugruntowana, analiza treści, analiza dyskursu itp.). Jednocześnie rozwój CAQDAS to implementacja nowych funkcjonalności, algorytmów czy technik analitycznych, np. automatyczne uczenie wzorców kodowania danych z wykorzystaniem języka NLP (Qualrus) czy techniki Text Mining (QDA Miner, Wordstat). Nie jest zaskoczeniem korzystanie w analizie danych jakościowych z tabel kontyngencji przy budowaniu typologii czy klasyfikacji, ale implementowanie w programach CAQDAS technik statystycznych czy algorytmów Text Mining może być dla niektórych badaczy zaprzeczeniem istoty badań jakościowych. Wykorzystywanie w analizie danych jakościowych technik statystycznych, np. analizy korespondencji, analizy skupień, skalowania wielowymiarowego, regresji czy reguł indukcyjnych wzbogaca proces eksploracji i odkrywania wiedzy, ukazując nowe obszary rozwoju socjologii jakościowej. Szybki rozwój technologii informacyjnych i dostępność dużych wolumenów danych tekstowych i webowych powoduje również, że wiele programów CAQDAS podąża obecnie w kierunku analiz online oraz eksploracyjnej metodologii drążenia danych tekstowych – Text Mining (Wiedemann 2013). Niezależnie od tego, jakie funkcjonalności są obecne w programach CAQDAS, badacz jakościowy powinien być na każdym etapie procesu badawczego krytyczny co do efektów stosowania nowych technologii czy funkcjonalności i mieć cały czas na uwadze fakt, że programy CAQDAS są tylko narzędziem w procesie analizy danych. A coraz większa ich uniwersalizacja pod kątem funkcjonalności pozwala na ich szerokie zastosowania, niezależnie od paradygmatu metodologicznego. Rozpatrując proces odkrywania wiedzy w danych jakościowych, można odnieść go do relacji między światem faktów empirycznych i konstrukcji teoretycznych. W relacji tej oprogramowanie komputerowe, NLP, procesy eksploracji danych, metody rozumowania naukowego czy algorytmy i techniki analityczne pełnią rolę pośredniczącą między tymi światami. Strukturę tych relacji przedstawia ilustr. 4.

W socjologii jakościowej, ze względu na wielość dostępnych danych tekstowych oraz w zasadzie nieograniczone możliwości ich gromadzenia, istnieje pole do wykorzystywania zarówno technik Text Mining, jak i podejścia CAQDAS w procesie odkrywania wiedzy. Oba rozwiązania opierają się zarówno na eksploracji i analizie nieustrukturyzowanych danych tekstowych, jak i wnioskowaniu „ugruntowanym" w tychże danych. Text Mining jest definiowany jako proces wydobywania

Ilustr. 4. Odkrywanie wiedzy w danych a relacja między światem empirycznym i teoretycznym

Źródło: opracowanie własne

informacji w zbiorach dokumentów poprzez identyfikację i poszukiwanie wzorów, regularności, struktur relacji w danych tekstowych (Feldman, Sanger 2006). Podobnie jest we wspomaganej komputerowo analizie danych jakościowych. Dominujący w programach CAQDAS paradygmat teorii ugruntowanej wymaga eksplorowania czy odkrywania wiedzy w danych jakościowych z otwartym umysłem, a identyfikowanie kategorii, koncepcji i konstrukcji, które wyjaśniają określone procesy społeczne nie powinno być w żaden sposób narzucone z zewnątrz (Glaser 1978, 1992; Glaser, Strauss 2009). Badacz jakościowy, podobnie jak analityk Data czy Text Mining w procesie drążenia danych powinien pozwolić, by kategorie czy konstrukcje analityczne „wyłoniły się" z analizowanych danych w procesie ciągłego porównywania zakodowanych treści, fragmentów czy dokumentów. Możliwość wykorzystania algorytmów drążenia danych tekstowych we wspomaganej komputerowo analizie danych jakościowych oraz automatyzacja wstępnego kodowania danych powstrzymuje badacza od narzucania jakiejkolwiek struktury, poza tą, która jest zawarta w samych danych tekstowych. Nie oznacza to oczywiście, że automatyczne tagowanie czy procedury analityczne wspomagające preprocessing danych tekstowych są lepsze od tradycyjnych procedur kodowania stosowanych przez badaczy związanych z paradygmatem teorii ugruntowanej. Text Mining, tak jak go opisujemy w tym artykule, stanowi rozwinięcie i uzupełnienie tradycyjnych metod analizy (Qualitative Analysis, Qualitative Content Analysis), a także całego procesu odkrywania wiedzy w teorii ugruntowanej. Pod względem eksploracji danych tekstowych Text Mining oraz jego techniki i algorytmy są bliższe logice i technikom analizy treści, w której w procesie kodowania z zastosowaniem klucza kategoryzacyjnego „wyłaniają się" kategorie analityczne. W Text Mining, podobnie jak w analizie treści, częstym efektem analizy są słowniki analityczne, zbiory słów kluczowych, wykorzystywane do klasyfikacji analizowanych dokumentów tekstowych. Analiza treści i Text Mining wykorzystują algorytmy

komputerowe dla zliczania słów kluczowych, stop listy itp., ale drążenie danych tekstowych idzie dalej, w kierunku odkrywania kontekstów znaczeniowych słów kluczowych czy wypowiedzi za pomocą przetwarzania języka naturalnego.

CAQDAS a Text Mining: podejścia alternatywne czy komplementarne?

Text Mining jako zbiór technik i algorytmów wspomagających procesy ekstrakcji informacji i indukcyjnego[25] poszukiwania wzorów i współzależności w zbiorach danych (odkrywania i reprezentacji wiedzy) stanowi kwintesencję podejścia eksploracyjnego w analizie danych tekstowych, powszechnie znanego jako Text Analytics. Jeżeli jednak tradycyjne analizy tekstu z wykorzystaniem programów CAQDAS, podobnie jak Text Mining, pozwalają na kompleksową eksplorację danych jakościowych w procesie odkrywania wiedzy, to pojawia się pytanie o ich wzajemne relacje na gruncie metodologii oraz analizy danych jakościowych. Czy podejścia te są względem siebie konkurencyjne, czy raczej komplementarne? Czemu Text Mining może służyć w analizie danych jakościowych? Co oferują procedury Text Mining? Jakie problemy Text Mining rozwiązuje?

W sensie metodologicznym drążenie danych dotyczy zarówno danych wywołanych, jak i danych zastanych[26]. W sensie analitycznym wymaga wiedzy oraz umiejętności integracji danych jakościowych i ilościowych, a także ich kompleksowej analizy. Dzięki rozwojowi informatyki, zaawansowanych algorytmów statystycznych, sztucznej inteligencji czy metod uczenia maszynowego Text Mining wzbogaca także schematy tradycyjnej eksploracyjnej analizy danych jakościowych, które są efektem rejestracji wypowiedzi, zdarzeń lub działań aktorów społecznych. Takie podejście umożliwia nie tylko kompleksowe zrozumienie zjawisk i procesów społecznych w socjologii jakościowej, lecz także w oparciu o odkryte wzory oraz regularności tworzenie analitycznych modeli klasyfikacyjnych lub predykcyjnych. W analizach Text Mining wykorzystuje się oprogramowanie

[25] Rozumowanie indukcyjne – w szerszym znaczeniu – polega na dokonywaniu obserwacji i eksperymentów, wyprowadzaniu na tej podstawie uogólnień oraz formułowaniu hipotez i ich weryfikacji. Zasada indukcji jest regułą pozwalającą na przejście od przypadków zaobserwowanych do twierdzeń ogólnych obejmujących także przypadki niezaobserwowane. W przypadku Text Mining indukcyjny charakter wnioskowania oznacza poszukiwanie relacji w danych oparte na algorytmach statystycznych i sztucznej inteligencji, a nie klasycznych kanonach indukcji Milla.

[26] W naukach społecznych, w praktyce badawczej nie wszystkie zjawiska możliwe są do uchwycenia przy pomocy rejestracji zachowań. W obszarze socjologii opis i analiza zjawisk wymaga wykorzystania nie tylko danych zastanych, lecz także danych gromadzonych w toku badań empirycznych w odniesieniu do określonego problemu. Dane generowane w toku takich badań są efektem konceptualizacji i operacjonalizacji, w wyniku której kształtują się: podejście do problemu, wymiary analizy, pytania i hipotezy badawcze.

komputerowe, techniki i algorytmy analityczne do znajdowania ukrytych dla człowieka prawidłowości zawartych w strukturze danych tekstowych, ze względu na jego ograniczone możliwości percepcyjne i czasowe. Wspomagana komputerowo analiza danych jakościowych wykorzystująca techniki Text Mining czy przetwarzanie języka naturalnego to swoiste *novum* na gruncie socjologii jakościowej. Zastosowanie Text Mining w obszarze CAQDAS podnosi zarówno wiarygodność wyników analizy danych jakościowych, jak i rangę badań jakościowych w socjologii i naukach społecznych. CAQDAS zyskuje bardziej wszechstronny charakter i ogromne możliwości analizy danych tekstowych w wymiarze lingwistycznym, syntaktycznym, semantycznym czy pragmatycznym. Na płaszczyźnie metodologicznej i analitycznej Text Mining i CAQDAS nie są aż tak odmienne, jakby się pierwotnie mogło wydawać. Różnice sprowadzają się głównie do wymiaru automatyzacji procesu eksploracji danych. Procedury analityczne w zakresie odkrywania wiedzy są podobne. Przegląd podstawowych różnic między CAQDAS a Text Mining przedstawia tabela 2.

W praktyce różnica między Text Mining a CAQDAS jest zauważalna w zakresie liczby przetwarzanych danych tekstowych. Jeśli oprogramowanie CAQDAS wykorzystuje się w pracy raczej z mniejszymi zbiorami danych jakościowych, to Text Mining pozwala na prowadzenie analiz, w których wielkość zbioru danych jest w zasadzie nieograniczona. Algorytmy Text Mining pozwalają na przeglądanie i analizę informacji, których liczba jest wręcz niewyobrażalna dla badacza jakościowego oraz wykonywanie obliczeń i analiz w niebywale krótkim czasie. Możliwości analizowania i rozumienia dużych wolumenów danych tekstowych są mniejsze ze względu na ich wielowymiarowość. Techniki i algorytmy analityczne Text Mining uzupełniają oraz wzbogacają nie tylko tradycyjne funkcjonalności oprogramowania CAQDAS, ale i nasze zdolności analityczne. W odróżnieniu od tradycyjnie stosowanego najczęściej w socjologii podejścia *a aprioryczne*go techniki CAQDAS i Text Mining odnoszą się do odkrywania relacji między zmiennymi w sytuacji, gdy nie ma określonych z góry oczekiwań ani założeń odnośnie natury tychże relacji. Hipotezy są generowane *a posteriori* z danych niż stawiane *a priori*. W typowym procesie eksploracyjnej analizy danych bierze się pod uwagę i porównuje wiele zmiennych, w wielu różnych kombinacjach i konfiguracjach, w poszukiwaniu istotnych zależności między nimi. Zależności te reprezentują modele drążenia danych tekstowych, budowane w oparciu o zaawansowane metody i techniki analityczne. Model analityczny powstaje jako efekt konfiguracji danych i zmiennych, niezależnie od wielkości zbioru czy zbiorów danych i liczby zmiennych. Text Mining, przy konstruowaniu takich modeli, bazuje na rozumowaniu indukcyjno-dedukcyjnym oraz indukcyjno-abdukcyjnym w obszarze skończonego zbioru dokumentów tekstowych. Modele reprezentują strukturę asocjacji empirycznych, które następnie się testuje i interpretuje. W CAQDAS zbiór

danych jakościowych jest także skończony, a wnioskowanie najczęściej opiera się na rozumowaniu indukcyjno-dedukcyjnym lub indukcyjno-abdukcyjnym. Jego efektem są zazwyczaj mapy kognitywne ukazujące relacje między dokumentami, kodami, kategoriami czy konceptami. Jeśliby przyjąć za Ann Lewins definicję CAQDAS odnoszącą się do tzw. jakościowej analizy danych jakościowych (Lewins, Silver 2007), to należałoby raczej wykluczyć zastosowanie technik i algorytmów Text Mining w obszarze wspomaganej komputerowo analizy danych jakościowych. Porównanie logiki analizy Text Mining i CAQDAS pokazuje, że na gruncie analitycznym i metodologicznym są one raczej epistemologicznie kompatybilne niż konkurencyjne. Podobnie jak w wielu jakościowych podejściach metodologicznych do procesu analizy danych jakościowych Text Mining „zachęca" badacza do otwartości w konstruowaniu modeli opisujących struktury relacji w danych. Innymi słowy można powiedzieć, że techniki i algorytmy drążenia danych tekstowych doskonale wkomponowują się w różne etapy wspomaganej komputerowo analizy danych jakościowych w procesie odkrywania wiedzy[27].

Tabela 2. Podstawowe różnice między metodologią CAQDAS a Text Mining

Wymiar	CAQDAS	Techniki Text Mining
1	2	3
Dominująca tradycja metodologiczna	Teoria ugruntowana	Analiza treści
Podejście analityczne	Analiza jakościowa danych tekstowych	Analiza jakościowa i ilościowa danych tekstowych
Procedury analizy danych tekstowych	Kodowanie tekstu	Tagowanie tekstu
Liczba dokumentów w zbiorze danych	Ograniczona	Nieograniczona
Integracja danych liczbowych i tekstowych	Zależna od metodologii	Niezależna od metodologii
Przetwarzanie języka naturalnego	Niestosowane	Automatyczne, półautomatyczne (nadzorowane)
Wzorce kodowania danych tekstowych	Ręczne (kontekstowe)/ semiautomatyczne	Uczenie reguł (automatyzacja kodowania)
Kodowanie, tagowanie	Nadawanie znaczeń oparte na regułach: semantycznych, pragmatycznych	Odkrywanie znaczeń w oparciu o reguły syntaktyczne/ semantyczne
Rola badacza-analityka w eksploracji danych	Konstruktor/interpretator	Rekonstruktor/interpretator

[27] Tym bardziej, że Text Mining w praktyce analitycznej można stosować do różnych danych tekstowych.

Tab. 2 (cd.)

1	2	3
Algorytmy i techniki analityczne	Typologie, analiza podobieństwa	Statystyka i uczenie maszynowe, klasyfikacja, grupowanie, reguły indukcyjne
Proces odkrywania wiedzy	Metoda ciągłego porównywania	Automatyczne generowanie wzorców
Analiza słownikowa danych tekstowych	Analiza słów kluczowych	Tezaurusy/słowniki analityczne
Walidacja procesu analizy danych	Ręczna	Automatyczna, nadzorowana i półautomatyczna

Źródło: opracowanie własne.

Podsumowanie

CAQDAS, Text Mining i nowoczesne technologie informatyczne pozwalają na rozwiązania metodologiczne, które automatyzują i wzbogacają analizę danych jakościowych. Jednakże w przeciwieństwie do popularnego przekonania analizy Text Mining nie jest zautomatyzowanym, pozbawionym refleksyjności działaniem. Podobnie jak w przypadku CAQDAS jest to proces iteracyjny, wymagający świadomego podejścia ze strony badacza do analizy[28]. W praktyce Text Mining to najczęściej metody półautomatyczne (nadzorowane), wymagające wiedzy i znajomości technik analitycznych.

Praca z programami CAQDAS uczy badacza rygoryzmu metodologicznego, przestrzegania procedur, dokładności i precyzji w procesie analizy danych jakościowych, a Text Mining otwiera na nowe obszary wiedzy, interdyscyplinarność i wymaga dodatkowych umiejętności analitycznych, co pozytywnie odbija się na jakości prowadzonych analiz i badań terenowych. Wsparcie wspomaganej komputerowo analizy danych jakościowych o zaawansowane techniki i algorytmy analityczne Text Mining powoduje, że w programach CAQDAS przecinają się różne paradygmaty metodologiczne: teoria ugruntowana – analiza treści czy mixed methods. Pod względem metodologicznym Text Mining jest w pewnym sensie odbiciem logiki teorii ugruntowanej. Pod kątem analitycznym jest podobny do analizy treści. Jednak chociaż oba podejścia wykorzystują algorytmy komputerowe w analizie danych tekstowych, to Text Mining idzie dalej. Charakteryzuje się unikalną zdolnością przetwarzania języka naturalnego oraz wykorzystywania w procesie analizy wiedzy zawartej w słownikach przedmiotowych

[28] Przykłady takiego podejścia znajdują się w artykule: Tomanek, Bryda (2014).

i tematycznych. Dzięki temu zastosowanie technik Text Mining w dziedzinie socjologii jakościowej i wspomaganej komputerowo analizie danych jakościowych prowadzi do pogłębiania wiedzy o mechanizmach działań i procesów społecznych. Sprzyja również integracji danych z wielu różnych źródeł, danych zastanych i danych pochodzących z terenowych badań jakościowych. Integracja danych prowadzi do systematycznego rozwoju i integracji wiedzy socjologicznej, a także poprawia jakość analiz i badań jakościowych. Dla badacza społecznego, a szczególnie badacza jakościowego, niezwykle ważne jest podejście od strony danych, odkrywania wiedzy z danych, budowanie wielowymiarowych modeli analitycznych, mechanizmów i działań społecznych, a w konsekwencji testowanie zależności i hipotez między zmiennymi w tych modelach poprzez stosowanie tradycyjnych metod i technik badań socjologicznych. Przekonanie o tym, że wiedza zawarta jest w danych, w sposobie ich zbierania i analizy, jest obecne w socjologii jakościowej od zawsze. Logika eksploracji dużych zbiorów danych tekstowych z wykorzystaniem Text Mining i przetwarzania języka naturalnego wnosi w obszar wspomaganej komputerowo analizy danych jakościowych nowe, niespotykane dotąd możliwości odkrywania relacji w różnych układach społecznych, a tym samym poszerzania i pogłębiania wiedzy socjologicznej.

Bibliografia

Berry Michael W. (ed.), (2004), *Survey of Text Mining: Clustering, Classification, and Retrieval*, Springer, New York.

Bryda Grzegorz (2014), *CAQDAS, Data Mining i odkrywanie wiedzy w 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: 26.05.2014].

Feldman Ronen, Sanger James (2007), *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, Cambridge.

Fielding Nigel G. (2012), *The Diverse Worlds and Research Practices of Qualitative Software*, "Forum Qualitative Sozialforschung", t. 13, nr 2; www.qualitative-research.net/index.php/fqs/article/view/1845/3369 [dostęp: 1.06.2014].

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.

Glaser Barney (1978), *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory*, Sociology Press, Mill Valley.

Glaser Barney (1992), *Basics of Grounded Theory Analysis: Emergence vs. Forcing*, Sociology Press, Mill Valley.

Hand David, Mannila Heikki, Smyth Padhraic (2005), *Eksploracja danych*, WNT, Warszawa.

Hearst Marti A. (1999), *Untangling Text Data Mining*, The 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland (invited paper), s. 3–10; http://people.ischool.berkeley.edu/~hearst/papers/acl99.pdf [dostęp: 1.06.2014].

Ho Yu Chong, Jannasch-Pennell Angel, Gangi 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, s. 730–744; www.nova.edu/ssss/QR/QR16-3/yu.pdf [dostęp: 1.06.2014].

Hotho Andreas, Nürnberger Andreas, Paaß Gerhard (2005), *A Brief Survey of Text Mining*, "LDV Forum – GLDV Journal for Computational Linguistics and Language Technology", no. 20 (1), s. 19–62; www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf [dostęp: 1.06.2014].

Johnson R. Burke, Onwuegbuzie Anthony J. (2004), *Mixed Methods Research: A Research Paradigm Whose Time Has Come*, "Educational Researcher", no. 33 (7), s. 14–26.

Konecki Krzysztof (2000), *Studia z metodologii badań jakościowych. Teoria ugruntowana*, PWN, Warszawa.

Larose Daniel T. (2006), *Odkrywanie wiedzy z danych: wprowadzenie do eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa.

Lewins Ann, Silver Christina (2007), *Using Software in Qualitative Research: A Step-by-Step Guide*, Sage Publications, Los Angeles.

Lofland John, Snow David, Anderson Leon, Lofland Lyn (2009), *Analiza układów społecznych: przewodnik metodologiczny po badaniach jakościowych*, przeł. Anna Rosińska-Kordasiewicz, Sylwia Urbańska, Monika Żychlińska, Wydawnictwo Naukowe Scholar, Warszawa.

Lula Paweł (2005), *Text Mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*, Statsoft, Kraków; www.statsoft.pl/Portals/0/Downloads/Text_mining_jako_narzedzie_pozyskiwania.pdf [dostęp: 1.06.2014].

Manning Christopher D., Raghavan Prabhakar, Schütze Hinrich (2008), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge; www-nlp.stanford.edu/IR-book/ [dostęp: 1.06.2014].

Ogden Charles Kay, Richards Ivor Armstrong (1923), *The Meaning of Meaning*, Harcourt: Brace & World, Inc., 8th ed., New York; http://courses.media.mit.edu/2004spring/mas966/Ogden%20Richards%201923.pdf [dostęp: 1.06.2014].

Porter Martin F. (1980), *An Algorithm for Suffix Stripping*, "Program", 14 (3), s. 130–137.

Prado do Hercules Antonio, Ferneda Edilson (2008), *Emerging Technologies of Text Mining: Techniques and Applications*, Hershey, New York.

Rijsbergen van Cornelis Joost (1979), *Information Retrieval*, Butterworths, London.

Rutkowski Leszek (2011), *Metody i techniki sztucznej inteligencji*, Wydawnictwo Naukowe PAN, Warszawa.

Salton Gerard M., Wong Andrew, Yang Chung Shu (1975), *A Vector Space Model for Automatic Indexing*, "Communications of the ACM", vol. 18, issue 11, s. 613–620.

Silverman David (2007), *Interpretacja danych jakościowych: metody analizy rozmowy, tekstu i interakcji*, Wydawnictwo Naukowe PWN, Warszawa.

Tomanek Krzysztof, Bryda Grzegorz (2014), *Odkrywanie wiedzy w wypowiedziach tekstowych. Metoda budowy słownika klasyfikacyjnego*, [w:] Jakub Niedbalski (red.), *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

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].
www.kdnuggets.com/ [dostęp: 1.06.2014].
www.provalisresearch.com/products/ [dostęp: 1.06.2014].

From CAQDAS to Text Mining
New Techniques in the Qualitative Data Analysis

Summary. The aim of this article is methodological reflection on the development of computer-assisted qualitative data analysis in the direction of Knowledge Discovery in Textual Databases and Text Mining. In our consideration we focus on the social sciences area, especially qualitative analysis. In the last few years the usage of computer-assisted qualitative data analysis in the field of qualitative sociology has become a fact. Environment of qualitative researchers are increasingly reaching for CAQDAS software in their research projects. Our experience show that possibility of working with various CAQDAS programs leads to increased methodological awareness, which moves to greater accuracy and precision in the process of qualitative data analysis. The idea of usage Text Mining techniques, Natural Language Processing and Data Mining methodology is a novelty in the field of qualitative sociology. Text Mining is a set of techniques, which are equipped with programs designed for automatic or semiautomatic extracting and analyzing informations from textual data. Text Mining involves the use of computer software in finding relationships in the unstructured, textual data, which are hidden for a human due to its limited perceptual and temporal abilities. If CAQDAS analytical algorithms are rather used to work with a smaller qualitative data sets, Text Mining techniques allows to analyze unlimited textual data sets. In this article we would like to approach the key elements of the process of Text Mining analysis and try to answer the question whether these approaches compete with each other, or are rather complementary?

Keywords: knowledge discovery in data, CAQDAS, Data Mining, Text Mining, grounded theory, Natural Language Processing (NLP), Knowledge Discovery in Textual Databases (KDT).