http://dx.doi.org/10.18778/7969-549-2.09

Krzysztof Tomanek
Uniwersytet Jagielloński

Jak nauczyć metodę samodzielności? O „uczących się metodach" analizy treści

Streszczenie. Systemy uczące się najpowszechniej stosowane są w analizach danych ilościowych (*quantitative data analysis* – QUAN). Rzadko się jeszcze zdarza, aby sięgali po nie badacze i analitycy zajmujący się danymi jakościowymi (*qualitative data analysis* – QUAL). Tymczasem rozwiązania z zakresu *machine learning*, osiągnięcia z obszaru *natural language processing*, zaawansowane statystyczne systemy uczące się mogą być stosowane w analizach danych tekstowych. Niniejszy artykuł ma na celu przybliżenie sposobu działania i stosowania metod uczących się (MUS). Opisane zostały podstawowe, dostępne w wybranych programach CAQDAS (ze szczególnym uwzględnieniem programu Qualrus), techniki wspierające opracowanie materiałów tekstowych. Charakterystyka ta dotyczy automatycznych i półautomatycznych metod kodowania. Podano także przykład zastosowania systemów uczących się do analiz tekstowych (transkrypcji z wywiadów FGI). Wskazano możliwości rozwoju metod uczących się. W podsumowaniu zaprezentowane zostały trudności związane ze stosowaniem omawianych metod.

Słowa kluczowe: systemy uczące się, Text Mining, kodowanie, kodowanie automatyczne, kodowanie półautomatyczne, strategie kodowania, analiza treści, system ekspercki, referencyjna teoria znaczenia, *machine learning*, *natural language processing*, NLP, CAQDAS, Qualrus, QDA Miner.

Wprowadzenie

Analityk danych jakościowych staje współcześnie przed wieloma dylematami. Jedne z nich dotyczą wyboru obszaru, do stosowania i testowania rozwiązań analitycznych (blogi, transkrypcje wywiadów indywidualnych lub wywiadów grupowych, prasa, Internet), drugie związane są z decyzją o zakresie danych branych pod uwagę. Jeszcze inny dylemat towarzyszy ambicji do rozwijania technik analitycznych. Takim obszarem, w którym obserwowany jest szybki rozwój technik QUAL, jest kodowanie materiałów tekstowych. Wśród dostępnych na rynku istnieją takie rozwiązania CAQDAS, które dysponują algorytmami potrafiącymi kodować dane tekstowe automatycznie. Niektóre z tych technik kodowania działają niczym czarne skrzynki. Analityk nie zna ich

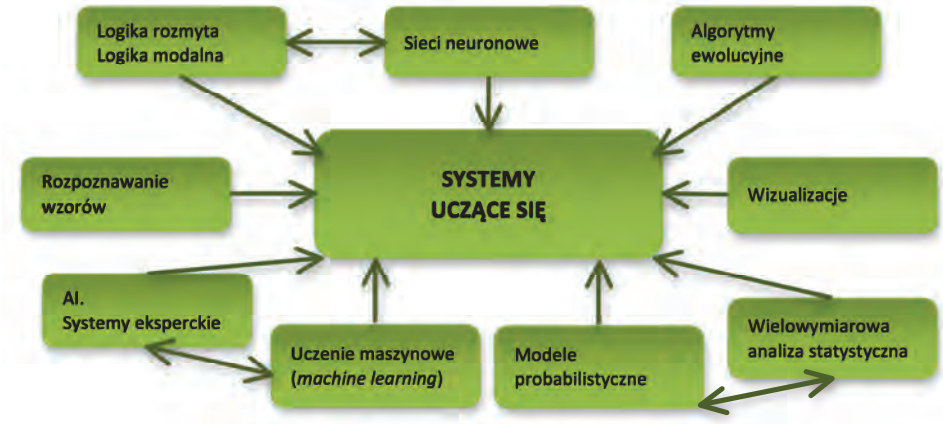
budowy, ani sposobu, w jaki przetwarzają one dane. Co prawda oprogramowanie takie pozwala np.: zdefiniować jednostkę analizy; określić słowa kluczowe; wskazać fragment tekstu, wykorzystywany jako wzorzec do kodowania (por. QDA Miner), ale nie daje to pełnej wiedzy o sposobie działania techniki analitycznej. Istnieją jednak rozwiązania CAQDAS, które oferują transparentne metody i techniki wspierające pracę z kodowaniem tekstów. Taka sytuacja oznacza, dla analityka, możliwość nie tylko zapoznania się z definicją algorytmu, ale także jego modyfikację. Mamy więc do czynienia z sytuacją, w której oprogramowanie niesie ze sobą możliwość samodzielnego budowania systemów uczących się kodowania tekstów (por. Qualrus). Daje to wiele nowych możliwości rozwoju metodologii analiz jakościowych. CAQDAS, które pozwala projektować metody i algorytmy, to niemal „nieograniczone" środowisko dla wyobraźni analityka. Omawiane tutaj rozwiązania – rozwijając propozycję Jakuba Niedbalskiego (Niedbalski 2013: 153–166) – można nazwać piątą generacją CAQDAS. Ten nowy etap ewolucji oprogramowania wspierającego analizy danych jakościowych posiada jeszcze jedną istotną cechę. Spełnia mianowicie istotne – z punktu widzenia poznania naukowego – kryterium transparentności metodologicznej w analizach jakościowych.

W niniejszym artykule zaprezentuję sposób, w jaki można wykorzystać otwarte środowisko analityczne do budowy systemów uczących się kodowania. Zacznę od syntetycznego opisu systemów uczących się. Pokażę możliwości związane z budową takiego systemu, wykorzystując możliwości, jakie daje w tym zakresie program Qualrus. Na zakończenie zaproponuję kilka sposobów rozwijania metod uczących się i dodam refleksję związaną z ograniczeniami charakteryzowanego tu podejścia.

Słowo o popularności systemów uczących się

Wzrost mocy obliczeniowej komputerów osobistych oraz pojemności ich pamięci stworzył w ostatnich latach możliwości zarówno gromadzenia olbrzymich ilości informacji, jak i ich przetwarzania. W konsekwencji obserwujemy szybki rozwój różnorodnych algorytmicznych metod odkrywania wiedzy ukrytej w danych. Ich liczba jest tak ogromna, że opis szczegółowych rozwiązań wymagałby opracowania encyklopedycznego. Dla uproszczenia na ilustr. 1 pokazuję klasy tych rozwiązań. Zaledwie kilka z nich zostanie omówionych w dalszej części artykułu.

Metody uczące się są powszechnie stosowane w bardzo wielu dziedzinach: od badań medycznych do przewidywania kursów giełdowych, od przemysłu do gier komputerowych. Oto kilka ich najczęstszych zastosowań:



Ilustr. 1. Obszary, z których korzystają i w których wykorzystywane są metody uczące się

Źródło: opracowanie własne na podstawie: Duch (1997); Bolc, Cytowski (1989, 1991); Cichosz (2000)

a) przetwarzanie języka naturalnego: tłumaczenie, odkrywanie wiedzy w wypowiedziach (Demski 2011), systemy dokonujące automatycznych streszczeń, ocena stopnia przystępności lub złożoności tekstu – *fog index* (Gunning 1952),

b) systemy eksperckie: doradztwo, systemy rekomendacji – *recommender systems* (Jannach i in. 2010; Ricci i in. 2011),

c) rozwiązywanie zadań i gry logiczne: logistyka, planowanie, prognostyczne zadania w ramach ekonomii, szachy (systemy uczące gry w szachy poprzez dostosowanie poziomu złożoności zadania do umiejętności gracza, systemy uczące się stylu gry przeciwnika w szachy, por. pojedynek Garri Kasparowa z Deep Blue (Pandolfini 1997)),

d) rozpoznawanie obrazów: rozpoznawanie sylwetek i twarzy, rozpoznawanie obiektów w ruchu, rozpoznawanie tekstów w obrazach i rozpoznawanie pisma ręcznego (OCR – Optical Character Recognition), diagnostyka medyczna (Tadeusiewicz, Korohoda 1997) i inżynieryjna, rozpoznawanie obrazów graficznych, rozpoznawanie obrazów dźwiękowych i mowy – *voice mining* (Manning, Raghavan, Schutze 2008),

e) optymalizacja wieloetapowych systemów decyzyjnych: budowa procesów optymalizujących jakość produkcji/zarządzania (Murty 2010),

f) sterowanie i robotyka: nawigacja, kierowanie pojazdem, sterowanie procesami produkcji (Kost, Łebkowski, Węsierski 2013),

g) inżynieria oprogramowania: budowa testów walidacyjnych oprogramowania, budowa inteligentnych interfejsów (Sacha 2010).

W większości obszarów, w których stosowane są MUS, ogólny model uczenia się wygląda podobnie choć istnieją różne wariacje tego schematu. Powszechnie stosowane są trzy uogólnienia procesu uczenia się. W ramach tych rozwiązań uczenie odbywa się:

a) z nauczycielem[1] (uczenie nadzorowane – to proces, w którym poza danymi zastanymi wykorzystywane są dodatkowe informacje uczące, np. podpowiedzi o zależnościach między tematem wypowiedzi a słowami charakterystycznymi dla danego tematu, wykorzystanie słowników klasyfikacyjnych itp.),

b) z krytykiem (uczenie kontrolowane – w tym procesie wykorzystywana jest tylko informacja o tym, jaki jest cel analizy, bez dodatkowych podpowiedzi od nauczyciela, np. informacja o trafności grupowania wypowiedzi tekstowych bez podpowiedzi o słowach kluczowych),

c) bez nadzorowania (takim procesem jest na przykład klasyfikacja wypowiedzi ze względu na ich podobieństwo, które może spełniać różnorodne kryteria takie, jak: długość ciągu tekstowego, podobieństwo semantyczne, podobieństwo profilu autora tekstu itp.).

Powyżej wspomniane zostały przykładowe obszary, w których stosowane są systemy uczące się. Przejdę teraz do pytania o obszary, w których system uczący się wspiera analityka w pracy z danymi tekstowymi. Oto kilka zastosowań MUS w pracy z tekstami:

a) rozpoznawanie metod argumentacji,

b) diagnoza sekwencji wypowiedzi prowadzących do napięć i konfliktów podczas FGI,

c) identyfikacja wypowiedzi tekstowych nacechowanych emocjonalnie i wywołujących emocje,

d) rozpoznawanie typowych sekwencji wypowiedzi w wywiadach IDI,

e) analiza schematów poznawczych w komentarzach oceniających,

f) identyfikacja problemów drażliwych w wywiadach FGI,

g) identyfikacja strategii radzenia sobie w sytuacjach codziennych,

h) klasyfikacja dużych zbiorów artykułów prasowych w oparciu o schematy kodowania,

i) diagnostyka specyfiki wypowiedzi zwiększających popularność postów na forach dyskusyjnych,

j) identyfikacja liderów opinii, szarych eminencji i grup opiniotwórczych na forach internetowych,

k) rozwój tematycznych słowników analitycznych w oparciu o reguły leksykalne (por. artykuł w tym tomie o metodzie budowy słownika klasyfikacyjnego (Tomanek 2014c).

[1] W szczególności są to statystyczne metody analizy danych. Znajdują one również zastosowanie w analizie danych tekstowych (por. Lula 2005; Kornacki, Ćwik 2005).

Po zarysowaniu obszarów tematycznych, w których wykorzystuje się MUS, warto zadać następujące pytania: Jak zbudowane są automatyczne i półautomatyczne metody analizy treści? Jak zastosować takie metody w analizie tekstów? Jak rozwijać i doskonalić metodę uczącą się? Jak interpretować wyniki uzyskane za pomocą MUS? Jakie są ograniczenia metod uczących się stosowanych w analizach treści?

System samouczący się w środowisku CAQDAS (Qualrus)

Skupię się teraz na opisie systemów uczących się, dostępnych w środowisku Qualrus. W oprogramowaniu tym pełnią one rolę tak zwanych „inteligentnych strategii kodowania" (ISK) materiału tekstowego. Ich zadaniem jest sugerowanie kodów, jakie mogą być stosowane dla niezakodowanego jeszcze tekstu lub zbioru tekstów. Sugestie te opierają się na różnych ideach metodologicznych i różnych rozwiązaniach praktycznych mających odzwierciedlać te idee. Wbudowane w Qualrusie algorytmy:

a) analizują logikę budowy tekstu,

b) szukają semantycznych zależności pomiędzy słowami i frazami występującymi w tekście,

c) odkrywają semantyczne zależności pomiędzy kodami użytymi wobec tych samych fragmentów tekstów,

d) odkrywają statystyczne zależności pomiędzy częściami mowy występującymi w tekście,

e) identyfikują statystyczne zależności pomiędzy kodami zastosowanymi do wcześniejszych partii tekstu (uczenie z zakodowanych już partii tekstu).

Wykorzystując ISK, Qualrus pozostawia analitykowi decyzję pomiędzy: zastosowaniem sugestii, jaką proponuje algorytm (automatyczne kodowaniem tekstu), a możliwością samodzielnej konstrukcji modelu kodującego (budowa metody uczącej się).

Zadania, które wykonują algorytmy Qualrusa, opierają się – jak już wspomniałem – na rozwiązaniach zaczerpniętych z systemów uczących się. W Qualrusie systemy te noszą następujące nazwy:

1) mini-systemy eksperckie,
2) uczenie maszynowe,
3) odkrywanie podobieństw obserwacji,
4) reguły semantyczne i asocjacyjne,
5) analiza NLP.

Metodom tym poświęcę teraz więcej uwagi. Podam przykłady ich zastosowania oraz wskażę wyniki, jakie przynosi ich zastosowanie. Na końcu wspomnę o ograniczeniach związanych ze stosowaniem tych metod w analizie wypowiedzi zapisanych w języku polskim.

Ad 1) Mini-systemy eksperckie

Systemy eksperckie to metody, które wykonują złożone zadania analityczne, specyficzne dla ekspertów w danej dziedzinie (Hayes-Roth, Waterman, Lenat 1983; Hess, Chen 2005). W analizie treści systemy te mają za zadanie odkrywanie wzorów w wypowiedziach, zbiorach tekstów.

Wyobraźmy sobie proste zadanie, jakie można postawić przed systemem eksperckim. Niech zadanie to polega na znalezieniu jakiegoś rodzaju zależności. Stosunkowo prostą regułę określającą sposób szukania zależności możemy zdefiniować tak: „Jeżeli X to Y" (Hayes-Roth, Waterman, Lenat 1983). Bardziej złożona reguła może wyglądać następująco: „Jeżeli N to M lub L". Zasady tego rodzaju mogą odwoływać się i czerpać z:

a) najprostszych modeli logicznych w rodzaju sylogizmów Arystotelesa,

b) złożonych systemów logiki w rodzaju logik modalnych (Haack 1997; Woleński 1996).

Bez względu na poziom złożoności reguły mogą być efektywne, nawet jeśli MUS, która je wykorzystuje, uwzględnia więcej niż jedną zależność postaci: „Jeżeli X to Y" lub „Jeżeli N to M lub L". Spójrzmy na kilka przykładów zastosowanych takich reguł do zadania związanego z analizą treści. Reguły prezentowane będą poniżej od najprostszej do najbardziej złożonej.

Przykład zadania, w którym stosowany jest system ekspercki, to zadanie klasyfikacyjne. Przed systemem eksperckim możemy postawić następujący problem: jak klasyfikować reakcje werbalne uczestników wywiadu grupowego w sytuacji, w której moderator kieruje w ich stronę komunikat w formie *zwierzenia*. Analiza dla tak postawionego problemu opiera się na:

a) zbudowaniu modelu, który kodował będzie w transkrypcji FGI:
   a. zwierzenia moderatora,
   b. reakcje uczestników wywiadu,

b) odkryciu zależności przyczynowo-skutkowej zachodzącej pomiędzy komunikatem moderatora i reakcjami uczestnika wywiadu,

c) sklasyfikowaniu reakcji werbalnych przez system uczący się.

Wyobraźmy sobie i zdefiniujmy reguły, za pomocą których system ekspercki będzie poszukiwał rozwiązań dla postawionego powyżej zadania.

Reguła 1: najprostsza z reguł szuka odpowiedzi na następujące pytanie: Czy po komunikacie moderatora wystąpiła jakakolwiek werbalna reakcja uczestników FGI?

Taka reguła zwróci informację w postaci 0–1 dla każdego wystąpienia komunikatu moderatora i sytuacji, w której zidentyfikowana zostanie jakakolwiek werbalna reakcja. Odwołując się do naszego zadania, zastosowanie reguły 1

da nam odpowiedź informującą, czy po zakomunikowaniu przez moderatora informacji w formie *zwierzenia* uczestnicy FGI zareagowali werbalnie na taką wypowiedź[2].

Ogólne pytanie, dla którego konstruowana jest reguła 1, brzmi: Czy po określonym komunikacje „K" wystąpiła reakcja „R"?

Reguła 2: trochę bardziej złożona reguła szuka odpowiedzi na następujące pytanie: Czy po komunikacie moderatora wystąpiła określona (już nie jakakolwiek) reakcja uczestników FGI (np. reakcja emocjonalna)?

Taka reguła zwróci informację w postaci 0–1, ale dla określonej werbalnej reakcji. W naszym zadaniu zastosowanie reguły 2 zwróci informację odpowiadającą na pytanie, czy po zakomunikowaniu przez moderatora *zwierzenia* uczestnicy focusa zareagowali emocjonalnie[3].

Reguła 2 przynosi odpowiedź na pytanie: Czy po określonym komunikacie „K_n" wystąpiła określona reakcja „R_n".

Reguła 3: bardziej złożona reguła udzieli odpowiedzi na pytanie: Czy po komunikacie moderatora wystąpiła określona (już nie jakakolwiek) reakcja emocjonalna i czy reakcja ta miała określoną (już nie jakąkolwiek) siłę?[4]

Taka reguła dostarczy wiedzy o sile emocjonalnej określonej reakcji, która wystąpiła po komunikacie (*zwierzeniu*) moderatora.

Reguła 3 da odpowiedź na pytanie: Czy po określonym komunikacie „K_n" wystąpiła określona reakcja „R_{ni}" o sile „S_i".

Jak podkreślałem wcześniej, zarówno proste, jak i zaawansowane reguły mogą być bardzo efektywne. Natomiast trafność ich wyszukiwania (do pewnego stopnia[5]) rośnie wraz z liczbą reguł zastosowanych w modelu analitycznym. Wyobraźmy sobie, że poszukujemy odpowiedzi na pytanie o rodzaj i siłę reakcji na *zwierzenia* moderatora wypowiedziane podczas FGI. Do rozwiązania tego zadania wykorzystujemy reguły klasyfikacyjne i słowniki klasyfikacyjne[6], które wbudowujemy w system ekspercki. Taki jeden mini-system ekspercki rozwiązuje zatem następujące zadania:

a) **identyfikuje**: *zwierzenia* moderatora oraz wypowiedzi będące reakcjami na *zwierzenia* moderatora,

---

[2] Taka analiza wymaga dookreślenia warunku brzegowego wyznaczającego ramy jednostki analitycznej, np. obszaru, w którym znaleziona wypowiedź kwalifikowana jest jako reakcja na bodziec moderatora. Inny sposób określenia warunku brzegowego to zdefiniowanie obszaru analitycznego jako zakresu tekstu objętego analizą, a więc tekstu od wypowiedzi moderatora w_n do kolejnej wypowiedzi moderatora w_{n+1}.

[3] Jest to przykładowy typ reakcji, jeden z wielu możliwych do zdefiniowania w takiej analizie.

[4] O analizie ładunku emocjonalnego w wypowiedziach por. Tomanek (2014a).

[5] Do momentu przeuczenia modelu, por. Wang 2006.

[6] Por. Tomanek (2014c).

b) **klasyfikuje**: wypowiedzi będące reakcjami na zwierzenia moderatora w ramach zdefiniowanych rodzajów wypowiedzi (np. reakcja oparta na empatii, interpretacja rozumiejąca, reakcja zadaniowa, zdziwienie itp.),

c) **określa**: siłę reakcji werbalnej.

Opisane powyżej reguły oraz model, w którym zostały wykorzystane, w roli jednostek analizy, wykorzystują:

a) części wypowiedzi: słowo, fraza, zdanie, paragraf, tekst zakwalifikowany przed dany kod,

b) kody: wystąpienie kodu, relacje kodu wobec kodu.

W Qualrusie analiza treści realizowana przez mini-systemy eksperckie posługuje się kilkoma rozwiązaniami analitycznymi. Są nimi:

a) **reguły predykcyjne** – oparte są na takich koncepcjach, jak: „X wywołuje Y" lub „X jest związane z Y"; te dwa zdania warunkowe pozwalają na analizę predykcyjną wystąpienia kodów lub słów,

b) **analiza współwystępowania** – to rozwiązanie stosowane jest wtedy, gdy określona kombinacja kodów zostaje uznana za specyficzną dla danego segmentu tekstu. Jeśli więc pojawia się tekst o podobnej specyfice, wtedy wynik analizy współwystępowania to *de facto* sugestia zastosowania kodów dla danego segmentu tekstu. Analiza ta czerpie ze sposobu, w jaki tekst był w przeszłości kodowany przez kodera,

c) **analiza sekwencji** – technika odkrywania sekwencji kodów w Qualrusie. Polega ona na przewidywaniu wystąpień kodów w danym segmencie tekstu w oparciu o sekwencje kodów występujących w segmencie go poprzedzającym. Analiza ta wymaga, aby tekst był dobrze ustrukturyzowany, a więc taki, w którym rozgraniczone są w sposób wyraźny porcje (segmenty) tekstów (np. wypowiedzi moderatora, wypowiedzi uczestników FGI itp.),

d) **analiza kontekstowa** – reguły analizy kontekstowej opierają się na wyszukiwaniu słów specyficznych dla zakodowanych fragmentów tekstu. Te specyficzne słowa są wykorzystywane są jako wskaźniki. Jeśli algorytm zidentyfikuje je w niezakodowanym tekście program Qualrus sugeruje wykorzystanie określonego kodu, dla którego wskazane słowa są specyficzne. Analiza kontekstowa stosowana jest również do całych dokumentów tekstowych. W takim przypadku sugestia użycia kodu pochodzi z wyników analizy przeprowadzonej na innych zakodowanych dokumentach. Jeśli dokument analizowany jest podobny np. pod względem słów kluczowych w nim występujących do dokumentów już zakodowanych, wówczas proponowane jest użycie kodu, który wystąpił w innym dokumencie najbardziej podobnym do dokumentu analizowanego.

Ad 2) Uczenie maszynowe

Strategie analityczne kryjące się pod nazwą *machine learning* (ML) to strategie uczące się (Winston 1975; Avron, Cohen, Feigenbaum 1990). W procesie uczenia się techniki te odwołują się do udokumentowanych wyników pracy analityka.

W obszarze kodowania można powiedzieć, że model ML analizuje decyzje, które analityk podejmował podczas procesu kodowania. ML w decyzjach tych identyfikuje to, co powtarzalne. Stosuje się tu zatem kryterium frekwencyjne. Celem tego procesu jest zbudowanie modelu opartego na zidentyfikowanych kryteriach decyzyjnych analityka. Kryteria te następnie będą stosowane do tekstów jeszcze nieanalizowanych. Qualrus wyposażony jest w dwa modele ML. Są to:

1) **Identyfikacja współwystępowania par kodów**. Ta metoda dokonuje analizy współwystępowania kodów, które wykorzystywał koder, a nie tych, które sugerował inteligentny system kodowania. Za każdym razem, kiedy dwa lub więcej kodów jest użytych dla określonego fragmentu tekstu, to informacja ta jest wykorzystywana do wyliczenia warunkowego prawdopodobieństwa wystąpienia każdego kodu z danej pary kodów współwystępujących. Odwołując się do rozważanego w poprzednim paragrafie przykładu, dzięki opisywanej tu analizie możemy dowiedzieć się:

a) jakie jest prawdopodobieństwo wystąpienia wypowiedzi o charakterze emocjonalnym wśród reakcji uczestników FGI na *zwierzenie* moderatora,

b) jakie jest prawdopodobieństwo, że wypowiedź nacechowana emocjonalnie poprzedzona jest *zwierzeniem* moderatora[7];

2) **Ocena wydajności metod adaptacyjnego uczenia się**. Ta metoda monitoruje poziom, z jakim koder wykorzystuje proponowane przez ISK sugestie dotyczące kodowania. Qualrus zapisuje częstotliwość, z jaką automatyczne sugestie systemu uczącego są akceptowane przez analityka. Jeśli analityk korzysta z sugestii i dokonuje kodowania zgodnie z nimi, wówczas wzrasta poziom trafności metod ISK. Wynik ten przekłada się na zwiększenie liczby sugestii kodowania przez model ML. Natomiast gdy stosunek akceptacji do odrzucenia sugestii systemu spadnie poniżej ustalonego przez analityka progu, wtedy metoda ML uzyskuje status nietrafnej i automatycznie zaprzestaje przesyłania sugestii. Metoda ta jest dobrym przykładem systemu uczącego się z nauczycielem.

Ad 3) Analiza podobieństwa obserwacji (*case-based reasoning* CBR)

Wnioskowanie o pojawiających się wypowiedziach na podstawie innych znanych wypowiedzi jest naturalną metodą stosowaną przez ludzi, ponieważ opiera się na wiedzy wyniesionej z przeszłych doświadczeń (Leake 1996; Schank 1982).

[7] Technika ta jest w Qualrusie parametryzowalna. Oznacza to, że możliwe jest samodzielne zdefiniowanie poziomu istotności wartości prawdopodobieństwa warunkowego oraz minimalnej liczby fragmentów tekstów, w których zaobserwowane musi być współwystępowanie kodów. Jeśli spełnione są warunki określone w obu tych parametrach, wówczas Qualrus będzie sugerował użycie kodu. Podobne rozwiązanie w ramach analizy sekwencji kodów oferuje QDA Miner.

W kontekście analizy tekstu metoda CBR bazuje na założeniu, że najprostszym sposobem na zakodowanie tekstu jest odwołanie się do już zakodowanych jego części. Technika analityczna kryjąca się pod frazą „odwołanie się" oznacza:

a) porównanie fragmentu tekstu niezakodowanego do fragmentów tekstu już zakodowanych (analiza pod kątem wystąpienia tych samych/podobnych słów),

b) porównanie zakodowanych za pomocą tych samych lub podobnych kodów fragmentów tekstów (analiza wystąpienia tych samych/podobnych kodów lub słów).

W przypadku (a) wynikiem analizy CBR jest propozycja zastosowania kodów do tekstu, który nie jest zakodowany. Posługując się przykładem, na który powoływałem się we wcześniejszych akapitach, możemy dla werbalnych reakcji na *zwierzenie* moderatora otrzymać takie sugestie kodów, jak: empatia, podejście zadaniowe, zdziwienie, silna reakcja emocjonalna.

W przypadku (b) porównanie zakodowanych fragmentów tekstów może przynieść wynik podobny do tego, jaki uzyskujemy w kodowaniu zogniskowanym. Porównanie fragmentów tekstów już zakodowanych może przynieść następujące przykładowe sugestie:

a) **pogłębienie kodowania**: zamiast kodu pozytywne emocje użycie kodu bardzo silne pozytywne emocje.

b) **użycie trafniejszego kodu**: zamiast kodu ciekawość użycie kodu ciekawość i współczucie.

c) **dodanie nowego kodu**: poza istniejącymi kodami (np. współczucie, zrozumienie, troska) użycie metakodu (np. empatia).

Ad 4) Reguły asocjacyjne i semantyczne

Celem kolejnej z metod samouczących się jest także podpowiadanie sposobu kodowania tekstu, ale tym razem podstawa dla tych podpowiedzi jest inna niż opisane do tej pory. Jest nią wynik analiz stosujących reguły asocjacyjne i semantyczne. Kluczowym jest więc zdefiniowanie tego, co rozumiemy i nazwiemy: powiązaniem koncepcyjnym (opartym na regułach asocjacyjnych), związkiem semantycznym (znaczeniowym). Qualrus umożliwia posłużenie się kilkoma rozwiązaniami. Oto one:

a) **odkrywanie związków pomiędzy koncepcjami** oparte jest na dwóch ideach zaczerpniętych z filozofii języka. Są to:

- ogólna teoria znaczenia (*systemic meaning of a concept*) – pozwala na łączenie modeli koncepcyjnych na podstawie relacji logicznych zachodzących pomiędzy koncepcjami (Kaplan 1964). Posłużę się przywoływanym wcześniej w tym tekście przykładem. Przyjmijmy, że termin „empatia" oznacza: zdolność do postawienia siebie na miejscu drugiej osoby (Elliot i in. 1997). Obserwując różnorodne

reakcje uczestników FGI na *zwierzenie* moderatora możemy zidentyfikować te, które będą świadectwem tego, co określiliśmy empatią. Działanie systemu uczącego się oparte na regułach asocjacji polega w tej sytuacji na analizie wypowiedzi objętych kodem empatia. Wyposażony w słowniki analityczne system uczący się zidentyfikuje w wypowiedziach takie elementy, jak: (a) emocje, (b) pytania, (c) skłonność do udzielania pomocy. Jeśli system uczący się posiada rozbudowany przez analityka model koncepcyjny pojęcia empatia, wówczas identyfikacja wskazanych trzech elementów pociągnie za sobą sugestię kodowania ich odpowiednio jako: empatię emocjonalną (emocje), empatię poznawczą (pytania), potwierdzenie hipotezy empatia–altruizm (chęć pomocy).

– referencyjna teoria znaczenia (*referential meaning of a concept*) – wyraża się w praktyce w stosowaniu definicji operacyjnych pojęć i terminów. W ten sposób definiowane znaczenie zostaje powiązane z jego empirycznymi wskaźnikami – czyli z: zachowaniami, przedmiotami, zjawiskami (Carnap 1956; Putnam 1975; Kripke 1980). W przykładzie dotyczącym *zwierzenia* i werbalnych reakcji na nie referencyjna teoria znaczenia jest podstawą do zdefiniowania wskaźników w postaci zachowań językowych, które analityk traktuje jako fakty językowe wymagające zakodowania.

b) **odkrywanie relacji semantycznych** odbywa się na dwa sposoby:

– pierwszy z nich oparty jest na poszukiwaniu związków semantycznych (*semantic networks*) pomiędzy wyrazami (Quillian 1968),

– druga strategia (*associative networks*) opiera się na poszukiwaniu związku pomiędzy fragmentami tekstu nie tylko w oparciu o relacje semantyczne, lecz także relacje referencyjne słowo–obiekt, obiekt–obiekt (Gonzalez, Dankel 1993). Technika ta stosowana jest w Qualrusie do poszukiwania relacji zachodzących pomiędzy kodami (np. relacje hierarchiczne). W omawianej sytuacji w FGI przykłady takich relacji mogą być następujące: empatia emocjonalna – jest wyrazem emocji, empatia poznawcza – jest formą empatii. Inne rodzaje relacji, jakie podpowiada system uczący w Qualrusie, odwołują się do: związku przyczynowo-skutkowego (zps), związku pomiędzy zdarzeniami/zjawiskami. Przykładem zps może być hipoteza: *zwierzenie* wywołuje empatię.

Ad 5) Analiza NLP

Uczenie maszyn rozumienia języka naturalnego jest zadaniem niebagatelnym (Barr, Feigenbaum 1981). Zarówno tłumaczenie automatyczne (Weaver 1955), jak i podejścia *ad hoc* do danych tekstowych spotykają dwie podstawowe trudności. Są nimi gramatyka (Chomsky 1957) i semantyka danego języka (Schank, Abelson 1977).

Techniki analizy języka naturalnego w Qualrusie próbują radzić sobie ze wspomnianymi problemami na dwa sposoby.

Po pierwsze mamy możliwość zbudowania koszyka słów (*bag of words* – BOW), który jest słowną reprezentacją modelu koncepcyjnego wyrażonego w CAQDAS za pomocą kodu. Taki koszyk słów posłuży systemowi uczącemu się do rozpoznania w języku naturalnym wypowiedzi, w których użyte słowa, frazy sugerują, że jest to fragment tekstu, który wymaga zakodowania. Dzięki BOW system NLP samodzielnie zakoduje tekst, dając analitykowi możliwość oceny poprawności kodowania (uczenie z nauczycielem). Jeśli chcielibyśmy zidentyfikować i zakodować różne typy emocjonalnych reakcji uczestników FGI na *zwierzenia* moderatora, możemy dla tego celu wykorzystać jeden z istniejących słowników do analizy sentymentu (Tomanek 2014b). Taki słownik posłuży systemowi NLP jako punkt odniesienia do kodowania fragmentów transkrypcji[8]. W praktyce analiz jakościowych trafne działanie takiego systemu obserwowano w analizach wywiadów medycznych (Brent, Thompson, Mirielli 1995), a także pokazując jego skuteczność w porównaniu do kodowania manualnego (Tomanek 2014b).

Po drugie mamy możliwość badania relacji pomiędzy kodami zastosowanymi do tego samego fragmentu tekstu. W takiej analizie stosowana jest metoda odkrywania relacji semantycznych. System uczący się identyfikuje znaczenia kontekstowe słów kluczowych, które były powodem użycia kodów dla tego samego fragmentu tekstu. Analiza kontekstowa podpowiada zależności, które mogą zachodzić pomiędzy tekstami i kodami. Zależności te mogą być oparte na następujących trzech relacjach: zależność hierarchiczna, relacja zawierania, podobieństwo semantyczne.

***

Prezentację metod uczących się dostępnych w CAQDAS (Qualrus) chciałbym zakończyć odpowiedzią na pytanie zamieszczone w tytule artykułu. Warto podkreślić, że istnieją metody, które odpowiednio zaprojektowane potrafią samodzielnie wykonywać różnorodne analizy na danych tekstowych. Zaakcentować jednak wpada, że analizy danych tekstowych wymagają stosowania złożonych modeli analitycznych korzystających z analiz QUAL i QUAN. Proces budowy takich modeli dla potrzeb analiz tekstowych jest możliwy w otwartych analitycznie środowiskach programistycznych. Takie środowisko umożliwia analitykowi

[8] W praktyce musielibyśmy jeszcze zdefiniować w Qualrusie to, co rozumiemy jako wypowiedź na *zwierzenie* moderatora. Może to być przykładowo: jedno lub 5 kolejnych zdań zapisanych po wypowiedzi moderatora. Może to być również ciąg znaków następujących bezpośrednio po wypowiedzi moderatora i zapisanych do momentu pojawienia się kolejnej wypowiedzi moderatora.

zbudowanie modelu dla tak wymagających w analizie języków, jak języki fleksyjne. Otwarte środowisko analityczne CAQDAS pozwala na radzenie sobie z takim wyzwaniem, ponieważ umożliwia wykorzystanie wiedzy z zakresu: językoznawstwa oraz filozofii języka, ale też logiki i statystyki. Ta nowa generacja programów pozwala analitykom: eksperymentować w zakresie budowy nowych metod, pracować na rzecz algorytmów bardziej efektywnych[9]. Wolność w projektowaniu nowych rozwiązań jest zapewniona przez posługiwanie się językiem programowania, którego dostarcza CAQDAS. Umiejętność ta daje swobodę budowania modeli, które wspierają analityka w odkrywaniu wiedzy „ukrytej w tekstach".

Podsumowanie

Metody uczące się otwierają przed analitykiem danych jakościowych nowe perspektywy. Wśród najbardziej owocnych można wymienić następujące możliwości:

a) wykorzystanie wiedzy z zakresu językoznawstwa, filozofii języka, logiki pierwszorzędowej, ale też logik modalnych, statystyki i matematyki w analizach tekstów pisanych w językach fleksyjnych,

b) identyfikacja schematów zachowań koderów i wykorzystanie ich jako wzorcowych praktyk kodowania dla celów uczenia metody samouczącej się,

c) adoptowanie i rozwijanie systemów uczących się do nowych problemów analitycznych, nowych materiałów zapisanych w tym samym języku,

d) rozpoznawanie relacji referencyjnych zachodzących między znaczeniami i obiektami wskazywanymi przez znaczenia. Ta technika analityczna jest wsparciem dla problemów badawczych, w których testowane są hipotezy o wieloznaczności analizowanych wypowiedzi i języków. Technika ta jest jednocześnie szansą na rekonstruowanie wielu ontologii, zawartych w tekstach,

e) identyfikacja zduplikowanych tekstów, wypowiedzi,

f) rozwój analiz, których celem jest rekonstruowanie profilu autora tekstu,

g) budowa metody uczenia się opartych na logikach „adaptywnych" (Nasieniewski 2008), które rozwijane mogą być niezależnie od materiału empirycznego, z którym analityk ma do czynienia w swojej pracy,

h) usprawnianie procesu kodowania treści poprzez budowę półautomatycznych i automatycznych systemów kodowania.

[9] Warto dodać, że nie wszystkie programy, które posiadają zaawansowane algorytmy pracujące z danymi tekstowymi, umożliwiają ich adaptowanie do nowych problemów (np. QDA Miner i Wordstat, MaxQDA – nie kwalifikują się do piątej generacji CAQDAS). Jako odrębną klasę należałoby też potraktować narzędzia, które umożliwiają stosowanie metod i technik z obszaru QUAL i QUAN (np. R Studio, Rapid Miner), ale rozwijają się szybciej w obszarze analiz ilościowych.

Podkreślić można, że systemy uczące w swojej pracy wymagają zarówno nauczyciela, który pełni rolę krytyka, jak i mentora. Opisane w tekście przykłady analiz dają możliwość rozwoju systemu uczącego się w obu tych schematach:

a) rozwój modelu ML: rozwój słowników analitycznych i BOW dzięki nowemu słownictwu pochodzącemu z tekstów klasyfikowanych przez algorytm ML,

b) rozwój relacji w ramach *associative networks*: rozwój takich relacji semantycznych, jak: symetria, przechodniość, jednoznaczność, wieloznaczność (przypisanie wielu znaczeń do jednego wyrazu, przypisanie jednego obiektu do wielu wyrazów, przypisanie wielu znaczeń do wielu wyrazów).

Środowisko CAQDAS, które nazwać można narzędziem piątej generacji, umożliwia szeroki rozwój metod analizy treści. Qualrus nie spełnia jeszcze wszystkich warunków takiego otwartego środowiska analitycznego. W szczególności w analizach treści wykorzystuje się tak zwane koszyki słów. W tym kontekście warto podkreślić, że brakuje w Qualrusie:

a) listy czasowników, przymiotników, rzeczowników,
b) tematycznych słowników analitycznych,
c) słowników klasyfikacyjnych w rodzaju Five Domains,
d) polskiego słownika synonimów,
e) polskiego thesaurusa,
f) listy referencyjnej wyrazów języka polskiego.

Stosowanie systemów uczących się stawia przed analitykiem danych jakościowych pewne ryzyko. Składają się na nie:

a) możliwość pominięcia w analizach półautomatycznych subtelności tkwiących w wypowiedziach np.: ironia, żart, parafraza, metafora, aforyzm,

b) możliwość błędnej interpretacji i klasyfikacji takich figur stylistycznych, jak: metafora, analogia (Weizenbaum 2008).

Na koniec warto wspomnieć o wyzwaniach, jakie stoją przed analitykami, którzy korzystają z metod uczących się w ramach analizy treści. Są to przede wszystkim:

a) analizy języka naturalnego nieustrukturyzowanego i analizy wypowiedzi spontanicznych (np. treści publikowane na forach dyskusyjnych),

b) analizy języków niszowych, slangowych, subkulturowych,

c) łączenie analizy tekstów i dźwięków/mowy (mariaż *text mining* i *voice mining).*

Bibliografia

Aronson Elliot, Akert Robin M., Wilson Timothy D. (1997), *Psychologia społeczna – serce i umysł*, Zysk i S-ka, Warszawa.

Avron Barr, Cohen Paul R., Feigenbaum Edward A. (1990), *The Handbook of Artificial Intelligence*, Addison–Wesley, Boston.

Barr Avron, Feigenbaum Edward A. (1981), *The Handbook of Artificial Intelligence*, HeurisTech Press, Los Altos, California.

Bolc Leonard, Jerzy Cytowski (1989, 1991), *Metody przeszukiwania heurystycznego*, t. I–II, PWN, Warszawa.

Brent Edward, Thompson Alan, Mirielli Edward (1995), *Disambiguating Verbal Comments in Social Interaction: A Computer Model of Meaning*, "Journal of Mathematical Sociology", vol. 20, s. 109–125.

Carnap Rudolf (1956), *Meaning and Necessity: a Study in Semantics and Modal Logic*, University of Chicago Press, Chicago.

Chomsky Noam (1957), *Syntactic structures*, The Hague, Mouton.

Cichosz Paweł (2000), *Systemy uczące się*, WNT, Warszawa.

Demski Tomasz (2011), *Maszyna do czytania, czyli text mining w odkrywaniu nadużyć ubezpieczeniowych*, Statsoft Polska, Kraków.

Duch Włodzisław (1997), *Fascynujący świat programów komputerowych*, Wydawnictwo Nakom, Poznań.

Gonzalez Avelino J., Dankel Douglas D. (1993), *The Engineering of Knowledge-Based Systems: Theory and Practice/Book*, Har/Dis: Prentice-Hall International.

Gunning Robert (1952), *The Technique of Clear Writing*, McGraw-Hill, Wisconsin–Madison.

Haack Susan (1997), *Logika modalna*, [w:] Jan Woleński (red.), *Filozofia logiki*, Wyd. SPACJA, Warszawa.

Hayes-Roth Frederick (2011), *The Meaning and Mechanics of Intelligence*, BiblioBazaar, London.

Hayes-Roth Frederick, Waterman Donald A., Lenat Douglas B. (1983), *Building Expert Systems*, Addison–Wesley, London.

Hess Jessica, Chen Wei-Fan (2005), *Expert System Applications in E-Learning: Present and Future*, [w:] Richards Gregory (ed.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, The Pennsylvania State University, Chesapeake.

Jannach Dietmar, Zanker Markus, Felfernig Alexander, Friedrich Gerhard (2010), *Recommender Systems: An Introduction*, Cambridge University Press, Cambridge.

Kaplan Abraham (1964), *The Conduct of Inquiry*, Chandler, San Francisco.

Kornacki Jacek, Ćwik Jan (2005), *Statystyczne systemy uczące się*, Wydawnictwo Naukowo-Techniczne, Warszawa.

Kost Gabriel, Łebkowski Piotr, Węsierski Łukasz (2013), *Automatyzacja i robotyzacja procesów produkcyjnych*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Kripke Saul (1980), *Naming and Necessity*, Basil Blackwell, Oxford.

Leake David B. (1996), *Case-based Reasoning: Experiences, Lessons, and Future Directions*, AAAI Press/MIT Press, Menlo Park, CA: Cambridge, MA.

Lula Paweł (2005), *Text mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*, Statsoft Polska, Kraków.

Manning Christopher D., Raghavan Prabhakar, Schütze Hinrich (2008), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge.

Murty Katta G. (2010), *Optimization for Decision Making* „International Series in Operations Research & Management Science", Springer, Berlin.

Nasieniewski Marek (2008), *Wprowadzenie do logik adaptywnych*, Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, Toruń.

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

Pandolfini Bruce (1997), *Kasparov and Deep Blue: The Historic Chess Match Between Man and Machine*, Fireside Chess Library, New York.

Putnam Hilary (1975), *The Meaning of 'Meaning'*, [w:] Keith Gunderson (ed.), *Language, Mind and Knowledge*, University of Minnesota Press, Minneapolis.

Quillian Ross (1968), *Semantic Memory*, [w:] Marvin Minsky (ed.), *Semantic Information Processing*, MIT Press, Cambridge, MA.

Ricci Francesco, Rokach Lior, Shapira Bracha, Kantor Paul (2011), *Recommender Systems Handbook*, Springer, Berlin.

Sacha Krzysztof (2010), *Inżynieria oprogramowania*, Wydawnictwo Naukowe PWN, Warszawa.

Schank Roger C. (1982), *Dynamic Memory: A Theory of Learning in Computers and People*, Cambridge University Press, New York.

Schank Roger, Abelson Robert P. (1977), *Scripts, Plans, Goals and Understanding*, Hillsdale, Lawrence Erlbaum, New York.

Tadeusiewicz Ryszard, Korohoda Przemysław (1997), *Komputerowa analiza i przetwarzanie obrazów*, Wydawnictwo Fundacji Postępu Telekomunikacji, Kraków.

Tomanek Krzysztof (2014a), *Analiza sentymentu: historia i rozwój metody w ramach CAQDAS*, [w:] Jakub Niedbalski (red.), *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Tomanek Krzysztof (2014b), *Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych*, „Przegląd Socjologii Jakościowej", t. X, nr 2, s. 118–137; www.qualitativesociologyreview.org/PL/Volume26/PSJ_10_2_Tomanek.pdf [dostęp: 1.06.2014].

Tomanek Krzysztof (2014c), *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.

Wang Haixun i in. (2006), *Suppressing model overfitting in mining concept-drifting data streams*, [w:] *KDD '06 Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, s. 736–741.

Weaver Warren (1955), *Translation*, [w:] Wiliam N. Locke, Andrew D. Booth (eds), *Machine Translation of Languages*, Technology Press of MIT and Wiley, New York.

Weizenbaum Joseph (2008), *Moglibyśmy mieć raj*, „Forum", styczeń, nr 28, s. 28–29.

Winston Patrick H. (1975), *The Psychology of Computer Vision*, McGraw-Hill, New York.

Woleński Jan (1996), *Momenty bytowe i modalności*, [w:] Jan Woleński (red.), *W stronę logiki*, Aureus, Kraków.

How to Learn the Method of Self-Reliance? The Learning Methods of Content Analysis

Summary. Learning systems are most commonly used in the quantitative data analyzes. It does not happens to often that those systems are used by researchers and analysts dealing with qualitative data. Meanwhile solutions taken from such a fields like: machine learning, natural language processing, statistical learning systems can be used in the text data analysis. This article aims to explain the mode of action and the use of learning methods applied to the textual data. Described are basic applications of learning system implemented in CAQDAS tools (with special emphasis on Qualrus). Highlighted are the possibilities of the use of learning systems for the analysis of text (e.g. transcription from FGI).

Keywords: learning systems, Text Mining, content analysis, automatic coding, semi-automatic coding, coding strategies, expert systems, the referential theory of meaning, machine learning, natural language processing, NLP, CAQDAS, Qualrus.