BOGDAN YUSKIV

SKOMPUTERYZOWANA ANALIZA TREŚCI

RZESZÓW 2011

Oryginał publikacji w języku ukraińskim znajduje się w rozdziale 4 pn. КОМП'ЮТЕРИЗОВАНИЙ КОНТЕНТ-АНАЛІЗ w książce: КОНТЕНТ-АНАЛІЗ, Історія розвитку і світовий досвід (Content analysis. History development and world practice), pod redakcją Juśkiw Bohdan Mikołajowycz, Równe 2006.

Tłumaczenie polskie i opracowanie graficzne zrealizowano w ramach projektu Internetowa Promocja Nauki

Publikacja dystrybuowana bezpłatnie.

Opracowanie wersji polskiej:
Projekt „Internetowa Promocja Nauki"

Uniwersytet Rzeszowski
Al. T. Rejtana 16 c
35-959 Rzeszów

www.inprona.pl

Tłumaczenie:
ROSYJSKI.COM.PL ,

Ul. Wojrowicka 50/2
54-436 Wrocław

Projekt graficzny:
Studio projektowe INVITRO s.c.

Ul. Mickiewicza 4
35-064 Rzeszów

www.agencjainvitro.pl

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

SPIS TREŚCI

Skomputeryzowana analiza treści 4

1.1. Pierwsze próby zmechanizowania i zautomatyzowania analizy treści 4

1.2. The General Inquirer i cechy drugiej generacji programów analizy treści 7

1.3. Trzecia generacja programów analizy treści 13

1.4. Współczesne technologie analizy treści i właściwości Text Mining 17

Wnioski 24

ZAŁĄCZNIK A. Podstawowe pojęcia i algorytm ilościowej analizy treści 24

1. Zasady ogólne i kluczowe pojęcia 24

2. Etapy analizy treści 25

3. Określenie problemów, pytań i hipotez badania 25

4. Określenie całokształtu badania 25

5. Formowanie wyboru 25

6. Wybór jednostek analizy 26

7. System kwantyfikacji (podliczenia ilości) 27

8. Tworzenie kategorii analizy 27

9. Zapewnienie pewnej pracy dla osób kodujących 29

10. Kodowanie treści 30

11. Analiza danych i interpretacja 30

12. Pewność 30

13. Błędy podczas przeprowadzania analizy treści 31

ZAŁĄCZNIK B. Programy analizy treści 32

1. Programy ilościowej analizy treści 32

2. Programy jakościowej analizy treści 35

3. Programy, które pomagają w przeprowadzaniu jakościowej analizy danych 36

ZAŁĄCZNIK C. Programy Text Mining 36

SKOMPUTERYZOWANA ANALIZA TREŚCI

Skomputeryzowana analiza treści miała inną podstawę techniczną niż jej poprzednik – ręczna analiza treści. Dzięki temu nie tylko wypracowała własną metodykę i technologię zastosowania, ale nadal rozwija się zgodnie ze swoją logiką.
To właśnie stanowi przedmiot naszego przeglądu.

1.1. Pierwsze próby zmechanizowania i zautomatyzowania analizy treści

Pierwsze analizy treści w dziedzinie humanistyki, przeprowadzone z wykorzystaniem środków technicznych, związane są z pracami badawczymi włoskiego duchownego Roberto Busy. Informacje na temat jego działalności znaleźć można w pracy J. Bradley'a i G.Rockwella[1] oraz na stronie internetowej Uniwersytetu Gregoriańskiego w Rzymie[2].

Ojciec R. Busa postawił sobie za zadanie stworzenie przy pomocy środków technicznych konkordancji pełnego zbioru listów włoskiego teologa XIII w. Tomasza z Akwinu. Praca habilitacyjna R. Busy, obroniona w 1946r., poświęcona została była zgłębieniu pojęcia „bytu" w rozumieniu dominikańskiego myśliciela. Po stworzeniu i przeanalizowaniu drukowanych indeksów wyrazów łacińskich praesens i praesentia, R. Busa doszedł do wniosku, że ich wykorzystywanie przez Tomasza z Akwinu było ściśle związane z przyimkiem in. Ponadto, R. Busa gotów był uwierzyć, że podobne wyrazy funkcjonalne dają wiele informacji dotyczących związku pomiędzy konceptualnym światem autora a wyrazami wykorzystywanymi przez niego do opisu tegoż świata. Okazuje się jednak, że R. Busa nie miał wystarczających zasobów badawczych do stworzenia ręcznie konkordancji ogólnych wyrazów łacińskich, jak in, sum lub et (w tłumaczeniu odpowiednio „w", czasownik pomocniczy „być", „i").
Mimo trudności, pod koniec roku 1940 postawił on przed sobą jeszcze bardziej skomplikowane zadanie, mianowicie, stworzenie „Index Thomisticus", który miał zawierać pełną konkordancję – 10,6 mln wyrazów Tomasza z Akwinu.

Oczywiście, realizacja tego zadania bez wykorzystania pewnego rodzaju urządzeń technicznych była niemożliwa.

Praca, rozpoczęta z wykorzystaniem perforatorów i maszyn sortujących, w praktyce została zakończona 33 lata później – dzięki wielkim uniwersalnym elektronicznym maszynom liczącym /EOM – przyp. tłum/ serii IBM. Wraz z innymi informacjami Indeks zawierał prawie 70000 stron. Powstały dwie konkordancje. Pierwsza, utworzona bezpośrednio przez maszynę, zawierała pełną listę odpowiedników dla wszystkich form słownych i otrzymała nazwę „nielematyzowanej". Drugą konkordancję tworzyły tzw. „lematyzowane" odpowiedniki, w wykazie których każdy wyraz występował tylko jeden raz w formie podstawowej, na przykład, rzeczowniki – tylko w liczbie pojedynczej, czasownik – w bezokoliczniku itd. Elektroniczna maszyna licząca nie była w stanie stworzyć „lematyzowanego" odpowiednika bez ludzkiej pomocy. Według oceny Busa, na całą pracę wykorzystano ponad 1 mln osobo/godzin, głównie na wprowadzanie i weryfikację danych, jak również lematyzację.

Warto podkreślić, że w roku 1992 ojciec R. Busa założył kierunek – lingwistyka i hermeneutyka, na Wydziale Filozofii Gregoriańskiego Uniwersytetu Papieskiego w Rzymie i kontynuował swoje prace w ramach projektu „Totius Latinitatis Lemmata"[3]. Ostatnie publiczne wspomnienie o jego słynnej pracy miało miejsce w 2002r., kiedy niestrudzony R. Busa miał prawie 90 lat. W powiadomieniu Katolickiej Agencji Informacyjnej AGNUZ z dn. 1 lutego 2002r.[4] była wzmianka o spotkaniu Ojca Świętego Jana Pawła II z grupą naukowców, którzy „skomputeryzowali" dzieła Św. Tomasza z Akwinu oraz z członkami Stowarzyszenia Komputeryzacji Analiz Leksykologicznych (CAEL), którzy sponsorowali wydanie „Index Tomisticus", z okazji zakończenia prac. Oprócz wersji drukowanej, składającej się z 56 tomów encyklopedycznych zbiór został również wydany na płytach CD. Ojciec Święty wsparł inicjatywę profesora R. Busy i międzynarodowego zespołu młodych uczo-

1 Bradley J., Rockwell G. TACTweb: workbook...

2 Roberto Busa: biography / Humanities and information science. Roma Univ. La Sapienza (24/11/95). – http://www.mediamente.rai.it/mmold/english/bibliote/biografi/b/busa.htm

3 Roberto Busa: biography...

4 Papież za wykorzystaniem komputera w służbie dla doktryny chrześcijańskiej/ Katolicka Agencja Informacyjna „AGNUZ". Watykan, 1 lutego 2002r. http://agnuz.info/print.php?year=2002&mounth1=February&day=1&files=101.txt&print=news116

nych w sprawie nowego projektu pn. „Leksykon Św. Tomasza z Akwinu", który ma obejmować współczesne tłumaczenie wszystkich terminów, wykorzystanych w pracach i listach średniowiecznego teologa. Podjęte próby zmechanizowania poszczególnych czynności w procesie analizy treści nie doprowadziły i nie mogły doprowadzić do istotnych zmian w zakresie jej wykorzystania i metodyki. Wraz ze wzrostem liczebności samych tekstów, zaistniała potrzeba wykonywania takiej analizy, która prowadzi do szybkiego i efektywnego przeszukiwania tekstów różnego rodzaju i o znacznej objętości. Pojawienie się techniki komputerowej stworzyło realne, choć na początku niejawne i nieuświadomione co do swoich skutków w przyszłości, możliwości przezwyciężenia „przeklętego problemu wymiarowości".

Pierwsze, niepewne kroki w kierunku automatyzacji analizy treści szybko przeszły w nieuzasadniony optymizm. Jednak dziś, po ponad czterdziestu latach od podjęcia pierwszych prób, mimo realnych osiągnięć, stało się jasne, że, jak się wypowiedział K. Neuendorf [5], automatyczna analiza treści, to *chimera. „(...) Metoda analizy treści, w której wykorzystywany jest komputer, najczęściej przedstawiana jest jako zmechanizowana lub zautomatyzowana (...)"* – stwierdza niemiecki badacz E. Mergenthaler[6]. Zatem, mowa tutaj jest jedynie o takim czy innym poziomie wykorzystania komputera przez człowieka w procesie analizy.

Pod pojęciem skomputeryzowanej analizy treści dalej będziemy rozumieć taką analizę materiałów tekstowych, w której wszelkie kroki wiodące do ustalenia właściwości (charakterystyk) treści i określenia ogółu konceptualnych wskaźników tekstu są wykonywane przy pomocy algorytmów, czyli jawnych operacji logicznych lub statystycznych.

J. Macnamara wyróżnia dwa rodzaje zastosowania programów komputerowych:

- zapamiętywanie i przechowywanie danych, na przykład: wyników kodowania lub notatek, analiz oraz raportów, m.in. w formie tabel, diagramów i wykresów – sporządzonych przez badacza ręcznie;
- automatyczny przegląd tekstów, rozpoznawanie i kodowanie wyrazów, fraz; ten rodzaj zastosowania może prowadzić do kodowania automatycznego i analizy lub połączenia przeglądu oraz kodowania automatycznego i ręcznego[7].

U. Kelle dzieli programy analizy tekstów na trzy generacje:

1. procesory tekstowe i systemy zarządzania bazami danych tekstowych,zabezpieczające ogólne funkcje zarządzania tekstami;
2. programy związane z techniką kodowania i odnajdywania w tekstach informacji, odpowiadających pewnym kryteriom, m.in. tworzenie różnego rodzaju indeksów i konkordancji, które są tradycyjne dla metody ręcznej;
3. programy, które mimo że opierają się na takich samych założeniach, co programy drugiej generacji, posiadają znacznie szersze funkcje i pozwalają istotnie zmieniać bądź wspomagać możliwości metody ręcznej analizy treści[8].

Naszym zdaniem, obecnie należy wyróżniać cztery generacje oprogramowania analizy treści (rys. 4.1). Kryterium takiego podziału powinny stanowić nie tyle ramy czasowe, ile funkcja, jaką pełnią zasoby programowe w procesie analizy. W dalszych rozważaniach szczegółowo przybliżymy generacje programów analizy treści.

Pierwszą generację trudno nazwać programami analizy treści – jest to jedynie wykorzystanie różnorodnych programów do celów realizacji analizy treści. H. Bernard i G. Ryan piszą: *„(...) podobnie do wczesnych edytorów tekstów i systemów sterowania bazami danych, pierwsza generacja procesorów tekstowych była projektowana jako pomoc dla nas w robieniu tego, co my i tak już robiliśmy (...)"*[9]

5 Neuendorf K.A. The Content Analysis Guidebook...– r.40.

6 Mergenthaler E. Computer-Assisted Content Analysis...– r.3-32. 117

7 Macnamara J.R. Media Content Analysis: Its Uses, Benefits & Best Practice Methodology / CARMA International (Asia Pacific).– Chippendale (Australia): CARMA, 2003.– R.8.– www.masscom.com.au/book/papers/media_content.html

8 Kelle U. Computer-Aided Qualitative Data Analysis...– r.34. 118

9 Bernard H.R., Ryan G. Text Analysis: Qualitative and Quantitative Methods...– r.625.

Rys. 4.1. Generacje programów analizy treści

PROGRAMY I GENERACJI

(LATA 50-TE – 60-TE XX W.)

- wąsko wyspecjalizowane programy do wykonywania poszczególnych obliczeń lub programy ogólnego użytkowania (procesory tekstów i arkuszy, systemy zarządzania bazami danych)
- przeznaczone do wykonania obliczeń liczbowych, analizy statystycznej, tworzenia podstawowych wykresów
- możliwość wykonywania poszczególnych funkcji zarządzaniem tekstem (przechowywanie, kopiowanie, tworzenie indeksów)

PROGRAMY II GENERACJI

(1966 R. – POŁOWA LAT 80-TYCH XX W.)

- specjalistyczne programy analizy ilościowej, ograniczone do opracowywania tekstów i te, które nie wychodzą poza granice technologii ręcznej
- główna uwaga skierowana jest na kodowanie, poszukiwanie słów kluczowych i fraz, przedstawianie informacji w formie różnych indeksów, konkordancji
- pozwalają na wykonanie ręcznego, zautomatyzowanego i automatycznego kodowania z wykorzystaniem słowników
- realizowane są różne strategie wyszukiwania, tworzenia tabel konkordancji, wykonywanie analizy kolokacji
- możliwość pracy z tekstami elektronicznymi
- wykorzystywane sporadycznie przez analityków ilościowych

PROGRAMY III GENERACJI

(POŁOWA LAT 80-TYCH – LATA 90-TE XX W.)

- specjalistyczne programy, ograniczone do opracowywania tekstów w ramach technologii jakościowej i ilościowej analizy
- wyróżniają się nadzwyczajną różnorodnością programów
- wychodzą poza granice technologii ręcznej i w znaczący sposób wspomagają możliwości analityczne badacza; program występuje jako swoisty ekspert
- realizowane są funkcje wszystkich etapów badań, zwiększa się ilość wykonywanych funkcji (strukturyzacja danych, wizualizacja wyników, tworzenie i weryfikacja hipotez, formułowanie wniosków i raportów), zachodzą zasadnicze zmiany w realizacji tych funkcji, które były realizowane wcześniej (kodowanie, konkordancja)
- są aktywnie stosowane przez badaczy ilościowych, zwróciły na siebie uwagę badaczy zajmujących się jakością

PROGRAMY IV GENERACJI

(OD POŁOWY LAT 90-TYCH XX W.)

- nie ograniczały się wyłącznie do obróbki tekstów
- to programy technologiczne, podstawą których jest analiza treści i które pracują w trybie czasu rzeczywistego
- realizowane są w formie systemów na dużą skalę, ze skomplikowanymi algorytmami matematycznymi i lingwistycznymi, dla których charakterystyczny jest rozwinięty interfejs graficzny, dostęp do różnorodnych źródeł danych, funkcjonują w konfiguracji klient – serwer

Biorąc pod uwagę możliwości pierwszych uniwersalnych elektronicznych maszyn liczących, badacze „zlecali" im wykonywanie obliczeń, przeprowadzanie analizy statystycznej, tworzenie prymitywnych wykresów już po tym, gdy przeczytane i zakodowane przez ludzi teksty zostały wprowadzone z arkuszy kodowania do maszynnośników. Początkowo wykorzystywane były specjalnie w tym celu napisane programy, nieco później pojawiły się bardziej uniwersalne programy: specjalistyczne pakiety przykładowych programów, systemy zarządzania bazami danych, procesory arkuszy itd.. Jeżeli chodzi o bezpośrednie opracowywanie tekstów, to należy powiedzieć, że programy pierwszej generacji pozwalały na przechowywanie i archiwowanie samych tekstów, tworzenie i drukowanie nieskomplikowanych indeksów oraz konkordancji, wyszukiwanie cytatów i wprowadzanie ich do druku w formacie na wzór tabel KWIC.

W tym miejscu warto zacytować Ł. Fedotową: *„(...) przykład z maszynową obróbką tekstów obrazuje niezwykle ważną dla zrozumienia istoty analizy treści myśl...Analiza treści jako metoda nie posiada magicznych właściwości – nie otrzyma się z niej więcej, niż w nią włożono. Jeśli coś, co jest znaczące, ważne, nadzwyczajne nie zostało przewidziane w procedurze, to nie uzyskamy tego w wyniku analizy, jakkolwiek bardzo skomplikowana i żmudna by ona nie była (...)"*[10].

1.2. The General Inquirer i cechy drugiej generacji programów analizy treści

Pojawienie się drugiej generacji programów skomputeryzowanej analizy treści jest kojarzone z Uniwersytetem Harwardzkim (USA). Grupa naukowców z tej uczelni, pod kierownictwem P. Stone'a, w 1961 roku rozpoczęła opracowywanie zasad systematycznej, skomputeryzowanej analizy treści. Już w 1966r. ukończono pierwszą wersję specjalistycznego programu ilościowej analizy treści tekstów The General Inquirer, który rozwiał tym samym mit, że uniwersalne elektroniczne maszyny liczące mogą być wykorzystywane wyłącznie do analizy statystycznej.

Sprawdzając możliwości programu, autorzy projektu wykorzystali wiele swoich i cudzych badań, stosując aparat kategoryjny poprzednich metod ręcznych. Przeprowadzono szereg analiz tekstów gazet, prac naukowych i publicystycznych, przemówień kandydatów na urząd prezydenta z ramienia Partii Demokratycznej i Partii Republikańskiej USA, dokumentów osobistych (listów, dzienników, autobiografii)[11]

Jeden z autorskich projektów, na którym był testowany system, stanowiło badanie 66 listów samobójców, z których 33 zostały faktycznie napisane przez osoby, które popełniły samobójstwo, pozostałe – przez osoby, które je tylko symulowały. The General Inquirer przeanalizowała teksty i w 91% przypadków wykryła listy rzeczywistych samobójców[12].

Proces analizy w The General Inquirer przebiegał następująco. Na etapie wstępnym kompilowany jest słownik kategorii analizy. Drugim etapem i nowym zadaniem komputera jest kodowanie tekstu, polegające na tym, że system przegląda każdy wyraz i porównuje go ze słownymi formami (kategoriami) słownika. Jeżeli forma słowna zostaje znaleziona, to trafienie dla odpowiedniej formy słownej powiększa się o jeden. W efekcie końcowym otrzymywany jest częstotliwościowy podział kategorii. W zależności od systemu, do danego algorytmu bazowego mogą być wprowadzane nowe zasady, na przykład, w celu obliczenia kontekstu wykorzystania wyrazów, usunięcia ich dwuznacznego rozumienia, wykrycia w tekście nie tylko wyrazów, ale i całych fraz.

Na trzecim etapie program przedstawia wyniki analizy. Mimo, że The General Inquirer jest narzędziem niedialogowym, to w efekcie końcowym prócz zwykłych tabel zawierających dane, uzyskuje się wydruk indeksów i konkordancji. Zazwyczaj informacje są przedstawiane w formacie KWIC (wyrazy kluczowe w kontekście). Ponadto, The General Inquirer daje możliwość przeprowadzenia nieskomplikowanej analizy statystycznej.

10 Ł. N. Fedotowa, Analiza treści – socjologiczna metoda badania środków komunikacji masowej...– s.111.

11 Stone Ph., Dunphy D., Smith M., Ogilvie D. The General Inquirer...

12 Ogilvie D.M., Stone P.J., Schneidman E.S. Some Characteristics of Genuine Versus Simulated Suicide Notes // The General Inquirer: A Computer Approach to Content Analysis / P.J. Stone, D.C.Dunphy, M.S.Smith, D.M.Ogilvie, eds.– Cambridge: M.I.T. Press, 1966.– r.527–535.

W razie potrzeby pozwala on na eksportowanie danych w formacie innych programów (pakietów statystycznych, arkuszy elektronicznych lub programów grafiki użytkowej). Jak zauważają badacze, np. K. Carley[13], podejście P. Stone'a dobrze sprawdza się w przypadku zadań na wzór analizy tematów związanych z wyznaczaniem kategorii analizy, ale jest ono niewystarczające w przypadku rozwiązywania zagadnień, w których istnieje potrzeba odnalezienia współzależności między pojęciami. Tym nie mniej, system stał się prototypem programów skomputeryzowanej analizy treści i był ogromnym osiągnięciem nauk społecznych. Program (oraz jego następcy) zademonstrował szerokie możliwości manipulowania tekstami, ich kodowania, wyróżniania kategorii analizy, poszukiwania odpowiedników itd..

Można zatem stwierdzić, że biorąc pod uwagę funkcjonalność, programy drugiej generacji nie wyszły jednak poza granice logiki technologii ręcznej analizy treści i praktycznie jej nie zmieniły.

Stanowiły one swego rodzaju narzędzie pomocnicze, które jedynie ułatwiało wykonywanie rutynowej pracy, ale należy podkreślić, że analityk wykonywał ją również przed pojawieniem się ww. programów. Zasadniczo dotyczyły one ilościowej analizy treści, jednak można było już wykonywać pewne elementy analizy jakościowej. W centrum uwagi programów znajdowało się kodowanie, a także poszukiwanie wyrazów lub fraz kluczowych oraz przedstawianie wyników poszukiwania w formie wydruku. Doświadczenie nabyte w realizacji powyżej opisanych elementów metody skomputeryzowanej analizy treści miało swoją teoretyczną kontynuację i doprowadziło do pojawienia się nowych odmian analizy treści.

Biorąc pod uwagę zasady i procedury kodowania, wyróżnia się:

- kodowanie ręczne;
- kodowanie zautomatyzowane, w którym komputer pomaga człowiekowi w znajdowaniu odpowiednich kategorii;
- kodowanie automatyczne bez człowieka konieczności interwencji ludzkiej.

W przypadku systemów komputerowych działających w trybie pakietowym, wykluczającym możliwość współpracy człowieka i programu w trakcie procesu, ręczny tryb pracy polegał na wprowadzaniu kodów, przygotowanych przez człowieka. Taki wariant pracy jest charakterystyczny dla programów pierwszej generacji. Programy analizy treści drugiej generacji, w których również niemożliwa była jednoczesna praca człowieka i urządzenia, oprócz ręcznego trybu pracy realizowały automatyczny tryb kodowania, głównie na poziomie wyrazów. Kodowanie zautomatyzowane bazuje na ścisłej współpracy człowieka z komputerem, dlatego z przyczyn czysto technicznych jego realizowanie w programach drugiej generacji nie było możliwe.

Skomputeryzowana analiza treści zaproponowała dwa, zasadniczo różne podejścia do kodowania automatycznego, które umownie nazywane są „a-priori" (lub dedukcyjne) i „a posteriori" (lub indukcyjne)[14]. Podejście „a-priori", zaproponowane przez P. Stone'a przy opracowywaniu The General Inquirer jest bardziej rozpowszechnione. Model analizy treści, realizowany przez podobne systemy, należy do kategorii instrumentalnej analizy treści. Podstawą badań jest w tym przypadku teoria i to ona wyznacza wszystkie części strukturalne: schemat klasyfikacji kategorii analizy, kolejność zasad kodowania tekstów, a także wnioski płynące z badań. Zasadniczo relewantność kategorii bazuje na rozumieniu kontekstu przez analityka, jego zainteresowaniach, intuicji, doświadczeniu i umiejętnościach, a także obranych celach

badania. Warto podkreślić, że w trakcie badania analityk może wprowadzać zmiany w schemacie klasyfikacyjnym – w zależności od nowego, głębszego zrozumienia tekstu po otrzymaniu pierwszych wyników, odnajdowania i poprawiania nieścisłości, błędów itd..

13 Carley K. Formalizing the social expert's knowledge // Sociological Methods and Research.–Vol.17.–1988.– R.165-232; Alexa M. Computer-assisted text analysis methodology in the social sciences / Zentrum für Umfragen Methoden und Analysen (ZUMA). ZUMA-Arbeitsbericht 97/07.– Mannheim (Germany): ZUMA, 1997.– 40 r. http://www.gesis.org/Publikationen/Berichte/ZUMA_Arbeitsberichte/97/97_ 07.pdf

14 Alexa M. Computer-assisted text analysis methodology in the social sciences...– 40 r.

Sformalizowane przedstawianie zasad i warunków kodowania w znacznej mierze realizowane było przez słowniki. *„(...) Już od pierwszych prób wykorzystania elektronicznych maszyn liczących do pracy z tekstem oczywiste stawały się zalety i wady opracowywania materiału tekstowego przy pomocy maszyn – zapewniały one adekwatność analizy dużych partii materiałów tekstowych, wymagały jednak znacznego nakładu sił, by przygotować programy do pracy – sporządzenie „słownika" z uwzględnieniem wszystkich wariantów pojęć synonimicznych, które trzeba będzie wyszukiwać w bezliku wyrazów przepuszczanych przez maszynę (...)*[15].

W praktyce słownik stanowi zbiór kilku tysięcy form słownych, należących do różnych kategorii. Kategorie te tworzą system, wyrażający istotę pewnego problemu, opisujący jakiś temat lub zestaw tematów. Zazwyczaj w słowniku wyznaczane są słowa dla 60-150 kategorii. Kategorie są dobierane w sposób indukcyjny, na podstawie tekstu, bądź dedukcyjnie – na podstawie bardziej ogólnych rozważań teoretycznych, które dyktują wybór kategorii. Dla każdej kategorii w zakresie rozszyfrowywania zadawane są swoiste „nośniki" treści w mowie rzeczywistej – wyrazy we wszystkich swoich formach lub formy słowne. *„(...) Stworzenie takiego słownika jest analogiczne do stworzenia „tezaurusa" – języka danej dziedziny wiedzy ludzkiej, kiedy to słowom kluczowym tejże dziedziny wiedzy odpowiada synonimiczny rząd ogólnie używanych wyrazów (...)"*[16].

Słownik komputerowy w istocie stanowi odmianę elektronicznej książki kodowej (codebook).

Przykładem słownika może być tak zwany Harwardzki Słownik Psychosocjologiczny (The Harvard Psychosocial Dictionary), stworzony specjalnie dla The General Inquirer i wykorzystywany w badaniach socjologicznych. Kategorie słownika formułowane były w sposób dedukcyjny, w oparciu o znane teorie H. Lasswell'a, T. Parsons'a, C. Osgood'a i in. W trakcie opracowywania kategorie były rewidowane i precyzowane. Jedną z najbardziej rozpowszechnionych jest wersja słownika z roku 1975, znana pod nazwą „Harvard IV". Obejmowała ona 182 kategorie, które pierwotnie zawierały 4000 form słownych (wejść do słownika), a według stanu na rok 1990 – 8500.

Największą pod względem ilości wejść jest kategoria „negative" (negatyw), na którą składa się 2291 form słownych. W odróżnieniu od swoich poprzedników, „Harvard IV" uwzględnia wieloznaczność wyrazów. Ponadto, użytkownicy mogą samodzielnie wprowadzać nowe kategorie z dowolną ilością odpowiednich form słownych. Dobrze znane są również następujące słowniki:

- Słownik Wartości Lasswella (The Lasswell Value Dictionary),
- Słownik Obrazów Regresywnych (The Regressive Imagery Dictionary),
- Stanfordzki Słownik Polityczny.

Przykładowo – Słownik Stanfordzki zawiera 6584 wyrazy o zabarwieniu emocjonalnym. Zostały w nim wydzielone wyrazy, charakteryzujące odczucia pozytywne (977 wyrazów), odczucia negatywne (1513 wyrazów), siłę (1391 wyrazów), słabość (579 wyrazów), aktywność (1218 wyrazów), pasywność (722 wyrazy). Oceniając każdy wyraz według intensywności emocjonalnej grupy ekspertów przeprowadzały „ważenie" według skali trybalnej. W ten sposób skonstruowano trójwymiarowe „pole semantyczne" obejmujące:

- odczucia pozytywne-negatywne;
- siłę-słabość;
- aktywność-pasywność[17].

Oprócz słowników ogólnych, często wykorzystywane są również słowniki specjalistyczne – stworzone dla konkretnej, jasno określonego zagadnienia. Za przykład może posłużyć Słownik Tematyczny Stanu Niepokoju (the Anxiety Theme Dictio-

15 Fedotowa Ł.N. Analiza treści – metoda socjologiczna badania środków komunikacji masowej...– s.108.

16 Fedotowa Ł.N. Analiza treści – metoda socjologiczna badania środków komunikacji masowej...– s.107.

17 Igołkin A. A., Media w warunkach globalizacji...– s.65.

nary) opracowany na Uniwersytecie w Ulm [18]. Innym przykładem słownika specjalistycznego jest Słownik Ikara, stworzony przez D. Ogilvy'ego w celu sprawdzenia hipotezy o istnieniu tzw. „kompleksu Ikara" – zjawiska psychologicznego, które odkryli psychologowie podczas obserwacji swoich pacjentów. D. Ogilvy weryfikował hipotezę analizując materiały tekstowe powstałe w różnych kulturach – wybrał 626 bajek pochodzących od 44 społeczeństw prymitywnych.

W celu przeprowadzenia analizy treści stworzony został słownik, który obejmował 74 kategorii i 2500 wyrazów. W trakcie opracowywania systemu kategorii wykorzystywano różne źródła: prace badaczy opisujące kompleks Ikara, historie osób, u których stwierdzono ten kompleks, notatki z obserwacji pacjentów i in.. Ponieważ pierwotnie kompleks Ikara był określany jako nastawienie na „lot, spadanie, ogień, wodę, nieśmiertelność i narcyzm", to hipotezę o istnieniu kompleksu na poziomie analizowanych bajek można było, zdaniem autora, uważać za udowodnioną, jeśli te tematy „są związane" z konkretnymi jednostkami folklorystycznymi lub na pewnym etapie pojawia się możliwość przewidzenia pojawienia się w tekście tematów pokrewnych. Wyniki analizy otrzymywano po obliczeniu korelacji pomiędzy 74 kategoriami i na skutek przeprowadzenia analizy faktorowej macierzy korelacyjnej[19].

Automatyczny wariant kodowania, wykorzystujący elementy sztucznej inteligencji, różni się od poprzedniego wariantu, m.in. sposobami, przy pomocy których rozwiązywane są problemy niejednoznaczności wyrazów. Systemy komputerowe ze sztuczną inteligencją próbują uwzględniać zarówno składnię, jak i leksykalne cechy wyrazów. Programy tej kategorii, np. późniejsze wersje The General Inquirer ze słownikiem „Harvard IV", zawierają reguły dzięki którym wśród różnych wariantów można uzyskać to znaczenie wyrazu, które odpowiada kontekstowi. Wraz z pojawieniem się programów analizy treści powstał jeszcze jeden problem: co właściwie należy kodować w tekście?

Czy kodować wszystkie wyrazy czy nie? Ze względu na powyższe zagadnienia wyodrębniono dwa kierunki analizy treści: wybiórczy i „totalny". Pierwszy, realizowany przez P. Stone'a, przewidywał kodowanie w tekście nie wszystkich wyrazów, a tylko nieznacznej części, wyrażającej istotę badanego problemu. Do swojego słownika P. Stone włączył 5 tys. wyrazów, jakie najczęściej występują w codziennym użyciu[20].

J. Laffal wyodrębnił i nazwał nowy rodzaj analizy treści, który otrzymał nazwę „totalny"[21]. Badacz podjął próby kodowania niemal każdego wyrazu tekstu, z wyjątkiem wyrazów funkcyjnych, najczęściej występujących w mowie – proponowana analiza obejmowała szeroki zakres rzeczowników, przymiotników, czasowników i in. Stworzony przez Laffal'a słownik początkowo zawierał 114 kategorii. Według informacji podawanych przez H. Bernarda i G. Ryana, zgodnie ze stanem z początku lat 90-tych XX w., słownik Laffala zawierał 43 tys. wyrazów, z których każdy skojarzony był z 1 - 5 kategoriami ze 168 możliwych[22]. Dla porównania: analiza proponowana przez P. Stone'a, wykorzystywała do kodowania około 10% tekstu, natomiast według J. Laffal'a pokrycie tekstu kategoriami wynosiło prawie 90%. Warto zauważyć, że „totalna" analiza treści znalazła najszersze zastosowanie w psychoterapii, gdzie główny nacisk kładziony jest na badanie mowy pacjentów. Alternatywnym do systemów, stworzonych na wzór P. Stone'a, ale bardziej „zaawansowanym", okazała się odmiana systemu zautomatyzowana analiza treści „a posteriori", która nie potrzebowała wstępnego tworzenia słownika, eliminując w ten sposób człowieka konieczność ingerencji człowieka nawet na poziomie formułowania kategorii. Tę odmianę analizy skomputeryzowanej H.P. Iker i N.I. Harway jeszcze w końcu lat 60-tych XX w. nazywali analizą, której *„(...) nie dotyka ludzka ręka (...)"*[23]. W odróżnieniu od podejścia „a priori", „kieruje się" ona najpierw danymi, a nie bliżej nieokreśloną teorią. W danym przypadku schemat kategorii analizy jest formułowany w wyniku przegląd-

18 Grünzig H. J., Themes of anxiety as psychotherapeutic process variables //Methodology in psychotherapeutic research / Minsel W.R., W.Herff (eds.). Proceedings of the 1st European Conference on Psychotherapy Research, Trier, 1981. – Lang, Frankfurt, 1983.– r.135-142.

19 Fedotowa Ł. N. Analiza treści – metoda socjologiczna badania środków komunikacji masowej...– s.108.

20 Mergenthaler E. Computer-Assisted Content Analysis...– r.8.

21 Laffal J. Total or Selected Content Analysis // International Conference on Computational Linguistics, Sínga-Síby-Kuragírd, Sweden, 1-4 September, 1969. – acl.ldc.upenn.edu/C/C69/C69-2401.pdf

22 Bernard H.R., Ryan G. Text Analysis: Qualitative and Quantitative Methods...–r.629.

23 Iker H.P., Harway N.I. A Computer Systems Approach Toward the Recognition and Analysis of Content // The Analysis of Communication Content / Gerbner G.A. et al. (eds.).– Wiley & Sons, 1969.

du badanego tekstu. Przy tym, od analityka na początku pracy nie są wymagane żadne dodatkowe informacje, wystarczy dostępność badanego tekstu. Faktycznie takie systemy realizują reprezentacyjny model analizy treści.

Obecnie, jak twierdzi P. Mochler i C.Zuell, automatyczne systemy takiego rodzaju analizy treści przeżywają okres renesansu[24]. Ich przykładem mogą być programy The Words, TextSmart, DICTION. System pracy The Words, opracowany przez H.P. Ikera i N.I. Hatway'a pod koniec lat 60-tych XX w., jest następujący. Początkowo tekst jest dzielony na odrębne segmenty, dla których tworzona jest tabela częstotliwości występowania wszystkich wyrazów, z wyjątkiem funkcjonalnych i z uwzględnieniem synonimów. Dla każdego segmentu wybiera się n wyrazów o najwyższej częstotliwości, które tworzą n mini kategorii. Następnie, na podstawie wszystkich segmentów, obliczana jest macierz wzajemnych korelacji pomiędzy tymi kategoriami, która poddawana jest analizie faktorowej. W wyniku tego określane są aktualne lub nieaktualne mini kategorie (lub tematy) tekstu.

Pragniemy podkreślić, że kodowanie automatyczne zapewnia większą dokładność, pewność i oszczędność zasobów w porównaniu z ręcznym. Jak zauważa M. Alexa, *„(...) przy kodowaniu ręcznym, kiedy to nie są jasno określone zasady kodowania, nie da się uniknąć rozbieżności pomiędzy sposobem myślenia różnych osób kodujących. Wymóg, dotyczący jawnego, jednowariantowego i dokładnego wyboru kategorii jest krytyczny w procesie kodowania. Pod tym względem kodowanie automatyczne ma niewątpliwe zalety w porównaniu z ręcznym. Kodowanie automatyczne jest wykonywane na podstawie jawnie sformułowanych reguł i jednoznacznych warunków (...)"*[25]. Głównym problemem kodowania automatycznego jest dwuznaczność wyrazów, wykorzystywanie metafor, zaimków, synonimów itd.. Innym, wyróżniającym składnikiem programów drugiej generacji jest realizacja strategii wyszukiwania danych. Wśród nich można wyróżnić tworzenie tabel konkordancji. O ich ważności wyraźnie świadczy chociażby fakt, że w samych nazwach wielu programów obecny jest wyraz „konkordancja". Przykładowo, program COCOA (Count and Concordance generation for the Atlas) jest częścią składową programu The Atlas. W roku 1978 Centrum Komputerowe Uniwersytetu Oksfordzkiego zamieniając COCOA stworzył OCP (The Oxford Concordance Program), a później Micro-OCP – dla mikrokomputerów. Dobrze znany system TACT (Text-Analysis and Concordance Tools) w swej nazwie również ma wyraz „konkordancja".

Głównym celem konkordancji jest skierowanie uwagi na bezpośrednie środowisko lingwistyczne wybranego wyrazu. Logika wyszukiwania polega na tym, że na początku badacz ustala potencjalnie interesujący wyraz, następnie znajduje odpowiadającą mu konkordancję, co daje możliwość wyznaczenia szablonów (patternów), charakterystycznych dla tego wyrazu i w których temu wyrazowi dedykowana jest w całości pełniona funkcja.

Istnieje kilka formatów konkordancji. Jednym ze sposobów przedstawienia kontekstu występowania wyrazów jest format KWOC (keyword-out-of-context – słowo kluczowe poza kontekstem) – wykaz wyrazów z określeniem jego miejsca. W danym przypadku słowo kluczowe jest przedstawiane z prawej lub z lewej strony kontekstu, a kontekst to forma całego zdania, które może zajmować kilka wersów. Bardziej rozszerzonym jest alternatywny format – KWIC (keyword-in-context – słowo kluczowe w kontekście), który zajmuje tylko jeden wers ze słowem kluczowym w środku tegoż wersu (jednakowa ilość wyrazów z prawej i z lewej strony słowa kluczowego). Wraz z konkordancją wykonuje się kolokację (collocate analysis) – analizę statystyczną występowania kombinacji wyrazów. Jak już mówiliśmy, kolokacja ma za zadanie wyznaczenie wyrazów, które występują w pobliżu zadanego wyrazu centralnego. Po wyborze wyrazu centralnego lub grupy takich wyrazów, połączonych wspólną ideą lub wspólnym obiektem, określane są wszystkie wyrazy, znajdujące się z lewej i z prawej strony wyrazów centralnych w granicach zadanej odległości. Ta odległość w sposób istotny zależy od języka, przykładowo – empirycznie udowod-

24 Mohler P., Zuell C. A Popperian Critique of Automatic Content Analysis // Journées Internationales d'Analyse Statistique des Données Textuelles / ZUMA.– 2000.– No 5.– http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/10/10.pdf

25 Alexa M. Computer-assisted text analysis methodology in the social sciences...– 40 r.

niono, że dla języka angielskiego znaczące są związki w granicach pięciu wyrazów z lewej i z prawej wyrazu centralnego. Opracowany w ten sposób kontekst tworzy mini tekst. Następnie przeprowadza się analizę częstotliwości występowania wyrazów w takich mini tekstach. Różnica pomiędzy oczekiwaną częstotliwością występowania tych czy innych wyrazów, a rzeczywistą częstotliwością występowania stanowi miarę „przyciągania" wyrazów przez centrum.

Potrzeba komputerowego opracowywania zrodziła konieczność stworzenia tekstów elektronicznych, które mogą w nieskomplikowany sposób i automatycznie być analizowane przez komputer. W latach 70-tych XX w. nie tylko zostały opracowane metody i propozycje teoretyczne dotyczące tworzenia archiwów tekstowych (np.[26]), ale również rozpoczęto praktyczną realizację tych idei. Do najstarszych i najbardziej znanych należy Oksfordzkie Archiwum Tekstowe, które obecnie zawiera obszerne, wysokiej jakości, dobrze udokumentowane zbiory literatury greckiej, łacińskiej i angielskiej. Od 1976 r. Archiwum Oksfordzkie zajmuje się rozpowszechnianiem tekstów w formie elektronicznej (zapoznać się z możliwościami archiwum można na cytowanej stronie internetowej[27]).

Archiwa elektroniczne zaczęto tworzyć w wielu uniwersytetach, urzędach, instytucjach. Wskutek tego dziś, w odróżnieniu od lat 60-tych i 70-tych XX w., kiedy nie było zbyt wielu dostępnych tekstów w formie elektronicznej a ich tworzenie było skomplikowanym i żmudnym procesem, dostęp do tekstów maszynowych stał się o wiele prostszy. Obecnie, istnieje bardzo wiele elektronicznych archiwów tekstowych z różnych źródeł, różnego przeznaczenia i w różnych językach. Dzisiejsze całotekstowe bazy danych zawierają nie tylko elektroniczne kopie oryginałów pierwszych źródeł (zeskanowane teksty, rękopisy, rysunki, audio-, video-, tele-, foto-obrazy), a także, niezależne od nich, zasoby interpretacji.

Szybki dostęp w trybie on-line umożliwia analitykom proste i nieskomplikowane wykorzystanie tekstów, w tym również przeprowadzenie analizy treści za pomocą komputerów.

Dostępność tekstów elektronicznych wpłynęła na opracowanie programów analizy treści. *„(...) Odtąd problemy, związane ze składaniem i edycją tekstów przestały być dominujące, zamiast tego badacze skupiają się na nowych drogach wykorzystania komputera do lepszego rozumienia tekstów; uzupełnieniem zwykłego wyszukiwania wyrazów i fraz w tekście jest możliwość szerszego wykorzystania wielowymiarowych metod statystycznych, co pozwala uzyskać więcej, niż zwykłe znalezienie czegokolwiek w tekście"*[28].

Na zakończenie pragniemy dodać, że możliwości programów analizy treści drugiej generacji zwróciły uwagę przede wszystkim tych badaczy, którzy zajmowali się dziedziną sformalizowanych metod badań tekstów. Natomiast badacze jakościowi nawet nie podejmowali prób wykorzystania ich w swojej pracy analitycznej. Tym nie mniej, liczne badania treści, z reguły, źródeł medialnych i głównie tekstów anglojęzycznych, pozwoliły wypracować metodologiczne zasady nowego podejścia do badań empirycznych w ramach nauk społecznych. Ł. N. Fedotowa zauważa, że w 1974 r. podczas roboczego spotkania dot. problematyki analizy treści we Włoszech przedstawiono nawet kilka projektów międzynarodowych, opracowanych z wykorzystaniem elektronicznych maszyn liczących, m.in. projekt międzynarodowego badania nagłówków gazetowych w celu określenia uwagi skierowanej przez drukowane media na wydarzenia lokalne, krajowe i światowe, porównywanie uwagi, skierowanej przez gazety amerykańskie i europejskie na problemy „Wspólnego Rynku", porównywanie relacji z prowadzonych działań wojennych w Nigerii przez gazety różnych krajów i in..[29]

Aktywność badaczy ww. zakresie była tak wysoka, że wielu naukowców okres lat 60- tych i 70-tych XX w. nazywa „złotym wiekiem" skomputeryzowanej analizy treści. Jednak optymizm i wysiłki włożone w skomputeryzowaną analizę treści w latach

26 DeWeese L.C. Computer content analysis of day-old newspapers: a feasibility study // Public Opinion Quarterly.– Vol. 41.– 1977.– PP.91-94; DeWeese L.C. Computer content analysis of printed media: a limited feasibility study // Public Opinion Quarterly.– Vol. 40.– 1976.– r.92-100.

27 The Oxford Text Archive.– http://ota.ahds.ac.uk/ota/index.html

28 Bradley J., Rockwell G. TACTweb: workbook...

29 Analiza treści – metoda socjologiczna badania środków komunikacji masowej...– s.110.

60- tych w latach 70-tych się nie zwiększyły. Głównie to wiąże się, zdaniem M. Alexa, z powolnym rozwojem techniki obliczeniowej, ograniczonym dostępem do elektronicznych maszyn liczących (dostęp jedynie w obrębie centrów obliczeniowych), a także brakiem wystarczającej bazy tekstów elektronicznych, nie mówiąc już o trudnościach związanych z przetwarzaniem w format maszynowy tekstów mówionych[30]. Zatem, zmniejszyła się ilość publikacji naukowych, badania teoretyczne nie rozwijały się, nastąpił zastój metodologiczny.

Co prawda, właśnie w tym dziesięcioleciu skomputeryzowaną analizę treści zaczęto stosować w psychologii i psychoterapii, bardziej rozpowszechniła się również w Europie. Taka sytuacja trwała aż do połowy lat 80-tych XX w.

1.3. Trzecia generacja programów analizy treści

Od połowy lat 80-tych XX w. obserwuje się znaczny postęp w rozwoju skomputeryzowanej analizy treści. Stymulowało go kilka przejść: w latach 80-tych – z wielkich elektronicznych maszyn liczących do komputerów personalnych, a następnie – z systemu operacyjnego MS DOS do MS Windows z jego interfejsem graficznym i przyjaznym środowiskiem dla użytkownika. Owe zmiany techniczne i technologiczne stworzyły możliwości do zapewnienia swobodnej współpracy człowieka z komputerem, który stał się rdzeniem całej pracy doświadczalnej, a nie jedynie administrowania danych tekstowych. Organicznym dopełnieniem był burzliwy rozwój Internetu, szerokie rozpowszechnienie elektronicznych archiwów/bibliotek, ogólna dostępność tekstów elektronicznych i możliwość dostępu do archiwów tekstowych poprzez Internet w trybie on-line.

Zaznaczmy od razu, że właśnie trzecia generacja programów analizy treści zwróciła uwagę badaczy jakościowych, którzy dostrzegli w nich ogromne, niezrealizowane dotąd możliwości niesformalizowanej obróbki tekstów. *„(...) Stało się jasne, że pomimo braku możliwości wykorzystania komputerów bezpośrednio do analizy interpretacyjnej tekstów, jednak mogą one, niewątpliwie, służyć znaczną pomocą w procesie interpretacji (...)”*[31].

Stało się to szczególnie jasne, gdy zaistniała potrzeba opracowywania znacznej ilości niestrukturalnych danych tekstowych. Nie zważano nawet na to, że wykorzystanie programów komputerowych w badaniach jakościowych jest bardzo ściśle związane ze specyfiką tego rodzaju badań. Oprogramowanie trzeciej generacji cechuje nadzwyczajna różnorodność. Wśród najbardziej znanych wówczas systemów były Atlas/ti, HyperResearch, Aquad, NUD'IST.

Mimo, że nową generację programów utworzono na tych samych zasadach, co ich poprzednicy, jednak pozwalały one na wykonywanie analizy treści, wychodzącej poza granice technologii ręcznej i znacznie wspierały możliwości analityczne badacza.

Zorientowane na badania jakościowe i ilościowe programy trzeciej generacji posiadają narzędzia do:

- formowania tekstów i stworzenia na ich podstawie całych projektów;
- badania częstotliwości i kontekstu wykorzystania wyrazów, m.in. udzielając odpowiedzi na pytania: jak często kategorie są przydzielane wyrazom lub segmentom tekstowym? jakie kategorie i jak często pojawiają się one razem? jakie związki istnieją pomiędzy kategoriami lub segmentami tekstowymi;
- tworzenia i wsparcia kategorii i schematów klasyfikacji;
- przydzielenia jednej lub więcej kategorii do rzędów symboli, wyrazów, fraz, zdań, paragrafów lub całych tekstów;
- przechowywania uwag („memo”) do tekstów, kodowania segmentów tekstów;
- otrzymywania różnych formatów przeglądu tekstów, części tekstów lub grup tekstów;
- eksportowania kodów do ich dalszej obróbki przez inne programy, a także formułowania raportów z przeprowadzonej analizy;

30 Alexa M. Computer-assisted text analysis methodology in the social sciences...– r.8.

31 Kelle U. Computer-Aided Qualitative Data Analysis...– r.36.

Tabela 4.1 Wykorzystanie kodów jako wskaźników segmentów tekstów

Numer dokumentu	Nazwa kodu segmentu tekstu	Nr pierwszego wersu segmentu w dokumencie	Nr ostatniego wersu segmentu w dokumencie
1	CZR	205	219
1	MGR	45	87
2	CZR	50	54
2	MGR	43	67
2	CZR	156	190

- wsparcia pracy zespołowej lub wspólnej w ramach projektu i połączenia w jeden kilku projektów[32].

Jeżeli chodzi o jakościową analizę treści, to w ramach programów realizowany był szereg wspominanych przez nas metod ręcznych, ale również i nowych możliwości. Funkcje programów analizy treści jakościowej można zgrupować w następujący sposób:

- zarządzanie danymi tekstowymi i towarzyszącymi im informacjami;
- systematyczna ocena fragmentów tekstu, na jej podstawie stworzenie kategorii analizy, a także budowanie związków pomiędzy kategoriami analizy;
- wyszukiwanie fragmentów tekstu na podstawie kryteriów ogólnych;
- weryfikacja hipotez interpretujących.

Pierwszym zadaniem postawionym przed programami jakościowej analizy treści, było zarządzanie niestrukturalnymi tekstowymi bazami danych. Wykorzystanie tradycyjnych systemów zarządzania bazami danych nie zapewniało potrzebnych operacji. Idea programów jakościowej analizy treści polegała na tym, że oprócz bazy danych zawierającej tekst podstawowy, tworzone były specjalne pliki lub bazy danych, zawierające adresy segmentów tekstu (np. numer zapisu początkowego i końcowego) oraz nazwy kodów skojarzonych z danym segmentem. W tab. 4.1 przedstawiony jest przykład takiego pliku.

Jeśli kodem CZR oznaczono kategorię „Republika Czeska" a MGR – „migracje międzynarodowe", to, jak widać z tabeli, w obu dokumentach tematy te są obecne. Między innymi kategoria „Republika Czeska" w pierwszym dokumencie występuje we fragmencie, zaczynającym się od 205 wersu, a kończy w wersie 219. Przy czym, w drugim dokumencie mowa jest o migracjach międzynarodowych właśnie w Republice Czeskiej, co świadczy o tym, że oba kody są obecne we fragmentach, mających wspólne wersy. Przy pomocy tego pliku można utworzyć osobno te segmenty tekstu, w których obecne są potrzebne kody.

Taki plik pomocniczy był wykorzystywany w celu wyszukania i wyróżnienia potrzebnych fragmentów tekstu. Uzupełniając te pliki nowymi zapisami, można ciągle rozszerzać bazę wyszukiwania, nie ingerując w sam tekst. Owa zasada obowiązywała w pierwszych pakietach programów Qualpro, The Ethnograph, Textbase Alpha. Rozwiązanie pierwszego zadania pozwoliło bez szczególnych problemów realizować pozostałe funkcje analizy: wyszukiwanie fragmentów tekstów; tworzenie konkordancji; wykonywanie kolokacji; sprawdzanie hipotez interpretujących drogą wyszukiwania segmentów z jednakowymi kodami; wprowadzanie, edycja i przechowywanie komentarzy teoretycznych do fragmentów tekstów itd.. Istotnym ich uzupełnieniem stały się różnorodne przedstawienia, w drodze zróżnicowanych sposobów wizualizacji, kategorii powiązanych. Zobrazowanie teorii w formie sieci powiązanych kategorii jest wspaniałym narzędziem do prezentacji struktury nowopowstałej teorii. Dzięki plikom pomocniczym wskaźników można prosto ustalać powiązania pomiędzy segmentami tekstów,

32 Alexa M., Zuell C. Commonalities, differences and limitations of text analysis Software: The results of a review / Zentrum für Umfragen Methoden und Analysen (ZUMA). ZUMA-Arbeitsbericht 99/06.– Mannheim (Germany): ZUMA, 1999.– r.2. http://www.gesis.org/Publikationen/Berichte/ZUMA_Arbeitsberichte/99/99_06.pdf

memo, kodami. Jako przykład posłużą nam chociażby: program NUD'IST, który pozwala tworzyć hierarchiczne i sieciowe struktury kategorii czy program Atlas/ti formujący różne, niehierarchiczne sieci. Szczegółowe informacje, dot. realizacji zaznaczonych algorytmów, można znaleźć w publikacji pod przypisem [33].

Zmiany zaszły we wszystkich etapach technologii badań. Przede wszystkim związane były z kodowaniem. Integracja kodowania ręcznego i automatycznego stała się nową technologią wielu programów, na przykład: PLCA (Program for Linguistic Content Analysis), MECA (Map Extraction, Comparison and Analysis). Od tego momentu stało się jasne, że tradycyjna różnica pomiędzy automatyczną i ręczną analizą ulegnie zatarciu. *„(...) Wsparcie lub jego brak dla przyjaznej i dostępnej technologii przestały ograniczać badacza, nie wymuszają korzystania z jedynego trybu analizy (...)"*[34]. Na połączeniu obu podejść najbardziej zyskuje analityk.

Zmieniła się taka funkcja, jak wykorzystanie konkordancji. Nabyła ona cech „dialogowych".

Pomimo, że częstotliwość występowania wyrazów nadal stanowiła dominujące źródło informacji i służyła jako jeden ze wskaźników potencjalnego zainteresowania kategorią analizy lub tematu, jednak odtąd, mając na ekranie wyrazy (kategorie) i ich częstotliwość, badacz miał możliwość operatywnego przeglądania kontekstu występowania danego wyrazu przy pomocy tabeli KWIC. Procedura ta zmieniła możliwości wykorzystania konkordancji, co wzmocniło siłę dowodu wniosków, podstawy których były wykonywane przez system. Taki tryb – poprzez system powiązanych okien jest stosowany w wielu programach, m.in. nawet w MS DOS w programie TACT.

Zwiększyła się ilość nowych funkcji programów. Pojawiła się możliwość weryfikacji hipotez, poszerzyły się możliwości interpretacyjne na skutek metodyki wyszukiwania wspólnego wchodzenia kodów (wyrazów), stworzenia modeli konceptualnych, łączących pojęcia w sieci semantyczne, analizy macierzy logicznej i kartograficznej. Wiele programów skupia uwagę na podliczeniach wskaźników pewności, na przykład, programy AGREE, Krippendorf's Ralpha 3.12a, PRAM (Program for Reliability Assessment of Multiple Coders) względnie poszczególne moduły pakietów programów statystycznych, m.in. SPSS i Simstat (patrz:[35]).

Powyższe możliwości stanowiły bazę dla utworzenia całego szeregu systemów programowych, które pozwalały na rozwiązywanie najróżniejszych specyficznych problemów w procesie analizy. Przykładowo – zbudowany na teorii „concept map ping" program the VBPro pozwala, drogą przedstawienia kartograficznego, na identyfikację dominujących tematów ich powiązań dla dużej ilości danych. Ten rodzaj analizy tekstowej jest wykorzystywany do analizy informacji medialnych. Inny rodzaj analizy przedstawia program Minnesota Contextual Content analysis (MCCA). Umożliwia on zmierzenie różnicy społecznej (odległości) pomiędzy statusami osób w danej instytucji, na przykład: lekarzami a pacjentami w szpitalu, menedżerami a innymi pracownikami firmy, uwzględniając stylistyczne szczegóły języka w procesie mowy, a także informacje kontekstualne.

W odróżnieniu od standardowej analizy treści, ta metoda uwzględnia cztery wymiary kontekstualne: tradycyjny, praktyczny, emocjonalny i analityczny. Wykonując analizę klasterową, system pozwala ocenić w sposób ilościowy stopień bliskości pomiędzy przedstawicielami różnych grup społecznych. Jednocześnie pojawiły się systemy analizy treści umożliwiające efektywną pracę w poszczególnych zakresach. Do systemów nowocześniejszych można zaliczyć programy służące do badania mediów, na przykład CARMA® (Computer Aided Research & Media Analysis), Precis™, Echo®Research, IMPACT™, Metrica, the Delahaye Medialink system. Oprócz wykonania samej analizy treści, programy te zawierają takie moduły, jak wbudowane media – bazy danych, zapewniające wprowadzanie, dostęp i podliczanie odpowiednich danych statystycznych, doty-

33 Kelle U. Computer-Aided Qualitative Data Analysis...– r.33-63.

34 Alexa M. Computer-assisted text analysis methodology in the social sciences...– r.9.

35 Lombard M., Snyder-Duch J., Bracken C.C. Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research Projects. 2004.– http://www.temple.edu/mmc/reliability/

cących różnych parametrów media-zasobów[36]. Z charakterystykami całego szeregu systemów analizy treści można zapoznać się ze źródeł przytoczonych w przypisie[37].

Nie zważając na oczywiste sukcesy, krytyka dot. skomputeryzowanej analizy treści praktycznie nie cichła. Główny jej kierunek wiąże się z tym, co J. Seidel nazywa *„ciemną stroną postępu technicznego"* [38]. Zarzuty są następujące:

- program dystansuje człowieka od samych danych, człowiek praktycznie nie odczuwa istoty algorytmu;
- wykorzystanie programów prowadzi do tego, że dane jakościowe będą analizowane ilościowo;
- wykorzystanie programów doprowadzi do zwiększenia jednorodności (jednostajności) w metodach analizy, co szczególnie negatywnie odbije się na badaniach jakościowych[39].

Do powyższych argumentów można dodać fakt, że komputer jedynie identyfikuje wyrazy, a od identyfikacji wyrazów do identyfikacji idei, które te wyrazy niosą, jest bardzo daleko.

Szczególną wagę zyskały problemy zapewnienia zgodności i pewności wyników:

- trudności w uwzględnieniu kontekstu;
- niezdolność programu do odróżnienia komunikacyjnych intencji wyrazów;
- niezdolność zapewnienia przez badacza wyczerpującego wprowadzenia na listę słów kluczowych dla pewnych kategorii;
- brak możliwości rozwiązania problemów odsyłaczy przed albo po wyrazach, występujących w dowolnym miejscu tekstu, zwłaszcza w odniesieniu do zaimków;
- niezdolność do wyznaczania przez programy granic jednostek analizy, przede wszystkim, w przypadku analizy jakościowej;
- na skutek określania licznych charakterystyk obliczanych przez programy, może zgubić się istota kategorii[40].

Zwraca się uwagę również na inne ograniczenia, m. in. tzw. „koszty" komputeryzacji, pod którymi R. Morris rozumie:

- koszt nabycia komputera z odpowiednią do pracy konfiguracją;
- czas i wysiłek wymagane do opanowania działania programów do analizy;
- czas i wysiłek zużyte na stworzenie słowników kategorii analizy;
- czas i wysiłek zużyte na testowanie, przegląd i potwierdzenie kategorii.

Odpierając powyższe argumenty zwolennicy skomputeryzowanej analizy treściukazują mocne strony tej dziedziny. Przede wszystkim fakt, że dzięki komputerowi analiza treści stała się dla badaczy zdecydowanie bardziej dostępna, niż kiedykolwiek, a możliwości analityka znacznie się zwiększyły. W związku z powyższym R. Morris pisze: „(...) choć komputery samodzielnie nie mogą podejmować decyzji, jednak ich zdolność do manipulowania tekstem poprzez wykorzystanie programów analizy treści umożliwia badaczowi spojrzenie na dane z innego punktu widzenia, w efekcie – do wyciągnięcia wniosków, dotyczących treści komunikacji (...)"[41]. Już sama możliwość zbudowania w procesie dialogu schematu konceptualnego analizy jest niezwykle istotna, ponieważ proces rekursywny przygotowania, uściślenia i zniszczenia zbędnych kategorii włącza badacza do procesu poznawczego, sprzyja lepszemu zrozumieniu i uświadomieniu przez niego istoty fragmentu rzeczywistości przedstawionej w tekście.

Można zatem stwierdzić, że programy analizy treści pełnią funkcję nie tylko jednego z narzędzi analizy, ale stają się peł-

36 Macnamara J. R. Media Content Analysis...– r.8.

37 Macnamara J. R. Media Content Analysis...– 24 r.; Alexa M. Computerassisted text analysis methodology in the social sciences...– 40 r.; Alexa M., Zuell C. Commonalities, differences and limitations of text analysis Software...– 29 r.

38 Seidel J. Method and Madness in the Application of Computer Technology to Qualitative Data Analysis // Using Computers in Qualitative Research / N. Fielding, R. M. Lee (editors).– London: Sage, 1991.

39 Barry C.A. Choosing Qualitative Data Analysis Software: Atlas/ti and Nudist Compared // Sociological Research Online.– 1998.– Vol. 3.– No.3.– http://www.socresonline.org.uk/socresonline/3/3/4.html (2004.05.14)

40 Morris R. Computerized content analysis in management research: a demonstration of advantages & limitations...

41 Morris R. Computerized content analysis in management research: a demonstration of advantages & limitations...

nowartościowym ekspertem, z racjami którego warto się liczyć. Takie działania, jak administrowanie tekstów, szybkie wykonywanie obliczeń, szybkie i wygodne wyszukiwanie i klasyfikacja danych, formułowanie i wyprowadzenie na zamówienie w zadanej formie wyników badań itd., zaliczają je do narzędzi analizy. Możliwości interpretacyjne, graficzne i statystyczne, zdolność do formułowania hipotez i przygotowywania wariantów wniosków zamieniają programy analizy treści w liczącego się eksperta, proponującego obiektywną ocenę wyników i wprowadzającego warianty rozwiązań. Ta zaleta uwidacznia się szczególnie w przypadku bardzo dużych zbiorach tekstów, kiedy wyniki są otrzymywane nie tylko prędko i w granicach dopuszczalnego budżetu, ale dodatkowo otrzymuje się wnioski i uogólnienia, których człowiek na taką skalę nie jest w stanie zrealizować, a pewność których gwarantują udoskonalone zasady kodowania.

W związku z tym symboliczne staje się utworzenie w różnych krajach całego szeregu ośrodków naukowo-badawczych, specjalizujących się w komputerowej analizie tekstów.

Oprócz wcześniej wymienionych, można do nich zaliczyć uniwersyteckie ośrodki naukowe Centre for Computer Assisted Qualitative Data Analysis Software (Surrey, Wielka Brytania), Centre for Social Anthropology and Computers (Kent, Wielka Brytania), dobrze znane ośrodki ZUMA – Zentrum für Umfragen Methoden und Analysen (Manheim, Niemcy), Qualitative Solutions and Research (La Trobe, Australia).

1.4. Współczesne technologie analizy treści i właściwości Text Mining

Pierwsze programy ilościowej analizy treści skupiały uwagę głównie na podliczeniu częstotliwości występowania pewnych charakterystyk tekstu. Większość współczesnych programów analizy treści również ograniczona jest koniecznością opracowywania tekstu, ale ich możliwości są o wiele szersze:

- przechowywanie danych i zarządzanie nimi (pozwala na przechowywanie w formie tekstowej lub w specjalnym formacie źródeł, a także przechowywanie różnorodnych materiałów audiowizualnych, jak: zdjęcia, diagramy, nagrania video i audio, łączność ze stronami internetowymi; dodatkowo analityk ma możliwość sporządzania, edytowania tekstów, automatycznej indeksacji i zapisywania własnych informacji);
- poszukiwanie danych (programy umożliwiają poszukiwanie danych tekstowych według wskazanych słów czy fraz, podliczanie częstotliwości występowania odpowiednich wyrazów, poszukiwanie informacji według zadanego kontekstu, a także różnych danych dodatkowych, np. daty, osoby przeprowadzającej wywiad, źródła danych itd.);
- kodowanie (proces kodowania staje się stosunkowo prosty; istnieje możliwość nadawania ważności pewnym danym i zaliczania ich do danej kategorii, wymieniania, łączenia i dzielenia kategorii, formułowania schematu konceptualnego do rozwoju teorii);
- rozwój i weryfikacja teorii (pozwalają na zastosowanie różnych modeli teoretycznych do tworzenia teorii i przedstawienia wyników);
- tworzenie raportów (umożliwia przygotowanie raportów dla różnych kategorii lub odtwarzanie odpowiednich fragmentów dokumentów w formie cytatów, tabel, przedstawień graficznych itd.; istnieje możliwość utworzenia w programie „dziennika", do którego można wpisywać komentarze, pomysły, po czym wyprowadzać je do druku lub do pliku)[42].

Wraz z pojawieniem się komputerowej analizy treści zmienił się sposób jej wykorzystania: dzięki możliwościom szybkiego opracowywania ogromnych zbiorów informacji, analiza treści organicznie wplata się w ogólną technologię, w ramach której jest stosowana. Podamy przykład prostego i technologicznego wykorzystania analizy treści w GAO, USA (United States General Accounting Office)[43]. Przede wszystkim – analiza treści

42 Lacey A., Luff D. Trent Focus for Research and Development in Primary Health Care: An Introduction to Qualitative Analysis.– Trent Focus, 2001.– http://www.trentfocus.org.uk/Resources/Qualitative%20Data%20Analysis.pdf

43 Spisak T. Content Analysis: A Methodology for Structuring and Analyzing Written Material / GAO (United States General Accounting Office).– Transfer Paper 10.1.3.– Washington, D.C. – March 1989.– r.6.– http://www.coe.uga.edu/~cwise/content_analysis/GAO_methodology.pdf

stała się dobrym narzędziem do analizy codziennej informacji. Notatki robocze, dokumenty, rozszyfrowania stenogramów spotkań, wstępnych ocen itd. z lokalnych przedstawicielstw GAO zawierają wiele cennych informacji, ale trudno je połączyć i uogólnić ze względu na ich różnorodność i niestrukturalność. W celu strukturyzacji materiałów tekstowych sporządzana jest ogólna lista poruszanych w nich problemów i podliczana częstotliwość ich występowania. Z tym dobrze radzi sobie analiza treści.

Uproszczenie i kategoryzacja również stanowią część analizy treści. Przykład: oceniając stopień dublowania pracy, analitycy zebrali informacje, dotyczące budżetów, wywiadów, raportów i w efekcie czego wyodrębniono 31 problemów, dotyczących lokalizacji budownictwa mieszkaniowego i rozwoju miast. Analogicznie, po rozpatrzeniu 38 raportów z dwóch biur, opracowany został system kategorii, który zaczęto stosować do każdego raportu, co umożliwiło wykrycie podobnych problemów w biurach GAO.

Jednak komputerowa analiza treści bardzo często doprowadza do bardziej radykalnych zmian, wskutek czego łańcuch technologiczny *źródło tekstu – tekst – analiza treści skraca się do źródło tekstu – analiza treści*, zatem, analiza treści staje się składnikiem pewnej technologii.

Przykładem takiego wykorzystania technologicznego mogą być systemy automatycznego rozpoznawania w czasie rzeczywistym tekstów audycji radiowych z automatycznym doborem niezbędnych informacji. Do podstawy tych systemów wbudowano analizę treści. Na przykład, system WASABI (Watson Automatic Stream Analysis for Broadcast Information) przetwarza w tekst wejściowy strumień dźwięków poprzez podsystem rozpoznawania mowy, wykorzystuje różne analizatory do identyfikacji elementów informacyjnych, generuje automatycznie zapytania na podstawie elementów informacyjnych i wybiera z bazy informacyjnej dane, relewantne do dyskursu potocznego[44].

Z innej strony, komputerowa analiza treści nie tylko stała się wcieleniem metodyki tradycyjnej, ograniczonej przez opracowywanie tekstu.

Ma on własną logikę rozwoju – technologiczną. Właśnie ta logika rozwinęła się w programach czwartej generacji. Do programów czwartej generacji zaliczać będziemy programy, które będąc wcieleniem analizy treści, „wbudowują" ją w inne technologie. Przykładem technologii tej generacji, która w znacznej mierze stworzona jest na zasadach analizy treści i jest stosowana już od połowy lat 90-tych XX w. jest technologia „wydobycia" danych lub Text Mining (pełna nazwa – Text Analysis and Knowledge Mining System). Text Mining – to algorytmiczne wykrycie na podstawie analizy statystycznej i lingwistycznej, a także sztucznej inteligencji nieznanych wcześniej związków i korelacji w już istniejących, niestrukturalnych danych tekstowych służących do przeprowadzenia analizy znaczeniowej, zapewnienia nawigacji i poszukiwania w tekstach niestrukturalnych z zamiarem otrzymania nowej, cennej informacji – wiedzy. Text Mining stanowi logiczną kontynuację i połączenie całego szeregu metodyk i metod, zwłaszcza technologii Data Mining, analizy treści, analizy statystycznej itd.

Nie tylko Text Mining, lecz i jego podstawa – Data Mining mają zbyt krótką historię. Wielu naukowców (m.in. przyt. w przypisie[45]) uważa, że biorą one swój początek od programów pozyskiwania informacji i zbliżonych do nich. Przykładem wczesnego programu pozyskiwania informacji, wspomnianym przez M. Dixon w przypisie[46], jest napisany w 1982r. program FRUMP. Wykorzystywał on szereg scenariuszy, które mogły skanować i opracowywać zbiory wiadomości, próbując na ich podstawie wykonywać opisy wydarzeń.
M. Dixon wylicza dwa ważne, pionierskie badania przeprowadzone z Text Mining. Przede wszystkim, są to prace grupy badawczej z Uniwersytetu Helsińskiego[47], która próbowała wy-

44 Brown E.W., Srinivasan S., Coden A., Ponceleon D., Cooper J.W., Amir A. Toward speech as a knowledge resource // IBM Knowledge Management.– Vol. 40.– Number 4.– 2001. – http://www.research.ibm.com/journal/sj/404/brown.html

45 Dixon M. An Overview of Document Mining Technology.– October 4, 1997.– http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dm.html; Wilks Y. Information extraction as a core language technology //Information Extraction: A Multidisciplinary Approach to an Emergine Information Technology.– Vol. 1299.– June 1997.– r.1-9.

46 Dixon M. An Overview of Document Mining Technology...

47 Ahonen H., Heinonen O., Klemettinen M., Verkamo A.I. Mining in the phrasal frontier // Proceedings of PKDD'97 / 1st European Symposium on Principles of Data Mining and Knowledge Discovery.– Norway.– Trondheim.– June 1997.

korzystać technologię Data Mining do niestrukturalnych, poprzednio nieopracowanych zbiorów tekstów. Po drugie, są to prace R. Feldmana[48], w których podstawę położono ustalenie znaczących dla tekstu pojęć (zamysłów) i określenie wzajemnego związku pomiędzy dokumentami i tymi pojęciami, czyli faktycznie – wykonanie klasyfikacji tekstu[49]. System Document Explorer, zaproponowany przez R. Feldmana, najpierw tworzy bazę danych na podstawie ogółu badanych dokumentów różnego rodzaju, w tym z Internetu, a następnie analizuje je, wykorzystując technikę pozyskiwania wiedzy i podejście graficzne. Zaznaczmy, że R. Feldman, w celu określenia technologii pozyskiwania informacji z dokumentów wykorzystuje zamiast Text Mining, jego zdaniem, dokładniejszy termin Knowledge Discovery in Textual Databases (pozyskiwanie wiadomości na podstawie tekstowych baz danych).

Pionierskimi pracami, w których rozpatrywana jest technika pozyskiwania informacji z zasobów HTML, stały się badania O.Etzioni[50].

Technologia Text Mining różni się od Data Mining tym, że w odróżnieniu od ostatniej, pracującej ze strukturalnymi bazami danych faktów, uzyskuje wzorce (szablony) ze zwykłych tekstów, przeznaczonych do czytania przez człowieka, a nie komputer. Jednocześnie, tak jak większość technologii kognitywnych, Text Mining, to nie tylko poszukiwanie środowisk dużych zbiorów gotowych, opracowanych przez osoby trzecie informacji, a przede wszystkim uzyskanie nieznanych wcześniej i nigdzie nie zapisanych informacji, a dokładniej – algorytmiczne wykrycie wcześniej niezauważonych związków zarówno w samych tekstach, jak i wskutek ich wspólnego czytania. Ponadto, niejednokrotnie na początku badań analityk sam dokładnie nie wie, która konkretnie informacja jest mu potrzebna i gdzie jej szukać.

Począwszy od połowy lat 90-tych XX w., technologia Text Mining, jako kierunek analizy niestrukturalnych danych, obrała za podstawę nie tylko metody klasycznego zdobywania wiedzy, ale także najnowsze osiągnięcie analizy treści: klasyfikację, klasteryzację, wydzielenie pojęć, faktów, szablonów etc.

Składnikami technologicznymi Text Mining są:

1. poszukiwanie informacji (wybór zapisów relewantnych lub tekstowych baz danych do dalszego opracowania);
2. przetwarzanie informacji (wyodrębnienie • –Ú Ü [wzorów na podstawie wybranych danych);
3. integracja informacyjna (połączenie komputerowego określenia informacji z możliwościami poznawczymi człowieka)[51].

Właśnie w przypadku realizacji drugiego składnika, Text Mining wykonuje takie rodzaje analizy, jak:

- wykrycie lub pozyskanie informacji (ta analiza poprzedza cały szereg rodzajów analizy komputerowej niestrukturalnych tekstów i służy jako podstawa do ich realizacji; w procesie analizy identyfikowane są frazy kluczowe i związki wewnątrz tekstu w drodze poszukiwania najpierw określonych kolejności wyrazów w tekście, które nazywane są szablonem lub wzorcem; wykrycie informacji jest nadzwyczaj korzystne w przypadku pracy ze znacznymi ilościami informacji);
- śledzenie kategorii lub tematów (określając i zachowując niektóre parametry użytkownika, np. jakiego rodzaju dokumenty przeglądał wcześniej, jakie tematy i pojęcia są obecne w dokumentach, interesujące dla użytkownika, system znajduje podobne materiały informacyjne i regularnie przekazuje je użytkownikowi; przykładem z zakresu biznesu może być śledzenie materiałów, zawierających informacje dot. konkurencji i jej produkcji lub poszerzenia produkcji własnej firmy);
- podsumowanie lub referowanie dokumentów (główna idea podsumowania polega na tym, aby wskutek wykluczenia nieistotnych szczegółów i skrócenia długości tekstu wy-

48 Feldman R., Klosgen W., Ben-Yehuda Y., Kedar G., Reznikov V. Pattern based browsing in document collections // Principles of data mining and knowledge discovery. – June 1997.– Vol. 1263.– r.112-122.

49 Dixon M. An Overview of Document Mining Technology... .

50 Etzioni O. The world-wide web: Quagmire or gold mine? // Communications of the ACM.– No 39(11).– November 1996.– r.65-68.

51 Kostoff R. Text Mining for Global Technology Watch / Office of Naval Research (ONR) Science & Technology.– 2001.– http://www.onr.navy.mil/sci_tech/special/technowatch/default.htm; Kostoff R. Information Extraction From Scientific Literature with Text Mining / Office of Naval Research (ONR) Science & Technology.– 2001.– http://www.onr.navy.mil/sci_tech/special/technowatch/default.htm

odrębnić główne momenty i ogólną treść; program posiada możliwości prostej identyfikacji osób, miejsc, zdarzeń, czasu, jednak najbardziej skomplikowana jest umiejętność analizowania składni powiadomienia i nadania mu interpretacji treści; takie efekty podsumowania są korzystne przede wszystkim, w tych przypadkach, kiedy użytkownik musi natychmiast ocenić, w jakim stopniu dokument odpowiada jego potrzebom i czy warto poświęcać czas na dalszą pracę z dokumentem);

- klasyfikacja lub kategoryzacja tekstu (główną ideą analizy jest zaliczenie obiektów ze zbiorów tekstowych do wcześniej określonych kategorii; klasyfikując obiekty, program nawet nie próbuje ich analizować, a rozpatruje jako podstawowy zestaw wyrazów, podlicza częstotliwość ich występowania i na tej podstawie identyfikuje główne tematy dokumentów; następnie możliwe jest grupowanie dokumentów w sieci internetowej i na stronach WWW, rozmieszczenie ich w określonych folderach, sortowanie wiadomości poczty elektronicznej, śledzenie i wybiórcze poszerzenie informacji tematycznej dla użytkowników itd.);
- klasteryzacja lub grupowanie (wydzielenie kompaktowych podgrup obiektów ze zbliżonymi właściwościami; ten rodzaj analizy jest bardzo podobny do klasyfikacji, jednak nie ma tu uprzednio określonych kategorii, są one formowane na podstawie samych danych; inną zaletą klasteryzacji jest fakt, że dokument może jednocześnie dotyczyć wielu kategorii: algorytm klasteryzacji dla każdego dokumentu formuje całą hierarchię kategorii i określa wytyczną, z pomocą której dokument można zestawić z każdą kategorią; zatem, na tej podstawie wykonywane są: tworzenie bazy odwołań od dokumentu do dokumentu, opartych na wagach i wspólnym używaniu zadanych słów kluczowych; referowanie dużej ilości dokumentów; wyznaczenie wzajemnie zależnych grup dokumentów; uproszczenie procesu przeglądu przy wyszukiwaniu niezbędnych informacji, wyłączenie dokumentów unikalnych ze zbioru; wykrycie duplikatów lub dokumentów o podobnej treści itd.);
- prognozowanie (przewidywanie na podstawie znaczenia jednych cech obiektu innych znaczeń);
- znajdowanie wyjątków (poszukiwanie obiektów, które wg swoich charakterystyk wyróżniają się na ogólnym tle ogólnej);
- poszukiwanie łącznych cech, pól, pojęć poszczególnych dokumentów (te narzędzia łączą dokumenty poprzez wspólne dla nich kategorie i pomagają użytkownikowi nie tylko w odnajdowaniu informacji, jak to ma miejsce w tradycyjnych metodach poszukiwania, ale bezpośrednio przechodzić od jednych pojęć do innych, związanych z pierwszymi i niekoniecznie w zakresie jednego dokumentu; ponadto, jeśli związki pomiędzy kategoriami X i Y, a także pomiędzy Y i Z, są dobrze znane, a z powodu nagromadzenia informacji badacz może nie dostrzec związku pomiędzy X i Z, m. in., kiedy jest on dopiero formowany lub jest bardzo słaby, to tego rodzaju mechanizm asocjacyjny jest bardzo cenny);
- wizualizacja danych (sposób przedstawienia treści całego zbioru dokumentów i realizacji mechanizmu nawigacyjnego do badania dokumentów i klas dokumentów; wizualizacja umożliwia przedstawienie dużych objętościowo dokumentów w formie „interaktywnych" przedstawień graficznych lub map, kiedy to w wyniku „współpracy" z nimi użytkownik ma możliwość ich przeglądania przy pomocy prostych środków wyszukiwania, skalowania danych, tworzenia różnorodnych podschematów; wizualizacja jest szczególnie korzystna w przypadku badania dużych zbiorów dokumentów i badania powiązanych konceptualnie dokumentów);
- odpowiedzi na pytania lub Q&A (próbuje znaleźć najlepszą odpowiedź na zadane pytanie, zazwyczaj zapisane naturalnym językiem; wykorzystywane są przy tym wszystkie odmiany analizy Text Mining, same pytania dotyczą natomiast takich kategorii, jak: ludzie, miejsca, wydarzenia i in., osobno opracowywane są algorytmy do FAQ – najczęściej zadawanych pytań; pragniemy podkreślić, że technologia przygotowywania odpowiedzi na pytania wykorzystywana jest przez wiele stron internetowych).

Przykład modelu[52] „wydobywania" danych przedstawiony jest na rys. 4.2. Mając zadany zbiór dokumentów, program Text Mi-

52 Fan W., Wallace L., Rich S., Zhang Z. Tapping into the Power of Text Mining // Communications of ACM.– February 16, 2005.– http://filebox.vt.edu/users/wfan/paper/text_mining_final_preprint.pdf

ning wybiera potrzebne egzemplarze dokumentów i wstępnie je opracowuje, sprawdzając formaty i zestawy symboli. Po tym następuje etap analizy tekstowej, w trakcie której za pomocą powtarzania pewnych działań wydobywane są informacje. Na podanym diagramie pokazane zostały tylko trzy metody, jednak jest ich znacznie więcej.

Oczywiście, nie są wykorzystywane wszystkie, a jedynie te, które są potrzebne do osiągnięcia celów. Informacje końcowe są zapisywane w bazie danych, skąd użytkownik otrzymuje potrzebne wiadomości.
Przytoczone w tab. 4.2 i 4.3 dane pokazują, jakie metody Text Mining są wykorzystywane przez różne programy komercyjne i w różnych sferach działalności.

Z punktu widzenia formatu przedstawienia danych, najpierw następuje oczyszczanie tekstu, podczas którego niestrukturalne dokumenty są przetwarzane w standaryzowaną formę pośrednią, a następnie wykonywane jest filtrowanie wiadomości, gdzie na podstawie formy pośredniej są wyprowadzane wzory lub wiadomości[53].

Forma pośrednia może mieć podstawę dokumentową lub konceptualną. W pierwszym przypadku wzory lub wiadomości są wyodrębniane na bazie dokumentów, w drugim – istotne informacje są zgodne z przedstawieniem konceptualnym lub obrazem problematycznej sytuacji (rys. 4.3).

Obecnie, wielu producentów oprogramowania proponuje własne technologie i programy Text Mining. Systemy Text Mining zazwyczaj są wykonywane w formie systemów na wielką skalę, ze skomplikowanymi, matematycznymi i lingwistycznymi algorytmami analizy, dla których charakterystyczne są: rozwinięty interfejs graficzny, bogate możliwości wizualizacji i manipulowania danymi, dostęp do różnych źródeł danych, funkcjonowanie w konfiguracji klient-serwer. Wg danych Ośrodka Polityki Technologicznej i Ocen (Technology Policy and Assessment Center – TPAC) Instytutu Technologii w Georgii, pod koniec XX w. w Internecie funkcjonowało ponad 70 systemów instrumentalnych Text Mining[54].

Opis możliwości funkcjonalnych poszczególnych systemów można znaleźć w Załączniku B, a także w pracach D. Landego[55]. Jednym z najbardziej rozwojowych kierunków uogólnienia strumieni informacyjnych jest monitoring treści. W najprostszej formie jego istotę można zdefiniować, jako stale wykonywaną w czasie analizę treści stałych strumieni informacji. Wśród zasad tworzenia systemu monitoringu można wyodrębnić: systemowość, liczbę adresów i ukierunkowanie przedmiotowe. Strumień tekstów badany jest tutaj na podstawie zadanych charakterystyk konfiguracyjnych (zestawów parametrów ilościowych lub wyrazów towarzyszących określonym tematom i pojęciom). Opracowywany jest on wielokrotnie, z dodawaniem charakterystyk, otrzymanych z samego strumienia. Podstawę metodologiczną badania stanowi analiza treści. W efekcie generowane są, a następnie przedstawiane wizualnie informacje uogólnione. Wraz z pojawieniem się systemów Text Mining monitoring treści otrzymał realną i potężną podstawę programową.

Do nowoczesnych i perspektywicznych kierunków wykorzystania Text Mining zalicza się także:

- wyszukiwanie informacji ogólnych i relewantnych na podstawie tekstowych baz danych;
- wyznaczanie infrastruktury zadanych dyscyplin i kierunków technologicznych oraz naukowych;
- wykonywanie strukturyzacji tematycznej pewnych zakresów działalności i związków pomiędzy tematami;
- wyznaczanie nowych kierunków badań, pojawianie się nowych idei w ramach pewnych dyscyplin i na styku tychże dyscyplin;
- prognozowanie rozwoju technologicznego[56].

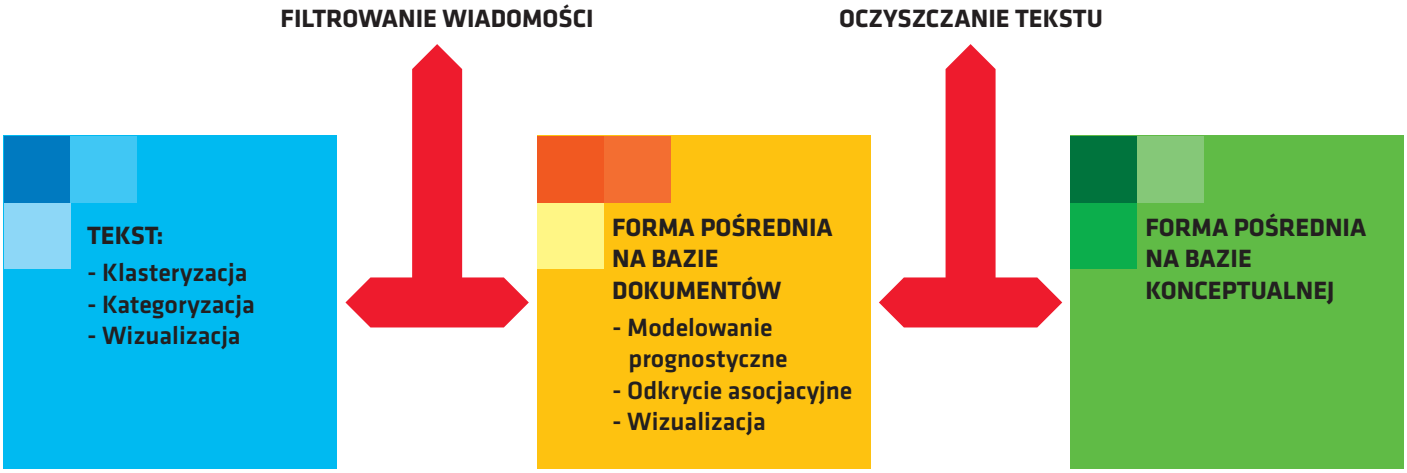
53 Tan A.-H. Text Mining: The state of the art and the challenges // The Pacic Asia Conference on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases.–1999.–PP.65-70.– http://citeseer.ist.psu.edu/tan99text.html

54 Text Mining: Review of TPAC Technologies for ONR // ASDL.– Aug. 2002.– http://www.asdl.gatech.edu/research_teams/pdf/2002/Text%20Mining%20Sum.doc

55 Lande D. Zdobywanie wiedzy // CHIP Ukraine. – 2003. – Nr10; Lande D. O oddzielaniu ziaren od plew // Mój Komputer Weekly. – Nr33(256). – 25.08.2003.

56 Text Mining: Review of TPAC Technologies for ONR...

Rys. 4.3. Schemat formowania nowych wiadomości w Text Mining



Jak określono w przypisie[57], na szczególną uwagę zasługuje tutaj innowacyjne prognozowanie.

Jeden z rodzajów prognoz bazuje na bibliometrii: podliczając ilość publikacji, patentów, odpowiednich wzmianek w wystąpieniach naukowców, można zmierzyć i zinterpretować kierunki rozwoju technologicznego. Takie pomiary są uogólniane w formie innowacyjnych wskaźników technologii, które mogą świadczyć o stadium rozwoju technologii, o innowacyjnym, kontekstualnym wpływie tej technologii na inne, rozwoju potencjału rynkowego itd.. Jeszcze jedna metoda prognozowania bazuje na kartografii wiadomości, co umożliwia identyfikację spokrewnionych grup technologii i środków, wzajemnych wpływów różnych grup technologii, lokalizowanie domen badawczych i ustalanie kręgu ich zainteresowań.

Nadzwyczajnie perspektywicznym kierunkiem Text Mining jest technologiczny wywiad konkurencyjny (Competitive Technological Intelligence). Jego znaczenie i wykorzystanie szczególnie wzrosło w latach 90-tych XX w., kiedy nasiliła się konkurencja technologiczna i firm – wówczas to uniwersytety oraz organizacje rządowe potrzebowały szczególnej wiedzy na temat nowych i perspektywicznych technologii. Znaczna część wyników jest otrzymywana przez wywiad technologiczny na podstawie prowadzenia poszukiwań z wykorzystaniem technologii Text Mining. O aktualności opracowań Text Mining świadczy chociażby ich stosowanie przez Służby Federalne i Agencje USA. Przykładowo – badanie GAO z maja 2004r.[58]wykazało, że ze 128 służb, które były badane, 52 wykorzystywały lub planowały wykorzystanie technologii Data Mining i Text Mining. Cele były tutaj bardzo zróżnicowane: zaczynając od potrzeby poprawy obsługi ludności, a na analizie oraz wykrywaniu działalności terrorystycznej i przestępczej kończąc. Analitycy GAO wykryli 199 wypadków zastosowań technologii pozyskiwania wiadomości, z których 68 związane są z planowaniem pracy i 131 – z działalnością operacyjną.

Przytoczmy zatem fakty. W marcu 2001r. w szeregu rosyjskich i ukraińskich wydawnictw internetowych[59] pojawiły się informacje o wykorzystaniu przez Resort Rozwoju Technologii Informacyjnych, będący częścią Dyrekcji ds. Nauki i Technologii CIA, USA, Text Mining do pracy z jawnymi źródłami informacji. Po-

57 Watts R.J., Porter A.L. Innovation Forecasting // Technology Policy and Assessment Center (TPAC) at Georgia Institute of Technology.– 2002.– http://www.tpac.gatech.edu/toa/inov.shtml

58 Data Mining. Federal Efforts Cover a Wide Range of Uses: Report to the Ranking Minority Member, Subcommittee on Financial Management, the Budget, and International Security, Committee on Governmental Affairs, U.S. Senate / GAO (United States General Accounting Office).– GAO-04-548.– Washington, D.C. – May 2004.– 71 r.– http://www.epic.org/privacy/profiling/gao_dm_rpt.pdf

59 Lande D. Zdobywanie wiadomości...; CIA rozpoczyna przesiew informacji // Magazyn sieciowy. Taśma wiadomości.– Wyd. z dn. 14.03.2001.– http://www.setevoi.ru/cgi-bin/srch.pl?id=579 ;
CIA wyciąga dane // Komputerinform€ ×Ł0Á&Á.– 2001.– Nr6.– http://www.ci.ru/inform06_01/p245moz.htm ;
Gordienko I. Zrozumieć i zmusić // Komputerra. – 10.04.2001.– http://www.ibusiness.ru/offline/2001/158/8585/print.html

szczególne publikacje krajowe odsyłają do pierwszego źródła – materiału na stronie dziennika „Washington post". Mowa była o zastosowaniu przez agencję wywiadowczą trzech systemów komputerowych – Oasis, FLUENT, Text Data Mining.

Pierwszy system związany jest z monitoringiem mediów, zarówno ze źródeł systematycznych, jak i przypadkowych, obejmujących wydania drukowane, materiały cyfrowe, prezentacje graficzne, informacje audio w 35 językach świata. Przykładowo – podczas pracy z informacją audio, system Oasis przetwarza materiały dźwiękowe na tekst, rozpoznając język, głosy męskie i żeńskie, głosy różnych osób i zapisuje je w formie dialogów. Poza tym, metodyka pozwala na wyodrębnienie ze strumienia tylko tych głosów i informacji, które założono w ustawieniach systemu. W momencie pojawienia się zaznaczonych materiałów internetowych system ograniczał się tylko do wersji angielskojęzycznej, choć przewidywane było również stworzenie wersji do rozpoznawania języka chińskiego, arabskiego i innych.

Technologia komputerowa FLUENT przeznaczona jest do wyszukiwania informacji w dokumentach tekstowych. Mając zadane słowa kluczowe w języku angielskim, system natychmiast tłumaczy je na szereg innych języków, szuka informacji w tekstowych bazach danych z dokumentami w różnych językach i po automatycznym przetłumaczeniu oddaje analitykowi wyniki takiego wyszukiwania. FLUENT pozwala na tłumaczenie na język angielski z języka chińskiego, koreańskiego, portugalskiego, rosyjskiego, serbsko-chorwackiego, ukraińskiego i innych.

Jeszcze jeden program, mianowicie Text Data Mining, umożliwia automatyczne tworzenie wizualnych obrazów dokumentów tekstowych, a także otrzymywanie danych, dotyczących wykorzystywania tych czy innych wyrazów.

Wymienione technologie CIA wykorzystuje do śledzenia nielegalnych operacji finansowych i handlu narkotykami.

Przytoczymy jeszcze jeden ciekawy kierunek zastosowania Text Mining. Są to tak zwane integratory wiadomości. W warunkach zwiększających się ilości strumieni informacji zwykłe wyszukiwarki internetowe nie pozwalają w sposób efektywny wyodrębnić potrzebne informacje z tych strumieni. Zdaniem D. Landego[60], optymalnym rozwiązaniem problemów orientacji w wiadomościach z Internetu jest wykorzystanie nowych służb sieciowych – integratorów wiadomości – integrujących strumienie informacji, realizując monitoring treści wiadomości w przestrzeni WWW, jako bazę do swojej pracy.

Przykładowo – Northern Light Technology jest klientem jednej z wielkich firm specjalizującej się w pozyskiwaniu informacji – COMTEX, integrującej zasoby sprawdzonych źródeł, wśród których są takie światowe agencje informacyjne, jak Associated Press, ITAR-TASS, Xinhua. Klientami COMTEX są też dziesiątki służb informacyjnych: OneSource, Screaming Media, Vertical Net, CompuServe i inne. Technologia monitoringu i syndykacji wiadomości internetowych przewiduje następujące etapy: „nauczanie" programów zbierania informacji w strukturze wybranych źródeł, skanowanie informacji, sprowadzenie ich do formatu wewnątrzsystemowego, klasyfikacja, klasteryzacja, przekazanie użytkownikom przez różne kanały, w tym e-mail, WWW, Wap, SMS.

Na zakończenie zaznaczymy, że technologia Text Mining znajduje się dopiero u progu swojej „kariery". Jednak nawet teraz wydaje się ona być bardzo perspektywiczną. Wielkie zasługi położyła tutaj analiza treści.

60 Lande D. Systemy monitoringu Internet-treści // Biznes-rejestr.– 2002.– Nr2(8). – http://www.biz-registr.com.ua

WNIOSKI

Tak więc, przeprowadzone badanie potwierdza, że analiza treści przeszła drogę od metody naukowej sformalizowanej analizy treści mediów do szeroko stosowanej metodyki zaawansowanej technologicznie.

Jak już to zostało przedstawione, w procesie kształtowania się analizy treści można wyodrębnić następujące etapy:

- jej narodziny (do lat 20-tych XX w. – charakteryzuje się metodologią opisową i metodologią intuicyjną, rozwijają się w różne podejścia do analizy i porównania tekstów w kontekstach interpretacyjnych, przede wszystkim środków komunikacji masowych, wczesna analiza gazet, analiza grafologiczna, analiza marzeń sennych);
- formowanie podstaw „klasycznej" analizy treści (lata 20-te – 40-te XX w. – rozwijają się systematyczne podstawy ilościowej analizy treści, znów jednak w ramach komunikacji masowej, rozwój następuje w zasadzie jednocześnie z teorią i praktyką propagandy);
- międzydyscyplinarne rozszerzenie i dyferencjacja (lata 50-te – 60-te XX w. – metodologia nie tylko rozwija swoje podstawy teoretyczne, lecz również znajduje drogę do różnych dyscyplin, przede wszystkim lingwistyki, psychologii, socjologii, nauk historycznych, sztuki itd.);
- rozwój podstaw teoretycznych i rozszerzenie praktycznego zastosowania (lata 70-te – 80-te XX w. – doskonalenie i zastosowanie różnych modelów związku, analiza komunikacji niewerbalnych, a także rozwój nowych odmian, m.in. jakościowej analizy treści, doskonalenie metodyki, przede wszystkim wskutek wykorzystania nowych możliwości techniki komputerowej);
- okres rozwoju „globalnego" (od lat 90-tych XX w. – okres, związany z wykorzystaniem analizy treści w praktyce działalności różnych podmiotów, począwszy od naukowców, którzy są zapoznani z jej metodyką i świadomie ją stosują, i kończąc na zwykłych użytkownikach Internetu – nieświadomych faktu, że wyszukiwanie informacji wykonują dla nich programy z wbudowanymi elementami analizy treści).

Charakteryzując niniejszy stan analizy treści, warto zauważyć, że obecnie nie nauka czy środki informacji masowej są największymi użytkownikami analizy treści, a instytucje państwowe i komercyjne, partie polityczne, ośrodki analityczne, osoby prywatne zainteresowane zdobyciem nowych wiadomości. Obecne badania treści związane są z przetwarzaniem dużych zbiorów tekstowych na podstawie technologii internetowych i technologii komputerowych pozyskiwania wiadomości na wzór Text Mining i Web Mining, skonstruowanych w znacznym stopniu na ideach analizy treści. Z nimi też związana jest perspektywa rozwoju analizy treści w najbliższych latach.

ZAŁĄCZNIK A. Podstawowe pojęcia i algorytm ilościowej analizy treści

Opis pojęć podstawowych i algorytmu ilościowej analizy treści w opracowywaniu i z uzupełnieniami autora jest przedstawiany zgodnie ze strukturą, zaproponowaną w pracy[61].

1. Zasady ogólne i kluczowe pojęcia

1) Systematyczność:

- treść jest analizowana zgodnie z wyraźnie opisanymi i kolejno zastosowanymi zasadami;
- wybór formuje się z wykorzystaniem należytych procedur, na przykład, wybór przypadkowy lub prawdopodobny, wybór wielostopniowy, które dość dobrze przedstawiają badany zbiór tekstowy;
- kodowanie i analiza są wykonywane wg jednorodnych zasad; kodowanie to przetwarzanie szczegółów treści na specyficzne wartości liczbowe;

2) obiektywność:

- sympatie osobiste i idiosynkrazja badacza nie powinny wpływać na wyniki badania;
- badanie, przeprowadzone przez jednego badacza, powinno być tak opracowane, abymogły je odtworzyć inne osoby;
- oznaczenia operacyjne i zasady klasyfikacji treści powinny być zrozumiałe i wystarczająco dobrze objaśnione, co umożliwi różnym kodującym osiągnięcie jednakowych wyników;

61 Department of Journalism Studies...

3) kwantyfikacja:

- dokładnie odtwarzanie podstawowej treści tekstów za pomocą liczb;
- często najważniejsze w analizie treści są dane liczbowe;
- liczbowe dane szacunkowe są gromadzone, i powinno być to wykonywane systematycznie i zgodnie ze ścisłymi zasadami, bez dopuszczania do subiektywnych uwag i tłumaczeń.

2. Etapy analizy treści

1. sformułować pytanie, badanie lub hipotezę;
2. określić całokształt na podstawie pytań;
3. określić wybór materiału do badania bezpośredniego;
4. wybrać i podać określenie jednostek analizy;
5. utworzyć system kategorii analizy treści;
6. ustalić system kwantyfikacji (ilościowej oceny kategorii);
7. przeprowadzić szkolenie kodujących i badanie eksperymentalne (wstępne);
8. zakodować teksty zgodnie z ustalonymi zasadami kodowania;
9. przeanalizować otrzymane dane;
10. opracować wnioski i przygotować propozycje.

3. Określenie problemów, pytań i hipotez badania

Ankieta służąca określeniu problemów i pytań badania:

– Na czym polega problem lub pytanie badania?

- Jaki fenomen (obiekt) będzie badany?
- Jaki efekt spodziewamy się uzyskać w wyniku analizy tekstów? Analiza treści – to nie jest teoria, lecz metoda. I nie może ona zadecydować za nas, co należy przeanalizować.
- Dowolna analiza treści potrzebuje podstawy teoretycznej. Czy istnieje możliwość odnalezienia gotowej teorii dla Państwa badania?
- Jak wiele informacji jest dostępnych na temat zjawiska lub tematu, który będzie badany?

Badacz musi zapoznać się z istniejącymi faktami.

- Co powinno opisywać badanie: rodzaje treści mediów, możliwe sposoby wpływania na czytelników, sposoby produkcji medialnych?
- Może, badanie próbuje powiedzieć coś na temat przedstawienia różnych grup socjalnych (np. mężczyzn i kobiet, grup etnicznych, wiekowych i zawodowych) względnie zachowań społecznych (np. gwałt, zdrowie, ochrona praw konsumentów) lub na tematy dot. instytucji (np. polityczne, biznesowe)?
- Może, badanie ma na celu podanie informacji dot. redakcyjnej polityki [danego – przyp. tłum.] wydania czy sposobu prezentowania wiadomości?

Zalecenia przy formułowaniu pytań do badania:

- podliczenie – nie jest celem w samym w sobie, dlatego należy unikać realizacji podliczenia dla podliczenia;
- cel końcowy analizy powinien być jasno sformułowany;
- należy kierować się jasno sformułowanymi pytaniami badania lub hipotezami;
- zapoznać się z odpowiednią literaturą; – zazwyczaj, źródła pytań lub hipotezy badania znajduje się w literaturze.

4. Określenie całokształtu badania

Zalecenia:

- podjąć decyzję dotyczącą granic badania (określić, co znajduje się w polu widzenia badacza);
- zidentyfikować treść analizy;
- określić schematy klasyfikacji treści;
- określić okres, który będzie obejmowało badanie.

5. Formowanie wyboru

Zalecenia ogólne:

- po określeniu całokształtu (całego zbioru tekstów) należy sformułować wybór, który tworzy pewna część tego całokształtu (wybór tekstów, które będą bezpośrednio badane);
- jeśli całokształt obejmuje stosunkowo niewielką ilość tekstów, to warto badać całość (tj. pełni on rolę wyboru);
- jeśli całokształt jest duży, to trzeba wybrać tę jego część, którą badacz realnie jest w stanie zbadać;
- najczęściej wykorzystuje się wybór wielostopniowy;
- opisy problemów są wybierane do analizy całkiem przypadkowo;
- w ramach całkowitego wyboru analizowane są wszystkie materiały źródłowe: każdy program telewizyjny, każdy nu-

mer dziennika, który ukazał się w ciągu określonego czasu. Czasami jest ona modyfikowana w „50 %". Całokształt wyboru jest określany jako „kompletny „ nie tylko wg parametrów czasu. Wielką rolę odgrywa także tematyczne ukierunkowanie wiadomości przedmiotu badania. Odpowiednio, analizowana jest albo cała treść z danej daty czy „dnia telewizyjnego", albo tylko poszczególne rubryki, treści, materiały, bezpośrednio dotyczące przedmiotu badania – np. stosunków międzynarodowych, reklamy, wartości politycznych, prezentacji ekologicznych itd. Całościowe wybory wykorzystywane są podczas obserwacji przebiegu kampanii przedwyborczych, mających określone granice czasu; dla przypadku analizy „operacyjnej" – omówienia w prasie światowej wizyt głów przywódców państw;

- w celu analizy wiadomości, tematyki poruszanej w mediach najczęściej stosuje się modele wyboru, zbliżone do typologicznych; np. w praktyce analizy treści przyjęto obserwować każdy piąty lub szósty dzień ukazywania się dziennika, uwzględniając periodyczność jego wydawania i zamieszczania informacji w zależności od dnia tygodnia;
- projektowane są również przypadkowe wybory, w których najpierw określany jest kierunek wyboru jednostek tekstowych lub eksperymentalnie oblicza się ich ilość, niezbędną do reprezentatywnego przedstawienia całego zbioru; taką ilość numerów dziennika można potem skompletować zgodnie z tabelą liczb przypadkowych.

Utworzenie wielostopniowego wyboru:

- początkowo określa się główne kanały lub źródła informacji, np. badane gazety;
- następnie wyznacza się ogół wiadomości wybranego źródła informacji: wybierane są ramy czasowe, w granicach których będą rozpatrywane materiały (za podstawę mogą być wzięte dane wydarzenia – od jednego wydarzenia do innego, np. miesiąc, pół roku, trzy lata lub pięć lat, „połowa lat 90-ch" i, ostatecznie, „cały okres powojenny");
- w ramach takiego okresu istnieje możliwość wyboru; doświadczenie pokazuje, że dwa przypadkowo wybrane dni każdego tygodnia miesiąca dają gwarancję całkowicie zbilansowanego wyboru dla tego miesiąca; można niejako „konstruować" tydzień na podstawie miesiąca: poniedziałek jest wybierany z przypadkowo wziętego tygodnia miesiąca, następnie wtorek itd., aż do momentu dopóki nie zostaną wybrane wszystkie dni tygodnia.

Ankieta do formowania wyboru:

- Jakie kanały telewizyjne, radiowe, wydania gazet i czasopism będą rozpatrywane?
- Jaka jest strategia tego wyboru?
- Jaki okres czasu będzie obejmować?
- Jakie specyficzne wydarzenia będą analizowane?
- Jakie części gazety będą badane?
- Jakie audycje telewizyjne lub radiowe są rozpatrywane?

Uwaga:

- dokonanie właściwego wyboru wiąże się z dylematem: z jednej strony – wybór konkretnego źródła najpierw ogranicza ramy interpretacji wniosków analityka, z drugiej – analityk powinien prawidłowo poszerzać wnioski o szczegóły i tendencje treści całości, czyli mówimy o dużej objętości i, być może, odmiennych jakościowo strumieni informacyjnych. Ponadto, wybór powinien uwzględniać kryteria ilościowe, m.in., nakłady gazet i czasopism, frekwencję w kinach itd.;
- opracowując zagadnienie tworzenia wyboru, warto pamiętać, że nie ma takich procedur, które można by zastosować do każdego postawionego zadania. Dla każdego badania to zagadnienie jest rozwiązywane indywidualnie, w zależności od celu posiadanych przez analityka dokumentów, wymaganego poziomu dokładności wyników;
- tworzenie wyboru kończy się sporządzeniem listy badanych dokumentów, a także tabeli, zawierającej informacje, dotyczące ilości wybiórczych informacji w formie dokumentów, ich podziału zgodnie z ramami czasowymi a także rodzajem dokumentów.

6. Wybór jednostek analizy

Zalecenia:

- jednostki analizy – są to niewielkie lub duże fragmenty tekstu, które osoba kodująca metodycznie obserwuje, odnajdując w nich odsyłacze do kategorii znaczeń lub ich wyrażone werbalnie cechy;

- jednostka analizy – to, co jest podliczane;
- w zapisie treści może to być wyraz lub symbol, temat (odrębne stwierdzenie, dotyczące przedmiotu badania), cały artykuł lub odrębna historia; do analizy programów telewizyjnych jednostka analizy może być symbolem, wydarzeniem, całym programem;
- operacyjne określenia jednostek analizy powinny być zadane wyraźnie i rzetelnie;
- jednostki analizy powinny być oczywiste tak, żeby można było je łatwo zidentyfikować i wykrywać; w celu wykrycia jednostek analizy potrzebna jest czasami określona praca wstępna.

Typologia jednostek analizy (wg K. Krippendorffa[62]):

- „fizyczne" (przedmioty z jasno określonymi granicami, fizycznymi, geometrycznymi lub czasowymi, np. egzemplarze książki, numery dzienników, egzemplarze plakatów lub ulotek, zdjęcia itd.);
- strukturalno-semiotyczne (główne elementy systemów semiotycznych; np. leksyka lub wskaźniki gramatyczne, jak partykuły przeczące lub wskaźniki takich kategorii, jak rzeczowniki odczasownikowe);
- pojęciowo-tematyczne (pojęcia, tematy, problemy, kryjące się za wyrazami, np. „kryminał", „wolność słowa");
- referencyjne i quasi-referencyjne (do nich zaliczane są określenia realnych osobistości, wydarzeń, miast, krajów, organizacji itd.; np. „nasz Kobzarz", „matka miast ruskich");
- proponowane i szacunkowe (np. za frazą „Ukraina dusi się bez inwestycji" stoi konkretne stwierdzenie faktu);
- makrostrukturalne (złożone konstrukcje pojęciowe, np. obecne w świadomości społecznej współczesnej Ukrainy: zmowa, orgia korupcji, rewolucja kryminalna, kraj durniów, walka o władzę);
- wyniki operacji konceptualnych (metafory, przykłady i analogie, np. metafora „wojenna": „wojna z biedą", „uderzenie w gubernatora", „atak ze strony opozycji", „publikacja pogromowa");
- poetyckie (dające się zmierzyć ilościowo środki wyrazu artystycznego, np. kalambury, aliteracje).

7. System kwantyfikacji (podliczenia ilości)

Rodzaje systemów kwantyfikacji (wykonane na podstawie praktyki badań):

- system liczenia „czas – przestrzeń" (za podstawę są brane wskaźniki zajmowanej powierzchni dla tekstów pisanych, np. ilość wierszy, akapitów, znaków, powierzchnia w centymetrach kwadratowych, wysokość standardowej szpalty gazetowej lub czas wiadomości audiowizualnych, jak czas wyświetlania wydarzenia);
- pojawienie się kategorii lub oznak w tekście (wykorzystuje się w przypadku, gdy jednostki analizy trudno sformalizować, np. w analizie treści informacji artystycznoobrazowej; przy danym systemie, po podziale tekstu na części, określana jest obecność jakiejś charakterystyki treści w każdej części, ponowne pojawienie się tej charakterystyki w tej samej części się nie liczy);
- częstotliwość występowania jednostek analizy (zapisywana jest nie tylko obecność /lub jej brak/ jednostki analizy, ale również ile razy ona występuje).

Uwagi:

- badacz ma podliczać częstotliwość występowania jednostek analizy w przekroju każdej kategorii;
- osoby kodujące powinny być poinstruowane odnośnie do wykorzystania skali ocen w procesie oceny pewnych właściwości symboli lub sytuacji;
- skala ocen umożliwia wykrycie głębokości i struktury treści, jednocześnie wprowadza pewien subiektywizm do analizy i wpływa na pewność wyników;
- częstotliwość występowania tematu w tym czy innym dokumencie może być wskaźnikiem jej znaczenia;
- podliczenie charakterystyk oceny tekstu (wszystkich „za" i „przeciw", dotyczących danego zagadnienia czy wydarzenia) umożliwia badanie nastawienia autora i wykrycie zamiarów, które podyktowały daną wiadomość.

8. Tworzenie kategorii analizy

Zalecenia ogólne:

- centralnym punktem dowolnej analizy treści jest system kategorii;
- kategorie analizy – to najogólniejsze pojęcia kluczowe, od-

62 Krippendorff K. Content Analysis: An introduction to its Methodology....

powiadające zadaniom badania, przy pomocy których opisywany jest przedmiot badania i zgodnie z którymi będą klasyfikowane jednostki analizy;
- wszystkie kategorie systemu powinny być wykluczające się wzajemnie, wyczerpujące i pewne, prawidłowo reprezentować tekst;
- jednostka analizy powinna zatem kojarzyć się tylko z jedną kategorią i w żadnym wypadku nie z kilkoma kategoriami jednocześnie;
- kategorie powinny obejmować wszystkie możliwe jednostki analizy w granicach wyboru;
- system kategorii analizy może zawierać typologię audycji lub typy materiału dziennikarskiego, typy osób lub zachowania osób;
- system kategorii ma być pewny w tym rozumieniu, że różne osoby kodujące powinny otrzymywać jednakowe wyniki;
- od wyboru kategorii w znacznej mierze zależy charakter otrzymanych wyników; dlatego przy ich opracowaniu badacz stoi przed koniecznością kilkakrotnego przejścia od schematu teoretycznego do danych, a od nich – znów do schematu, aby wykryć relewantny, do celów badań system kategorii.

Ankieta służąca do określenia kategorii analizy:
- Czy analiza będzie się zajmowała policzeniem powierzchni na stronach gazet, która odnosi się do materiałów pewnego typu?
- Czy analiza będzie określać i mierzyć specyficzne wyrazy lub frazy w tekście?
- Czy w centrum analizy znajdą się pewne specyficzne kategorie zachowania na ekranie?
- Czy analiza będzie podliczać pojawienie się programów, gazet, informacji, szczegółów samych informacji? Jaki poziom podliczenia będzie miał miejsce?

Czynniki, uwzględniane podczas określania kategorii:
- aktorzy (politycy, sportowcy, przemysłowcy, naukowcy, przestępcy itd.);
- źródła (partie polityczne, eksperci, pewne instytucje itd.);
- kontekst (gdzie przeprowadza się nagranie wywiadu – w domu, w pracy, w studiu itd.);
- zapewnienie wiadomości źródłowych (faktyczne dane, oceny, obliczenia, zalecenia itd.);
- obecność źródła (częściowo określane jest kontekstem i w wyniku opracowania informacji, przedstawionej przez media);
- podmioty, tematy i problem (jeśli, przykładowo – badany jest problem gwałtu w mediach, to stawiane jest pytanie: jakie rodzaje gwałtu (fizyczne, werbalne, szkoda) są przedstawiane?);
- zasób słownikowy lub wybór leksyki (częstotliwość, z którą pewne wyrazy są wykorzystywane; wykorzystywanie terminów emocjonalnych, np. te same osoby można nazwać różnie: obrońcy wolności, terroryści, gangsterzy);
- sposób mierzenia znaczeń (np. materiał kodowany jest jako: „pozytywny", „negatywny", „neutralny').

Schemat (model) konceptualny:
- kategorie mogą zostać podzielone na bardziej szczegółowe – podkategorie;
- ogół kategorii i podkategorii tworzy schemat konceptualny analizy treści;
- schemat konceptualny – nie jest abstrakcją teoretyczna, lecz konstrukcją treści, adaptowaną do przedstawienia obrazu przedmiotu, obecnego w rzeczywistości tekstowej; ważne jest, by była ona przydatna do pełnego i jednocześnie oszczędnego opisu przedmiotu, uwzględniając teoretyczne wyobrażenia o nim i doświadczenie jego wersji tekstowych.

Sposoby opracowania schematu konceptualnego:
- kategorie są dobierane w drodze eksperymentu na podstawie wyobrażeń o przedmiocie, np. eksperci nazywają najaktualniejsze problemy społecznie – ekonomiczne i ten wykaz służy jako znamionowa skala kategorii w analizie treści współczesnej prasy. Jej celem jest wyjaśnienie jaki obraz rzeczywistości prasa proponuje czytelnikom;
- kategorie są brane z badanych tekstów; jest to tak zwana droga poszukiwań, do której należy się zwrócić wtedy, gdy istnieje potrzeba wysuwania hipotez, dotyczących przedmiotu badania;
- kategorie są wybierane przez badacza na bazie pragma-

tycznej, czyli badacz opiera się na własnych teoretycznych wyobrażeniach o przedmiocie, analizie historycznokulturalnej i społecznej przedstawionej w tekstach sytuacji, praktyce analizy treści.

Pewność kategorii:

- kategorie powinny spełniać wymogi pewności (im bardziej jasno wyznaczone są kategorie, tym mniej problematyczne staje się zaliczenie części treści do określonych kategorii);
- najbardziej radykalnym sposobem zwiększenia stopnia pewności kategorii jest jej wyczerpujące oznaczenie (przeliczenie wszystkich elementów części treści); jednak takie przypadki zdarzają się rzadko i wyczerpujące oznaczenia możliwe są dla stosunkowo wąskiej klasy prostych zadań);
- przy ocenie pewności kategorii warto pamiętać, że wąska granica ich oznaczenia doprowadza do rozbieżności pomiędzy osobami kodującymi przy określonej przynależności tego czy innego dokumentu do pewnej kategorii; można to wyeliminować poprzez zwiększenie (uogólnienie) kategorii, jednak znaczne zwiększenie może doprowadzić do zmniejszenia stopnia dyferencjacji badanego zjawiska i badacz nie zauważy istotnych dla celów badania rozbieżności;
- nadmierna ilość wykorzystywanych kategorii prowadzi także do analitycznych pomyłek; ciężko jest określić optymalną ilość: chęć uzyskania maksymalnie dokładnego opisu przedmiotu poprzez zestaw pojęć może obrócić się w zbyt rozdrobniony schemat kategorialny, nie dostosowany do klasyfikacji treści tekstów w wygodnej i przekonującej formie;
- wysoka pewność kategorii zazwyczaj właściwa jest dla prostych form analizy treści; przy skomplikowanych występuje bardziej subtelna dyferencjacja kategorii ze zmniejszeniem ich pewności, jednak pomaga to w otrzymaniu szerszych i głębszych informacji dotyczących badanego obiektu; decyzja co do współzależności potrzeb i znaczenia, jest podejmowana na podstawie postawionych przed badaczem zadań.

9. Zapewnienie pewnej pracy dla osób kodujących

Zalecenia:

- badanie powinno wyraźnie określać, jakie jednostki treści będą podliczane (wyrazy, ilość cm2 itd.);
- każdą jednostkę, która ma być określana, należy oznaczyć; jeśli kategorie są dzielone na podkategorie, to badacz musi określić każdą podkategorię i sposób jej rozpoznawania;
- należy także określić schemat liczbowy, który wskazuje dla każdej kategorii, jaki kod liczbowy jest przydzielany i w którym przypadku;
- badacz powinien opracować klasyfikator analizy i instrukcję do kodowania, która zadaje szablony kodowania dla każdej kategorii.

Klasyfikator analizy:

- klasyfikator analizy treści stanowi dokument metodyczny, przeznaczony do objaśnienia procesów wyodrębnienia i rejestracji cech informacji tekstowej;
- klasyfikator analizy treści – to tabela ogólna, w której zestawione są wszystkie kategorie i podkategorie analizy i jednostki analizy;
- podstawowe przeznaczenie klasyfikatora – maksymalnie wyraźny zapis, w jakich jednostkach analizy jest wyrażana każda kategoria, wykorzystywana w badaniu; klasyfikator można upodobnić do ankiety socjologicznej, gdzie kategorie analizy występują w roli pytań, a jednostki analizy – odpowiedzi;
- klasyfikator analizy jest podstawowym dokumentem metodologicznym, który określa treść wszystkich innych narzędzi.

Etapy projektowania klasyfikatora:

- formułowanie zadań i projektowanie klasyfikatora;
- tworzenie kategorii analizy;
- terminologiczne uściślenie zapisanych cech kategorii analizy w języku tekstu dokumentu, usunięcie dwuznaczności przecięcia się dwóch lub więcej kategorii;
- określenie przypadków nieterminologicznego zapisu oznak cech kategorii analizy;
- ustalenie systemu kwantyfikacji;
- konstruowanie narzędzi do pomiaru i rejestracji jednostek analizy (kart rejestracyjnych, wierszy itd.);

- tworzenie ogólnego schematu logicznego kategorii analizy i systemu kwantyfikacji;
- tworzenie ogólnego schematu klasyfikacji w kolejności optymalnej do rejestracji cech i ich dalszego opracowania;
- projektowanie druków analizy na podstawie klasyfikatora;
- akceptacja projektów druków i ich uściślenie;
- opracowanie programu opracowywania druków analizy;
- zatwierdzenie klasyfikatora, jego powielanie.

10. Kodowanie treści

Uwagi ogólne:

- proces zaliczenia jednostki analizy do kategorii analizy nazywany jest kodowaniem;
- osoby, zajmujące się kodowaniem, nazywane są osobami kodującymi;
- do kodowania, zazwyczaj, wykorzystywana jest niewiele osób kodujących;
- rzetelne szkolenie osób kodujących jest istotne i gwarantuje uzyskanie najpewniejszych wyników;
- osoby kodujące powinny być dobrze zapoznane z metodą kodowania i klasyfikatorem treści;
- aby prace zostały należycie wykonane osoby kodujące muszą być dobrze przeszkolone;
- potrzebne jest badanie eksperymentalne w celu sprawdzenia pewności interkodowania;
- kodowanie wykonuje się na drukach kodujących, które umożliwiają osobom kodującym klasyfikowanie treści; druki są wypełniane poprzez naniesienie pewnych znaków w konkretnych miejscach naprzeciw uprzednio określonych wskaźników (kodów kategorii).

11. Analiza danych i interpretacja

Uwagi ogólne:

- do analizy danych może być wykorzystywana statystyka opisowa, np. procenty, średnie arytmetyczne, mody, mediany;
- bardziej skomplikowana statystyka wykorzystywana jest do sprawdzenia związków pomiędzy zmiennymi, najpopularniejszą z nich jest statystyka chi-kwadrat;
- do zbadania pewnych rodzajów danych mogą być stosowane: analiza korelacyjna i dyspersyjna;
- badacz powinien być nakierowany na wszechstronne podliczenie natury treści w formie opisowej, ale także rozumieć, że istnieją związki pomiędzy różnymi rodzajami treści; wyniki są oceniane według pytań wstępnych badania lub hipotez.

Przedstawienie i interpretacja wyników:

- znaczną rolę w rozumieniu wyników analizy treści odgrywa przygotowanie metodyczne: zrozumiała i prawidłowa metodyka badania, pewna procedura sformalizowanego czytania;
- do otrzymania wyważonych i prawidłowych wniosków służy wizualne przedstawienie wyników badania; w praktyce najczęściej wykorzystywane są różne środki graficzne: diagramy, wykresy, schematy, grafy;
- interpretacja treści wyników zależy od celów analizy; jest ona procesem twórczym, którego wyniki w wielu przypadkach określają: kwalifikacje i intuicja analityków.

12. Pewność

Zalecenia ogólne:

- pojęcie pewności jest krytyczne dla analizy treści;
- do osiągnięcia założonego poziomu pewności zalecane są następujące kroki:
  1. określenie kategorii z maksymalną dokładnością; nieokreślone lub dwuznacznie określone kategorie stanowią źródło o niskiej pewności;
  2. gruntowne szkolenie osób kodujących; przed rozpoczęciem zbioru danych należy się przekonać, że osoby kodujące znają zasady kodowania i umieją je stosować; należy przyjąć też odpowiednią ilość czasu na omówienie i rozwiązanie wszelkich problemów;
  3. przeprowadzenie eksperymentalnego badania; należy wybrać z całości pewne teksty i przekazać je osobom kodującym do zakodowania.

Wskaźnik pewności:

- obliczenie procentu zgodności osób kodujących co do ocen lub kodowania kategorii stanowi najprostszą metodę weryfikacji pewności;
- ta miara pewności jest określana jako stosunek: % faktycznej zgody π = 1 – % oczekiwanej zgody
- założony poziom niezbędnej zgodności zależy od rodzaju

prowadzonego badania; zwykle standardowym jest minimalny poziom w wysokości 80 procent;

- przykład: normatywne (oczekiwane) wyniki ocen reklamy w czasopismach wg siedmiu kategorii powinny być następujące:
  - reklama wizerunku – 30 procent;
  - reklama sprzedaży – 20 procent;
  - reklama zaangażowania w promocję – 20 procent;
  - reklama-przestroga – 15 procent;
  - wezwanie utylitarne – 10 procent;
  - inne wezwania – 5 procent.

Najpierw obliczany jest procent oczekiwanej zgodności dwóch osób kodujących: jest to suma kwadratów procent w każdej kategorii: % oczekiwanej zgody = (0.3) 2 + (0.2) 2 + (0.2) 2 + (0.15) 2 + (0.1) 2 + (0.05)2 = .20 Jeśli osoby kodujące zgadzają się co do 90 procent z tego, co było oczekiwane, to π może być obliczane następująco: π = 0,19 0– –0 ,02,02 0 = 0,875 .

Ostateczna weryfikacja pewności kodowania:

- po zadowalającym zakończeniu szkolenia osób kodujących i przeprowadzeniu badania eksperymentalnego, kolejne sprawdzenie pewności wykonywane jest podczas pracy z głównym założeniem;
- zazwyczaj zaleca się 10-25 procent wielkości założenia opracowywać z ponownym kodowaniem dublującym;
- do sprawdzenia pewności interkodowania wykorzystuje się kilka indeksów;
- indeks Olsti jest obliczany następująco:

Pewność = N21 *+ MN 2, gdzie M – to ilość decyzji osób kodujących, co do których są oni zgodni, a N1 i N2 oznacza ilość rozwiązań kodujących (np. ilość zakodowanych artykułów, jeśli jednostka analizy – to artykuł), które podjął odpowiednio pierwszy i drugi kodujący. Wówczas, o ile obydwie osoby kodujące kodowały po 50 jednostek analizy i zgodziły się co do 35, to otrzymujemy Pewność = 520 *+ 3 550 = 0,70.

13. Błędy podczas przeprowadzania analizy treści

Podczas wykorzystywania analizy treści badacze często popełniają błędy. Oto wykaz najbardziej rozpowszechnionych:

- analiza dokumentów poprzedza opracowanie programu badań;
- analizowane są dokumenty, nie mające związku z hipotezami badania, np. są zgodne z tematem badania wyłącznie w odniesieniu do nazwy;
- nie sprawdzono wiarygodności dokumentów;
- nie uwzględniono ich autorstwa;
- nie w pełni uwzględniono ich przeznaczenie;
- kategorie analizy nie są wyznaczone z taką dokładnością, która pozwala wyraźnie odróżniać jednostki znaczeń tekstu dokumentu;
- kategorie analizy nie są subordynowane i nie są doprowadzone do zgodności z tymi definicjami i terminami, które je operacjonalizują;
- kategorie analizy nie są zgodne z treścią i językiem tekstu analizowanego dokumentu;
- jednostki analizy charakteryzują kategorie tylko z zewnątrz, a nie zgodnie zprzesłaniem treści;
- jednostki analizy nie umożliwiają identyfikacji treści dokumentu wg kategorii analizy;
- analizowane są dokumenty przy braku pełnego zestawu narzędzi metodycznych;
- klasyfikator sporządzono z naruszeniem praw logiki;
- osoby kodujące nie otrzymały należytego przygotowania metodycznego;
- instrukcja do rejestracji i kodowania jest niedostateczna, sporządzona przez badacza, który nie przeprowadził wstępnej aprobaty narzędzi;
- wyniki analizy treści nie zostały ponownie sprawdzone poprzez informację, otrzymaną innymi metodami.

ZAŁĄCZNIK B. Programy analizy treści

Wykaz i opis programów analizy treści sporządzono na podstawie szeregu prac K. Neuendorfa[63], M.Alexy[64] i grupy badawczej z Wielkiej Brytanii ESRC[65].

1. Programy ilościowej analizy treści

CATPAC

http://www.terraresearch.com/

CATPAC czyta pliki tekstowe i formuje różnego rodzaju raporty, począwszy od zwykłej tabeli częstotliwości, uporządkowanej wg częstotliwości lub alfabetycznie kończąc krótkim raportem podsumowującym „główne idee" tekstu. Program przedstawia wzory wykorzystania słów i wykonuje analizę klasterową lub dyspersyjną. Dodatkiem do niego jest program, który przedstawia dane wyników badania w formie grafiki dwu - i trójwymiarowej.

Computer Programs for Text Analysis

http://unix.dsu.edu/~johnsone/ericpgms.html

Stanowi całokształt programów E.Johnsona, z których każdy jest przeznaczony do realizacji jednej lub kilku funkcji, np. analiza charakterystyk zewnętrznych bohaterów sztuki (program ACTORS), otrzymanie KWIC (program CONCORD), znajdowanie cytatów w tekstach (program DIALOG), porównanie słownika dwóch tekstów (program IDENT). Ogólnie, programy przeznaczane są do analizy literackiej tekstów.

Concordance 2.0

http://www.rjcw.freeserve.co.uk/

Pozwala formować pełne konkordancje tekstów o dowolnym wymiarze, ograniczając się jedynie wolnym miejscem na dysku. Można także operatywnie otrzymać konkordancje dla poszczególnych wyrazów, a także formować konkordancje WWW.

Diction 5.0

http://www.sagepub.com/

Diction 5.0 posiada szereg wbudowanych słowników, które klasyfikują i znajdują dokumenty wg pięciu cech semantycznych (działalność, optymizm, pewność, realizm i unifikacja) i 35 cech bardziej szczegółowych (m.in. zawziętość, odpowiedzialność, dwojakie nastawienie, pobudzenie, komunikacja).

Po tym, jak tekst użytkownika zostaje przeanalizowany, Diction 5.0 porównuje wyniki dla każdej z 40 kategorii słownika z „normalnym rzędem szacunkowym", sformowanym na podstawie analizy przez program ponad 20 tys. tekstów. Użytkownicy mogą porównać swój tekst albo z profilem normatywnym 20 tys. tekstów, albo z 6 odrębnymi rodzajami tekstów (biznes, wydarzenia codzienne, rozważania, dziennikarstwo, literatura, polityka, edukacja), które mogą, z kolei dzielić się na 36 podrodzajów (np. sprawozdania finansowe, czaty komputerowe, muzyka liryczna, czołówki gazet, nowele lub krótkie opowiadania, debaty polityczne, badania opinii społecznej). Poza tym, Diction 5.0 oblicza częstotliwości bezwzględne i względne wyrazów lub oceny standaryzowane i wyprowadza je wg częstotliwości występowania lub alfabetycznie. W wyniku analizy dodatkowej program może tworzyć słowniki użytkownika.

DIMAP (DIctionary MAintenance Programs)

http://www.clres.com/

Celem DIMAP jest opracowanie słownika. Program zawiera różne narzędzia do lingwistyki komputerowej i procesorów języka naturalnego. Przy pomocy DIMAP można tworzyć, zarządzać, edytować, obsługiwać słowniki, szukać wyrazów i porównywać słowniki użytkownika z leksyką tradycyjną. Program zawiera również moduł analizy tekstowej MCCA, opisany niżej.

63 Neuendorf K.A. The Content Analysis Guidebook Online. Computer Content Analysis Programs. – http://academic.csuohio.edu/kneuendorf/content/cpuca/ccap.htm

64 Alexa M. Computer-assisted text analysis methodology in the social sciences...– 40 p.; Alexa M., Zuell C. Commonalities, differences and limitations of text analysis Software...– 29 r.

65 Assessment and Development of New Methods for the Analysis of Media Content // An ESRC Research Methods Programme: Research Bulletin 01.– May 2003.– www.lboro.ac.uk/research/mmethods

General Inquirer (Internet version)

http://www.wjh.harvard.edu/~inquirer/

Ten poważny i wciąż jeszcze wykorzystywany program odrodził się w World Wide Web. Wersja dialogowa General Inquirer pozwala łatwo i szybko wykonać nieskomplikowaną analizę komputerową tekstu: wchodząc na stronę General Inquirer i wprowadzając tekst do odpowiedniego bloku, można od razu otrzymać wyniki analizy. General Inquirer koduje i klasyfikuje tekst, wykorzystując słownik Harvard IV-4 do oceny takiej cechy, jak np. męstwo, wykonuje trójwymiarową ocenę semantyczną wg metody C. Osguda, wyznacza poziom emocjonalności języka, orientację poznawczą tekstu i in.. Na zakończenie każdej analizy program także zwraca statystykę kumulatywną, m.in. proste częstotliwości dla wyrazów.

HAMLET

http://www.apb.cwc.net/homepage.htm

Główna idea programu HAMLET to znalezienie w pliku tekstowym wyrazów z zadanej listy, a także podliczenie częstotliwości w granicach jakiejkolwiek określonej jednostki kontekstu lub łączenie wyrazów w granicach zadanego zakresu wyrazów. Indywidualne częstotliwości wyrazów, częstotliwości wspólnego występowania wyrazów, wyrażone w terminach wybranej jednostki kontekstu, przedstawiane są w macierzy podobieństwa, która może zostać następnie poddana tradycyjnej analizie klasterowej.

INTEXT/TextQuest – Text Analysis Software

http://www.intext.de

INTEXT jest to program, przeznaczony do analizy tekstów o kierunku humanistycznym i ogólnym. Formuje on indeksy, konkordancje, tabele KWIC i KWOC, przeprowadza analizę poczytności, indywidualną analizę strukturalną, formuje listy wyrazów, kolejności wyrazów, możliwych przestawień wyrazów, analizuje stylistykę itd.. Program TextQuest jest wersją Windows programu INTEXT.

LIWC (Lingustic Inquiry and Word Count software)

https://www.erlbaum.com/shop/tek9.asp?pg=products&specific=1-56321-208-0

LIWC posiada serię 68 wbudowanych słowników, pomagających realizować wyszukiwanie plików tekstowych i podliczenie, jak często wyrazy odpowiadają każdemu z 68 wcześniej ustalonych parametrów. Słowniki zawierają oceny lingwistyczne, kategorie słowne konstrukcji psychologicznych i indywidualne kategorie problemowe. Program umożliwia także użytkownikom tworzenie własnych słowników. Jest on szczególnie korzystny dla psychologów, którzy badają zapisy rozmów z pacjentami.

MCCA

http://www.clres.com/

MCCA analizuje tekst, tworząc listy częstotliwościowe i alfabetyczne, tabele KWIC, a także wykonuje kodowanie na podstawie wbudowanych słowników. Wbudowane słowniki specjalizują się w wyszukiwaniu pewnych szczegółów tekstu, na przykład, rodzaj działalności, slang, cechy humoru itd.. Wyniki są wyprowadzane w formie okien. Można łatwo uporządkować okna, przechodzić od jednego do innego. MCCA także pozwala analizować nagrania wywiadów, różnych słuchowisk, występów w telewizji, pracy grup dyskusyjnych, gier biznesowych, związanych z udziałem dużej ilości osób.

MECA (Map Extraction Comparison and Analysis)

MECA zawiera 15 szablonów służących do analizy tekstu. Wiele z nich przeznaczonych jest do tworzenia map kognitywnych i kładą one uwagę zarówno na samych pojęciach kognitywnych, jak i związkach pomiędzy nimi. Szablony wykorzystywane są do realizacji klasycznej analizy treści, np. program wykonuje podliczenie ilości pojęć do każdej mapy.

MonoConc
http://www.ruf.rice.edu/~barlow/mono.html
MonoConc przeznaczony jest do formowania konkordancji. Wyniki mogą na różny sposób być porządkowane i przedstawiane w konfiguracjach, zadanych przez użytkownika. Program także oblicza częstotliwości wyrazów do zadanego ogółu tekstów.

ParaConc
http://www.ruf.rice.edu/~barlow/parac.html
ParaConc – dwujęzyczny/wielojęzyczny program konkordancji, stosowany podczas komparatystycznych badań językowych tekstów.

PCAD 2000
http://www.gb-software.com/
PCAD 2000 wykonuje analizę treści nagrań przemówień i innych tekstów, stosując podejście Gottschalka-Glesera do oceny umysłowego i emocjonalnego stanu osób. Do ocenianych parametrów należą niepokój, wrogość, odtrącenie społeczne, nadzieja, depresja itd.. Program umożliwia porównanie otrzymanych ocen ze wskaźnikami normatywnymi do różnych grup demograficznych subiektów. Może on wykonywać niektóre klasyfikacje diagnostyczne, kierując się „Instrukcją diagnostyczną i statystyczną rozstrojów psychicznych" (DSM-IV), opracowaną przez Amerykańskie Stowarzyszenie Psychiatryczne.

PROTAN (for PROTocol ANalyzer)
PROTAN – zautomatyzowany system analizy treści. Pierwsze pytanie, na które próbuje odpowiedzieć brzmi: jak wygląda tekst. Do osiągnięcia celu PROTAN wykorzystuje szereg słowników semantycznych, które są częściami składowymi systemu. Inne zadanie: tekst wyjaśnienie przesłanek tekstu.

SALT (Systematic Analysis of Language Transcripts)
http://www.waisman.wisc.edu/salt/index.htm
Program został opracowany głównie po to, żeby pomóc lekarzom w wyodrębnieniu i udokumentowaniu specyficznych problemów językowych pacjentów. Wykonuje on szereg rodzajów analizy, m.in. związanych z wymową (np. niepełne wypowiedzi, niewyraźność, niewerbalność), obliczeniem długości wypowiedzi, ilości i czasu trwania pauz, prędkości mowy, analiza częstotliwości różnych zestawów wyrazów (np. zaprzeczenia, spójniki, słowniki użytkownika).

SWIFT (Structured Word Identification and Frequency Table)
Swift w procesie dialogu wykonuje zorientowaną na wyrazy kluczowe analizę krótkich tekstów. Jest aktualizowany bezpłatnie, system operacyjny – MS DOS.

TABARI (Text Analysis By Augmented Replacement Instructions)
TABARI jest spadkobiercą programu KEDS. Jest on stosowany do analizy krótkich informacji, wiadomości, różnych raportów służbowych. Koduje on dane obejmujące wydarzenia międzynarodowe (osobliwy jest tutaj fakt, że przedstawiają one wzajemne działania pomiędzy uczestnikami), wykorzystując szablony rozpoznawania i zwykłą analizę gramatyczną. Autorzy opracowali cały szereg słowników do kodowania wydarzeń. Przykładowo – schemat kodowania KWEIS może określić, kto działa przeciwko komu, np. Irak przeciwko Kuwejtowi. Kiedy taki temat pojawia się w materiale dziennikarskim, program może automatycznie kodować agresora, ofiarę i działanie, a także datę wydarzenia.

TACT (Text Analysis Computing Tools)
http://www.chass.utoronto.ca:8080/cch/tact.html
TACT jest systemem analiz i poszukiwania tekstu dla MS-DOS, co pozwala na realizację zapytań w językach europejskich podczas opracowania tekstowych baz danych. Został on opracowany przez zespół programistów, projektantów i naukowców – fachowców w dziedzinie analizy tekstowej. Ponadto, system zawiera dodatkowy program TACTweb, który jest środkiem do połączenia TACT z World Wide Web. Wykorzystując format WWW, użytkownicy otrzymują dostęp do wielu usług dialogowych TACT.

TEXTPACK 7.0

http://www.social-sciencegesis.de/en/software/textpack/index.htm

Program TEXTPACK, projektowany pierwotnie do analizy ankiet z zapytaniami otwartego typu, został rozszerzony w ostatnich latach na skutek włączenia analizy treści, analityki literackiej i lingwistycznej.

Obecnie podlicza on częstotliwości występowania wyrazów, listy alfabetyczne, tabele KWIC i KWOC, odsyłacze krzyżowe, wykonuje porównanie wyrazów w dwóch tekstach, a także kodowanie zgodnie z opracowanymi przez użytkownika słownikami. Wyniki mogą być eksportowane w formacie programów analizy statystycznej. Wersja oprogramowania Windows pomyślnie wykorzystuje zalety interfejsu systemu operacyjnego.

TextSmart by SPSS Inc.

http://www.spss.com/spssbi/textsmart/

Oprogramowanie to projektowane było przede wszystkim w celu analizy ankiet z pytaniami typu otwartego, wykorzystuje ono analizę klasterową i technikę skalowania wielowymiarowego do automatycznej analizy słów kluczowych oraz grupuje teksty w kategorie. Program może „kodować" teksty, nie wykorzystując stworzonych przez użytkowników słowników. TextSmart ma przyjemny i łatwy w użytkowaniu interfejs systemu Windows, umożliwiający szybkie sortowanie list wyrazów alfabetycznie i wg częstotliwości, przedstawianie informacji w formie kolorowych diagramów, wykresów, map, schematów, tabel.

VBPro

http://excellent.com.utk.edu/~mmmiller/vbpro.htm

Program formuje częstotliwości i alfabetyczne listy wyrazów, tabelę KWIC, zakodowane rzędy wyrazów, wykorzystując utworzone przez użytkownika słowniki. Pakiet programów zawiera model VBMap, przeznaczony do tworzenia map wielowymiarowych, które mierzą i przedstawiają stopień wspólnego występowania wyrazów w tekście lub wspólnego występowania fragmentów tekstu. Program pracuje w MS DOS.

WordStat v3.01

http://www.simstat.com/wordstat.htm

Ten program, będąc dodatkiem do programu analizy statystycznej Simstat, zawiera kilka narzędzi badawczych, np. analizę klasterową i wielowymiarowe modele statystyczne, do analizy ankiet z zapytaniami typu otwartego i tekstów innego rodzaju. Umożliwia on także przeprowadzenie kodowania na podstawie słowników użytkownika, podlicza częstotliwość występowania wyrazów, formuje alfabetyczne listy wyrazów, tabele KWIC , przeprowadza porównanie pomiędzy podgrupami danych. Różnice między podgrupami lub zmiennymi liczbowymi (np. wiek, data publikacji) mogą być przedstawione wizualnie na diagramach w formie dwu- lub trójwymiarowej grafiki. Na szczególną uwagę zasługuje narzędzie specjalne, umożliwiające użytkownikowi tworzenie ogólnego schematu kategorii, korzystając z leksykalnej bazy danych WordNet i innych słowników (jeden słownik angielski i pięć słowników innych języków).

2. Programy jakościowej analizy treści

ATLAS/ti

To oprogramowanie przeznaczone jest do interpretacji treści tekstu, zarządzania tekstem, pozyskiwania wiedzy konceptualnej z tekstu, czyli tworzenia teorii. Zakres wykorzystania: nauki społeczne, ekonomia, badania rynku, oświata, kryminalistyka, zarządzanie jakością, tworzenie baz danych, teologia. Oprogramowanie pracuje z MS DOS i MS Windows.

Code-A-Text

Code-A-Text – pakiet programów, napisany do pomocy w szkoleniu psychoterapeutów. Najpierw zakładało się, że ma on ułatwić analizę rozmów terapeutycznych, podczas których klinicyści, wykładowcy i naukowcy próbują zrozumieć idee i struktury, założone w tekstach. Obecnie Code-A-Text jest stosowany do analizy innych rodzajów tekstów, m.in. memo (przypisów), odpowiedzi na zapytania ankiet, metaplanów. Wkrótce ma zostać uruchomiona wersja programu, wspomagająca kodowanie plików dźwiękowych.

3. Programy, które pomagają w przeprowadzaniu jakościowej analizy danych

The Ethnograph v4.0
Program zapewnia badania jakościowe i analizę danych, ułatwia zarządzanie i analizę tekstowych baz danych, m.in. nagrań wywiadów, prace grup dyskusyjnych, krótkich spotkań, memo, dzienników, innych dokumentów.

Kwalitan 4.0
Kwalitan – jest to program do wspierania analizy danych jakościowych, m.in. protokołów wywiadów, obserwacji, istniejących materiałów pisemnych, gazet, rocznych sprawozdań instytucji, starożytnych rękopisów itd.. Faktycznie, Kwalitan – jest programem do sterowania bazami danych o specjalnym przeznaczeniu. Został opracowany do realizacji procedur w ramach „teorii ugruntowanej".

NUD*IST
Program pomaga badaczom w manipulowaniu nieliczbowymi niestrukturalnymi danymi drogą indeksacji, poszukiwania i tworzenia „teorii". Między innymi, automatyzuje żmudną pracę „autokodowania" tekstu.

winMAX
Pomaga w przeprowadzeniu jakościowej analizy danych.

ZAŁĄCZNIK C. Programy Text Mining

Wykaz i opis programów Text Mining sporządzono na podstawie prac D.Landego[66].

Intelligent Miner for Text
http://www3.ibm.com/software/data/iminer/fortext/
Produkt firmy IBM, Intelligent Miner for Text, stanowi zestaw odrębnych jednostek, które uruchomiane są z rzędu komend niezależnie od siebie. System ten jest jednym z najlepszych narzędzi „głębokiej" analizy tekstów. System składa się z następujących jednostek zarządzania informacjami:

- Language Identification Tool (automatyczne rozpoznawanie języka, w którym został sporządzony dokument);
- Categorisation Tool (jednostki klasyfikacji, czyli automatycznego zaliczenia tekstu do jakiejś kategorii; informacją wejściową na etapie nauki może być wynik pracy jednostki Clusterisation Tool);
- Clusterisation Tool (jednostka klasteryzacji, czyli podziału dużej ilości dokumentów na grupy wg bliskości stylu, formy, różnych charakterystyk częstotliwości wykrytych słów kluczowych);
- Feature Extraction Tool (jednostka określenia nowego wykrycia w dokumencie nowych słów kluczowych, np. imion, nazw, skrótów, na podstawie analizy wcześniej zadanego słownika);
- Annotation Tool (jednostka „wykrytej treści" tekstów i sporządzania na ich podstawie referatów lub adnotacji).

IBM Intelligent Miner for Text łączy olbrzymi zestaw narzędzi, bazujących głównie na mechanizmach poszukiwania informacji (information retrieval), co jest cechą szczególną całego produktu.
System zawiera szereg składników bazowych o samodzielnym znaczeniu poza granicami technologii Text Mining, tj. wyszukiwarkę informacyjną Text Search Engine, jednostki skanowania przestrzeni WWW – Web crawler, Net Question Solution – roz-

66 Lande D. Zdobywanie wiedzy...; Lande D. Monitoring treści New Media.– http://uaport.net/files/?dwl-yal01;
Lande D. O oddzielaniu ziaren od plew...; Lande D., Litwin A. Fenomeny współczesnych strumieni informacyjnych // Sieci i biznes.– Nr1.– 2001.

wiązanie do wyszukiwania na lokalnej stronie WWW lub na kilku intranet-/Internet-serwerach, Java Sample GUI – zestaw interfejsów Java Beans do administrowania i organizowania wyszukiwania na bazie Text Search Engine. Intelligent Miner for Text jako produkt IBM został włączony do zespołu „Information Integrator for Content" dla systemu zarządzania bazami danych DB2 jako środek Information Mining („głębokiej analizy informacji").

TextAnalyst

http://www.megaputer.com/

Firma rosyjska Megaputer Intelligence, znana dzięki swojemu systemowi PolyAnalyst klasy Data Mining, opracowała także system TextAnalyst, który rozwiązuje następujące zadania Text Mining:

- tworzenie sieci semantycznej dużego tekstu;
- przygotowanie streszczenia tekstu;
- wyszukiwanie w tekście;
- automatyczna klasyfikacja i klasrteryzacja tekstów.

Tworzenie sieci semantycznej – to wyszukiwanie pojęć kluczowych tekstu i ustalenie związków pomiędzy nimi. Stworzona sieć umożliwia nie tylko zrozumienie, o czym jest tekst, ale również wykonanie nawigacji kontekstowej. Przygotowanie streszczenia –to wyodrębnienie w tekście zdań, w których częściej, niż w innych, występują znaczące dla tego tekstu wyrazy. W 80% przypadków jest to zupełnie wystarczające do otrzymania ogólnego wyobrażenia o tekście. Do wyszukiwania informacji w systemie przewidziano wykorzystanie pytań w języku naturalnym. Wg zapytań tworzona jest unikalna sieć semantyczna, która we współpracy z siecią dokumentu umożliwia wyodrębnienie potrzebnych fragmentów tekstu. Klasteryzacja i klasyfikacja są wykonywane za pomocą standardowych metod pozyskiwania danych. System Text-Analyst rozpatruje Text Mining jako odrębny aparat matematyczny, który osoby opracowujące oprogramowanie mogą wbudować w swoje produkty, nie bazując na platformie wyszukiwarek lub systemu zarządzania bazami danych. Główna platforma do zastosowania systemu to MS Windows 9x/2000/NT.

WebAnalyst

http://www.megaputer.com/products/wa/index.php3

System WebAnalyst – jest również produktem MegaputerIntelligence i stanowi intelektualne rozwiązanie skalowane klient-serwer dla firm, które chcą zmaksymalizować efekt analizy danych w środowisku WWW. Serwer WebAnalyst funkcjonuje jako system w roli eksperta gromadzenia informacji i zarządzania treścią strony WWW. Moduły WebAnalyst rozwiązują trzy zadania:

- gromadzenie maksymalnej ilości informacji na temat odwiedzających stronę i zasobów, o które pytali;
- badanie zgromadzonych danych;
- generowanie personalizowanej treści na podstawie wyników badań.

Rozwiązanie tych zadań ogółem powinno, zdaniem opracowujących system, maksymalizować ilość nowych odwiedzających strony WWW i utrzymać istniejącą, a nawet zwiększyć ich popularność. WebAnalyst także może integrować możliwości Text Mining bezpośrednio do strony WWW instytucji. Pozwala to zaproponować zindywidualizowany, zautomatyzowany i celowy marketing, wyszukiwanie automatyczne i realizację sprzedaży krzyżowej oraz rozszerzyć zbiór ustawianych przez użytkownika danych. W istocie WebAnalyst jest serwerem intelektualnym dodatków sprzedaży elektronicznej. Platforma techniczna – analogiczna jak w przypadku TextAnalyst.

Text Miner

http://www.sas.com/technologies/analytics/datamining/textminer/

Firma amerykańska SAS Institute wyprodukowała system SAS Text Miner służący do porównania pewnych wersów gramatycznych i słownych w zapisanym tekście. Text Miner jest wyjątkowo uniwersalny, ponieważ może pracować z dokumentami tekstowymi różnych formatów – w bazach danych, systemach plików i nawet w sieci WWW. Text Miner zapewnia opracowywanie logiczne tekstu w środowisku potężnego pakietu SAS Enterprise Miner. Umożliwia to użytkownikom wzbogacanie procesu analizy danych, integrując niestrukturalne informacje tekstowe z istniejącymi danymi strukturalnymi, takimi jak: wiek, dochód, cechy wydatków itd..

Program Text Miner pozwala na określenie, na ile prawdziwy jest ten czy inny dokument tekstowy. Wykrycie nieprawdziwości dokumentów jest wykonywane drogą analizy tekstu i wykrycia zmiany stylu listu, co może mieć miejsce przy próbie stworzenia lub ukrycia informacji. Do wyszukiwania zmian wykorzystywana jest zasada, polegająca na wyszukiwaniu anomalii i trendów wśród zapisów baz danych bez wyjaśniania ich treści. Przy tym, do Text Miner włączono ogromny zestaw dokumentów o różnym stopniu prawdziwości, których struktura jest odbierana jako szablony. Każdy dokument jest „przepuszczany" przez wykrywacz kłamstw, analizowany i porównywany z tymi szablonami, po czym program nadaje dokumentowi taki lub inny indeks prawdziwości. Program może być szczególnie przydatny w instytucjach otrzymujących dużo korespondencji elektronicznej, a także w organach ochrony prawa, do analizy wskaźników prawdziwości zeznań wraz z wykrywaczami kłamstw, których działanie bazuje, w odróżnieniu od Text Miner, na obserwacji stanu emocjonalnego osoby.

SemioMap

http://www.entrieva.com/entrieva/index.htm

SemioMap jest produktem firmy Entrieva, stworzonym w 1996r. przez badaczasemiotyka Claude Vogel'a. W maju 1998r. program został wyprodukowany jako kompleks przemysłowy SemioMap2.0 – pierwszy system Text Mining, który pracował w architekturze klient-serwer. System SemioMap składa się z dwóch komponentów – serwera SemioMap i klienta SemioMap. Praca systemu jest wykonywana trójfazowo:

- indeksowanie (serwer SemioMap automatycznie czyta zbiory tekstu niestrukturalnego, wyciąga kluczowe słowa lub pojęcia i tworzy na ich podstawie indeks);
- klasteryzacja pojęć (serwer SemioMap określa związki pomiędzy wyciągniętymi frazami i tworzy z nich, na podstawie ich wspólnego występowania, sieć leksykalną – „mapę pojęciową");
- przedstawienie graficzne i nawigacja (wizualizacja map związków, zapewniająca szybką nawigację po słowach kluczowych i związkach między nimi, a także możliwość szybkiego dostępu do konkretnych dokumentów).

SemioMap wspiera rozmieszczenie materiału w różnych „folderach" i utworzenie odrębnej bazy danych dla każdego folderu. Związki pomiędzy pojęciami, które określa SemioMap, opierają się na wspólnym występowaniu fraz w akapitach pierwotnego zbioru tekstowego. Centralnym zespołem SemioMap jest ekstraktor leksykalny – program, który wyciąga frazy z ogółu tekstu i określa wspólne występowanie tych fraz (ich wzajemne związki). Ekstraktor leksykalny bazuje na opatentowanej technologii SemioLex. Realizuje ona teoretyczne idee semantyczne formowania komunikacji językowych Claude Vogel'a.

InterMedia Text, Oracle Text

http://technet.oracle.com/products/text/content.html

Środki Text Mining, począwszy od Text Server w składzie systemu zarządzania bazami danych Oracle 7.3.3 i cartridge interMedia Text w Oracle8i, są nieznanym składnikiem produktów Oracle. W Oracle9i środki te rozwinęły się i otrzymały nową nazwę „Oracle Text" – kompleks programowy, zintegrowany w systemach sterowania bazami danych, co umożliwia efektywną pracę z zapytaniami, dotyczącymi tekstów niestrukturalnych. Przy tym, opracowywanie tekstu jest łączone z funkcjami, umożliwiającymi użytkownikowi pracę z relacyjnymi bazami danych. Między innymi, podczas pisania załączników służących do opracowywania tekstu pojawiła się możliwość wykorzystania SQL.

Głównym zadaniem, na które skierowane są zasoby Oracle Text, jest wyszukiwanie dokumentów wg ich treści – wyrazów lub fraz, które w razie potrzeby są zestawiane z wykorzystaniem operacji typu boolowskiego. Wyniki wyszukiwania są stopniowane wg relewantności, z uwzględnieniem częstotliwości wyrazów zapytania w wyszukanych dokumentach. W celu zwiększenia pełności wyszukiwania Oracle Text udostępnia szereg środków do rozszerzenia zapytania, wśród których można wyodrębnić trzy grupy:

1. rozszerzenie słów zapytania o wszystkie formy morfologiczne wskutek otrzymania informacji, dotyczących morfologii języka;
2. Oracle Text dopuszcza poszerzenie słów zapytania o wyra-

zy bliskoznaczne wskutek podłączenia tezaurusa – słownika semantycznego;

3. poszerzenia zapytania o wyrazy, zbliżone pisownią i brzmieniem – „szczęśliwy traf" i wyszukiwanie wyrazów pokrewnych. "Szczęśliwy traf" warto zastosować do poszukiwania pomyłkowo napisanych wyrazów, a także w tych przypadkach, kiedy pojawiają się wątpliwości co do prawidłowej pisowni, np. nazwiska, nazwy instytucji itd.

System Oracle Text umożliwia wykonanie analizy tematycznej tekstów w języku angielskim. W trakcie opracowywania tekst każdego dokumentu poddawany jest analizie lingwistycznej i statystycznej, wskutek czego określane są jego tematy kluczowe, tworzone streszczenia tematyczne, a także streszczenie ogólne – referat.

Wszystkie środki mogą być wykorzystywane wspólnie, co wsparte jest językiem zapytań w połączeniu ze składnią tradycyjną SQL i PL/SQL do wyszukiwania dokumentów. Oracle Text umożliwia pracę ze współczesnymi relacyjnymi systemami zarządzania bazami danych w kontekście złożonego wyszukiwania wielocelowego i analizy danych tekstowych. Możliwości opracowywania informacji tekstowej w języku rosyjskim w Oracle Text są ograniczone. W celu rozwiązania tego problemu firma Garant-Park-Internet opracowała moduł Russian Context Optimizer (RCO), przeznaczony do korzystania wraz z interMedia Text (lub Oracle Text). Oprócz wsparcia morfologii rosyjskojęzycznej, RCO zawiera środki niedokładnego wyszukiwania („szczęśliwy traf"), analizy tematycznej i referowania dokumentów.

Autonomy Knowledge Server

http://www.autonomy.com/tech/whitepaper.pdf

Architektura systemu firmy Autonomy, znanej ze swoich opracowań, dotyczących statystycznej analizy treści, łączy intelektualną analizę składni z szablonami ze złożonymi metodami analizy treści i służącymi do rozwiązania zadań klasyfikacji automatycznej i organizacji odsyłaczy krzyżowych. Główna zaleta systemu Autonomy to potężne algorytmy intelektualne zrealizowane na podstawie obróbki statystycznej. Te algorytmy bazują na teorii informacyjnej Claude Shannon'a, prawdopodobieństwach Bayes'a i sieciach neuronowych.

Koncepcja adaptacyjnego modelowania probablistycznego APCM umożliwia systemowi Autonomy identyfikację szablonów w tekście dokumentu i automatyczne określanie podobnych szablonów dla innych dokumentów. Zaleta systemu Autonomy Knowledge Server to możliwość analizy tekstów oraz identyfikacji koncepcji kluczowych w granicach dokumentów drogą analizy korelacji częstotliwości i związków pomiędzy terminami tekstu. Składnik systemu Agentware wykorzystuje unikalną technologię analizy szablonów (nieliniową, adaptacyjną cyfrową obróbkę sygnału) z dokumentów i określenia charakterystyk zawartych w tekstach. APCM umożliwia identyfikację unikalnych sygnatur tekstu, a także tworzenie agentów koncepcji, przy pomocy których wyszukiwane są zapisy o podobnej treści na stronach WWW, w wiadomościach, archiwach poczty elektronicznej i innych dokumentach. Ponieważ system nie bazuje na z góry założonych wyrazach kluczowych, może on pracować z dowolnymi językami. Centrum systemu agentów Autonomy jest mechanizm opinii dynamicznych DRE na podstawie technologii obróbki szablonów, która wykorzystuje metody sieci neuronowych. Wykorzystuje on koncepcję adaptacyjnego modelowania probablistycznego do realizacji czterech głównych funkcji: określenia koncepcji, utworzenia agenta, nauki agenta i standardowego wyszukiwania tekstu. DRE przyjmuje zapytania w języku naturalnym lub terminy, związane z operatorami typu boolowskiego, i oddaje listę dokumentów, uporządkowanych wg relewantności zapytania. Ten mechanizm jest podstawą wszystkich produktów systemu agentów od Autonomy.

Galaktika-ZOOM

http://zoom.galaktika.ru/content.htm

System Galaktika-ZOOM jest produktem rosyjskiej korporacji „Galaktika". Główne przeznaczenie systemu to wyszukiwanie intelektualne wg słów kluczowych z uwzględnieniem morfologii języków rosyjskiego i angielskiego, a także formowanie

zbiorów informacyjnych do konkretnych aspektów zapytania. Obszary wyszukanych informacji mogą osiągać setki gigabajtów. System jest zorientowany na duże obiekty informacyjne – wiadomości i artykuły w mediach, wydawnictwa branżowe, dokumentację normatywną, korespondencję handlową, materiały wewnętrznego obiegu dokumentów przedsiębiorstwa, informacje z Internetu. Przekazuje on narzędzia do analizy obiektywnych związków treści wybranych danych i formowania „obrazu" problemu – modele wielowymiarowe w strumieniu informacji w formie stopniowanej listy istotnych wyrazów z danej problematyki. Znaczną uwagę zwraca się tutaj na określanie tendencji dynamiki rozwoju badanego problemu. System zawiera konwertory często wykorzystywanych formatów: prosty tekst, RTF, DOC, HTML. Galaktika-ZOOM funkcjonuje w środowisku OC Windows 2000.

InfoStream

http://infostream.com.ua

Objęcie i uogólnienie dużych dynamicznych zbiorów informacji, ciągle generowanych w Internecie, wymaga jakościowo nowych podejść. Jednym z nich jest monitoring treści. W celu otrzymania przekrojów jakościowych i ilościowych taki monitoring powinien być przeprowadzany stale, w ciągu nieokreślonego z góry czasu. Aby rozwiązać to zadanie, na Ukrainie, w Ośrodku Informacyjnym „ELWISTI" została opracowana technologia InfoStream™. Środki programowo-technologiczne InfoStream obejmują trzy podstawowe składniki (centra):

- gromadzenia i obróbki informacji;
- organizowania interaktywnego dostępu do baz danych;
- monitoringu treści.

Jądrem mechanizmu obróbki treści InfoStream jest pełnotekstowa wyszukiwarka informacyjna InfoReS. Technologia umożliwia stworzenie pełnotekstowych baz danych i realizację wyszukiwania informacji, formowanie tematycznych kanałów informacyjnych, automatyczne przeprowadzanie rubrykacji informacji, tworzenie przeglądów, tabel związków pojęć na podstawie informacji internetowych, histogramów podziału rozpatrywanych znaczeń odrębnych pojęć, a także dynamiki ich występowania w sieci. Technologia InfoStream umożliwia obróbkę danych w formatach MS WORD (DOC, RTF), PDF, a także wszystkich formatach tekstowych (tekst zwykły, HTML, XML). Systemy na bazie InfoStream funkcjonują na platformach OC FreeBDS, Linux, Solaris.

Uniwersytet Rzeszowski
al. Rejtana 16c, 35-959 Rzeszów

tel. +48 17 872 13 43

biuro@inprona.pl www.inprona.pl

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego